Casualty Actuarial Society
E-Forum, Fall 2015

The CAS *E-Forum*, Fall 2015

The Fall 2015 edition of the CAS *E-Forum* is a cooperative effort between the CAS *E-Forum* Committee and various other CAS committees, task forces, or working parties.

This *E-Forum* contains five independent research papers and six reserves call papers, which were created in response to a call for non-technical papers on reserves issued by the CAS Committee on Reserves. Some of the call papers will presented at the CAS Annual Meeting, held November 15-18, 2015, in Philadelphia.

CAS *E-Forum*, Fall 2015

Table of Contents

For information on submitting a paper to the *E-Forum*, visit http://www.casact.org/pubs/forum/.

Interpolation Hacks and their Efficacy

Lynne Bloom, FCAS, MAAA

Abstract
Actuaries are consistently faced with the decision of how to interpolate loss development factors. Methods vary from linear to more theoretical. This paper explores how various methods hold up to actual data and each other by estimating errors in reserve prediction when using paid loss development, incurred loss development and Bornhuetter Ferguson methods. It also lays out a variety of methods for actuaries to use. Lastly, this paper adds an additional process to account for unique situations such as seasonal fluctuations in claims activity. Along with this paper, I have included a practical tool programmed with interpolation formulae and the seasonal method.

Keywords. Interpolation, Development, Quarterly Reserving

1. INTRODUCTION

As Actuaries, we are challenged with producing estimates which are assumed to be accurate given our vast background and industry knowledge. In practice, the documentation of our thought process is a crucial part of third-party assessment of our work product. The hindsight accuracy of the estimate is something we seldom evaluate.

One of the crucial assumptions we make is the selection of development factors and how to interpolate them. Although practices vary widely, it is something that in my experience is not well documented. In fact, in many instances, third party software is relied upon to determine the interpolated amount.

Extrapolated development factors, such as those used for a 9 month old accident year, are more frequently inconsistent and poorly documented. Actuarial practice also varies in determining how to treat the exposure growth portion of a development factor and how to document this treatment.

While differences in judgement regarding a loss development selection and ancillary differences in judgment due to the nuances of a particular interpolation/extrapolation method may not seem material in the context of a reserve review when compared to other more substantial judgments, they do have an impact. More importantly, they impact the quality of our documentation.

Interpolation is heavily relied upon for interim reserves studies and year-end studies where the practicalities of timing only allow for a third quarter in-depth review. Interpolation is also relied upon to form opinions of actual versus expected loss emergence.

Methodology for interpolation varies from sophisticated curve fits, to shortcut methods, to straight linear. This paper will examine various methods (known to the author) and how they

compare to actual results produced with sample data that covers various lines of business. This paper will examine the relative degrees of error each method might be expected to produce and the overall effect on reserves estimation.

In addition, the paper will address a special situation where there is specific knowledge of development patterns which are expected to vary on a seasonal basis.

1.1 Research Context

These general concepts are covered by other authors (Flannery, Press, Teukolsky and Vetterling) in "Numerical Recipes" and more recently in Variance Magazine by Joseph Boor in "Interpolation Along a Curve."

Richard Sherman also explored these concepts in "Extrapolating, Smoothing and Interpolating Development Factors."

1.2 Objective

The objective of this paper is to provide options, easy to follow formulae, and tools for the purpose of interpolation, along with context regarding the efficacy of various methods. The hope for this paper is to be a useful reference source for actuaries and students familiarizing themselves with actuarial methods and shortcuts.

1.3 Outline

The remainder of the paper proceeds as follows:

Section 2: Background

Section 3: Interpolation Methods and Formulae

Section 4: Extrapolation Methods and Formulae

Section 5: Handling of Exposure Growth

Section 6: Testing of Methods and Results for Interpolation

Section 7: Testing of Methods and Results for Extrapolation

Section 8: Seasonal Adjustment Method

Section 9: Conclusions

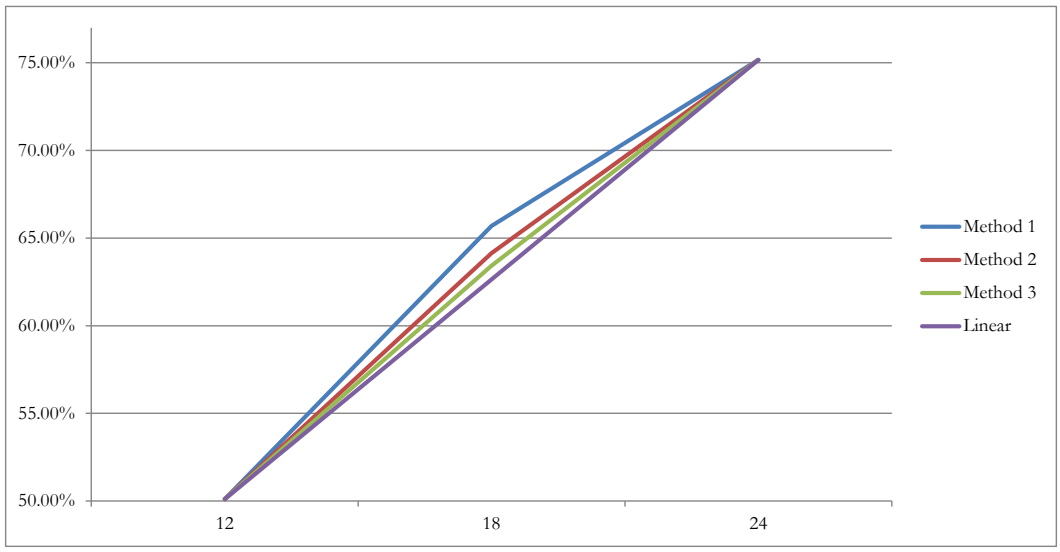
2. BACKGROUND

When we think of the concept of interpolation, it is extremely simple – how will losses paid or reported vary over the course of a year (or other specified period)? The easiest concept to grasp is linear interpolation, which assumes development proportional to time over the period. However, sometimes actuaries are more comfortable with the assumption that more will be reported or paid earlier rather than later (and sometimes vice versa). This assumption gives rise to alternative methods, including curve fitting methods.

Since performing a power curve or Weibull regression inside of a spreadsheet can be cumbersome, actuaries have developed many "shortcut" methods and formulae which mimic the effects of a curve regression and consequently mimic the effect that more losses emerge sooner over the interpolation period. We can see this graphically using a simple example. Suppose annual development is as follows:

Ages	12	24	36
	12 - 24	24 - 36	36 - 48
Selected Result	1.500	1.200	1.050
FacToUlt	1.996	1.331	1.109
Percent of Ult	50.11%	75.16%	90.19%

If we apply various interpolation methods which assume more losses emerge sooner and compare the implied development to that indicated by linear interpolation, we would observe the following:



The corresponding 18 month development factors would be as follows:

Age	18
Method 1	1.522
Method 2	1.559
Method 3	1.577
Linear	1.597

As can be seen above, since linear interpolation assumes steady development, the factor at 18 months is higher than that given by the other methods (which assume accelerated development earlier in the period). These methods are used for illustration and all methods will be given in detail in Section 3.

Although the above demonstrates the general goal of interpolation, in practice, we seldom evaluate the results of one method versus others. The following sections will outline several methods. While the derivations of some of the formulae are quite obvious, some of the shortcut formulae have been passed down from actuary to actuary. It is beyond the scope of this paper to understand the derivation of each method; rather this paper will evaluate the efficacy of each given certain assumptions, which that individual practitioner might make. Since the data evaluated in this paper is far from exhaustive, the link between assumptions and accuracy of the each method is important. The choice of the curve should be driven by what the actuary assumes about the true shape of the curve. This paper will also not explore all possible curve fitting methods, but only some of the more common ones to compare to other shorter methods.

Lastly, based on the same notion that the assumptions about emergence are important, the actuary may use these methods for accident year or policy year methods equally. Some of the observation about early maturities made in the sections to follow would obviously apply for a longer period of time when using policy year data.

3. INTERPOLATION METHODS AND FORMULAE

The following is a list of methods I will explore:

1. Linear
2. Inverse Power Curve on Remaining Development (CDF-1)

3. Weibull
4. Inverse Power Curve on Total Development (CDF)
5. Exponential Curve on Remaining Development (CDF-1)
6. Exponential Curve on Total Development (CDF)
7. Logarithmic Proportions Shortcut (Shortcut 1)
8. Exponential Weighted Shortcut (Shortcut 2)

The formulae included below will contain the following terms. Since in practice most of these formulae will be utilized in Excel, I have used shorthand geared toward excel functions.

LDF_T – Incremental Loss Development Factor at age T

CDF_T – Cumulative Loss Development Factor at age T

T – Development Age in Months

PR_T or PP_T – Percent Reported or Percent Paid or $1/ CDF_T$

EXP (Value) – e^{Value}

* - x or multiplication

3.1 Linear Method

The most commonly used method is the Linear Method, which as stated above assumes that the percent paid or reported grows at a constant rate with time. For the purpose of demonstrating the methods, I will assume that we are interpolating between 12 and 24month factors in all of our examples. I will also use the following Paid Development Factors:

$CDF_{12} = 1.996$

$CDF_{24} = 1.331$

$PP_{12} = 50.11\%$

$PP_{24} = 75.16\%$

I will also suppose I am interpolating to 15 months. To derive a linear interpolation estimate, I use the following formula.

$PP_{15} = PP_{12} * (24 – 15)/(24 – 12) + PP_{24} * (15 – 12) / (24 – 12) =$

$CDF_{15} = 1.774$

3.2 Inverse Power Curve Regression on Remaining Development (IVP Decay)

An Inverse Power Curve Regression assumes that development and loss emergence behave in such a way that interim CDFs can be expressed as follows:

$$CDF_T - 1 = a * T^{-b}$$

In more qualitative terms, it is assumed that the remaining development at any point in time varies inversely with time.

Translating this into a convenient linear regression results in the following equation:

$$\ln(CDF_T-1) = b * \ln(1/T) + \ln(a)$$

In Excel, the function to find the 15 month CDF would be as follows:

EXP(TREND(ln(CDF_{12}-1):ln(CDF_{24}-1),ln(1/12):ln(1/24),ln(1/15)))+1

Using the values stated in the linear example, the resulting value for CDF_{15} = 1.698

Note that I do not try to interpolate between points any wider than the two adjacent development points, as fitting a large curve is more complex and often results in aberrant values. The theoretical considerations for best fit are outside the scope of this paper.

3.3 Weibull Method

The Weibull Method assumes that development and loss emergence behave in such a way that interim CDFs can be expressed as follows:

$$1 - PP_T = EXP(-a \times T^b)$$

Translating this into a convenient linear regression results in the following equation:

$$\ln(-\ln(1-\%PP_T)) = \ln(a) + b \ln(T)$$

In excel, the function to find the 15 month CDF would be as follows:

1/[1 – EXP{-EXP(TREND(ln(-ln(1-1/CDF_{12})):ln(-ln(1-1/CDF_{24})),ln(12):ln(24),ln(15)))}]

Using the values stated in the linear example, the resulting value for CDF_{15} = 1.722

Note the difference between the above curves is merely the form of the equation. The basic principal is the same: remaining development varies inversely with time. This paper does not lay out every possible combination of type of curve and dependent variable, but rather some of the more commonly used ones.

3.4 Inverse Power Curve Regression on Total Development (IVP)

The Inverse Power Curve Regression on Total Development assumes that development and loss

emergence behave in such a way that interim CDFs can be expressed as follows:

$$CDF_T = a * T^{-b}$$

In more qualitative terms, it is assumed that the total development at a point in time varies inversely with time.

Translating this into a convenient linear regression results in the following equation:

$$\ln(CDF_T) = b * \ln(1/T) + \ln(a)$$

In excel, the function to find the 15 month CDF would be as follows:

EXP(TREND(ln(CDF_{12}):ln(CDF_{24}),ln(1/12):ln(1/24),ln(1/15)))

Using the values stated in the linear example, the resulting value for CDF_{15} = 1.752

Note that this method will not create errors when the CDF is less than or equal to 1.000. While this is an advantage, in practice, I often have formulae default to linear values (and have done so in the practical tool) when CDFs are less than one as the differences in small factors are less material. When CDFs are large, the method tends to produce much higher values than the regression on remaining development.

3.5 Exponential Curve Regression on Remaining Development (Expo Decay)

Exponential Regression assumes that development and loss emergence behave in such a way that interim CDFs can be expressed as follows:

$$CDF_T - 1 = a * EXP(bT)$$

In more qualitative terms, it is assumed that the remaining development at a point in time varies inversely with time.

Translating this into a convenient linear regression results in the following equation:

$$\ln(CDF_T-1) = b * T + \ln(a)$$

In excel, the function to find the 15 month CDF would be as follows:

EXP(TREND(ln(CDF_{12}-1):ln(CDF_{24}-1),12:24,15))+1

Using the values stated in the linear example, the resulting value for CDF_{15} = 1.756

This has some properties of the Weibull curve and some properties of the inverse power curve and acts as a variation.

3.6 Exponential Curve Regression on Total Development (Expo)

Exponential Regression assumes that development and loss emergence behave in such a way that interim CDFs can be expressed as follows:

$$CDF_T = a * EXP(bT)$$

In more qualitative terms, it is assumed that the total development at a point in time varies inversely with time.

Translating this into a convenient linear regression results in the following equation:

$$\ln(CDF_T) = b * T + \ln(a)$$

In excel, the function to find the 15 month CDF would be as follows:

EXP(TREND(ln(CDF_{12}):ln(CDF_{24}),12:24,15))

Using the values stated in the linear example, the resulting value for CDF_{15} = 1.803.

3.7 Logarithmic Proportions Shortcut (Shortcut 1)

This shortcut will produce results which are generally about midway between linear results and curve fitted results.

In excel, the formula to find the 15 month CDF would be as follows:

CDF_{15} = CDF_{12}^((ln(CDF_{24})/ln(CDF_{12}))^((15-12)/(24-12)))

Using the values stated in the linear example, the resulting value for CDF_{15} = 1.740.

This formula is easier to program into a spreadsheet than regressions and provides a directionally similar result. Regressions require the logarithm to be made in a separate cell first which is cumbersome when dealing with multiple development points.

3.8 Exponential Weighted Shortcut (Shortcut 2)

This shortcut will produce results which are generally higher than Shortcut 1, but lower than linear. The results tend to hover near the exponential regression as well.

In excel, the formula to find the 15 month CDF would be as follows:

CDF_{15} = 1/ln(EXP(1/CDF_{12})*(24-15)/(24-12) + EXP(1/CDF_{24})*(15-12)/(24-12)))

Using the values stated in the linear example, the resulting value for CDF_{15} = 1.755.

This formula is easier to program into a spreadsheet than regressions and provides a directionally similar result.

3.9 Summary of Values from Various Methods

The following table summarizes the results of the methods extended out to further maturities:

Development Factor Selection Ages	12	24	36	48	60	72
	12 - 24	24 - 36	36 - 48	48-60	60-72	72-84
Selected Result	1.500	1.200	1.050	1.025	1.020	1.010
FacToUlt	1.996	1.331	1.109	1.056	1.030	1.010
Percent of Ult	50.11%	75.16%	90.19%	94.70%	97.07%	99.01%
Interim Ages		15	27	39	51	63
Linear		1.774	1.267	1.095	1.049	1.025
IVP Decay		1.698	1.239	1.090	1.047	1.022
Weibull		1.722	1.248	1.092	1.048	1.023
IVP		1.752	1.262	1.094	1.049	1.025
Expo Decay		1.756	1.250	1.092	1.048	1.023
Expo		1.803	1.271	1.095	1.049	1.025
Logarithimic Proportions Shortcut 1		1.740	1.248	1.092	1.048	1.023
Exponential Weighting Shortcut 2		1.755	1.264	1.095	1.049	1.025

It is obvious that our choice of method matters less as accident years mature. In this particular example the Exponential regression of total development actually provides a development factor that is higher than linear. This implies that more losses will emerge in the latter part of the year than would be indicated proportionally with time. Later in this paper we will explore how actual data relates to this assumption.

4. EXTRAPOLATION METHODS AND FORMULAE

Probably the only concept more elusive than interpolation methods is extrapolation methods. All of the methods are either linear or based on shortcuts and the theoretical bases for these methods are more tenuous than for interpolation methods.

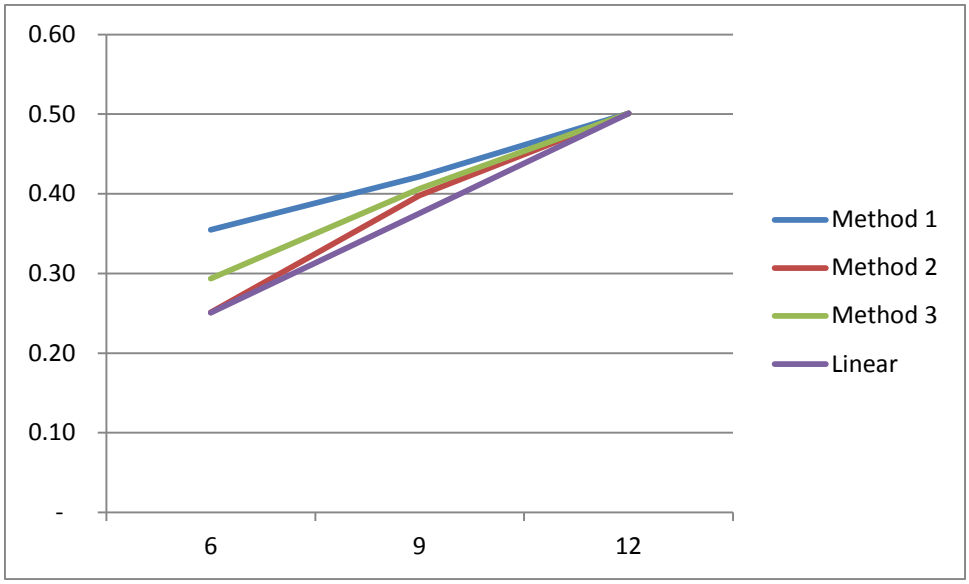
Development prior to 12 months or prior to the first known development point is complicated by a more rapid rate of growth as well as exposure growth. An extrapolation formula needs to consider both of these factors and the actuary should document each piece separately. All of the shortcuts provided mimic the general concept that since more losses are reported or paid closer to the time of the accident, the development will be less than linear within the 12 months. In other words, more than half of losses reported by 12 months on losses occurring within the first six months will be reported as of six months. While this may not always be the case, it is the concept

behind these shortcuts.

Section 5 will deal more directly with exposure growth and how it should be handled.

For example, suppose that as of 12 months, we expected reported losses to grow by an additional 40% or that we have a reported development factor of 1.40. Assuming that all premium is earned as of 12 months in time, the development factors pulled from triangles would not include any exposure growth. When evaluating a development factor as of 6 months, we might assume an extrapolation curve that assigns a value of 1.80. The curve would take into account the expected additional development growth, but not necessarily the exposure growth. Assuming even earning of premium throughout the year, the exposure growth factor is 2.00. Therefore the factor to apply to losses at 6 months in order to get a full year of losses would be 1.80 x 2.00 or 3.60. It is beyond the scope of this paper to analyze why specific formulae do not account for exposure growth. In section 7, we will test the adequacy against actual data using our assumption that most actuarial shortcuts do not include an exposure growth component. The following formulae all assume extrapolation without exposure growth.

I will explore several shortcut methods in additional to linear extrapolation. A graph of the percent reported implied by the various formulae might look as follows with most methods assuming the percent reported (or paid) is greater than that implied by the linear method:



Note, as we see in our examples, that when development factors are very high, the reverse is true and shortcut methods produce higher development than linear.

4.1 Linear Method

The easiest method to apply is the Linear Method, which, as stated above, assumes that the percent paid or reported grows evenly with time. For the purpose of demonstrating the methods, I will assume that we are extrapolating from a 12 month factor in all of our examples. I will also use the following Paid Development Factors:

$CDF_{12} = 1.996$

$PP_{12} = 50.11\%$

I will also suppose I am interpolating to 6 months. To derive a linear extrapolation estimate, I use the following formula.

$PP_6 = PP_{12} * (6/12)$

$CDF_6 = 3.992$

4.2 The Plus 12 Method (Method 1)

This method raises the base development factor to a power which increases as the number of months decreases by using subtraction. The formula is as follows:

$CDF_6 = CDF_{12}\text{^}((12+12 - 6)/12)$ where the first 12 in the exponent represents the age of the base factor and the second 12 is always present.

$CDF_6 = 2.819$.

4.2 The Power Ratio Method (Method 2)

This method raises the base development factor to a power which increases as the number of months decreases by using a ratio. The formula is as follows:

$CDF_6 = CDF_{12}\text{^}(12/6)$

$CDF_6 = 3.983$

This method tends to reach uncommonly high values when applied to smaller maturities.

4.3 The Natural Log Method (Method 3)

This method uses the natural log of the remaining development, applies a ratio based on the extrapolation month, and converts it back using Euler's number, e.

$CDF_6 = 1/(1-EXP(\ln(1-1/ CDF_{12})*(6/12)))$

$CDF_6 = 3.405$

This method tends to be more stable at lower maturities.

4.4 Summary of Values from Various Methods

The following table summarizes the results of the methods:

Ages	12
	12 - 24
Selected Result	1.500
FacToUlt	1.996
Percent of Ult	50.11%
Interim Age	6
Linear	3.992
Method 1	2.819
Method 2	3.983
Method 3	3.405

5. HANDLING OF EXPOSURE GROWTH

As mentioned earlier, varying practices exist with regards to exposure growth and unfortunately many actuaries are unaware of whether their extrapolation method accounts for it. The use of reserving software has created, to some degree, a "black box" that obscures the derivation of early maturity loss development factors. To be fair, with the use of exposure based methods such as Bornhuetter Ferguson or Cape Cod, most actuaries feel that factors for immature periods are immaterial to an analysis. While this is true, it is preferable to have extrapolation methods explicitly used and documented.

It is difficult to extrapolate a factor for a short accident period. This is further complicated by the existence of exposure growth in exposure based methods. Essentially, there are two ways to look at a short period: it can be viewed as a short period on its own or as a fraction of the full year. From the former viewpoint, we would use factors (utilizing the methods above) which do not include exposure growth. From the latter viewpoint, we would adjust our factors for exposure growth and then scale the final ultimate produced. For the loss development method, it seems arbitrary to make a distinction as the two answers will never be different. However, for the Bornhuetter Ferguson (BF) method, the assumption can make a difference in the final answer (sometimes a large one).

Consider the example and factors from Section 4. Suppose we chose Method 3. Further

	Earned Premium	Paid Losses	CDF	LDM	Proration	Ultimate	Loss Ratio
Partial Year Method	50	10	3.405	34.05	100%	34.05	68.1%
Full Year Method	100	10	6.811	68.11	50%	34.05	68.1%

suppose that earned premium for the full year is 100 and that paid losses as of 6 months are 10. As mentioned above, the assumption of partial year or full year makes no difference to the Loss Development Method (LDM):

Now assume that the Initial Expected Loss Ratio is 60%. The following shows the results of the BF Method under each assumption:

	Earned Premium	Paid Losses	CDF	IELR	BF	Proration	Ultimate	Loss Ratio
Partial Year Method	50	10	3.405	60.0%	31.19	100%	31.19	62.4%
Full Year Method	100	10	6.811	60.0%	61.19	50%	30.60	61.2%

In the second example, the differences in loss ratio are not due to inaccuracies in either method. The difference is driven by the assumed maturity of the year. The BF method assigns more weight to the loss development method based on the maturity of the accident year as measured by the inverse of the loss development factor. Since the partial year method uses a smaller development factor, the loss development method receives more weight. Given the shortened period relative to the full year, one could argue that it is more mature (i.e., that the average accident date is earlier than a full year) and that the partial year method is therefore preferable. In reality, the relative weighting assigned to the ELR and the loss development method is subjective and many actuaries prefer to give less weight to a loss development method based on highly leveraged and extrapolated factors. In any event, awareness is important as the weighting can have material effects on results.

6. TESTING OF METHODS AND RESULTS FOR INTERPOLATION

The sample data was based on actual triangles from two different insurance companies. The lines of business underlying the data include 30 different lines and sublines with a mix of property and casualty. Lines were matched into groups based on their development properties (using 12 month factor and length of tail to group them). I began with quarterly loss triangles for many lines of business on a paid and incurred basis. Accident years within our data span from 2003 to 2014. For some lines the latest evaluation is December 31, 2014 and for others is June 30 or September 30 of

2014. First, I calculated quarterly development factors using several averages including simple, weighted, and simple excluding high and low. Then I created 4 separate annual triangles, incepting at 3, 6, 9 and 12 months respectively. Since only the weighted average ties back between quarterly and annual triangles and since I wanted to isolate the error solely due to interpolation method, I made the selections for these triangles equal to their quarterly triangle equivalent (by type of average).

I applied the interpolation methods to each triangle as described in Section 3. Then I projected the quarterly results by accident year using the most recent data and evaluation. Each result was then compared to the interpolated result based on various annual triangles. I evaluated the interpolated results using interpolations from 3, 6, and 9 months prior to the latest quarterly date. To illustrate, if the latest data was evaluated as of 12 months, I would use factors from my 3 month annual triangle to determine the 9 months prior interpolation error. The 6 month annual triangle interpolated to 12 months gave the 6 month prior error and the 9 month annual triangle gave the three month prior error. The errors were also calculated on a paid and incurred basis and for all three averages.

Error was measured in terms of IBNR for the incurred triangles and total reserves for the paid triangles. The percent error was calculated as a percent of total IBNR or reserve. Therefore percent error for paid losses would equal:

[Ultimate losses derived from Interpolated method – Ultimate losses derived from quarterly triangle factors] / [Ultimate losses derived from quarterly triangle factors – Paid losses at latest evaluation].

6.1 Results for Short Tailed Lines of Business

I grouped lines where the incurred development factor at 12 months was not greater than 2.0. These lines were primarily various types of personal auto business. The average incurred loss development factor at 12 months was roughly 1.09 and typical development dropped off at about 84 months. The average paid loss development factor at 12 months was approximately 1.5.

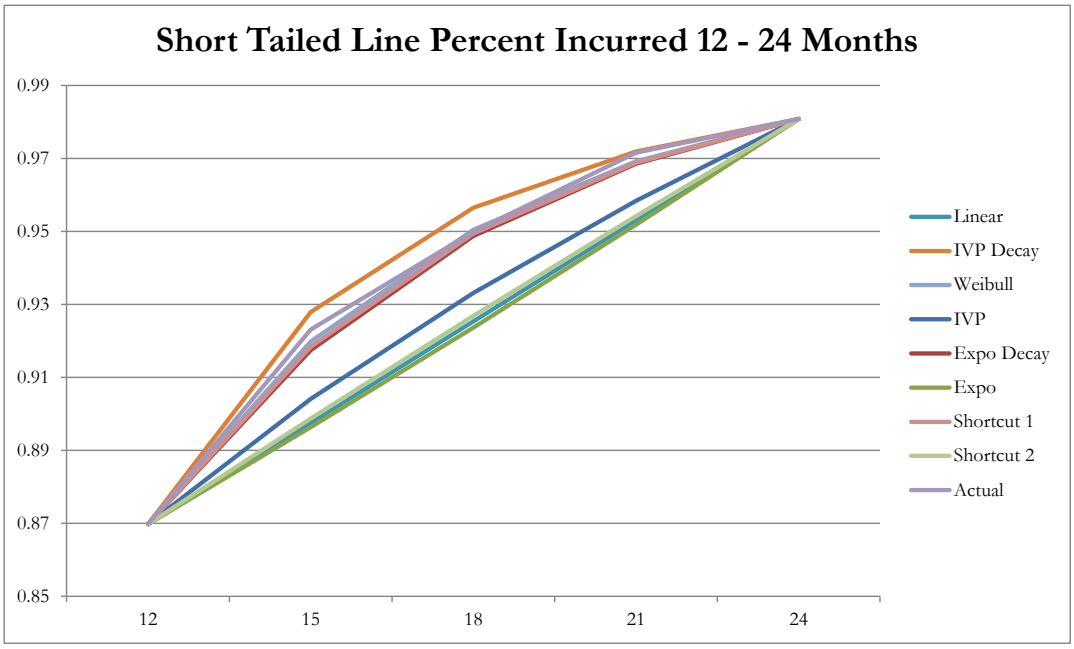
6.1.1 The Factors

Looking at the group as a whole, the following table displays the actual and estimated incurred factors from various methods measured on interpolating from various points during the year (using weighted average factors).

Short Tailed Lines - Incurred Development Factors

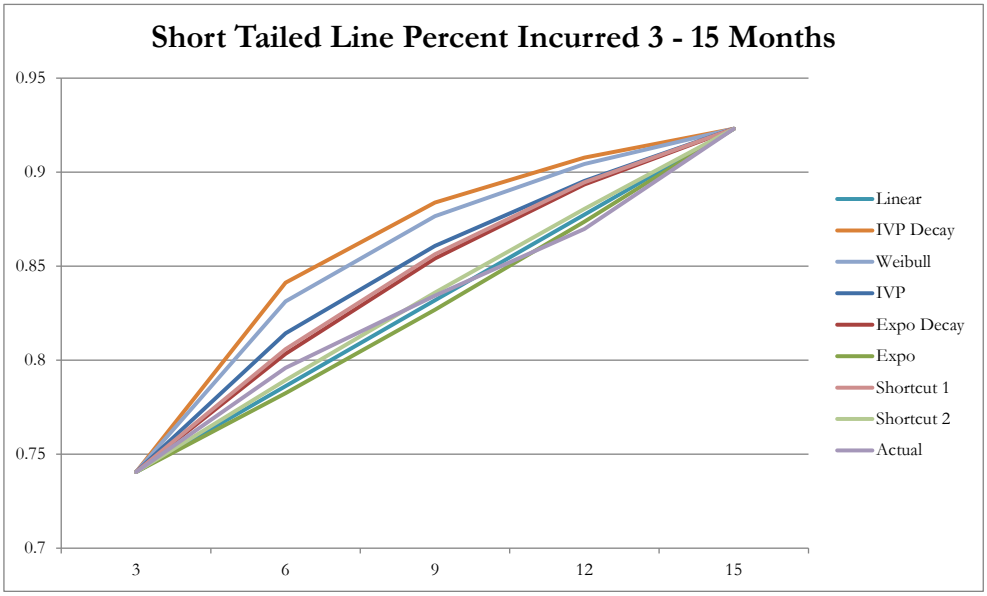
Accident Year	Age	From Quarterly "Actual"	Linear	IVP Decay	Weibull	IVP	Expo Decay	Expo	Shortcut 1	Shortcut 2
					Using factors interpolated from 9 months prior to date					
2006	108	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2007	96	1.000	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001
2008	84	1.002	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001
2009	72	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002
2010	60	1.006	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005
2011	48	1.005	1.004	1.004	1.004	1.004	1.004	1.005	1.005	1.004
2012	36	1.012	1.012	1.011	1.011	1.011	1.011	1.012	1.012	1.012
2013	24	1.029	1.032	1.027	1.028	1.030	1.028	1.032	1.029	1.031
2014	12	1.090	1.096	1.063	1.067	1.079	1.077	1.105	1.076	1.090
					Using factors interpolated from 6 months prior to date					
2006	108	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2007	96	1.000	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001
2008	84	1.002	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001
2009	72	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002
2010	60	1.006	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005
2011	48	1.005	1.005	1.004	1.004	1.004	1.004	1.005	1.005	1.005
2012	36	1.012	1.012	1.012	1.012	1.012	1.012	1.013	1.012	1.012
2013	24	1.029	1.033	1.028	1.029	1.031	1.029	1.033	1.028	1.032
2014	12	1.090	1.097	1.061	1.067	1.083	1.077	1.103	1.077	1.093
					Using factors interpolated from 3 months prior to date					
2006	108	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2007	96	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2008	84	1.002	1.001	1.001	1.001	1.001	1.001	1.001	1.001	1.001
2009	72	1.002	1.002	1.002	1.002	1.002	1.002	1.003	1.002	1.002
2010	60	1.006	1.006	1.006	1.006	1.006	1.006	1.006	1.006	1.006
2011	48	1.005	1.005	1.004	1.005	1.005	1.005	1.005	1.005	1.005
2012	36	1.012	1.012	1.011	1.012	1.012	1.012	1.012	1.012	1.012
2013	24	1.029	1.031	1.029	1.030	1.031	1.030	1.032	1.030	1.031
2014	12	1.090	1.098	1.071	1.078	1.090	1.083	1.100	1.082	1.096

An examination of actual quarterly incurred factors between 12 and 24 months for a typical line in this group reveals the differences between actual factors and our interpolation methods.



Note that the shape of the curve of methods indicates that most methods anticipate more losses emerging in the beginning of the period versus the latter part of the period. Also all methods are higher than linear in terms of percent reported. The actual data agrees with the majority of the methods in the accelerated emergence of losses. This would suggest that linear interpolation overstates estimates for short tailed lines. Early maturities are shown in these graphs because visually it is easier to see the shape. From the table above, one can see that choice of method matters less once factors are less than 1.05.

However looking further back at the period between 3 and 15 months we see the following results:



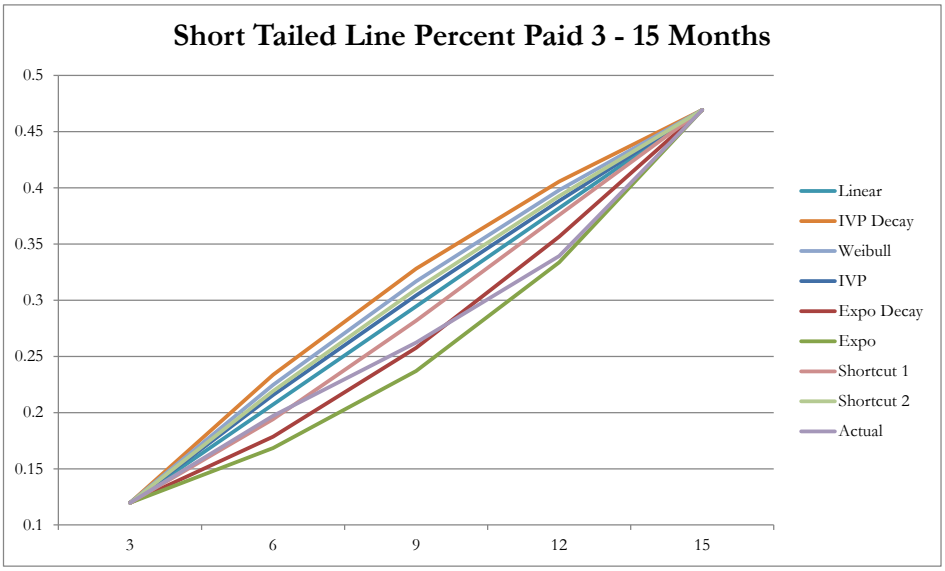
Although the methods are interpolating along the same general curve, the actual results are much nearer to linear. Note to avoid confusion about these factors, these factors are not extrapolated even though they are lower than 12 months. These are interpolated factors from a 3, 15, 27, etc. triangle. They use the same methods as other interpolated factors. They are added here to show how the shapes may differ during this time period.

The table of paid factors (using weighted average factors) is as follows:

Short Tailed Lines - Paid Development Factors

Accident Year	Age	From Quarterly "Actual"	Linear	IVP Decay	Weibull	IVP	Expo Decay	Expo	Shortcut 1	Shortcut 2
					Using factors interpolated from 9 months prior to date					
2006	108	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002
2007	96	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003
2008	84	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005
2009	72	1.009	1.009	1.009	1.009	1.009	1.009	1.009	1.009	1.009
2010	60	1.017	1.018	1.017	1.017	1.018	1.017	1.018	1.017	1.018
2011	48	1.033	1.036	1.033	1.034	1.036	1.034	1.037	1.034	1.036
2012	36	1.074	1.080	1.073	1.075	1.079	1.075	1.082	1.075	1.078
2013	24	1.175	1.193	1.176	1.183	1.195	1.189	1.212	1.183	1.185
2014	12	1.502	1.419	1.372	1.389	1.420	1.485	1.570	1.428	1.396
					Using factors interpolated from 6 months prior to date					
2006	108	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002
2007	96	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003	1.003
2008	84	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005
2009	72	1.009	1.009	1.009	1.009	1.009	1.009	1.009	1.009	1.009
2010	60	1.017	1.018	1.017	1.017	1.018	1.017	1.018	1.017	1.018
2011	48	1.033	1.037	1.033	1.034	1.036	1.034	1.037	1.034	1.036
2012	36	1.074	1.081	1.072	1.075	1.080	1.075	1.083	1.074	1.079
2013	24	1.175	1.194	1.174	1.182	1.195	1.188	1.210	1.182	1.187
2014	12	1.502	1.414	1.365	1.390	1.429	1.470	1.550	1.410	1.385
					Using factors interpolated from 3 months prior to date					
2006	108	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002	1.002
2007	96	1.003	1.003	1.003	1.003	1.003	1.003	1.004	1.003	1.003
2008	84	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005	1.005
2009	72	1.009	1.009	1.009	1.009	1.009	1.009	1.009	1.009	1.009
2010	60	1.017	1.018	1.017	1.017	1.018	1.017	1.018	1.017	1.018
2011	48	1.033	1.036	1.033	1.034	1.036	1.034	1.036	1.034	1.036
2012	36	1.074	1.079	1.073	1.075	1.078	1.075	1.080	1.074	1.078
2013	24	1.175	1.189	1.174	1.180	1.188	1.183	1.196	1.179	1.184
2014	12	1.502	1.447	1.408	1.434	1.467	1.479	1.532	1.442	1.421

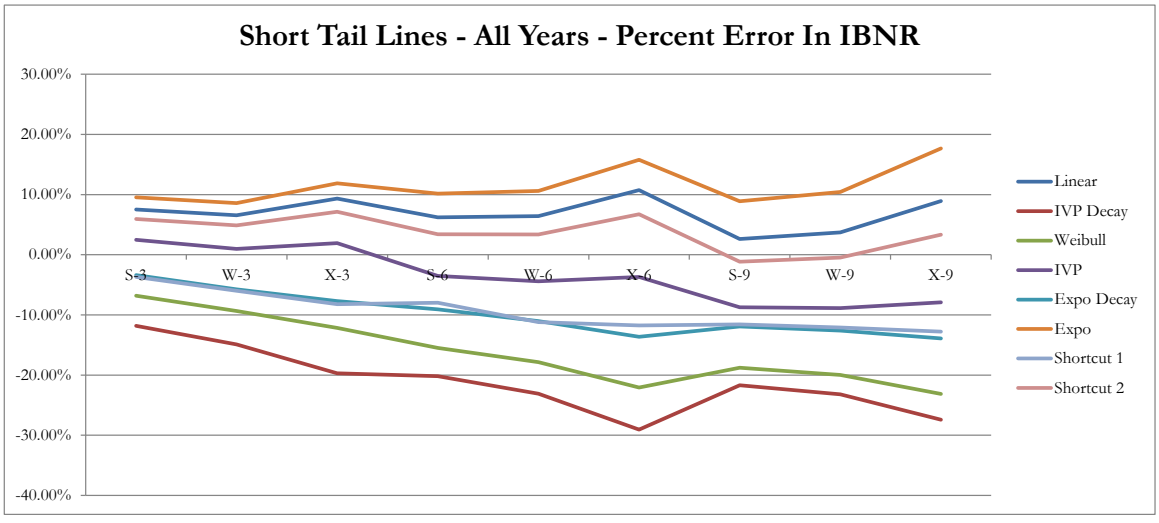
Looking at the paid factors for the 3 – 15 month period for a typical line, the curve reverses itself:



Both the method results and the actual data support the notion that prior to 15 months on a paid basis, less losses are paid earlier in the period and more losses are paid later in the period.

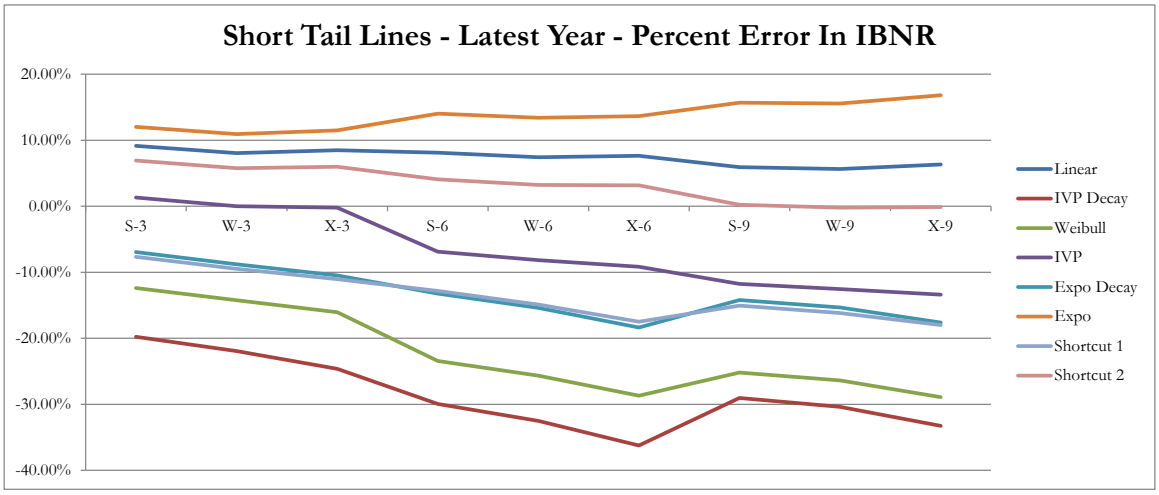
6.1.2 The Errors

The actual errors in each method for all years combined are as follows for incurred factors:

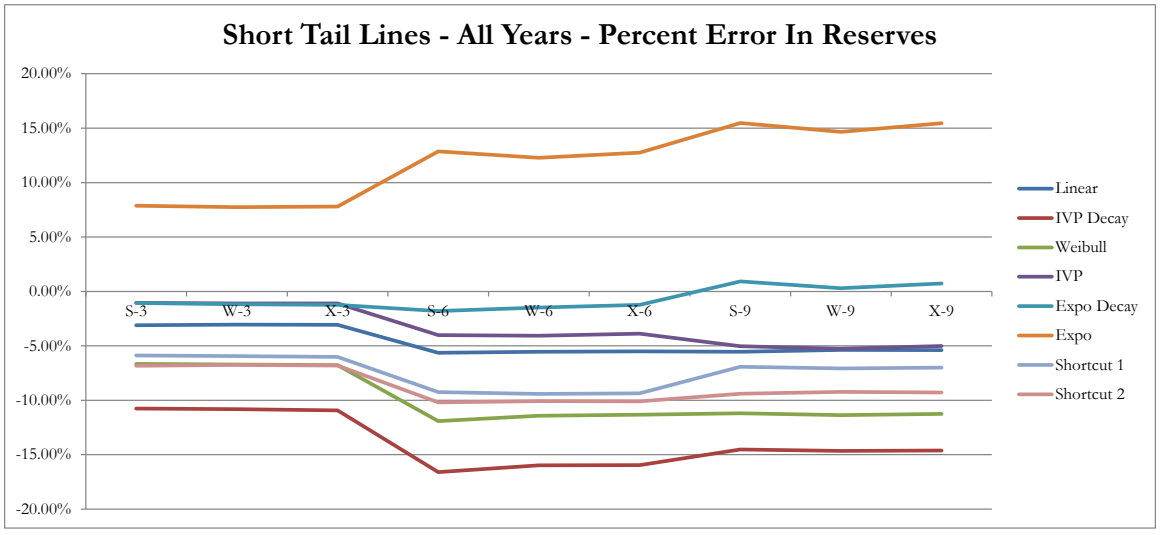


In this graph, S – 3 represents the method performed on simple average factors 3 months prior to evaluation date. W represents weighted factors and X represents simple average excluding high and low factors. Seemingly the most appealing method for short tailed incurred business based on this graph would be Shortcut 2. Many of the curves understate reserves and the linear method overstates reserves.

Looking at the latest year only, the pattern is more exaggerated:



For paid development, the error in all years is illustrated in the following graph. Where the factors are higher, in this case, the exponential method seems to work best:



A complete set of graphs pertaining to short tailed lines is provided in Appendix A.

6.2 Results for Medium Tailed Lines of Business

The medium tailed business had an average 12 month incurred development factor of approximately 7.00 and paid development factor of approximately 25.00. It consists primarily of claims made liability. The pattern becomes negligible after 96 months.

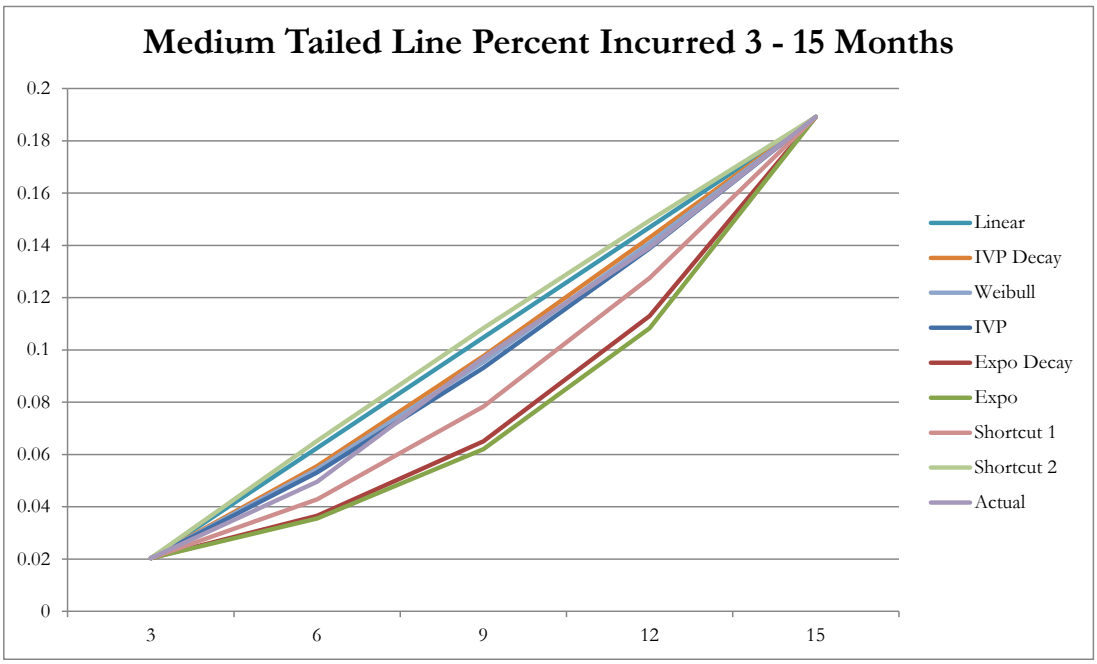
6.2.1 The Factors

The following table displays the results for weighted average factors:

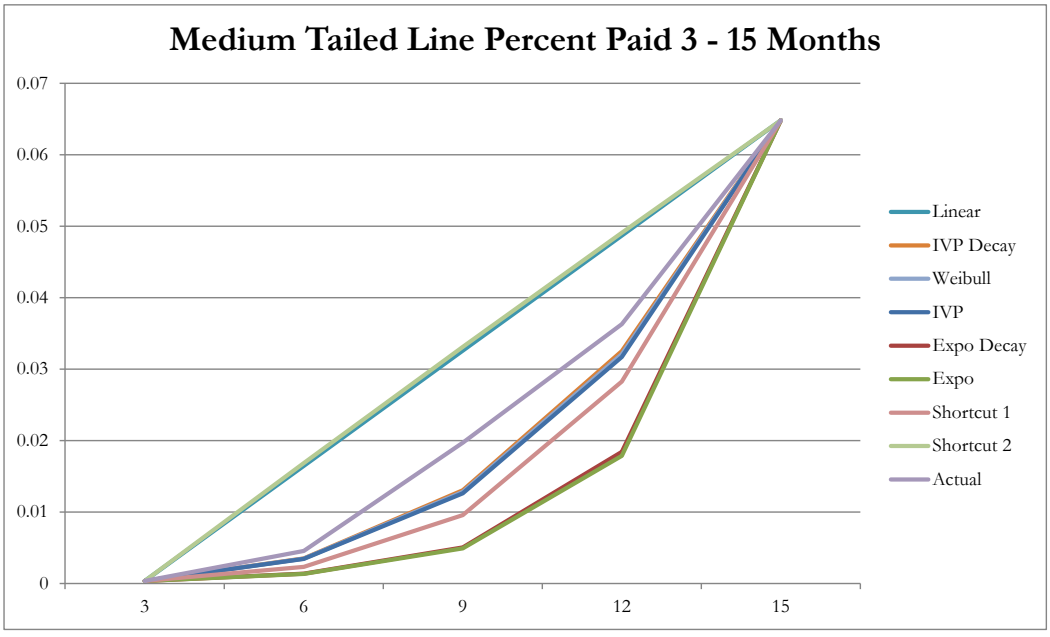
Medium Tailed Lines - Incurred Development Factors

Accident Year	Age	From Quarterly "Actual"	Linear	Regression	Weibull	IVP Whole	Expo	Expo whole	Shortcut 1	Shortcut 2
					Using factors interpolated from 9 months prior to date					
2006	102	1.006	1.007	1.006	1.006	1.007	1.006	1.007	1.006	1.007
2007	90	1.016	1.028	1.021	1.022	1.028	1.021	1.028	1.021	1.028
2008	78	1.121	1.099	1.091	1.093	1.099	1.092	1.100	1.092	1.098
2009	66	1.191	1.214	1.209	1.210	1.213	1.211	1.215	1.211	1.213
2010	54	1.335	1.347	1.338	1.341	1.345	1.342	1.349	1.341	1.345
2011	42	1.609	1.582	1.571	1.574	1.579	1.579	1.587	1.576	1.578
2012	30	2.092	2.065	2.035	2.053	2.077	2.076	2.125	2.053	2.043
2013	18	4.014	3.962	3.953	3.997	4.047	4.249	4.375	4.089	3.902
2014	6	40.322	-	-	-	-	-	-	-	-
					Using factors interpolated from 6 months prior to date					
2006	102	1.006	1.004	1.004	1.004	1.004	1.004	1.004	1.002	1.004
2007	90	1.016	1.037	1.021	1.025	1.037	1.022	1.038	1.022	1.037
2008	78	1.121	1.110	1.102	1.104	1.110	1.103	1.111	1.103	1.110
2009	66	1.191	1.209	1.202	1.204	1.208	1.204	1.210	1.203	1.208
2010	54	1.335	1.351	1.340	1.344	1.349	1.345	1.354	1.344	1.349
2011	42	1.609	1.581	1.565	1.570	1.577	1.575	1.588	1.572	1.576
2012	30	2.092	2.128	2.087	2.107	2.133	2.134	2.181	2.111	2.104
2013	18	4.014	3.953	3.932	4.000	4.076	4.268	4.426	4.081	3.855
2014	6	40.322	-	-	-	-	-	-	-	-
					Using factors interpolated from 3 months prior to date					
2006	102	1.006	1.006	1.006	1.006	1.006	1.006	1.006	1.004	1.006
2007	90	1.016	1.059	1.041	1.047	1.058	1.043	1.059	1.042	1.058
2008	78	1.121	1.123	1.119	1.120	1.122	1.120	1.123	1.119	1.122
2009	66	1.191	1.216	1.210	1.211	1.215	1.212	1.216	1.211	1.215
2010	54	1.335	1.361	1.353	1.355	1.359	1.356	1.362	1.355	1.359
2011	42	1.609	1.584	1.573	1.577	1.582	1.581	1.588	1.578	1.581
2012	30	2.092	2.184	2.152	2.168	2.186	2.186	2.218	2.170	2.167
2013	18	4.014	4.096	4.059	4.109	4.162	4.266	4.360	4.160	4.019
2014	6	40.322	31.981	35.997	36.814	37.648	54.642	56.336	46.673	30.635

Similar to our short tailed paid curve, the incurred curve for 3 – 15 months shows development higher than linear (or percent incurred lower than linear interpolation would suggest):



The actual data falls closer to linear. On a paid basis, the actual data follow the curves but only to a limited degree:



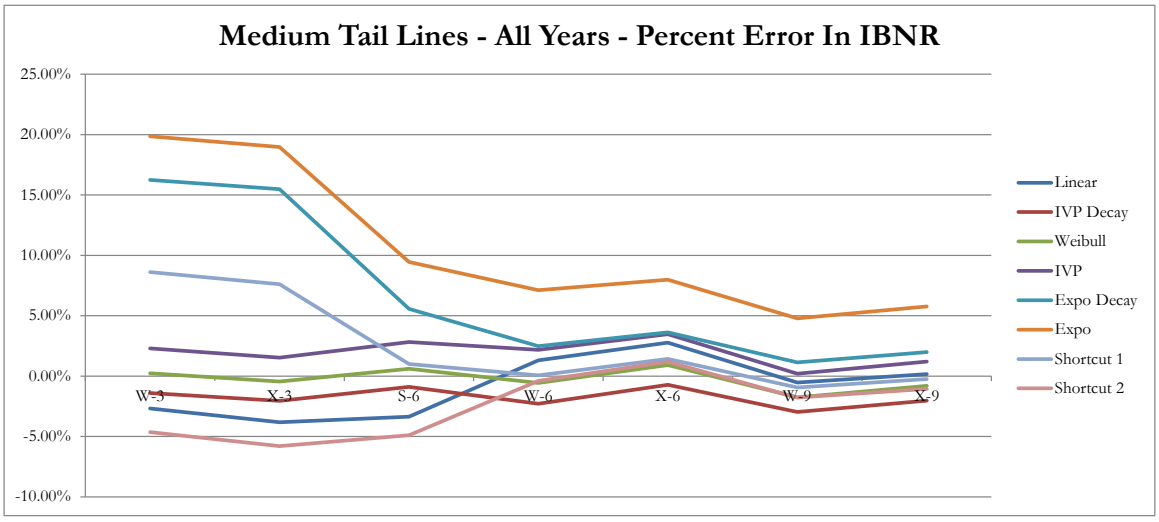
The following is a table of paid weighted average results:

Medium Tailed Lines - Paid Development Factors

Accident Year	Age	From Quarterly "Actual"	Linear	IVP Decay	Weibull	IVP	Expo Decay	Expo	Shortcut 1	Shortcut 2
					Using factors interpolated from 9 months prior to date					
2006	102	1.062	1.071	1.069	1.070	1.071	1.070	1.071	1.070	1.071
2007	90	1.104	1.120	1.115	1.117	1.120	1.116	1.121	1.116	1.120
2008	78	1.215	1.240	1.231	1.234	1.240	1.233	1.243	1.232	1.238
2009	66	1.442	1.430	1.427	1.428	1.429	1.429	1.431	1.429	1.430
2010	54	1.618	1.634	1.620	1.626	1.636	1.629	1.646	1.624	1.627
2011	42	2.214	2.246	2.232	2.242	2.254	2.256	2.281	2.242	2.233
2012	30	3.718	3.643	3.669	3.699	3.734	3.785	3.861	3.701	3.603
2013	18	10.286	9.291	10.176	10.280	10.389	11.561	11.854	10.558	9.183
2014	6	438.308	-	-	-	-	-	-	-	-
					Using factors interpolated from 6 months prior to date					
2006	102	1.062	1.070	1.065	1.066	1.069	1.065	1.070	1.065	1.070
2007	90	1.104	1.134	1.129	1.130	1.134	1.130	1.135	1.130	1.134
2008	78	1.215	1.242	1.233	1.236	1.241	1.235	1.244	1.234	1.240
2009	66	1.442	1.403	1.394	1.397	1.402	1.398	1.406	1.397	1.401
2010	54	1.618	1.614	1.605	1.608	1.612	1.612	1.619	1.609	1.612
2011	42	2.214	2.069	2.044	2.059	2.077	2.071	2.105	2.056	2.052
2012	30	3.718	3.474	3.493	3.541	3.594	3.632	3.737	3.527	3.408
2013	18	10.286	9.171	10.401	10.565	10.735	11.931	12.310	10.780	8.954
2014	6	438.308	-	-	-	-	-	-	-	-
					Using factors interpolated from 3 months prior to date					
2006	102	1.062	1.064	1.062	1.062	1.063	1.062	1.064	1.062	1.063
2007	90	1.104	1.137	1.130	1.132	1.137	1.131	1.138	1.131	1.137
2008	78	1.215	1.206	1.204	1.204	1.205	1.205	1.206	1.205	1.206
2009	66	1.442	1.403	1.388	1.394	1.403	1.393	1.408	1.390	1.399
2010	54	1.618	1.620	1.614	1.616	1.618	1.618	1.623	1.617	1.619
2011	42	2.214	2.168	2.147	2.160	2.176	2.169	2.197	2.156	2.154
2012	30	3.718	3.703	3.703	3.735	3.769	3.795	3.857	3.731	3.653
2013	18	10.286	9.616	10.309	10.422	10.537	11.135	11.345	10.561	9.403
2014	6	438.308	121.292	565.378	572.677	580.003	1,448.769	1,471.631	851.796	118.463

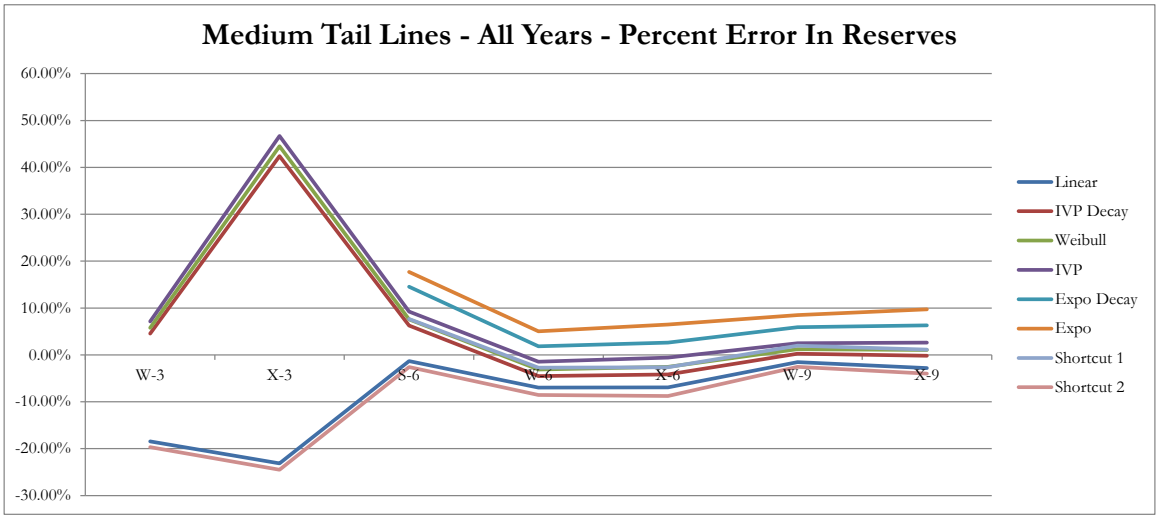
6.2.2 The Errors

The errors for all years on an incurred basis are fairly small except for the exponential curves which overestimate the liability:



Note that the reason the errors are smaller when estimated from 9 months prior is because the latest year is not estimated in the 9 month prior (or 6 month prior) scenario due to data limitations and therefore errors are smaller than the 3 month prior scenario.

Paid data shows a similar effect:



Note that values that were extraordinarily high (greater than 50% error) are shown as blank so as not to distort the graph. These types of large errors only occur with overestimation and in this case with the exponential methods.

Graphs which isolate the latter years are included in Appendix A.

6.3 Results for Long Tailed Lines of Business

The long tailed business had an average 12 month incurred development factor of approximately 15.00. Paid data was unavailable. The pattern has a tail of 1% at 126 months. It is mainly comprised of high layer property lines.

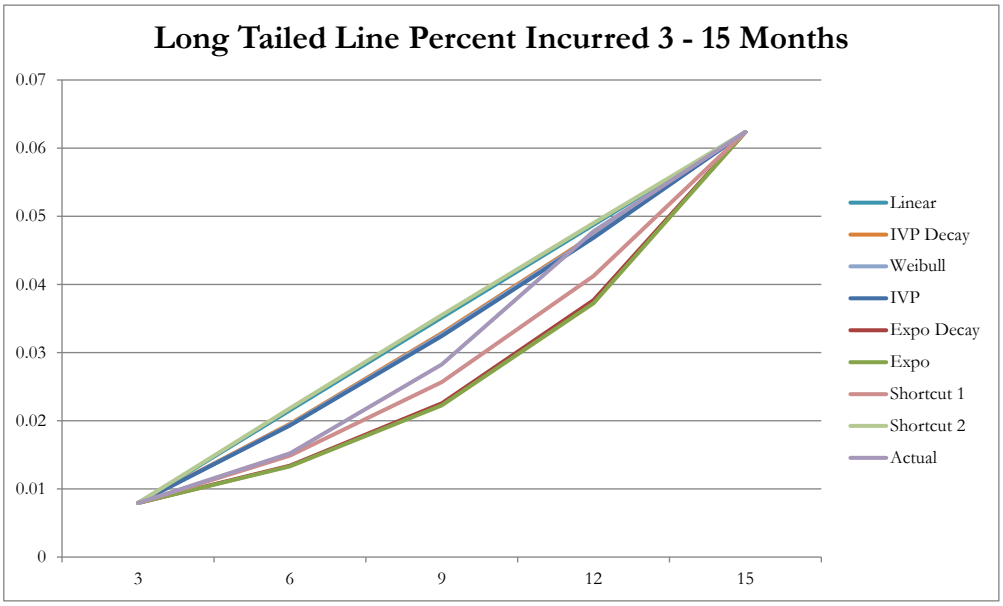
6.3.1 The Factors

Long Tailed Lines - Incurred Development Factors

Accident Year	Age	From Quarterly "Actual"	Linear	IVP Decay	Weibull	IVP	Expo Decay	Expo	Shortcut 1	Shortcut 2
				Using factors interpolated from 9 months prior to date						
2006	105	1.117	1.112	1.109	1.110	1.112	1.110	1.112	1.110	1.112
2007	93	1.203	1.196	1.189	1.191	1.196	1.191	1.198	1.190	1.195
2008	81	1.359	1.346	1.342	1.343	1.345	1.343	1.347	1.343	1.345
2009	69	1.522	1.506	1.500	1.502	1.505	1.503	1.509	1.502	1.504
2010	57	1.773	1.767	1.759	1.762	1.767	1.766	1.775	1.763	1.764
2011	45	2.305	2.276	2.264	2.273	2.285	2.286	2.309	2.273	2.265
2012	33	3.679	3.532	3.531	3.549	3.569	3.605	3.648	3.560	3.507
2013	21	7.146	6.874	6.928	6.960	6.995	7.307	7.391	7.118	6.826
2014	9	32.831	-	-	-	-	-	-	-	-
				Using factors interpolated from 6 months prior to date						
2006	105	1.117	1.110	1.107	1.108	1.110	1.108	1.110	1.108	1.110
2007	93	1.203	1.205	1.194	1.198	1.205	1.196	1.207	1.195	1.203
2008	81	1.359	1.350	1.344	1.346	1.349	1.346	1.352	1.346	1.349
2009	69	1.522	1.526	1.518	1.521	1.526	1.523	1.530	1.521	1.524
2010	57	1.773	1.797	1.786	1.790	1.796	1.795	1.805	1.791	1.792
2011	45	2.305	2.336	2.318	2.330	2.345	2.346	2.374	2.329	2.320
2012	33	3.679	3.615	3.610	3.634	3.662	3.703	3.757	3.647	3.578
2013	21	7.146	6.923	6.996	7.042	7.092	7.430	7.533	7.212	6.846
2014	9	32.831	28.282	33.502	33.905	34.318	53.458	54.753	43.627	27.802
				Using factors interpolated from 3 months prior to date						
2006	105	1.117	1.108	1.105	1.106	1.108	1.105	1.109	1.105	1.108
2007	93	1.203	1.198	1.192	1.194	1.197	1.193	1.199	1.193	1.197
2008	81	1.359	1.334	1.328	1.330	1.334	1.330	1.336	1.329	1.333
2009	69	1.522	1.510	1.504	1.507	1.510	1.508	1.513	1.506	1.509
2010	57	1.773	1.764	1.756	1.759	1.763	1.762	1.768	1.759	1.761
2011	45	2.305	2.290	2.275	2.283	2.293	2.293	2.310	2.283	2.280
2012	33	3.679	3.635	3.630	3.654	3.679	3.701	3.746	3.655	3.599
2013	21	7.146	7.036	7.089	7.126	7.166	7.373	7.443	7.231	6.968
2014	9	32.831	30.369	33.421	33.700	33.981	40.888	41.407	37.789	29.838

The incurred factors based on weighted averages were as follows:

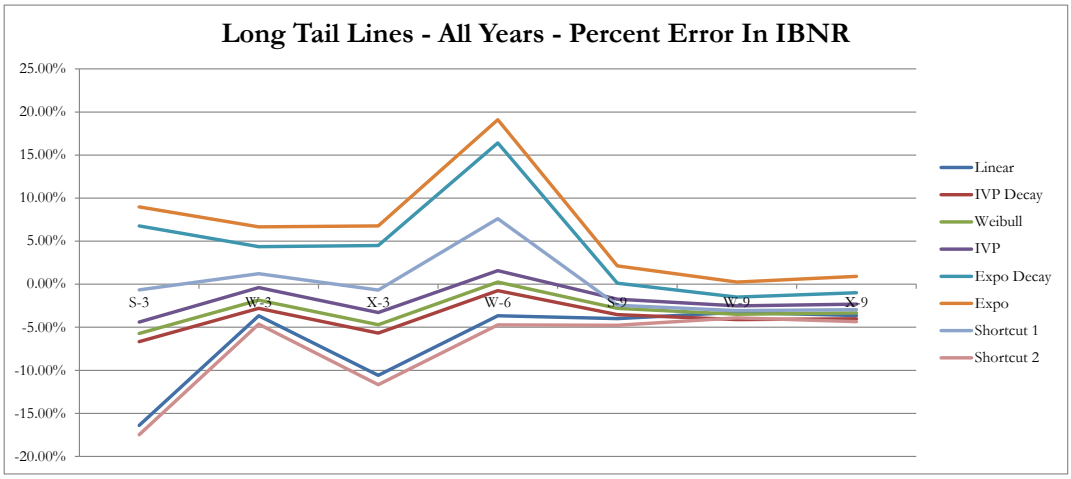
As seen below, the curve of percent reported between 3 and 15 months shows a reverse arc with



actual results falling between this arc and the linear method:

6.3.2 The Errors

The graph of results shows that Shortcut 1 behaves closest to actual data:



6.4 Results for Very Long Tailed Lines of Business

The very long tailed business had an average 12 month incurred development factor greater than 20.00 and an average paid development factor nearing 90.00. The age to age paid factors are around 3% at 120 months. This data set is mainly comprised of casualty lines.

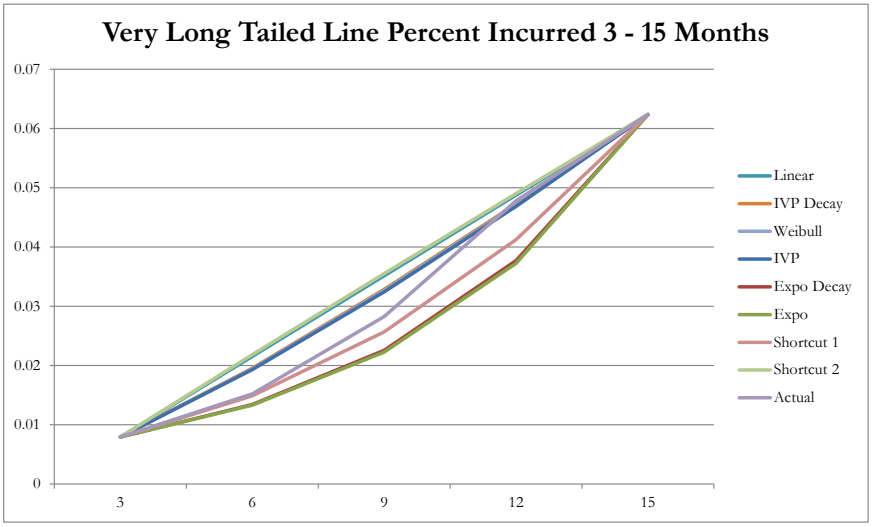
6.4.1 The Factors

The factors for weighted averages on an incurred basis were as follows:

Very Long Tailed Lines - Incurred Development Factors

Accident Year	Age	From Quarterly "Actual"	Linear	IVP Decay	Weibull	IVP	Expo Decay	Expo	Shortcut 1	Shortcut 2
					Using factors interpolated from 9 months prior to date					
2006	105	1.064	1.029	1.020	1.022	1.029	1.020	1.030	1.020	1.029
2007	93	1.138	1.125	1.110	1.114	1.125	1.111	1.127	1.110	1.123
2008	81	1.313	1.310	1.305	1.307	1.310	1.307	1.312	1.307	1.309
2009	69	1.479	1.476	1.471	1.473	1.475	1.474	1.478	1.473	1.474
2010	57	1.753	1.714	1.703	1.708	1.715	1.712	1.724	1.707	1.709
2011	45	2.388	2.277	2.264	2.274	2.286	2.286	2.309	2.273	2.265
2012	33	3.884	3.536	3.535	3.553	3.574	3.611	3.654	3.565	3.510
2013	21	7.367	7.130	7.363	7.416	7.473	7.900	8.040	7.559	7.060
2014	9	47.131	-	-	-	-	-	-	-	-
					Using factors interpolated from 6 months prior to date					
2006	105	1.064	1.040	1.023	1.027	1.040	1.023	1.041	1.023	1.040
2007	93	1.138	1.134	1.121	1.125	1.134	1.123	1.136	1.122	1.132
2008	81	1.313	1.294	1.281	1.286	1.295	1.284	1.298	1.282	1.291
2009	69	1.479	1.480	1.475	1.476	1.479	1.478	1.482	1.477	1.479
2010	57	1.753	1.749	1.734	1.741	1.750	1.745	1.761	1.739	1.742
2011	45	2.388	2.342	2.323	2.336	2.352	2.352	2.380	2.335	2.326
2012	33	3.884	3.653	3.651	3.677	3.706	3.748	3.805	3.687	3.614
2013	21	7.367	7.713	8.124	8.211	8.302	8.830	9.023	8.359	7.580
2014	9	47.131	37.923	40.563	40.828	41.097	59.057	59.851	51.933	37.528
					Using factors interpolated from 3 months prior to date					
2006	105	1.064	1.051	1.051	1.051	1.051	1.051	1.051	1.051	1.050
2007	93	1.138	1.127	1.121	1.123	1.126	1.122	1.127	1.122	1.126
2008	81	1.313	1.276	1.264	1.268	1.276	1.267	1.278	1.265	1.274
2009	69	1.479	1.454	1.448	1.450	1.453	1.451	1.456	1.450	1.452
2010	57	1.753	1.755	1.743	1.749	1.755	1.751	1.763	1.747	1.750
2011	45	2.388	2.332	2.317	2.326	2.336	2.336	2.355	2.325	2.321
2012	33	3.884	3.805	3.809	3.835	3.863	3.887	3.939	3.833	3.764
2013	21	7.367	8.406	8.676	8.739	8.804	9.127	9.245	8.844	8.286
2014	9	47.131	40.685	45.744	46.044	46.345	56.352	56.912	52.049	40.113

As seen below, the curve of percent reported between 3 and 15 months shows a reverse arc with actual results falling between this arc and the linear method:



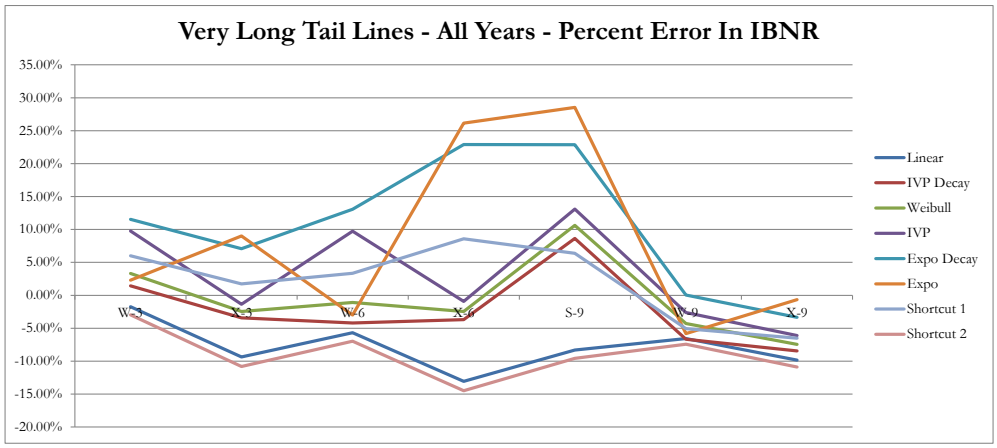
Paid Factors are as follows:

Very Long Tailed Lines - Paid Development Factors

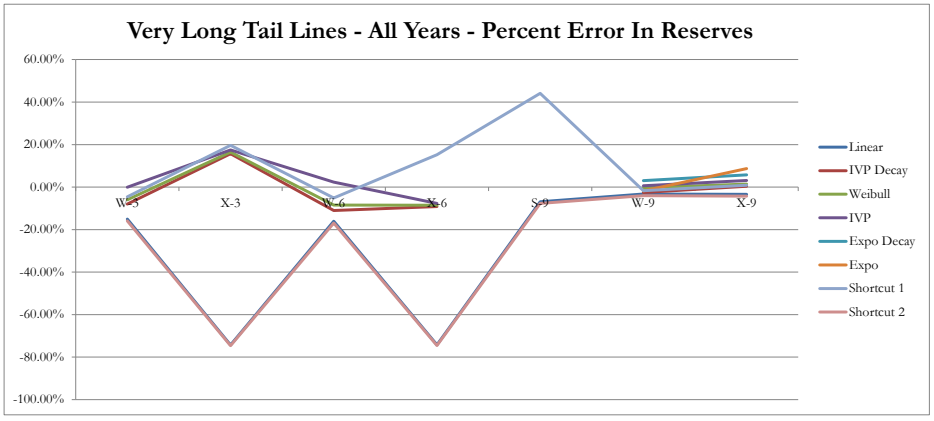
Accident Year	Age	From Quarterly "Actual"	Linear	IVP Decay	Weibull	IVP	Expo Decay	Expo	Shortcut 1	Shortcut 2
				Using factors interpolated from 9 months prior to date						
2006	105	1.105	1.123	1.116	1.118	1.123	1.117	1.124	1.116	1.122
2007	93	1.225	1.244	1.239	1.241	1.243	1.241	1.245	1.240	1.243
2008	81	1.371	1.397	1.390	1.393	1.398	1.393	1.401	1.391	1.395
2009	69	1.684	1.694	1.683	1.689	1.697	1.690	1.705	1.686	1.688
2010	57	2.265	2.280	2.272	2.282	2.293	2.289	2.312	2.277	2.269
2011	45	3.561	3.497	3.512	3.528	3.547	3.563	3.601	3.522	3.475
2012	33	6.944	6.549	6.682	6.711	6.743	6.883	6.952	6.735	6.507
2013	21	23.245	16.879	18.972	19.078	19.186	21.128	21.404	19.488	16.771
2014	9	369.532	-	-	-	-	-	-	-	-
				Using factors interpolated from 6 months prior to date						
2006	105	1.105	1.126	1.118	1.120	1.126	1.119	1.127	1.119	1.126
2007	93	1.225	1.231	1.225	1.227	1.231	1.227	1.232	1.226	1.230
2008	81	1.371	1.386	1.376	1.380	1.386	1.380	1.390	1.378	1.383
2009	69	1.684	1.670	1.656	1.664	1.673	1.665	1.682	1.660	1.664
2010	57	2.265	2.260	2.246	2.261	2.280	2.269	2.304	2.250	2.241
2011	45	3.561	3.489	3.496	3.516	3.538	3.554	3.597	3.511	3.460
2012	33	6.944	6.343	6.512	6.556	6.602	6.752	6.845	6.570	6.273
2013	21	23.245	17.116	20.421	20.595	20.772	23.056	23.445	20.948	16.886
2014	9	369.532	139.914	217.088	217.822	218.561	411.127	413.753	315.033	139.326
				Using factors interpolated from 3 months prior to date						
2006	105	1.105	1.108	1.095	1.099	1.108	1.096	1.109	1.096	1.107
2007	93	1.225	1.229	1.221	1.224	1.228	1.223	1.230	1.222	1.227
2008	81	1.371	1.372	1.366	1.368	1.371	1.368	1.373	1.368	1.370
2009	69	1.684	1.664	1.651	1.658	1.666	1.658	1.672	1.654	1.658
2010	57	2.265	2.272	2.259	2.271	2.284	2.276	2.300	2.263	2.258
2011	45	3.561	3.496	3.499	3.516	3.535	3.542	3.576	3.509	3.470
2012	33	6.944	6.612	6.752	6.790	6.829	6.927	6.999	6.789	6.542
2013	21	23.245	19.497	22.361	22.504	22.649	24.103	24.374	22.717	19.208
2014	9	369.532	157.185	272.604	273.555	274.508	382.104	384.055	326.841	155.917

6.4.2 The Errors

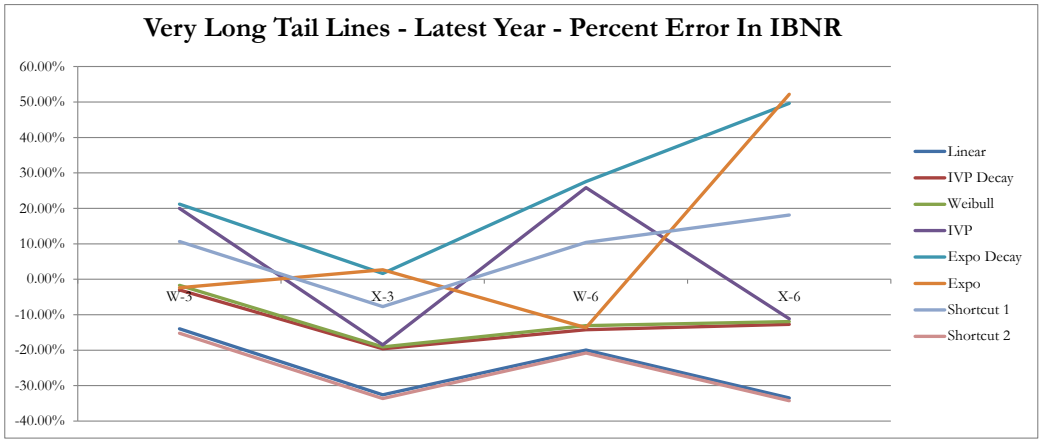
The error pattern is similar to long tailed lines for incurred losses:



Paid results are more erratic particularly for simple averages, which do not seem to perform well with interpolation:



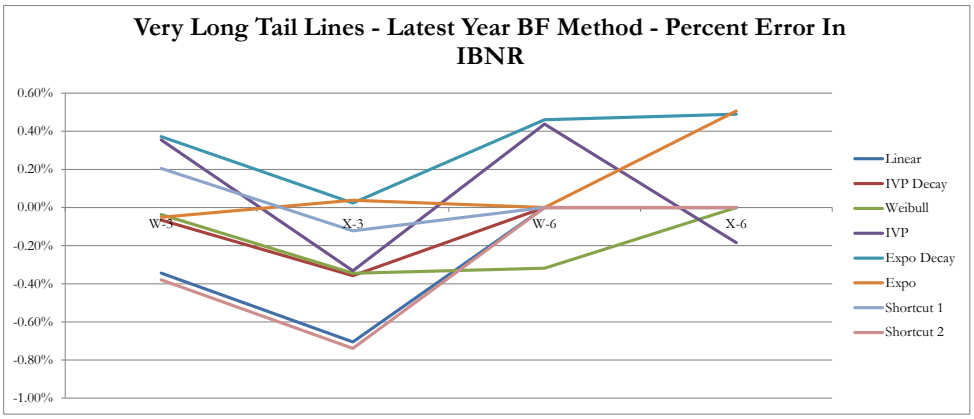
A graph of the latest accident year shows that only Shortcut 1 performs reasonably well in terms of being close to zero error and not underestimating amounts:

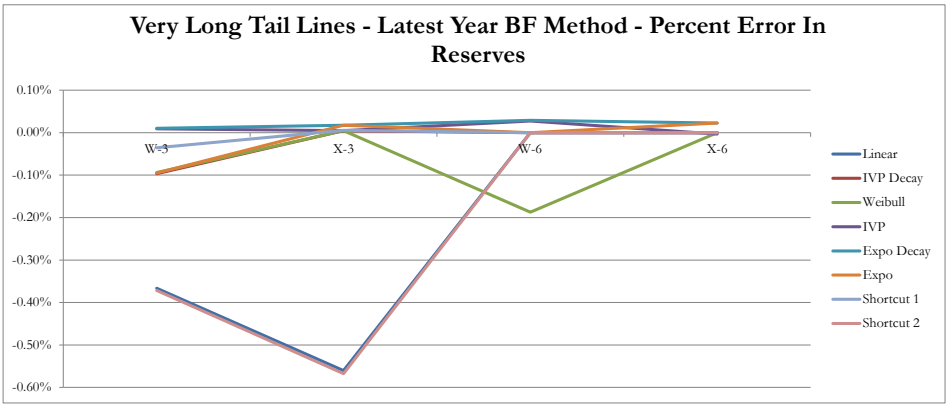


6.5 The Impact of using Exposure Based Methods

It is well known that many actuaries use exposure based methods, such as Bornhuetter Ferguson (BF) in the more recent accident years. To the extent that this is true, the magnitude of errors might be less significant. To test this effect, I used an all year initial expected loss ratio for my data based on a Cape Cod calculation and tested the differences between using this method with the quarterly data versus using it with the interpolated annual data. Note that I used the same initial expected loss ratio for both quarterly and annual data in order to isolate the reserves changes that would be caused by interpolation alone. It is assumed that the practitioner has a reasonable initial expected loss ratio estimate that does not rely on interpolation.

As can be expected the use of the BF method has very little impact on the shorted tailed and medium tailed lines of business. Percent errors decreased very little and sometimes even increased for all years since the latest year had the most impact. I tested the BF on the very long tailed lines of business since there was no exposure data for the long tailed line. The BF method reduced the errors to nearly zero for both paid and incurred data:





7. TESTING OF METHODS AND RESULTS FOR EXPTRAPOLATION

Using the annual triangles as described above, I extrapolated factors from the 6, 9 and 12 month triangles using the earliest CDF. Each one was used to estimate earlier quarters such that the 12 month factor was used to estimate a 3, 6 and 9 month factors and so forth.

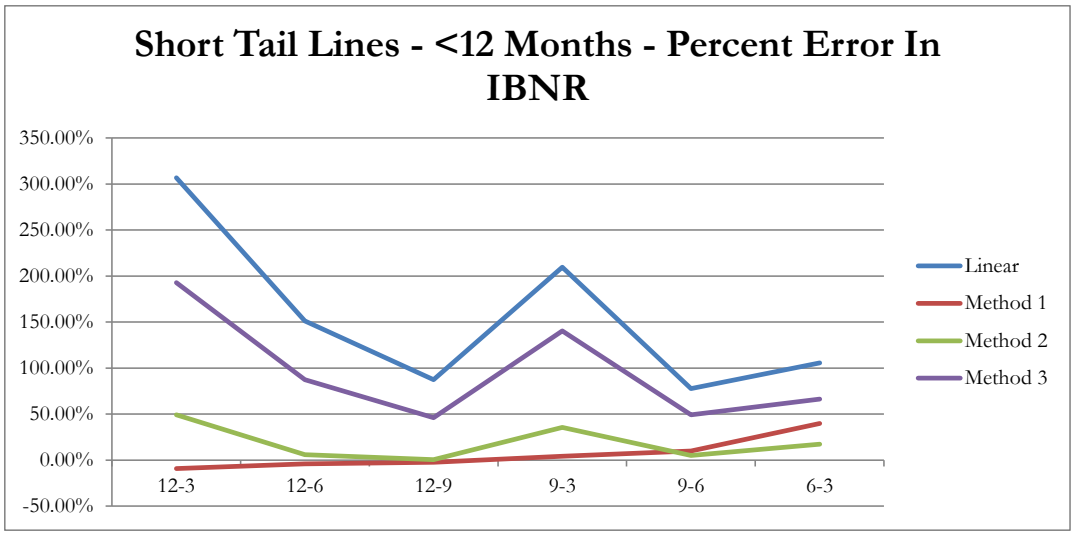
I applied the interpolation methods as described in Section 4. Then I projected the quarterly results by accident year using the most recent data and quarterly factors. Each result was then compared to the extrapolated estimate. The errors were also calculated on a paid and incurred basis and for all three averages.

Error was measured in terms of IBNR for the incurred triangles and total reserves for the paid triangles. The percent error was calculated as a percent of total IBNR or reserve. Therefore percent error for paid losses would equal:

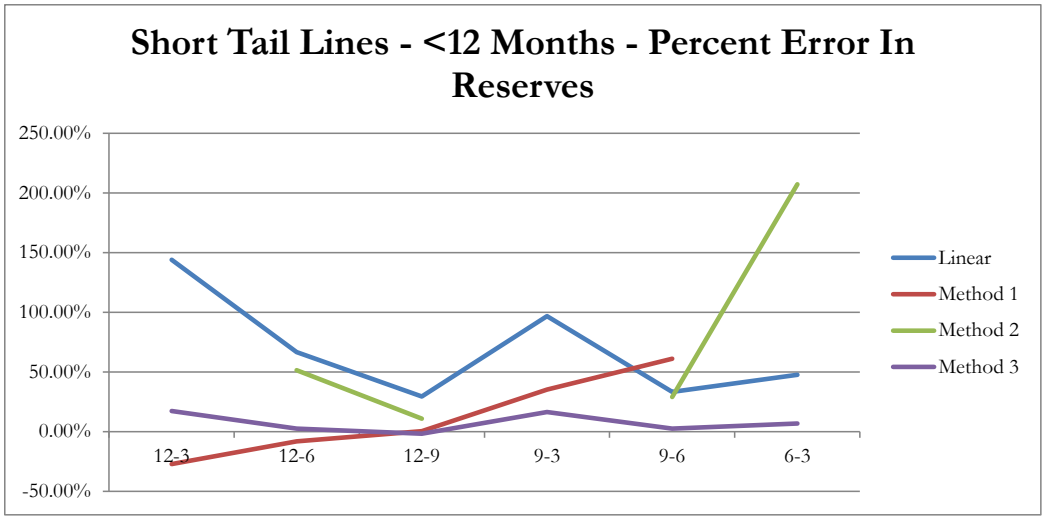
[Ultimate losses derived from Extrapolated method – Ultimate losses derived from quarterly triangle factors] / [Ultimate losses derived from quarterly triangle factors – Paid losses at latest evaluation].

7.1 Results for Short Tailed Lines of Business

Results were extremely volatile, particularly for the Linear method and Method 3 (which tends toward linear). Note in the graph below "12-3" indicates a 3 month factor estimated from a 12 month factor and "6-3" indicates a 3 month factor estimated from a 6 month factor and so on.



The 12-9 factors in general tend to be more accurate across all methods. Paid results are more volatile (results over +500% not shown) but once again the 12-9 gives better results than other maturities:

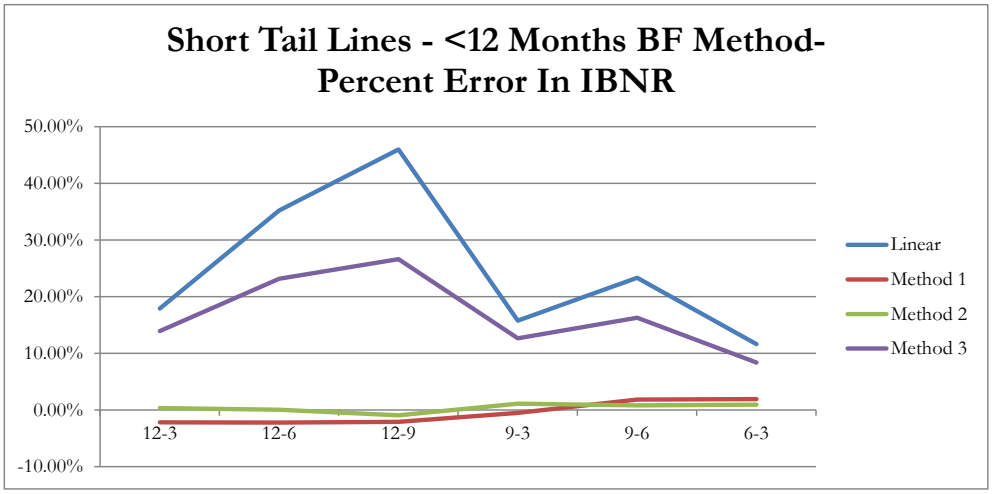


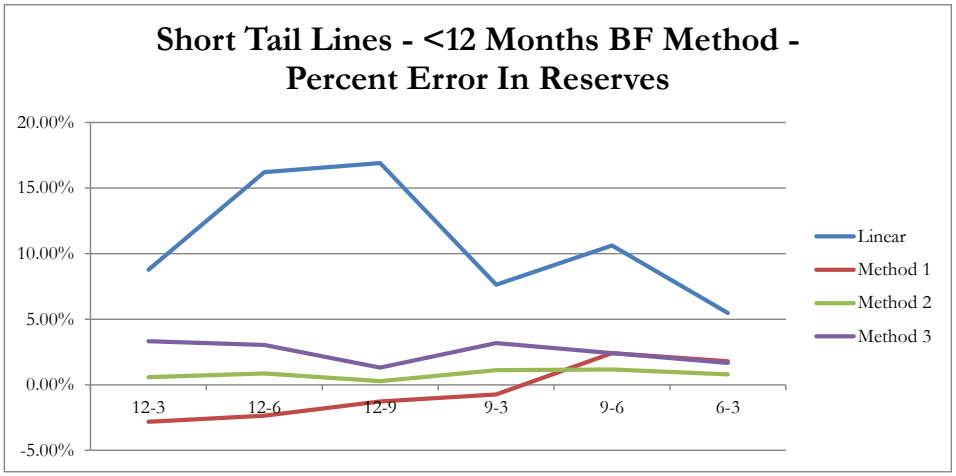
The overall factors for weighted averages looked as follows:

Short Tailed Lines

Accident Month	*Paid Development Factors* From Quarterly "Actual"	Linear	Method 1	Method 2	Method 3	*Incurred Development Factors* From Quarterly "Actual"	Linear	Method 1	Method 2	Method 3
					Using factors interpolated from 12 months					
9	2.163	2.508	2.168	2.290	2.143	1.497	1.931	1.486	1.499	1.726
6	3.523	5.205	3.321	4.827	3.588	2.319	4.314	2.264	2.398	3.470
3	8.458	19.204	6.432	127.700	9.753	4.975	17.168	4.615	6.930	12.642
					Using factors interpolated from 9 months					
6	3.523	4.368	5.065	4.261	3.588	2.319	3.342	2.448	2.386	2.966
3	8.458	15.684	11.087	62.655	9.689	4.975	13.310	5.138	6.390	10.550
					Using factors interpolated from 6 months					
3	8.458	12.008	86.524	23.925	8.977	4.975	9.179	6.562	5.661	7.614

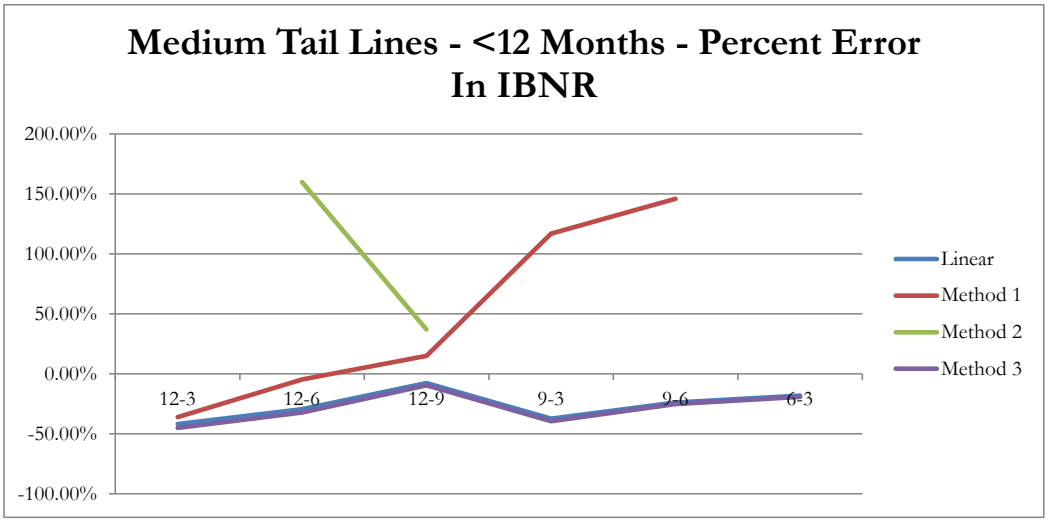
Using the BF approach as described in Section 6, incurred and paid results become much more stable with Methods 1 and 2 having error close to zero:

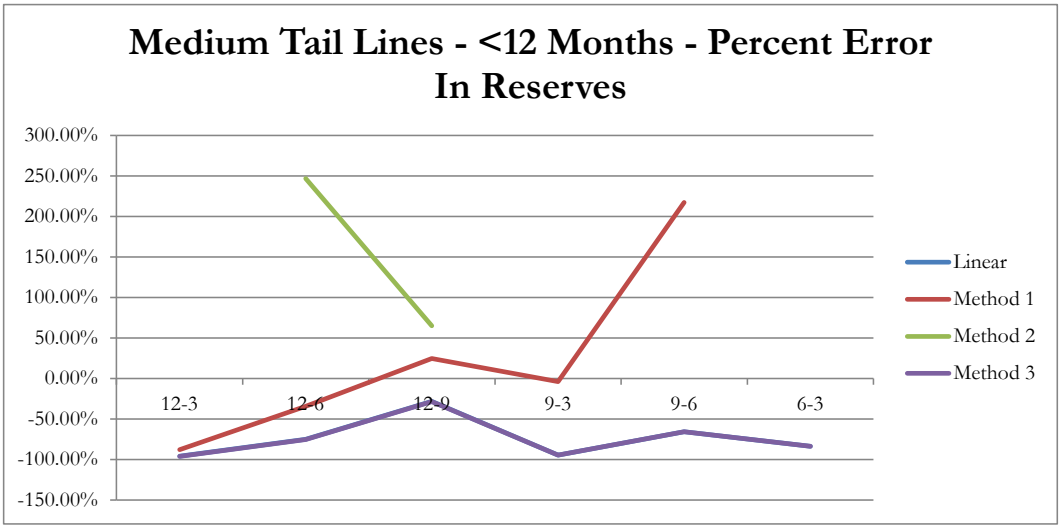




7.2 Results for Medium Tailed Lines of Business

Results are more volatile on paid and incurred bases as the development factors increase. In this case the Linear method and Method 3 perform better but underestimate the liability.





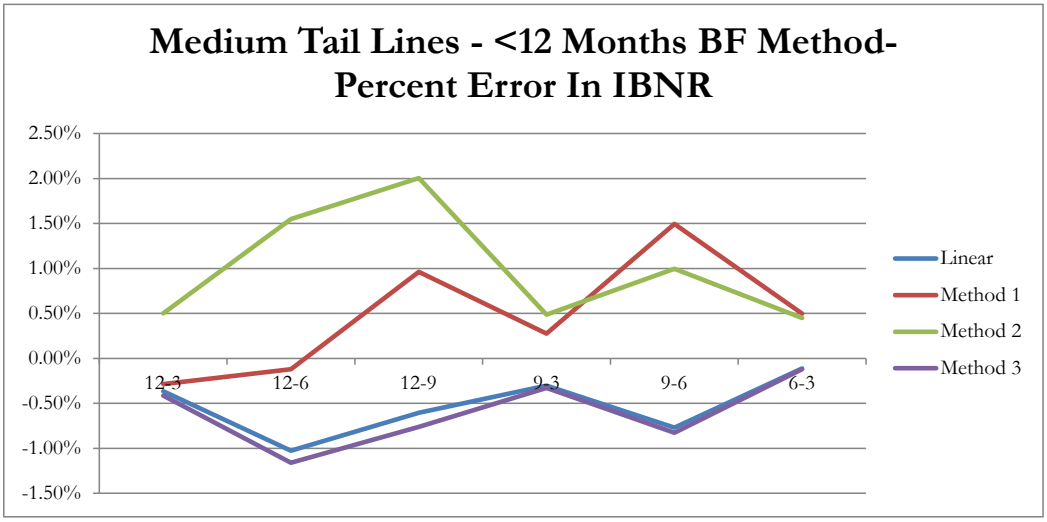
Once again, the 12-9 factors generally tend to be more accurate across all methods.

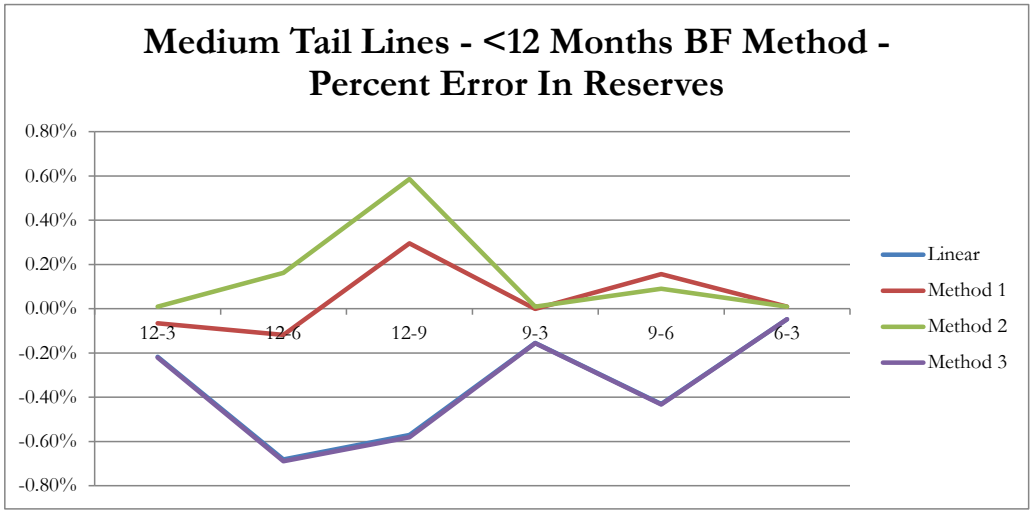
The overall factors looked as follows:

Medium Tailed Lines

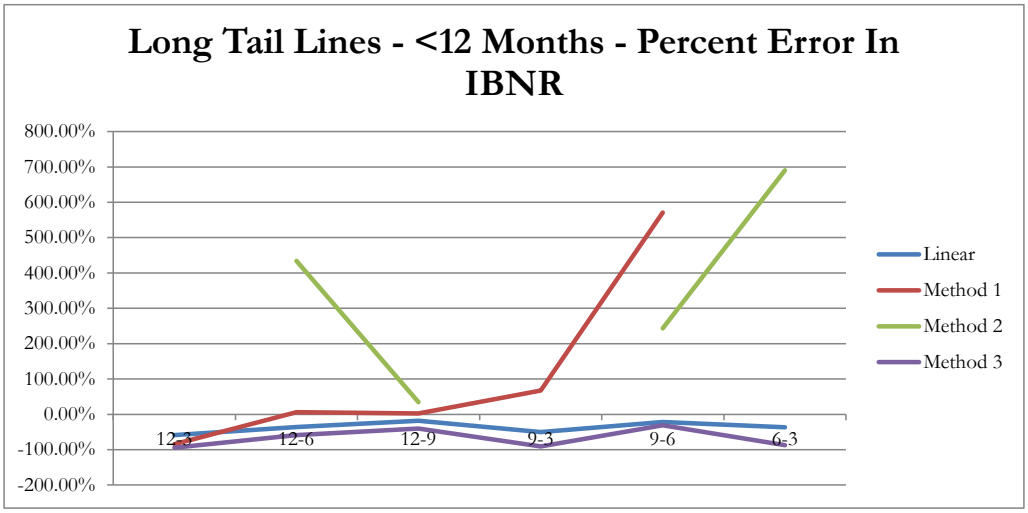
Accident Month	***Paid Development Factors*** From Quarterly "Actual"	Linear	Method 1	Method 2	Method 3	***Incurred Development Factors*** From Quarterly "Actual"	Linear	Method 1	Method 2	Method 3
				Using factors interpolated from 12 months						
9	67.597	48.971	84.143	110.923	48.746	13.753	12.768	15.676	18.475	12.536
6	438.308	110.185	289.149	1,517.584	109.175	40.322	28.727	38.493	103.157	27.690
3	10,674.545	440.739	1,324.852	2,303,061.379	434.693	196.797	114.909	126.029	10,641.442	108.722
				Using factors interpolated from 9 months						
6	438.308	152.094	1,388.929	721.963	151.591	40.322	30.944	97.749	66.253	30.429
3	10,674.545	608.375	10,281.108	521,230.571	604.357	196.797	123.774	425.557	4,389.411	119.684
				Using factors interpolated from 6 months						
3	10,674.545	1,753.233	2,844,029.625	192,114.122	1,751.231	196.797	161.289	7,300.351	1,625.874	159.263

Using the BF approach as described in Section 6, incurred and paid results become much more stable with all methods having error close to zero:





7.3 Results for Long Tailed Lines of Business



Results are very similar to medium tailed lines of business with a higher degree of error.

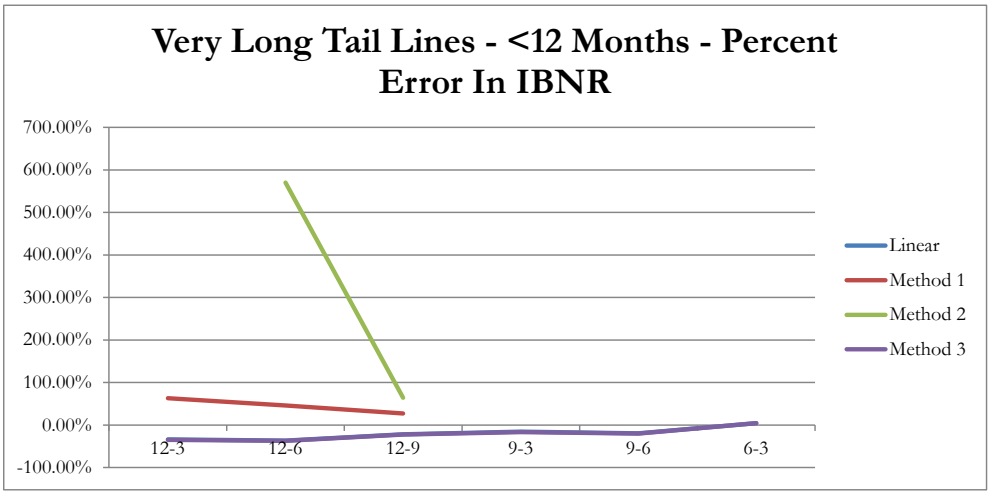
The factors for weighted averages looked as follows:

Long Tailed Lines
Incurred Development Factors

Accident Month	From Quarterly "Actual"	Linear	Method 1	Method 2	Method 3
		Using factors interpolated from 12 months			
9	32.831	27.002	39.980	50.154	26.776
6	94.059	60.756	118.391	461.404	59.739
3	590.278	243.022	467.445	212,894.074	236.937
		Using factors interpolated from 9 months			
6	94.059	73.869	416.804	244.366	73.363
3	590.278	295.476	2,425.167	59,714.879	291.439
		Using factors interpolated from 6 months			
3	590.278	376.236	60,671.471	8,847.073	374.225

7.4 Results for Very Long Tailed Lines of Business

Results are very similar to medium and long tailed lines of business with a higher degree of error. In fact, the paid graph has very few points. All Graphs are shown in Appendix A.

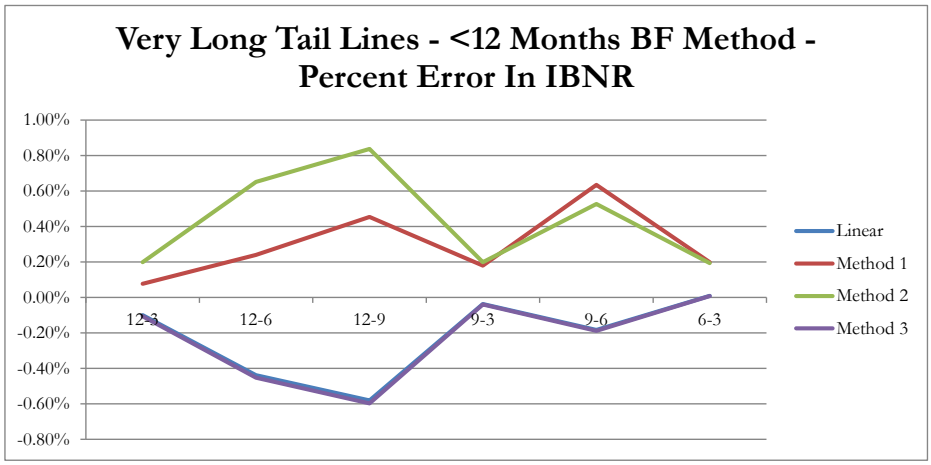


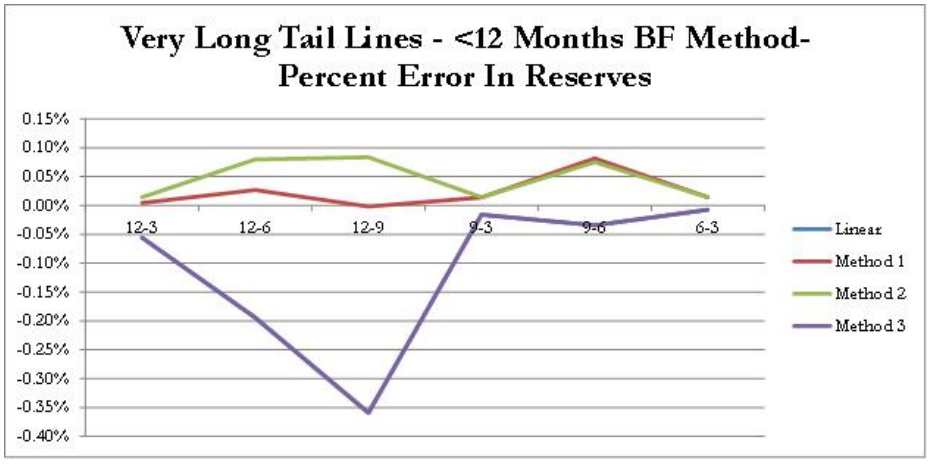
Factors for weighted averages were as follows:

Very Long Tailed Lines

Accident Month	***Paid Development Factors*** From Quarterly "Actual"	Linear	Method 1	Method 2	Method 3	***Incurred Development Factors*** From Quarterly "Actual"	Linear	Method 1	Method 2	Method 3
				Using factors interpolated from 12 months						
9	369.532	159.124	367.080	533.846	158.901	47.131	37.181	59.634	76.829	36.955
6	1,165.974	358.029	1,693.623	16,023.066	357.026	131.394	83.657	191.290	874.810	82.645
3	7,125.067	1,432.115	10,418.643	256,738,649.772	1,426.101	502.602	334.628	818.151	765,292.557	328.566
				Using factors interpolated from 9 months						
6	1,165.974	831.447	23,562.370	9,227.840	830.947	131.394	106.045	761.457	420.324	105.541
3	7,125.067	3,325.789	307,246.366	85,153,035.921	3,321.785	502.602	424.180	4,998.025	176,671.942	420.155
				Using factors interpolated from 6 months						
3	7,125.067	4,663.895	32,825,164.845	1,359,494.738	4,661.894	502.602	525.578	139,935.418	17,264.505	523.570

Once again, BF methods render the errors in development factors immaterial:





8. SEASONAL ADJUSTMENT METHOD

This section deals with the situation where the actuary has specific knowledge of company practices which may change the view of how interpolation should occur. For this example we will assume that the company has unusually high payments during the fourth quarter of every year due to extra efforts to close claims in that quarter. To start we will assume that the company knows that payments are 50% higher in Q4 than they would be without such efforts. An alternative assumption will be addressed following the main scenario.

Any of our interpolation methods can be used and adapted for this situation. In this example, I use Shortcut 1. I start by interpolating factors to each quarter as usual, but I extend the calculations to each quarter of the year even though I am most interested in the CDF at fourth quarter after the unusually high payments since I don't want ultimate losses to be overstated or fluctuate wildly from quarter to quarter.

Using the selected interpolation method I set up a table (more detail given in Appendix B):

			3Q 2014			4Q 2014		
Accident Year	Maturity in Months	Paid CDF 2nd Quarter	Maturity	Interpolated Paid CDF	Incremental Percent Paid	Maturity	Interpolated Paid CDF	Incremental Percent Paid
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
					1/(4) - 1/(2)			1/(7) - 1/(4)
2004	126	1.001	129	1.000	0.001	132	1.000	-
2005	114	1.003	117	1.002	0.001	120	1.002	0.001
2006	102	1.008	105	1.006	0.002	108	1.005	0.001
2007	90	1.018	93	1.015	0.003	96	1.012	0.003
2008	78	1.038	81	1.032	0.006	84	1.026	0.005
2009	66	1.064	69	1.057	0.007	72	1.050	0.006
2010	54	1.108	57	1.095	0.011	60	1.083	0.010
2011	42	1.191	45	1.165	0.019	48	1.143	0.017
2012	30	1.334	33	1.290	0.026	36	1.252	0.024
2013	18	2.068	21	1.780	0.078	24	1.580	0.071
2014	6	2.843	9	2.596	0.033	12	2.390	0.033
2014 Full Year	6	5.686	9	3.462	0.113		2.390	0.130

For the next step, I calculate the percentage of total yearly payments assumed paid in each quarter by the selected interpolation method. I use the relative values from columns (5), (8), (11) and (14) above.

	1Q 2015			2Q 2015		
Accident Year	Maturity	Interpolated Paid CDF	Incremental Percent Paid	Maturity	Interpolated Paid CDF	Incremental Percent Paid
	(9)	(10)	(11)	(12)	(13)	(14)
			1/(10) - 1/(7)			1/(13) - 1/(10)
2004	135	1.000	-	138	1.000	-
2005	123	1.001	0.000	126	1.001	0.000
2006	111	1.004	0.001	114	1.003	0.001
2007	99	1.010	0.002	102	1.008	0.002
2008	87	1.022	0.004	90	1.018	0.004
2009	75	1.044	0.005	78	1.038	0.005
2010	63	1.073	0.009	66	1.064	0.008
2011	51	1.124	0.015	54	1.108	0.013
2012	39	1.219	0.021	42	1.191	0.019
2013	27	1.438	0.063	30	1.334	0.054
2014	15	2.216	0.033	18	2.068	0.032
2014 Full Year		2.216	0.033		2.068	0.032

I then "restate" these percentages by assuming that the 4th quarter will be 50% higher than what is shown above. The other three quarters are renormalized to the new remainder. For example

	Percent of Year Paid in				
Accident Year	3Q2014	4Q 2014	1Q 2015	2Q 2015	Total
	(15)	(16)	(17)	(18)	(19)
2004					
2005	36.0%	27.4%	20.8%	15.8%	100.0%
2006	34.7%	27.2%	21.3%	16.7%	100.0%
2007	32.9%	27.0%	22.1%	18.1%	100.0%
2008	32.1%	26.8%	22.4%	18.7%	100.0%
2009	29.6%	26.3%	23.3%	20.7%	100.0%
2010	29.4%	26.3%	23.4%	20.9%	100.0%
2011	29.4%	26.3%	23.4%	20.8%	100.0%
2012	28.7%	26.1%	23.7%	21.5%	100.0%
2013	29.4%	26.7%	23.6%	20.3%	100.0%
2014	25.3%	25.2%	24.9%	24.5%	100.0%

Column 15 would be restated as follows.

(15)/ [(15)+(17)+(18)]*[1-(16)*1.5]

The restated percent paid for Q4 would be simply:

(16) *1.5.

Finally, the restated CDF for Q4 is given as:

1/[Percent paid at 2Q 2014 +sum(restated percent in 3Q and 4Q 2014)*[expected paid for full calendar year]]

This method can be adapted for other seasonal situations using paid or incurred losses. In addition, in the situation where the percent increase is a rough estimate, the company's own Q4 data can be used to calibrate a percentage that fits. More detail is provided in Appendix B.

However, if the knowledge about fourth quarter payments reflects a percentage higher than the payments *in other quarters* as opposed to simply a percentage higher than *it would be otherwise*, the last restated percent paid for Q4 should be given as:

(16)/[(16)*1.5 + 1 – (16)] x (1.5)

In reality, the percentage increase and the choice of which assumption is more appropriate will be very hard to ascertain. However, using actual emergence to calibrate the adjustment over time in the absence of full quarterly triangles should add more value to the interpolated factors.

9. CONCLUSIONS

The appropriateness and accuracy of various interpolation and extrapolation methods varies greatly with the development characteristics of the line of business. Sophisticated methods don't seem to provide much advantage over simple shortcuts. For short tailed lines or lines with development factors less than 2.00 at 12 months, Shortcut 2 seems to perform relatively well, whereas Shortcut 1 seems to perform better on paid data or once development is greater than 2.00 at 12 months. Shortcut 1 also seems to perform well once the second year of development is reached. Exponential curves seem to regularly overstate reserves by large amounts.

Weighted average development factors also seem to work much better and are not prone to unusual swings which may distort interpolation methods. However, in practice development factors are often selected judgmentally so this may be hard to follow when interpolating.

Extrapolated values, especially for long tailed lines, are predictability overstated and distorted. However the BF method seems to mitigate this risk almost entirely. For the Development method, Methods 1 and 2 seem to perform the best without understating reserves on shorter tailed lines whereas Method 3 performs well on longer tailed lines.

Since the data used is not exhaustive but more a sampling of typical quarterly triangles, the practitioner can use this paper to decide how each of these formulae are applicable to the underlying characteristics of individual company data.

Acknowledgment

The author acknowledges Stephanie Celona for extensive spreadsheet work and fantastic editorial review.

Supplementary Material

The Appendices to this paper and a practical tool are available electronically at the CAS website at (fill in later). The practical tool demonstrates interpolation and extrapolation methods as well as the seasonal adjustment method.

Appendix A

The graphs in the body of the paper as well as some additional graphs are included in this appendix.

Appendix B

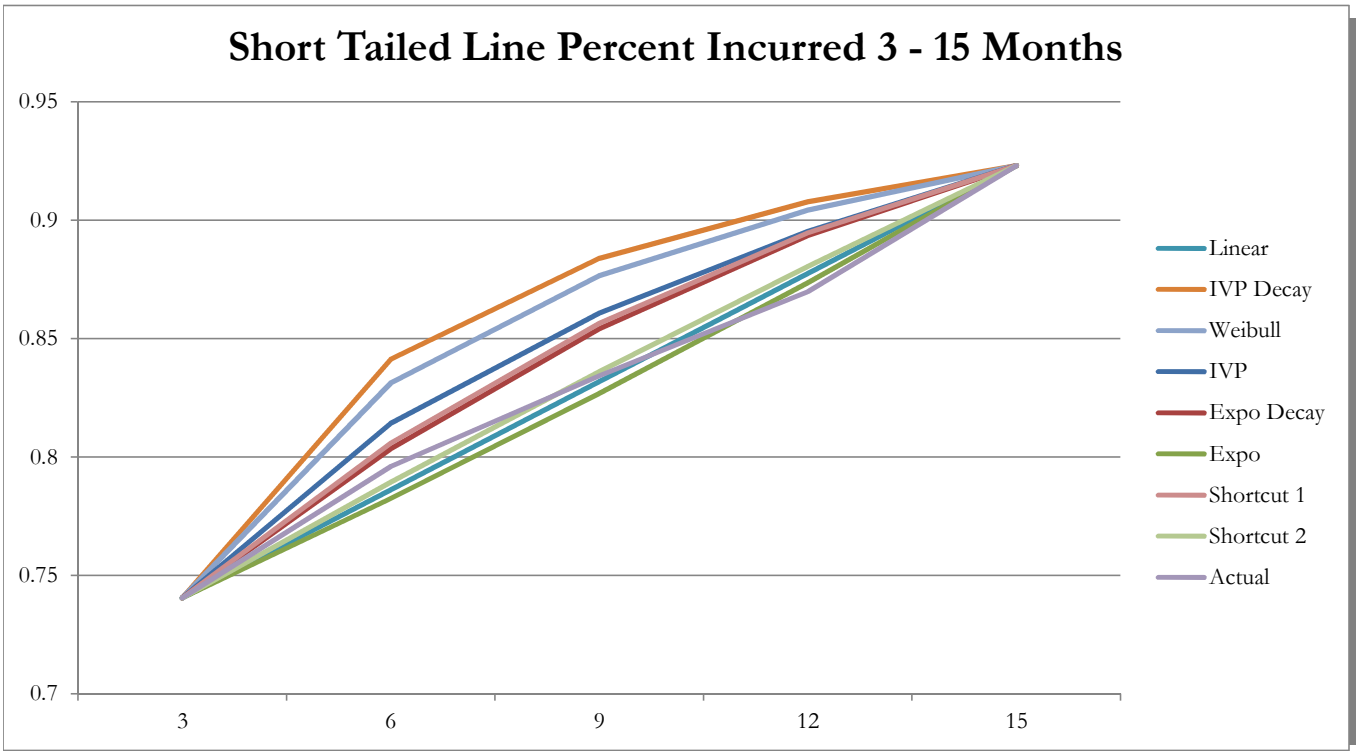
Details of the Seasonal Adjustment Method are provided here.

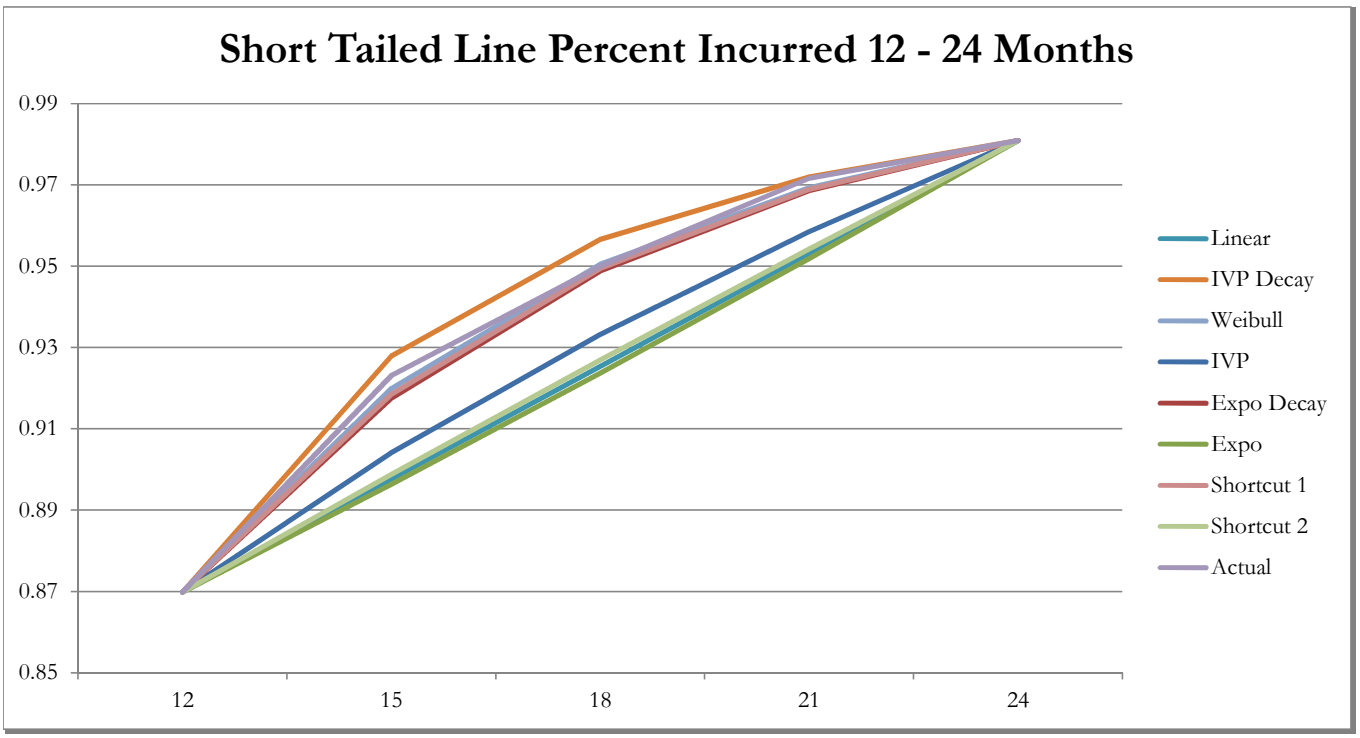
Biography of the Author

Lynne Bloom, FCAS, MAAA, is a Director at PwC in Philadelphia, PA. She has a B.B.A. in Finance from the Wharton Business School at the University of Pennsylvania. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. Lynne is the chairman of the CAS Research Oversight Committee and Vice President of CAMAR.

Short Tail Line Incurred Graphs

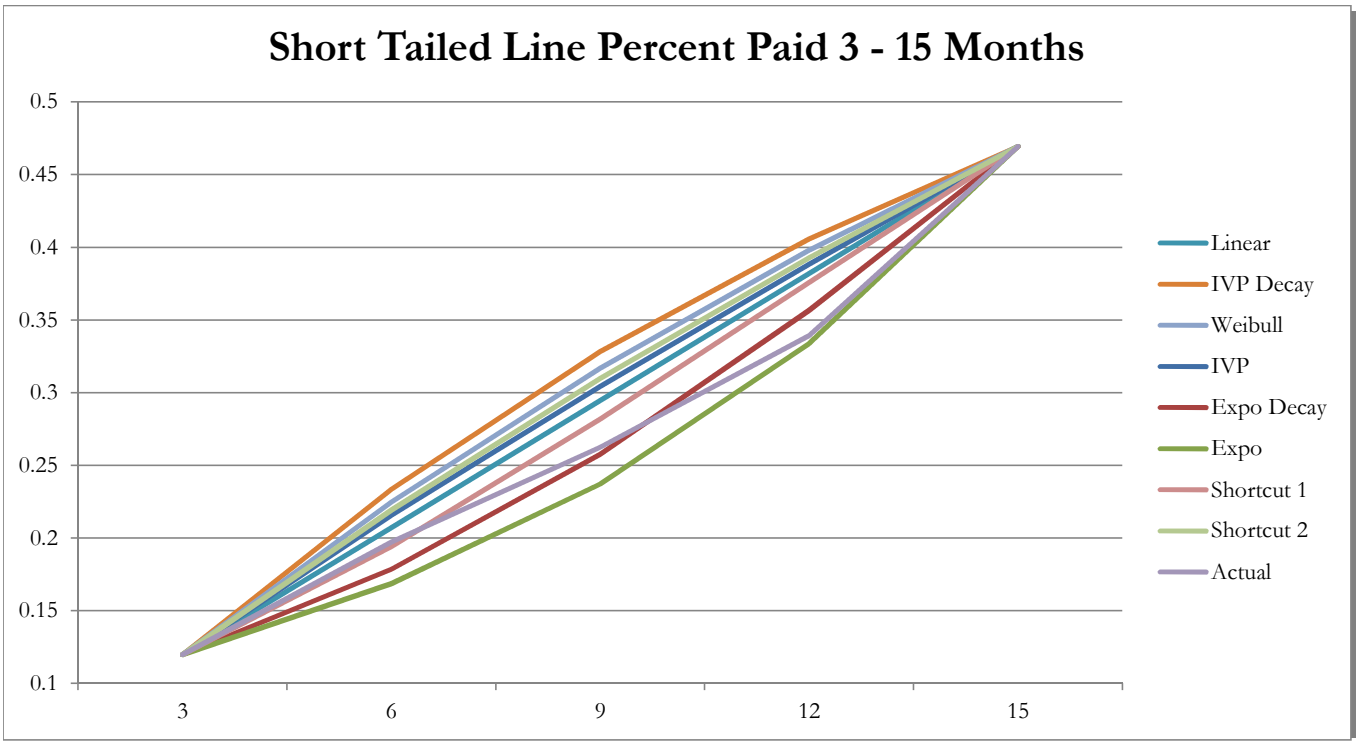
Appendix A
Sheet 1

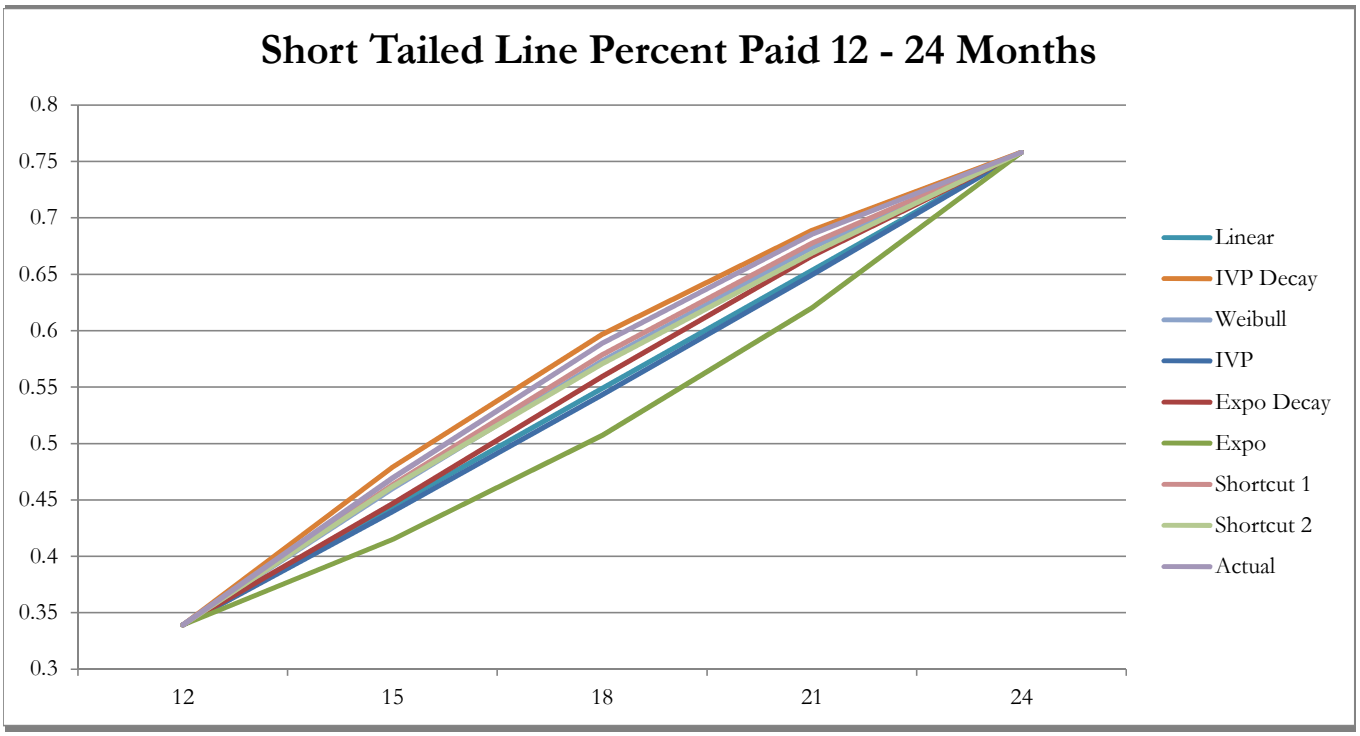




Short Tail Line Paid Graphs

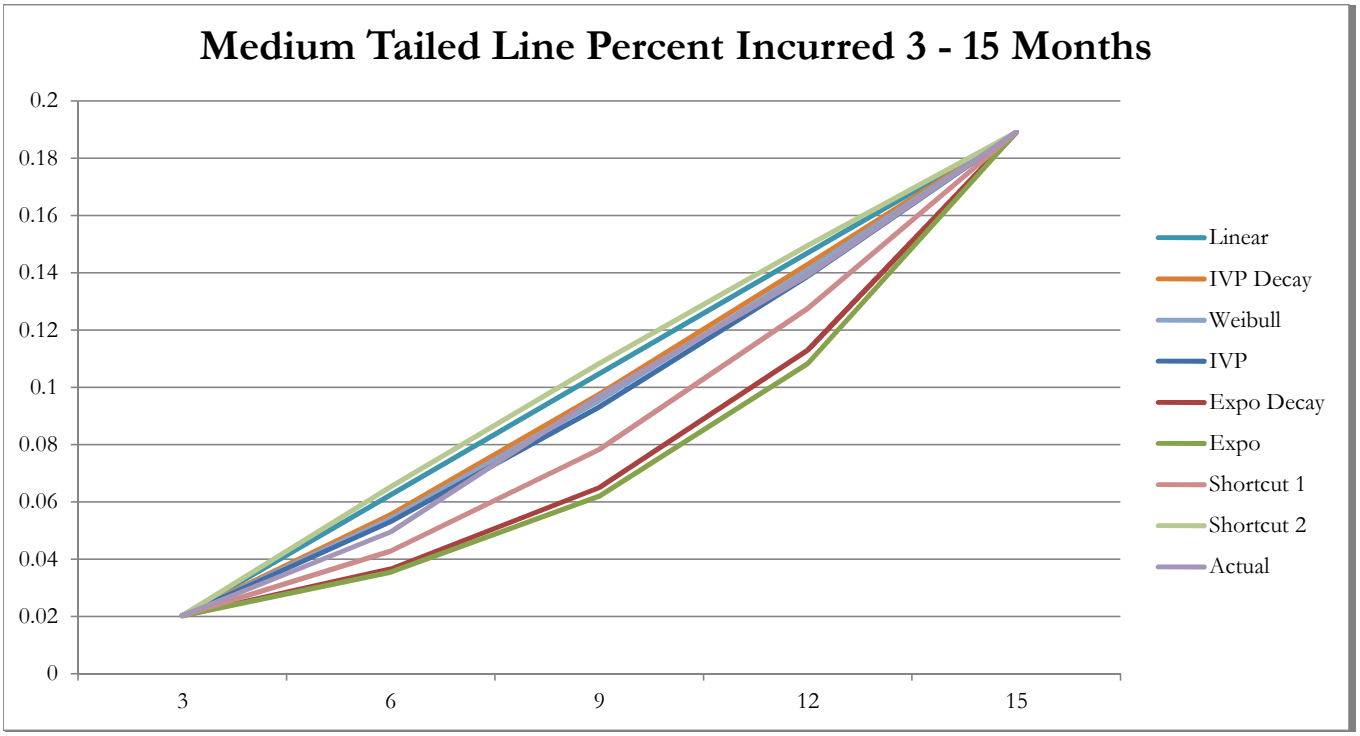
Appendix A
Sheet 2

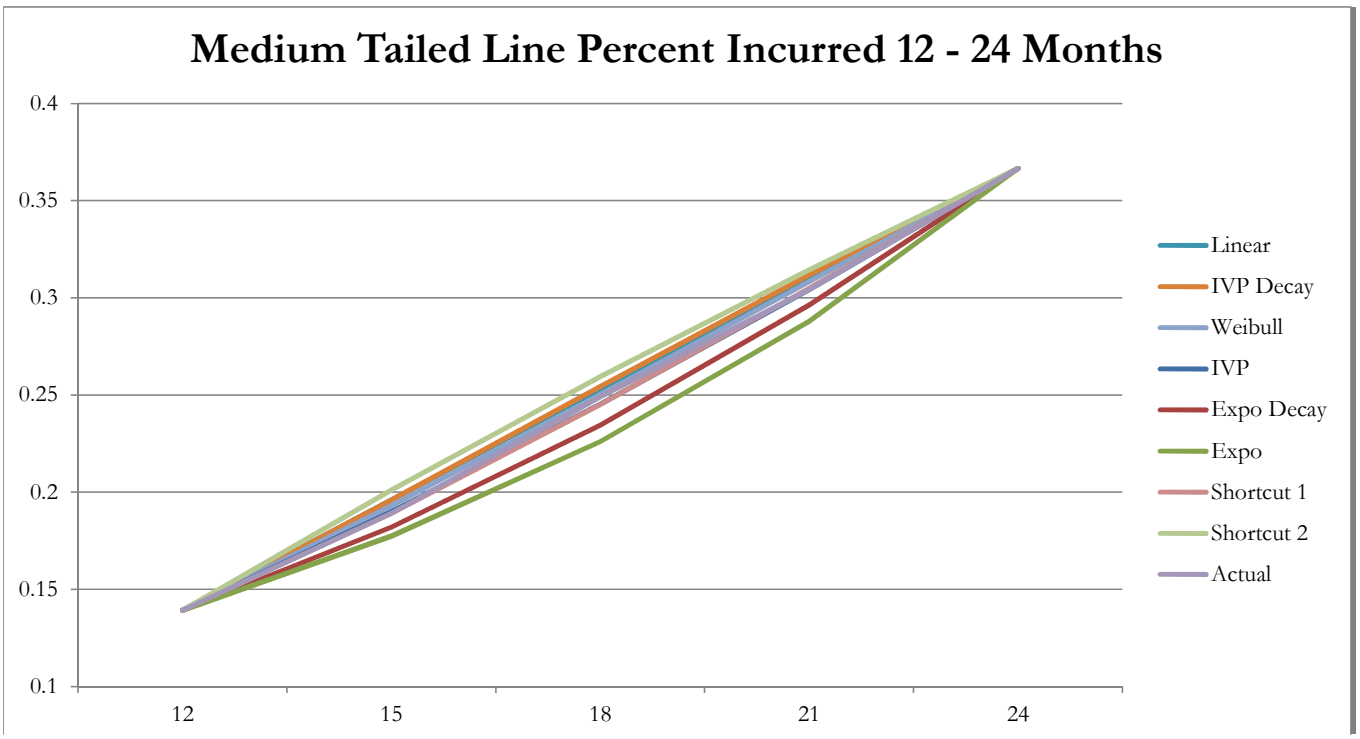




Medium Tail Line Incurred Graphs

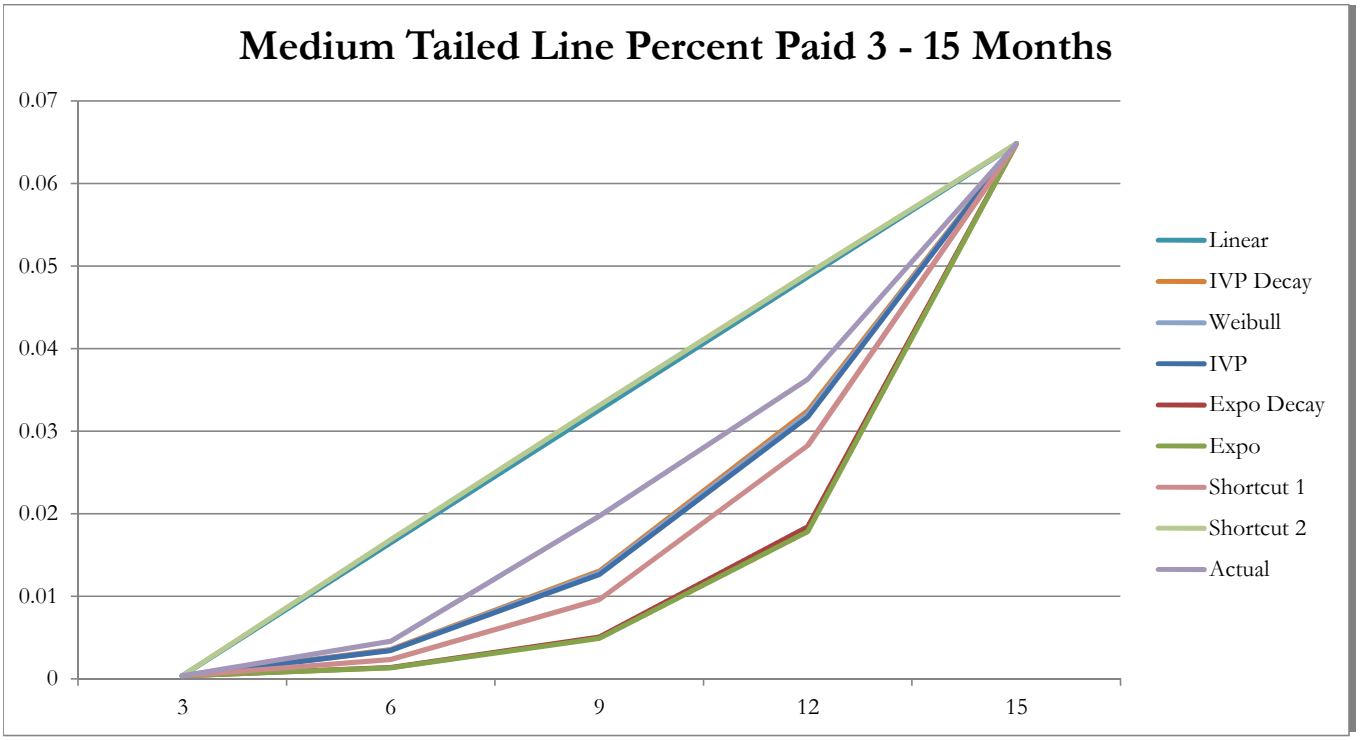
Appendix A
Sheet 3

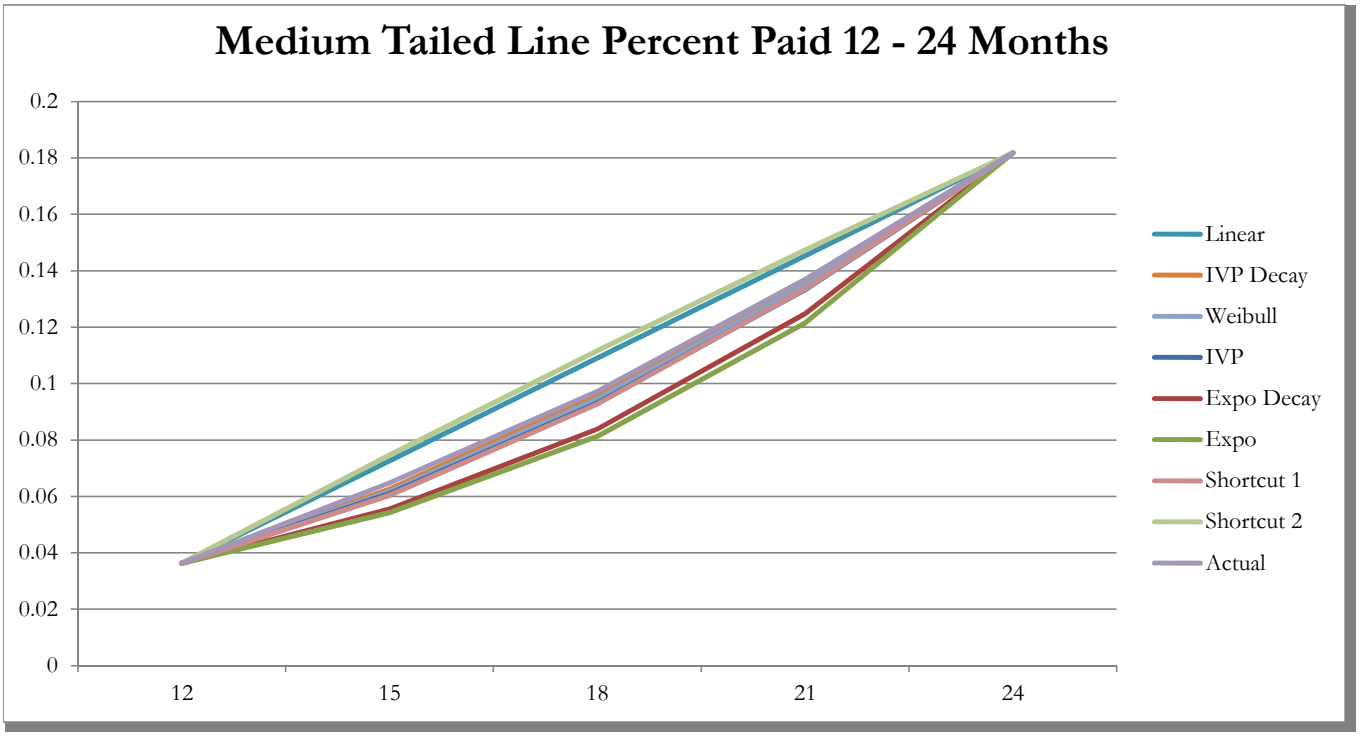




Medium Tail Line Paid Graphs

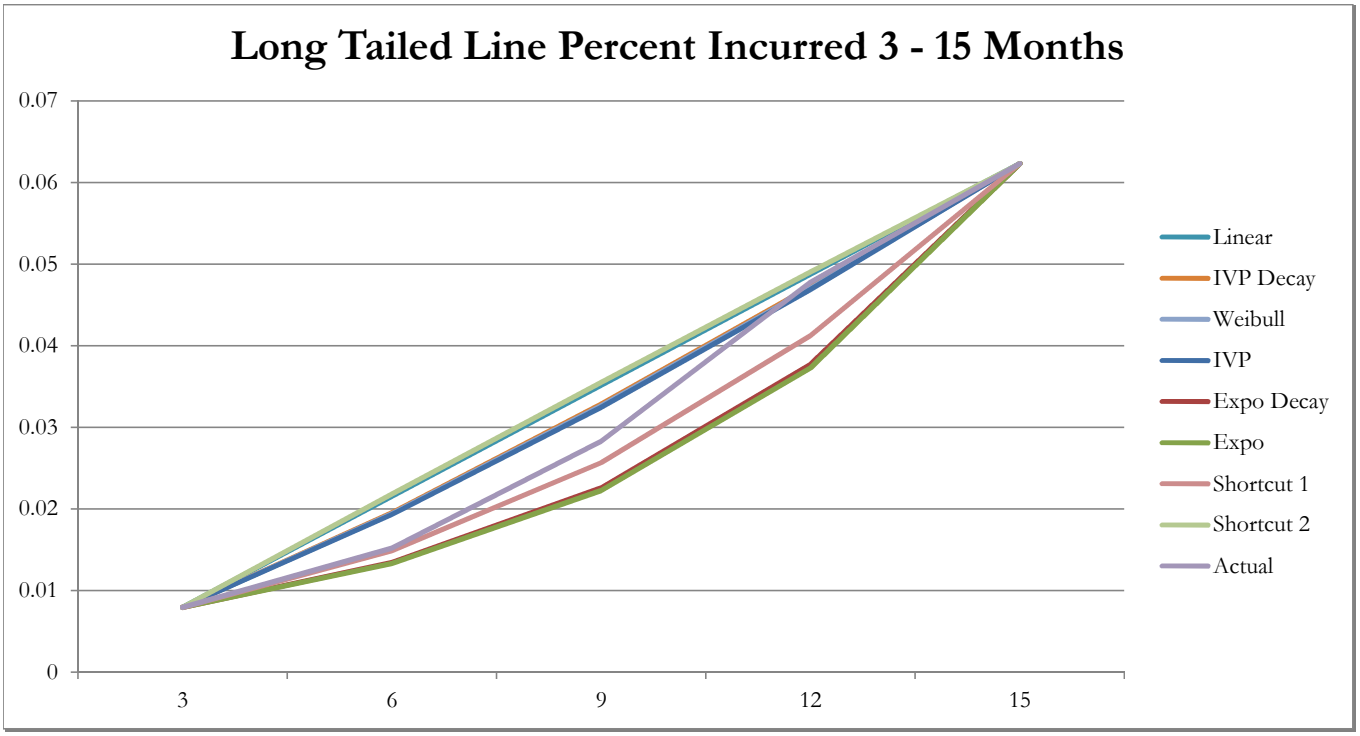
Appendix A
Sheet 4

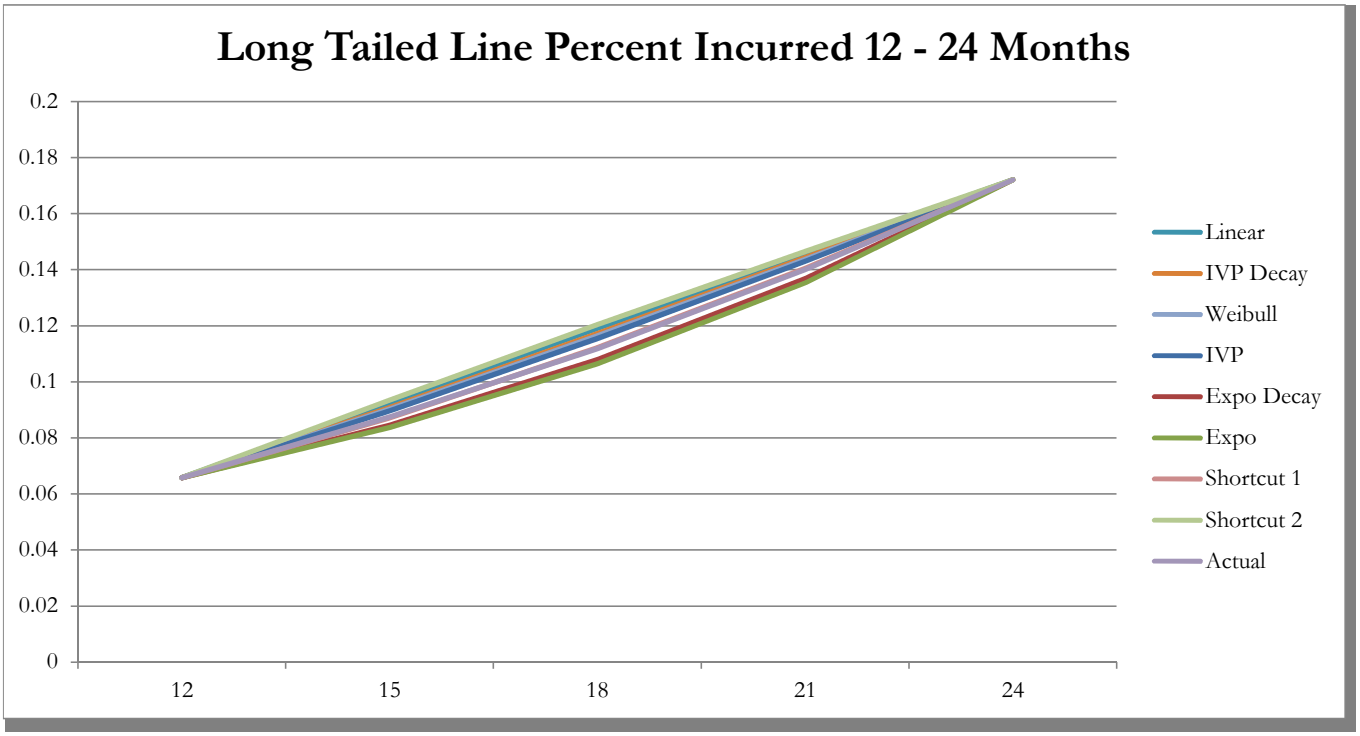




Long Tail Line Incurred Graphs

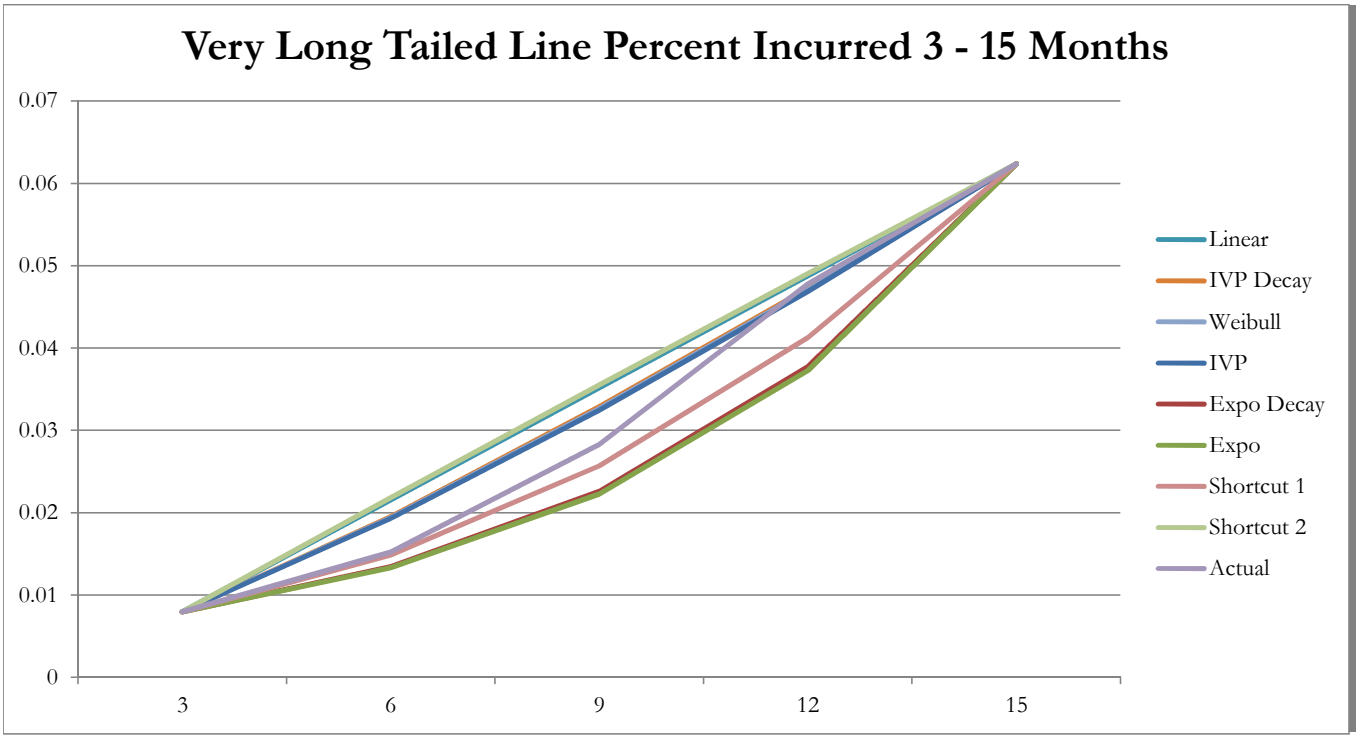
Appendix A
Sheet 5

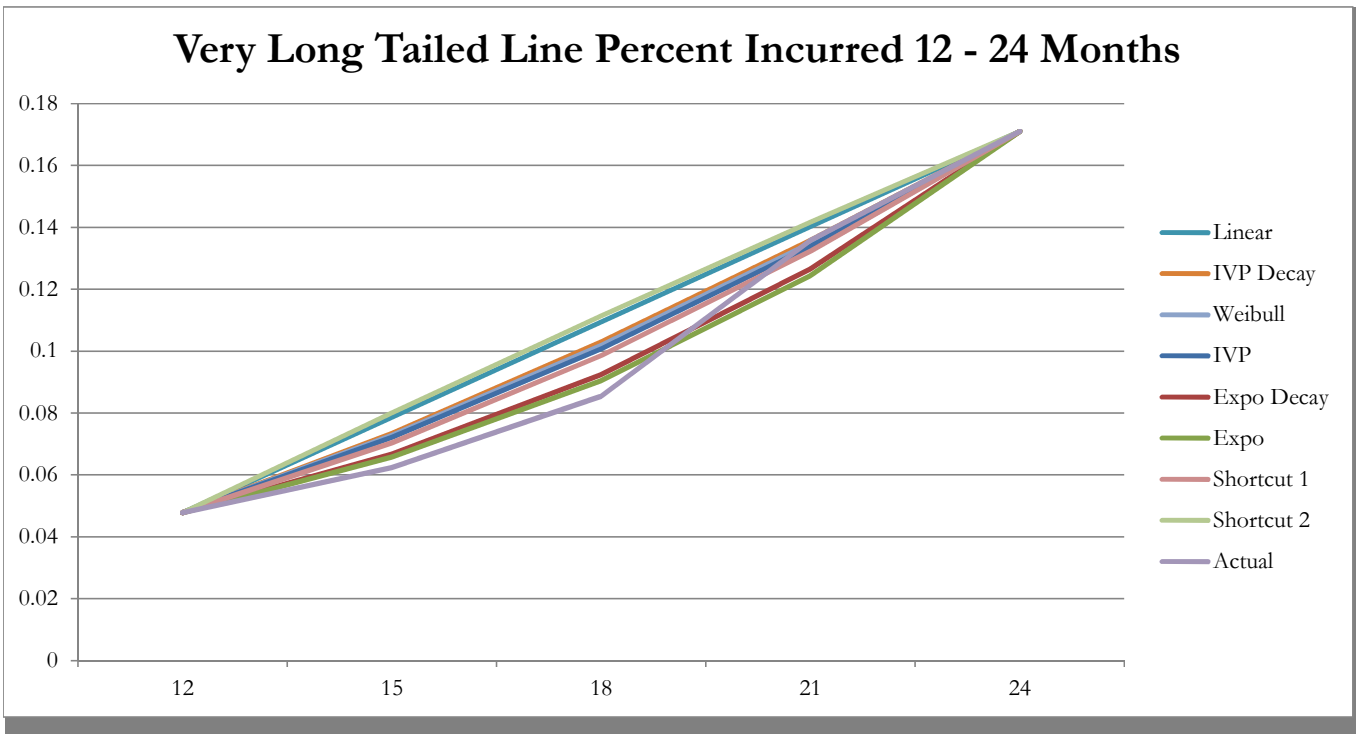




Very Long Tail Line Incurred Graphs

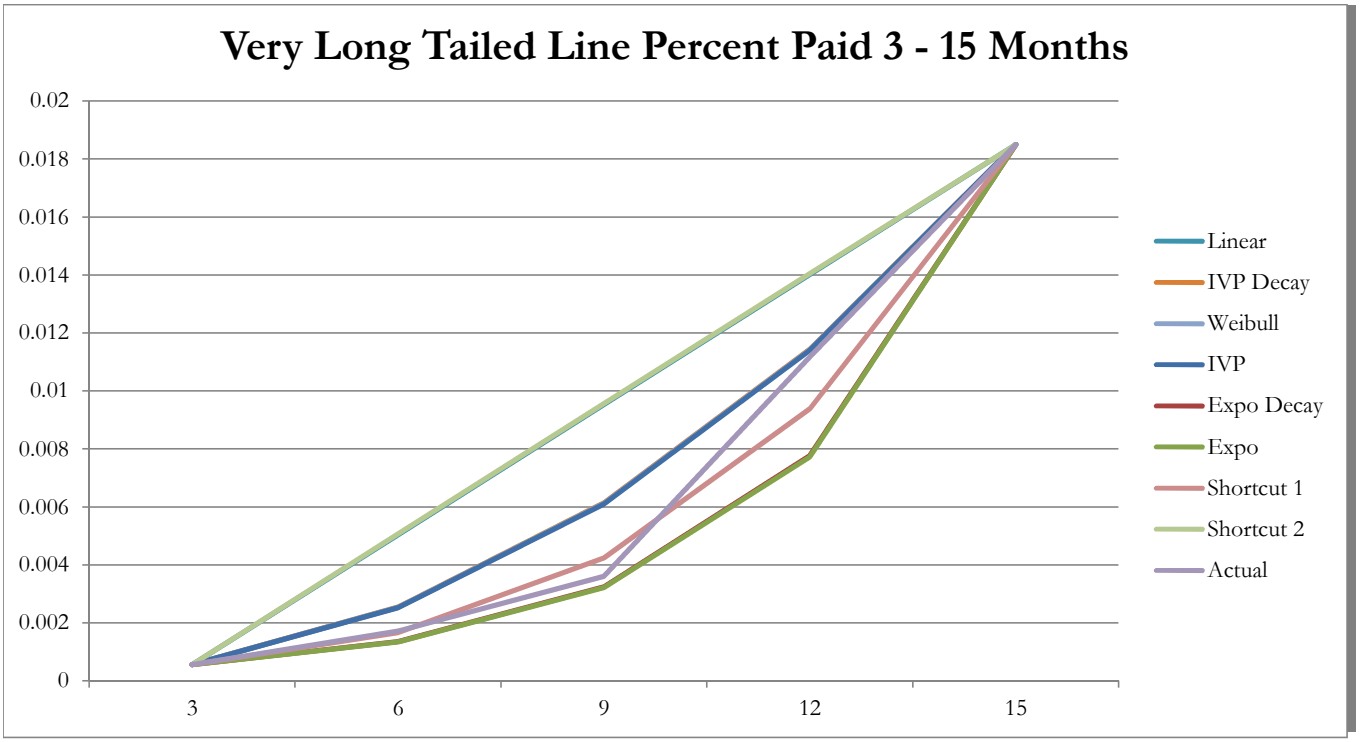
Appendix A
Sheet 6

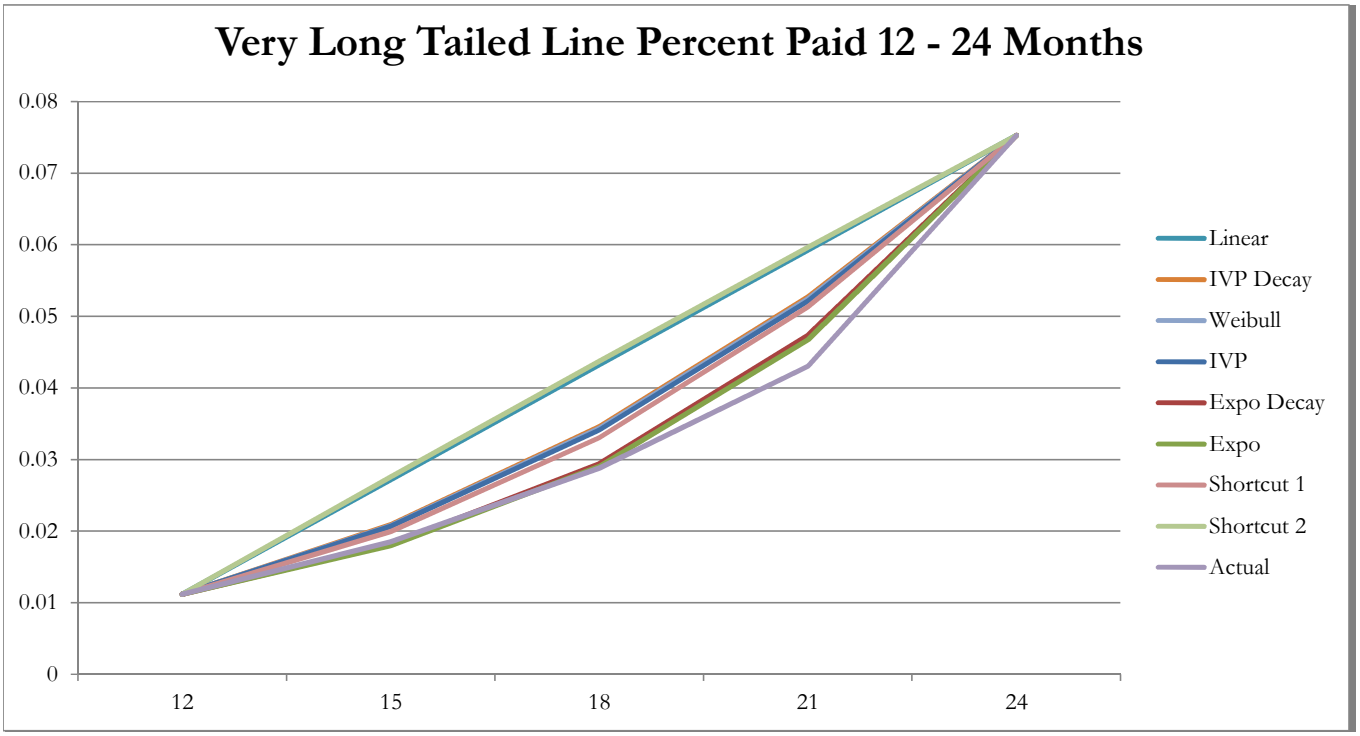




Very Long Tail Line Paid Graphs

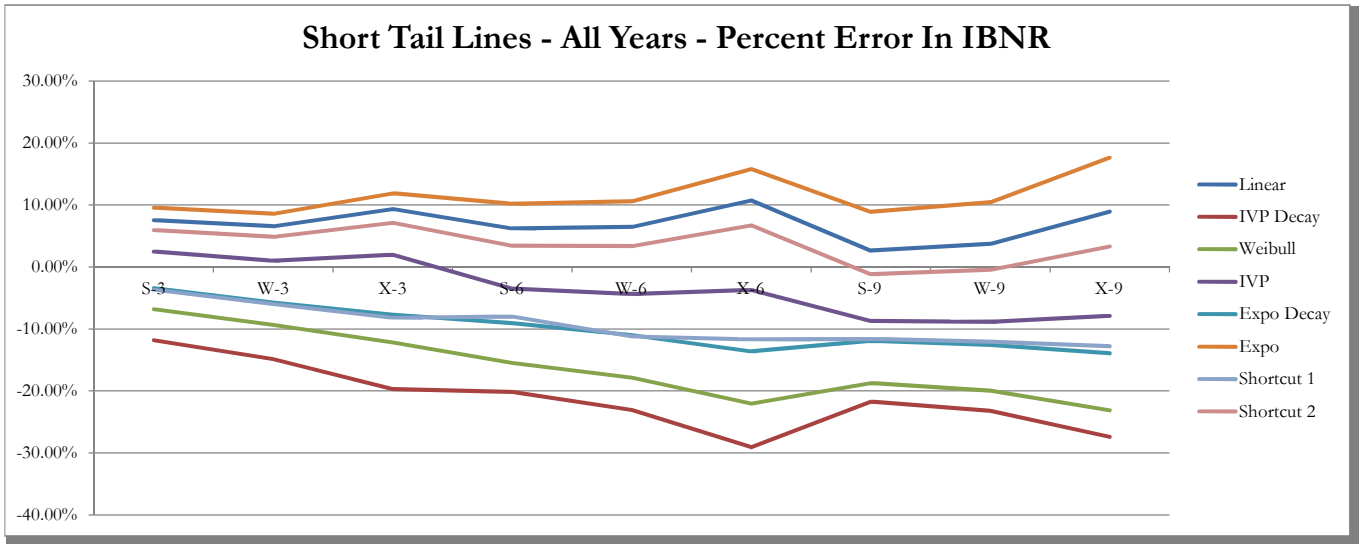
Appendix A
Sheet 7

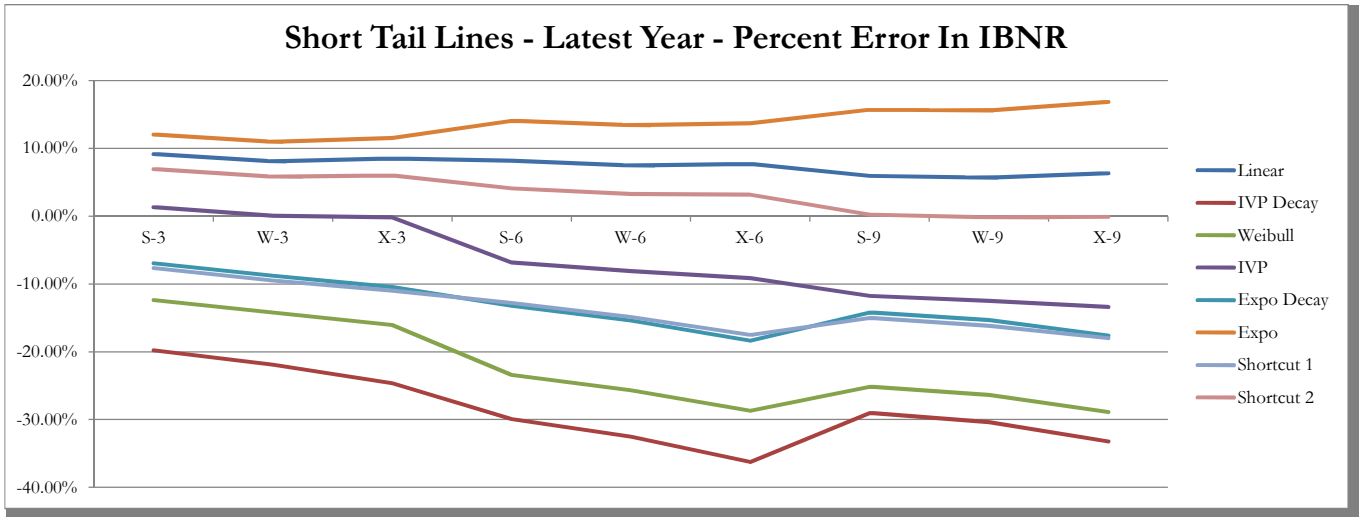


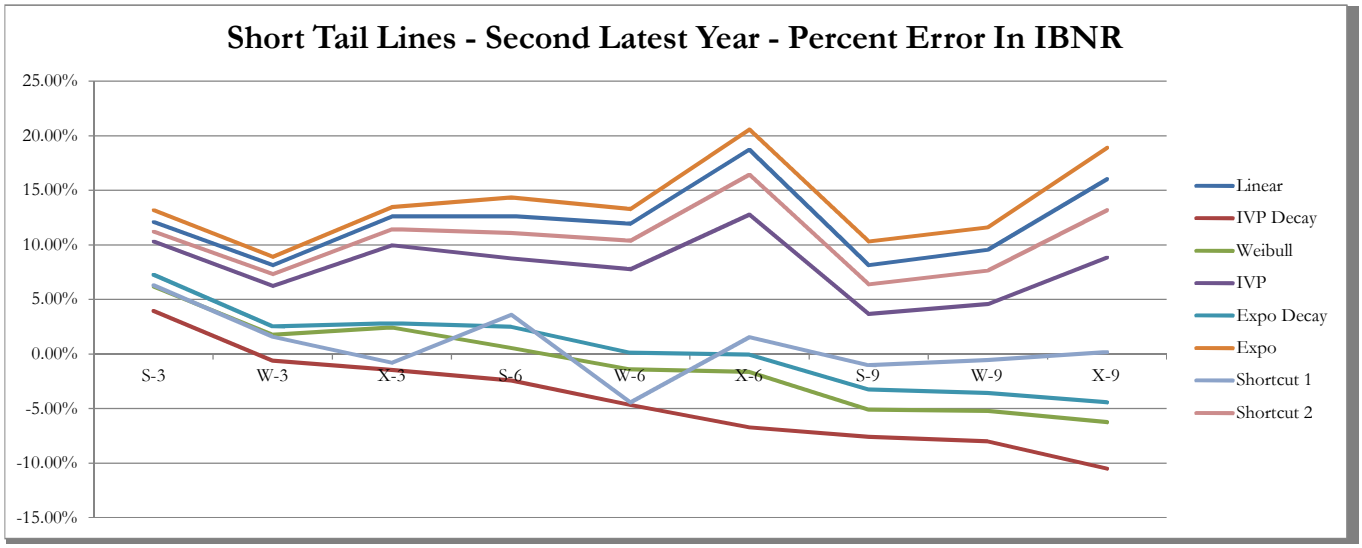


Short Tail Lines Incurred Error

Appendix A
Sheet 8

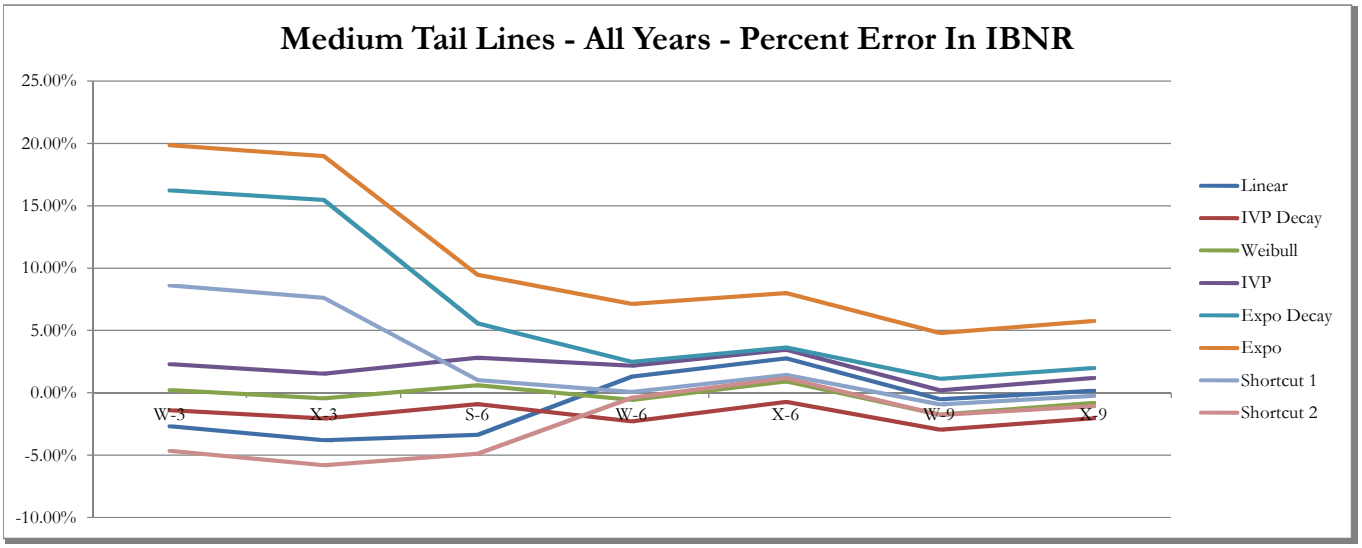


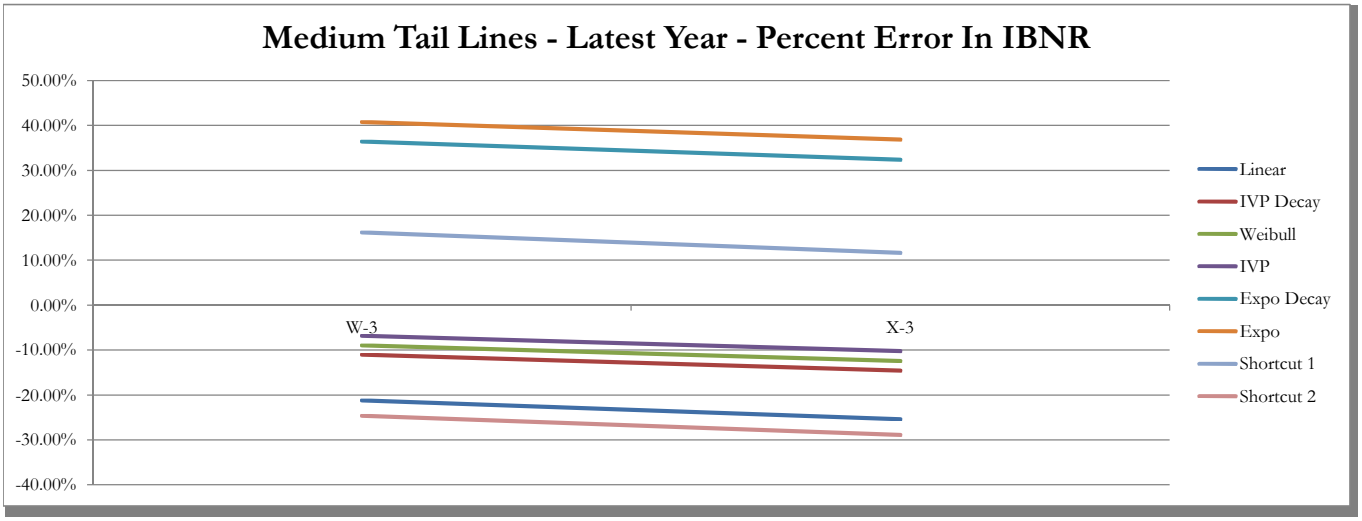


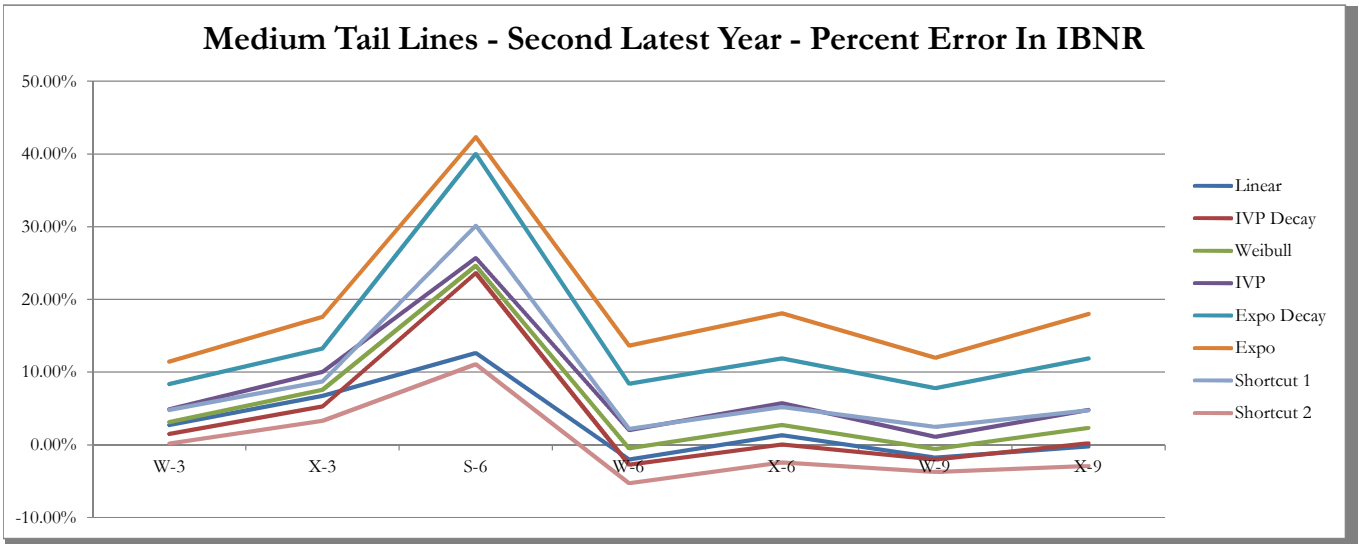


Medium Tail Lines Incurred Error

Appendix A
Sheet 9

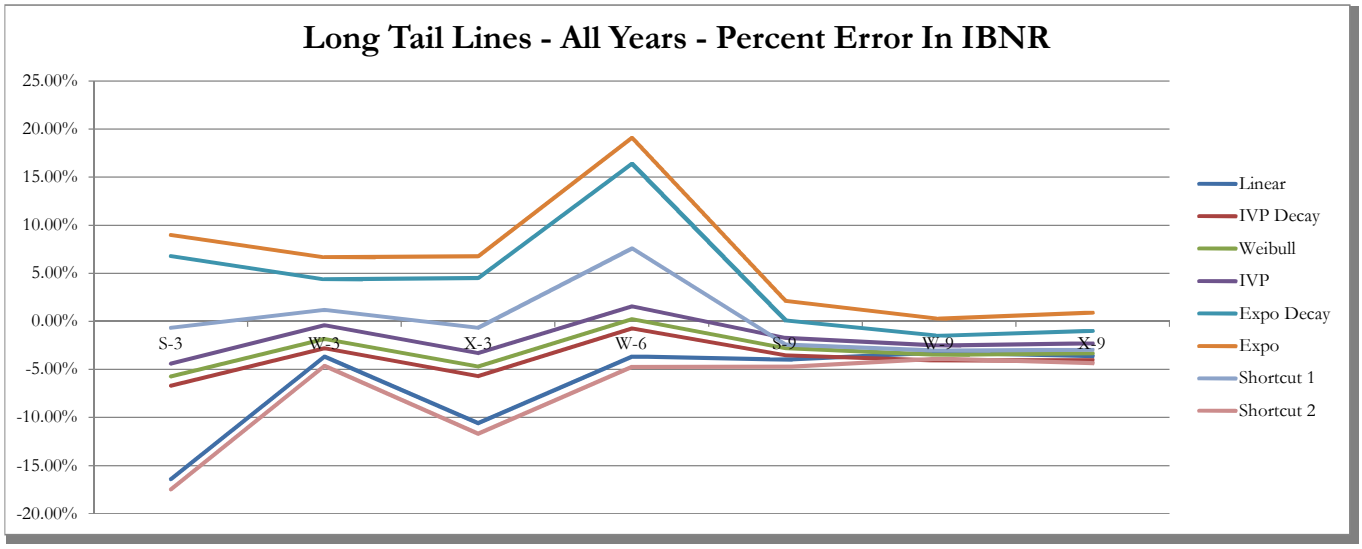


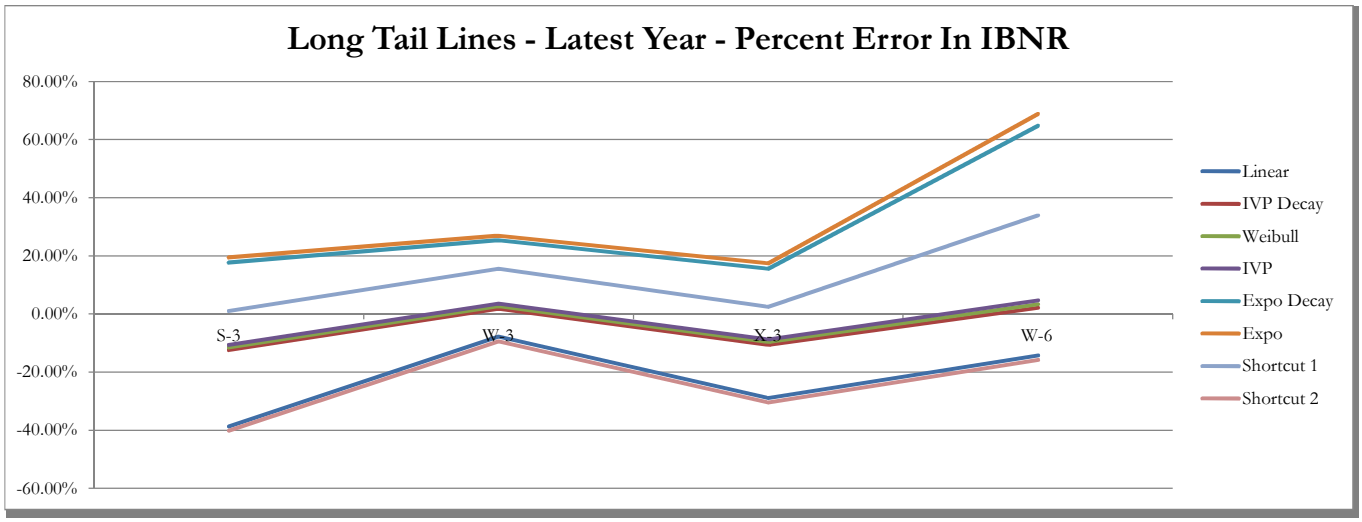


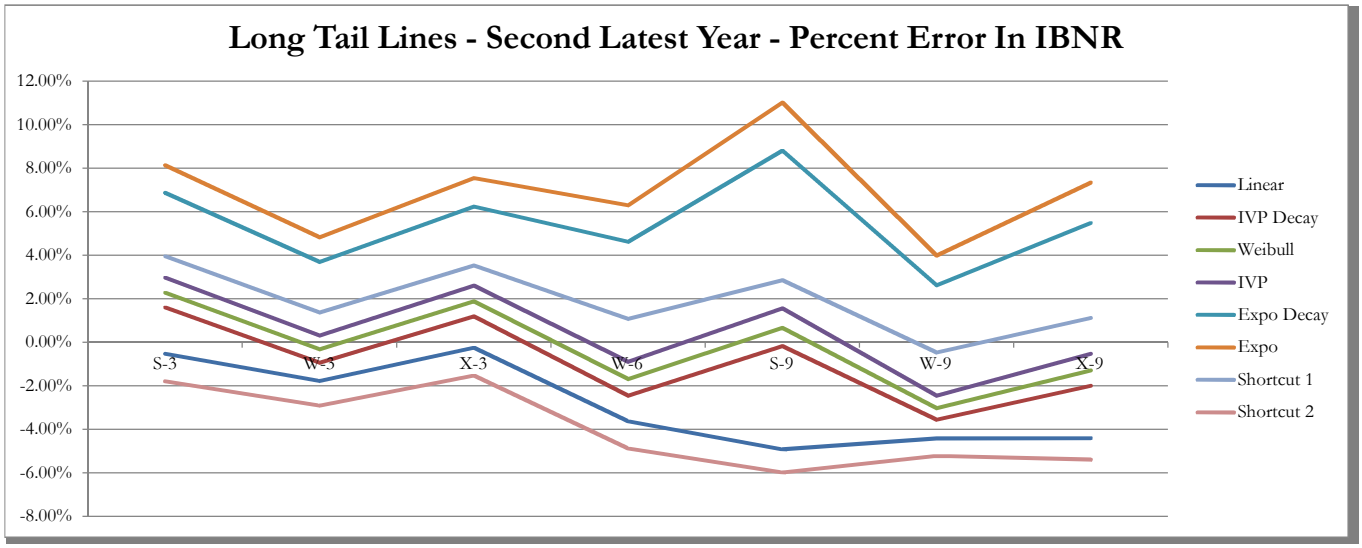


Long Tail Lines Incurred Error

Appendix A
Sheet 10

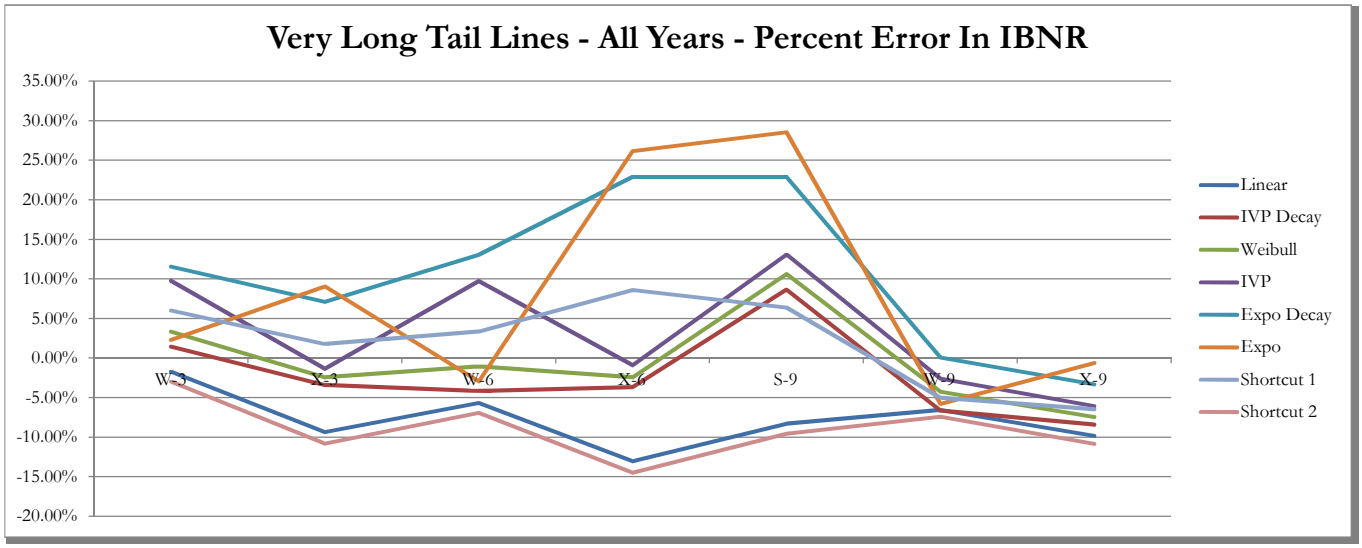


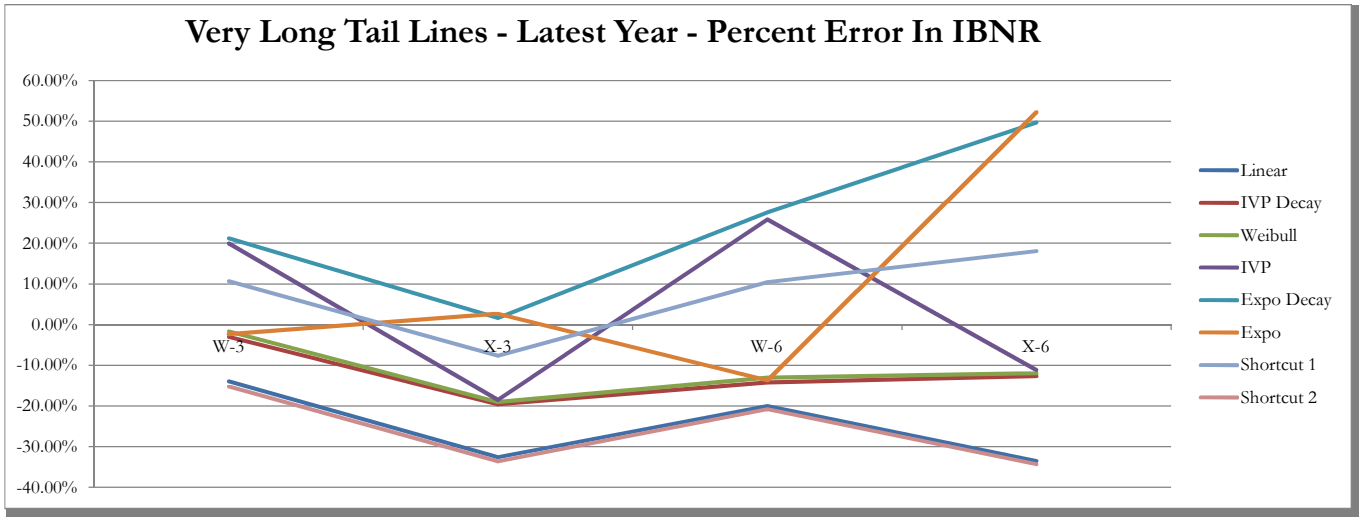


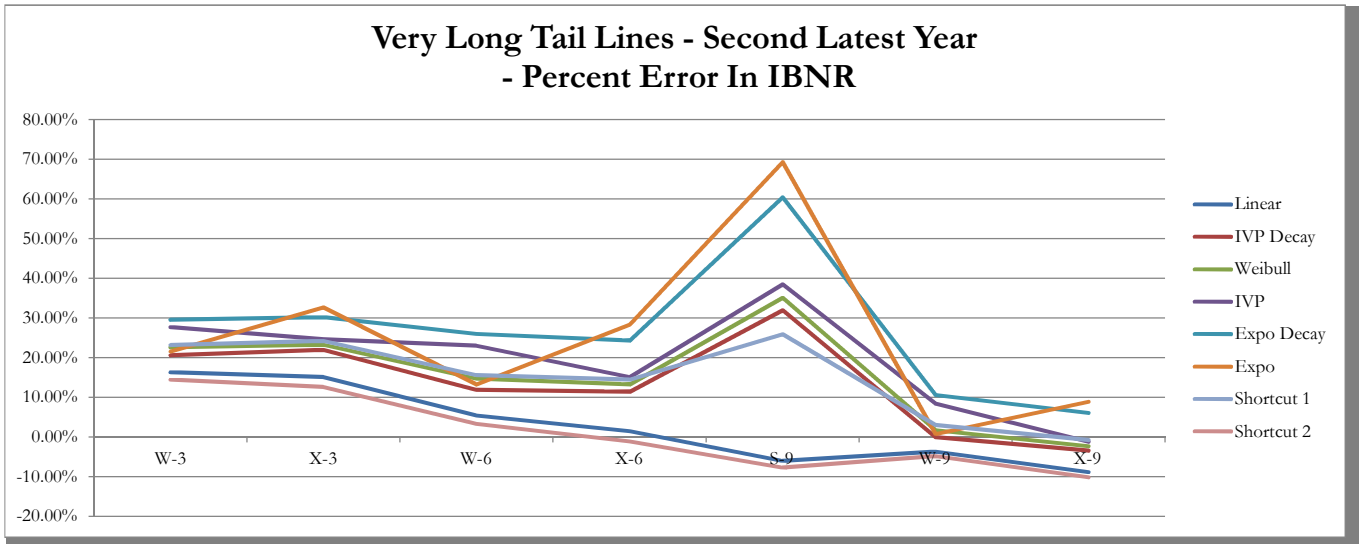


Very Long Tail Lines Incurred Error

Appendix A
Sheet 11

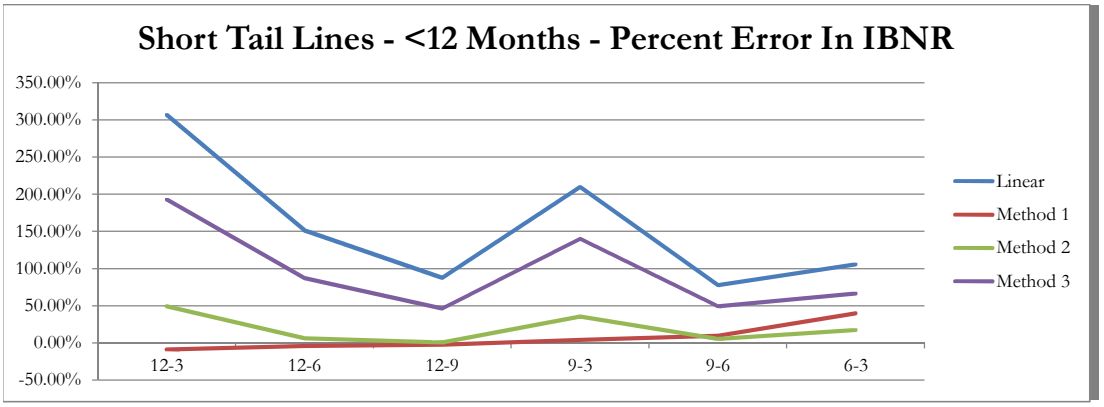


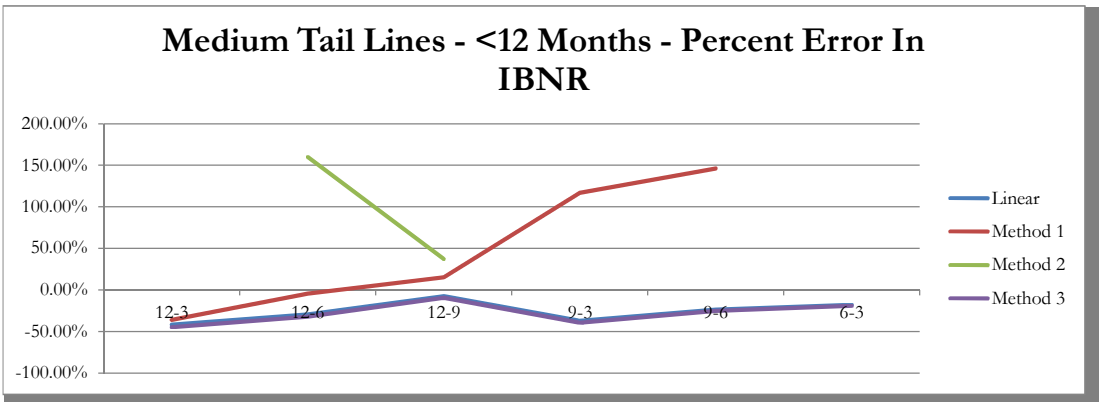


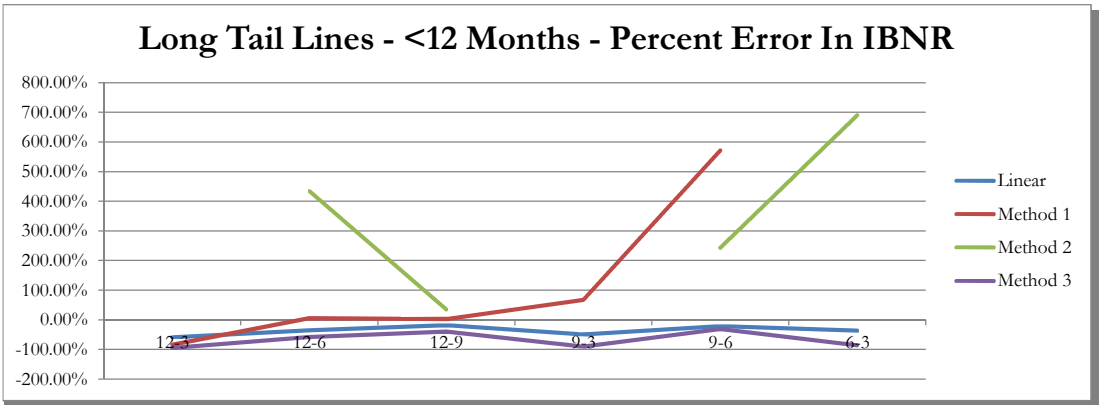


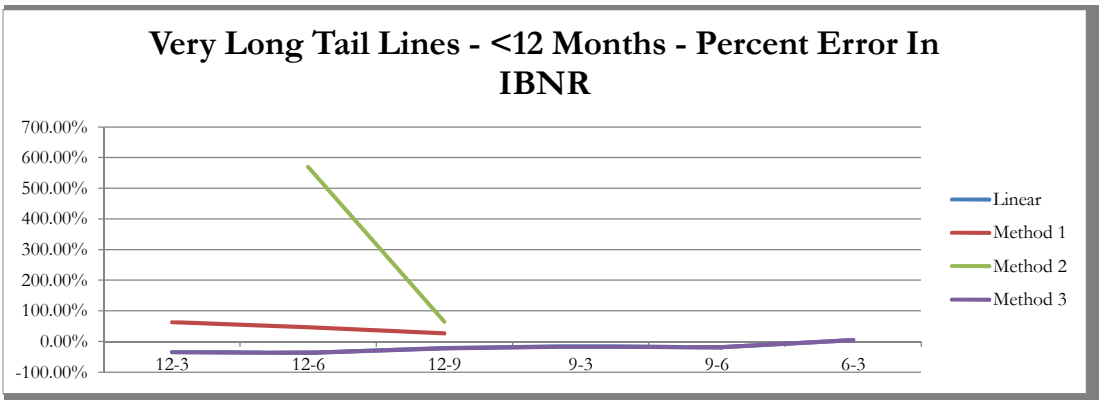
Incurred Extrapolation Errors

Appendix A
Sheet 12



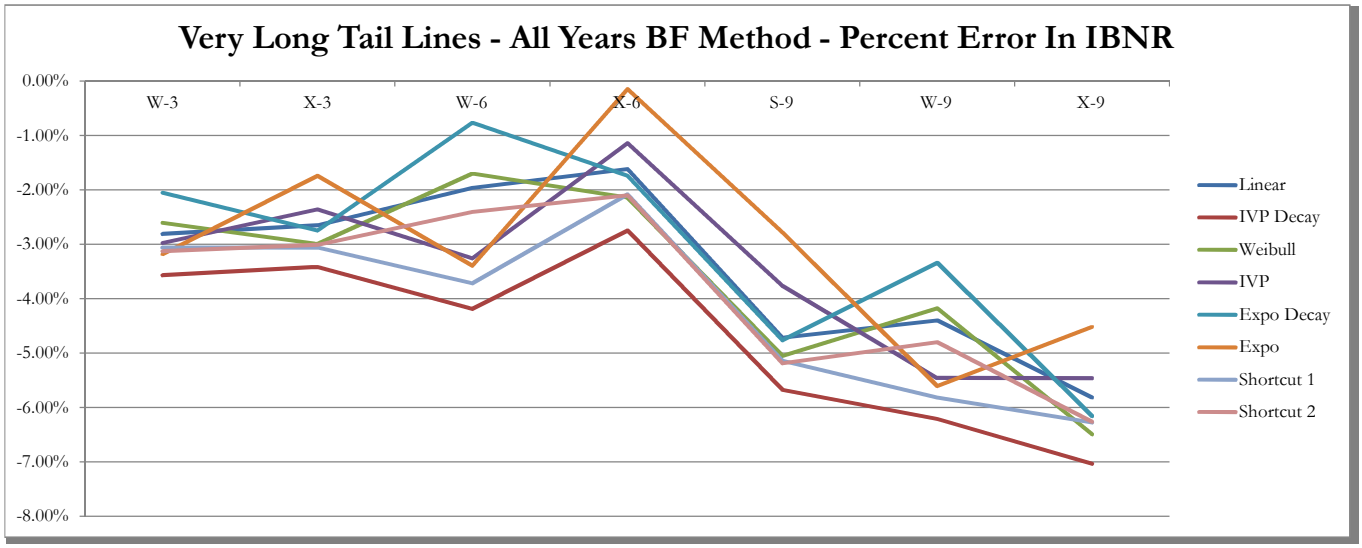


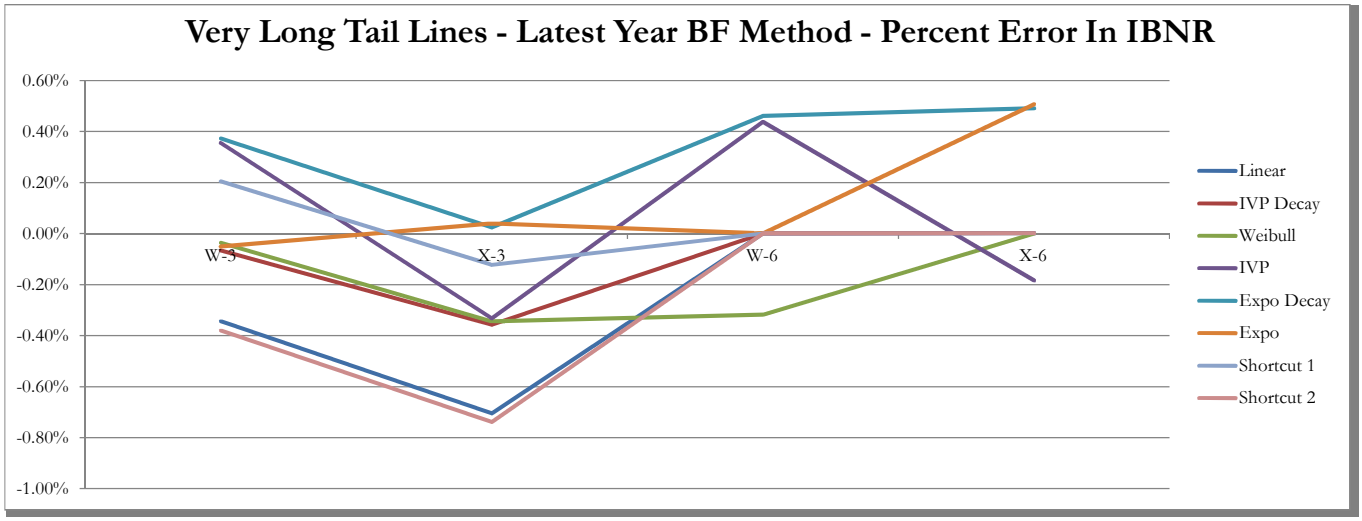


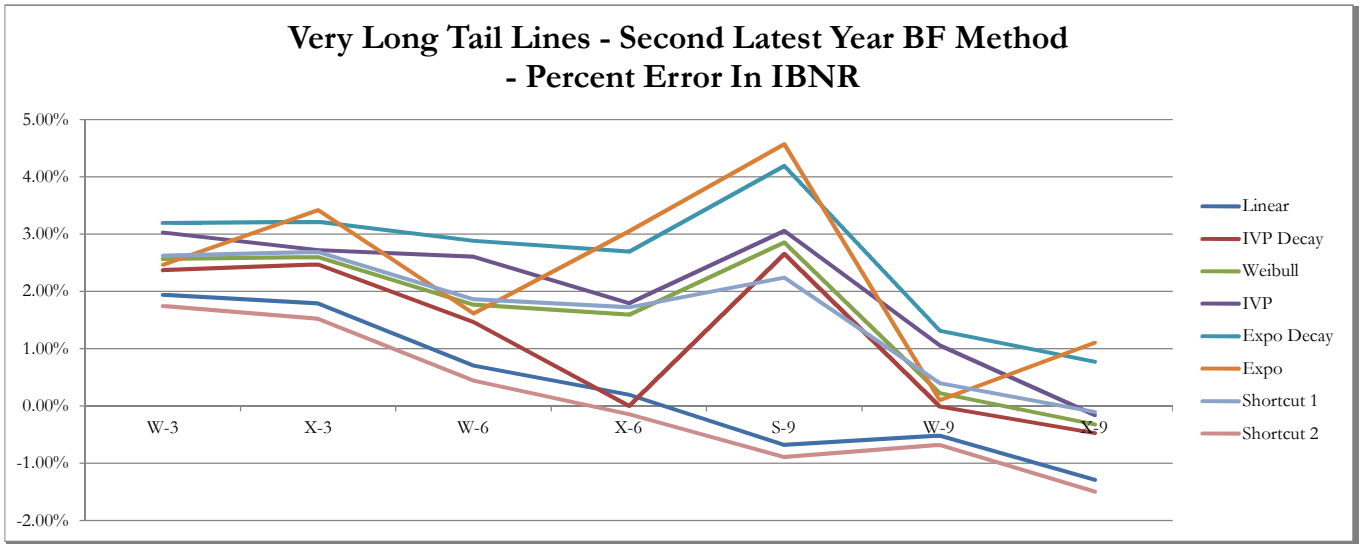


Very Long Tail Lines BF Error

Appendix A
Sheet 13

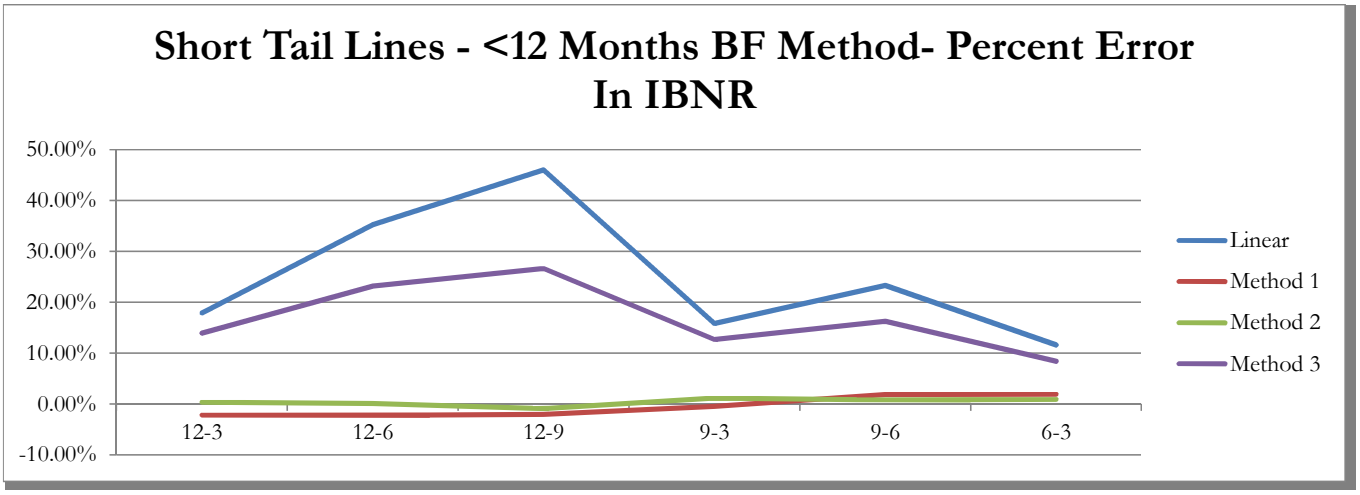


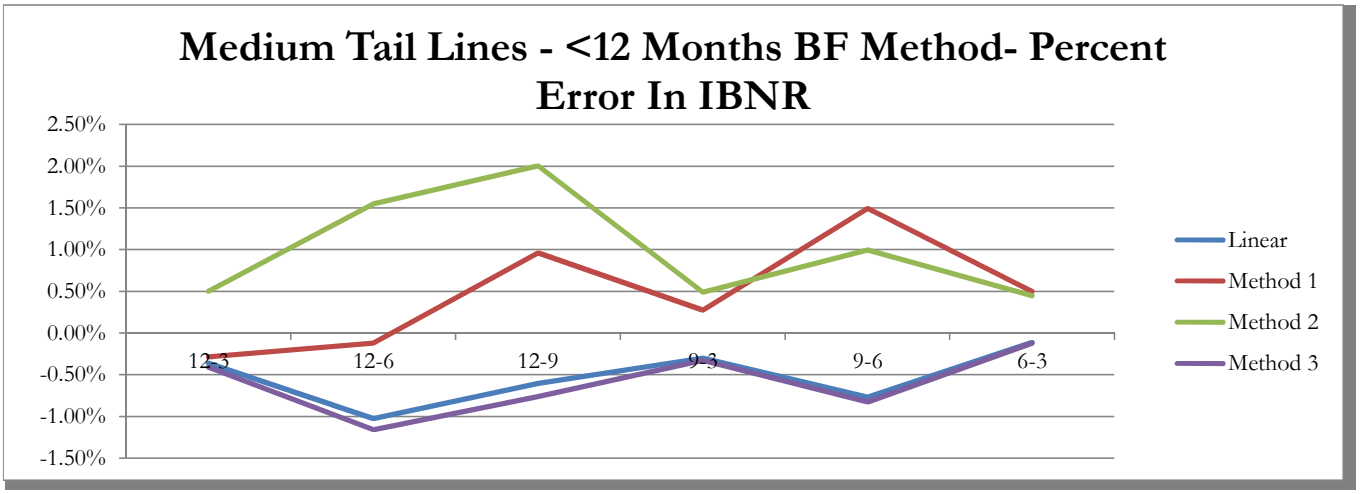


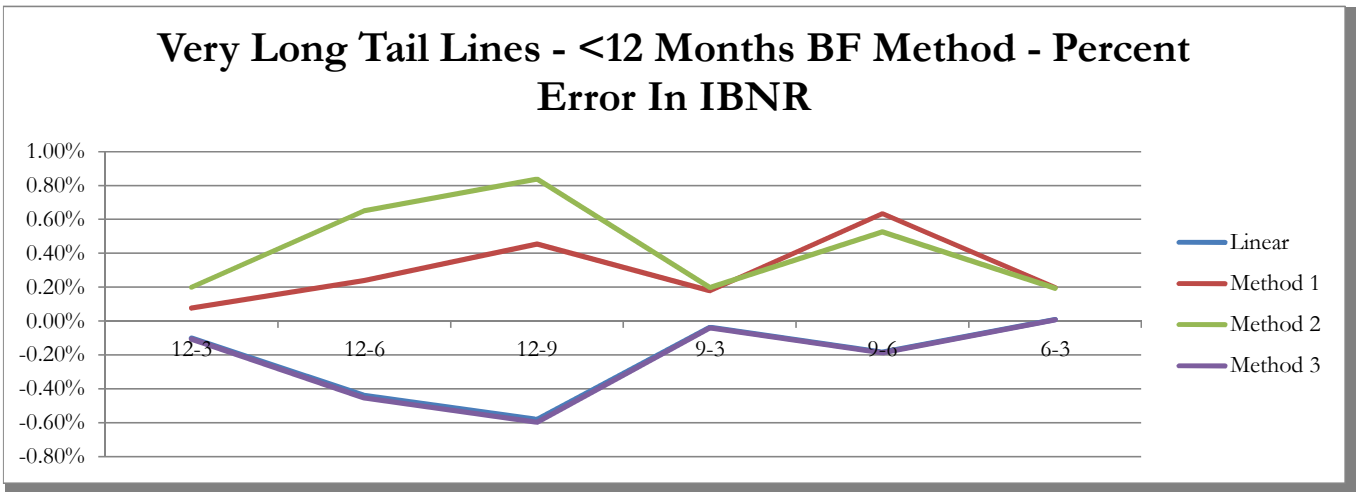


Incurred BF Extrapolation Errors

Appendix A
Sheet 14

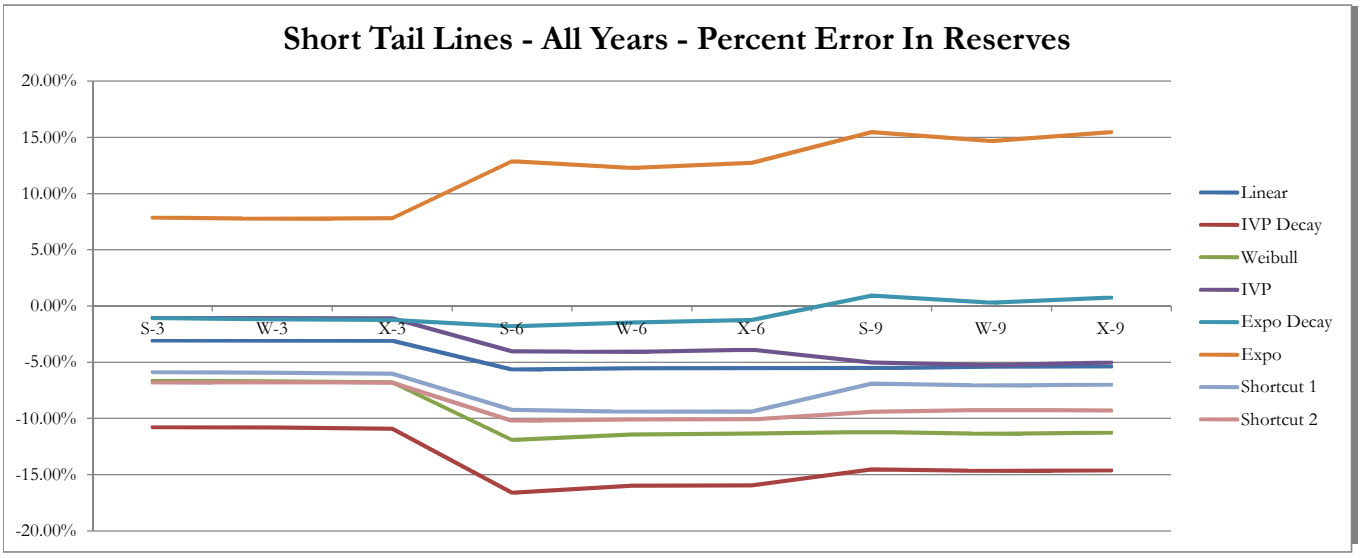


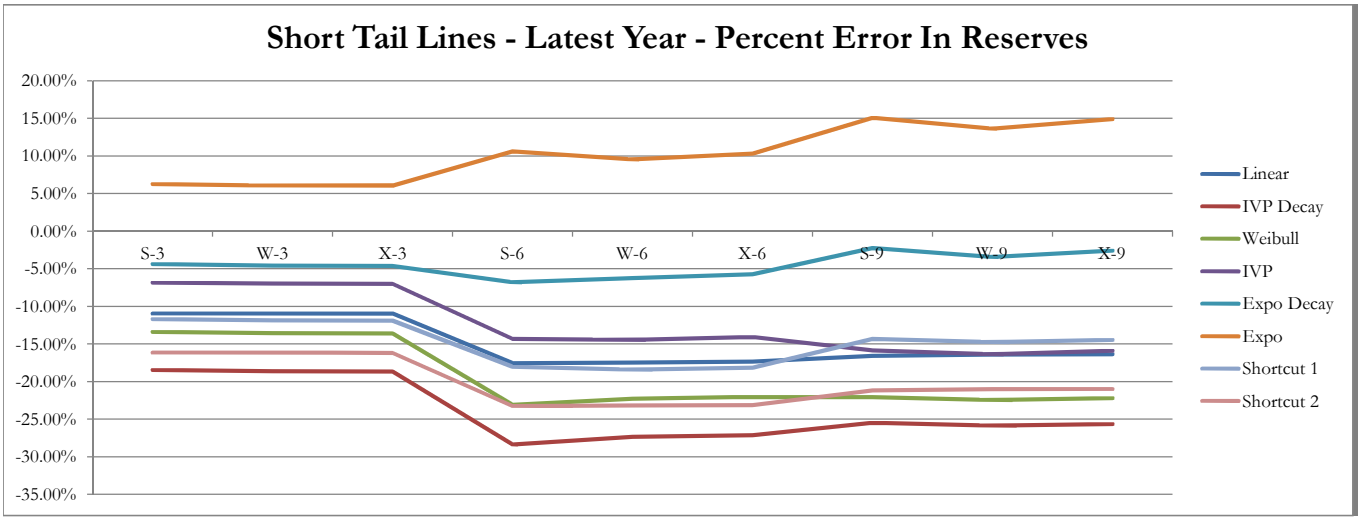


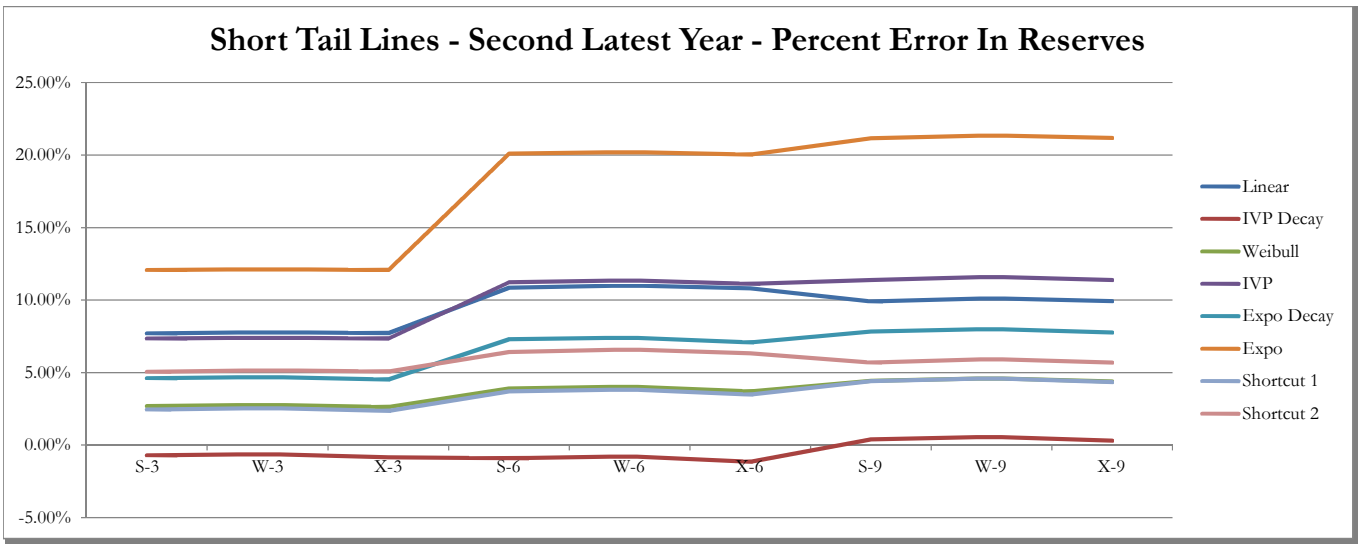


Short Tail Lines Paid Error

Appendix A
Sheet 15

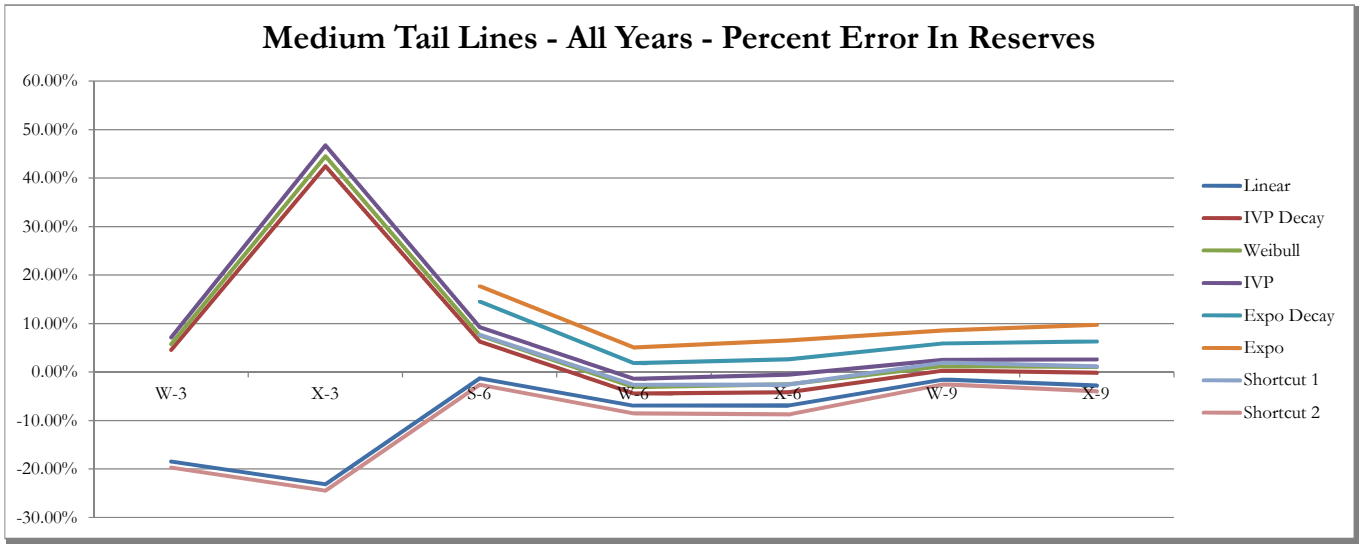


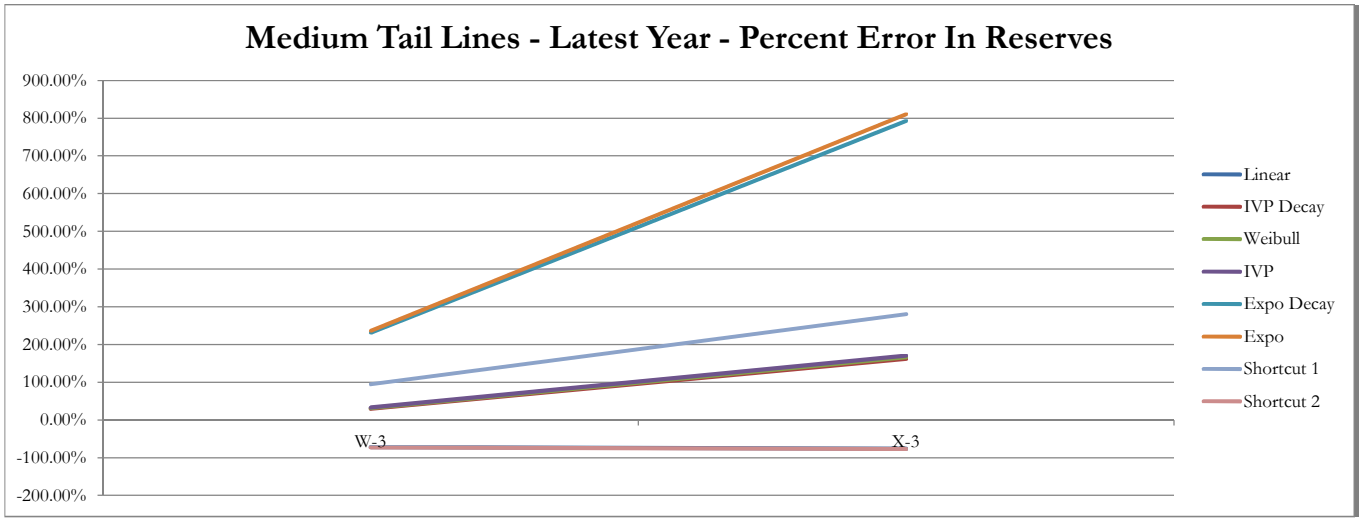


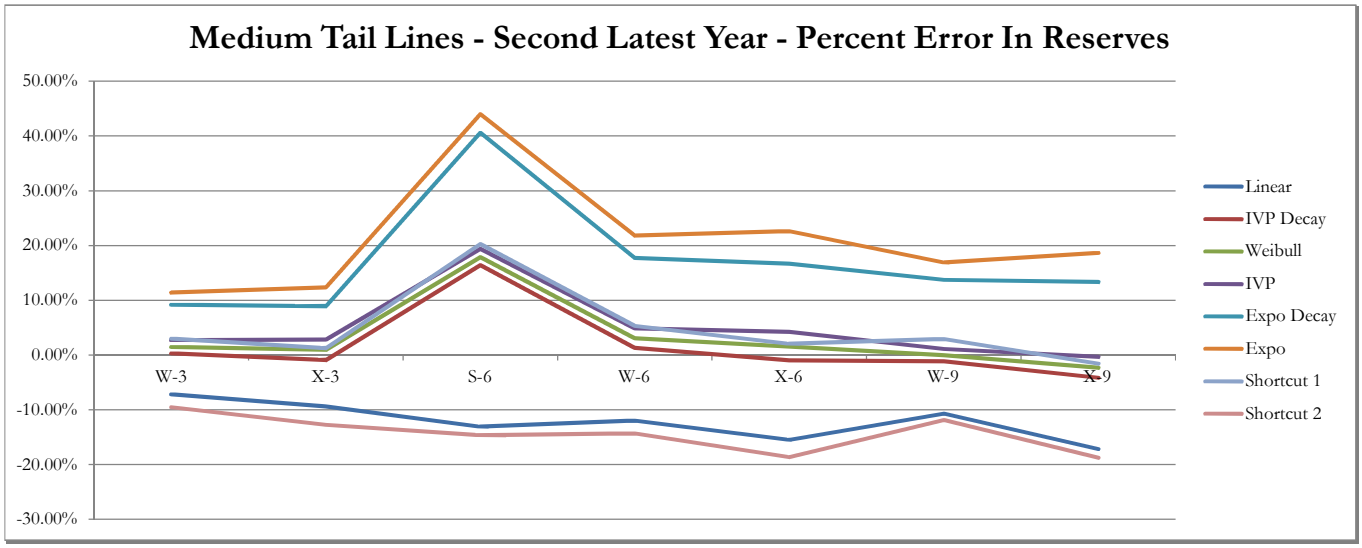


Medium Tail Lines Paid Error

Appendix A
Sheet 16

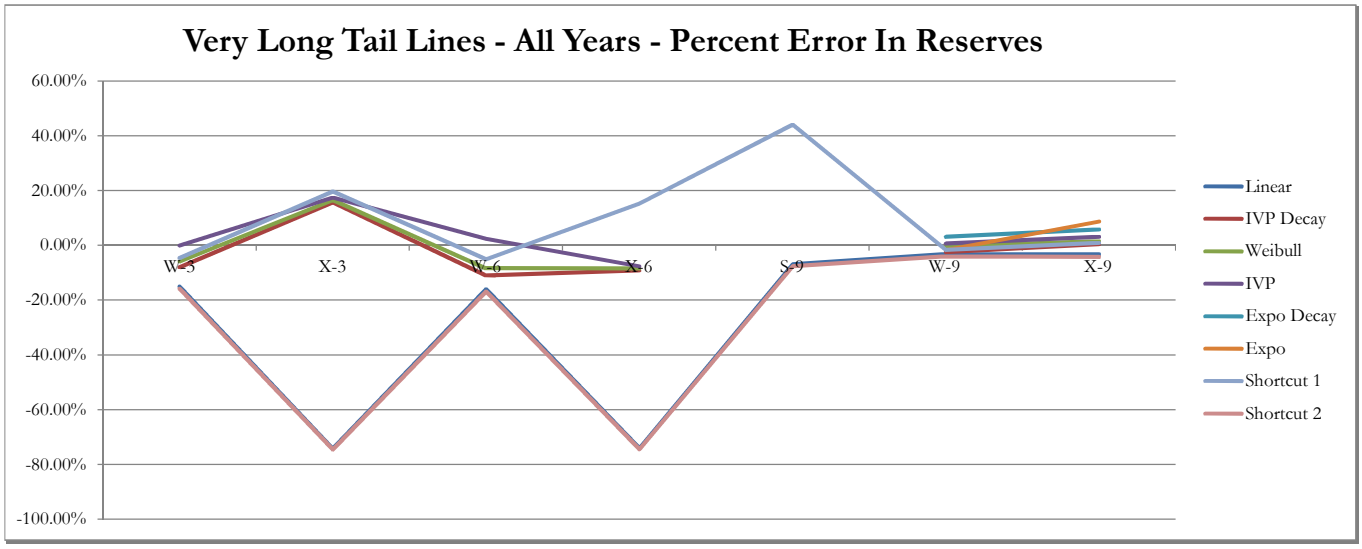


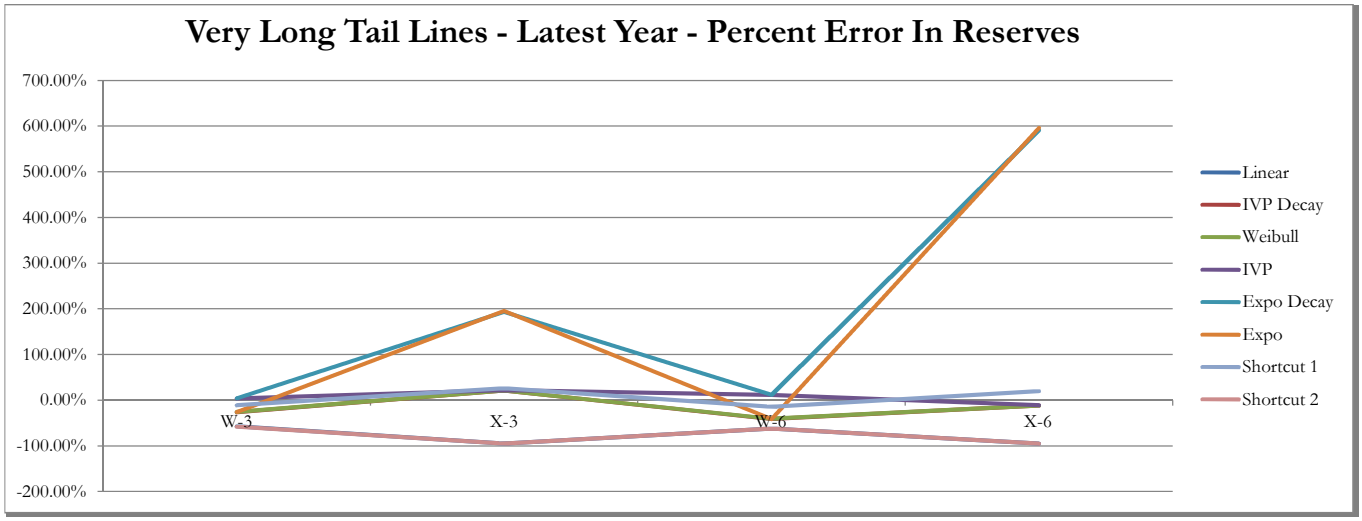


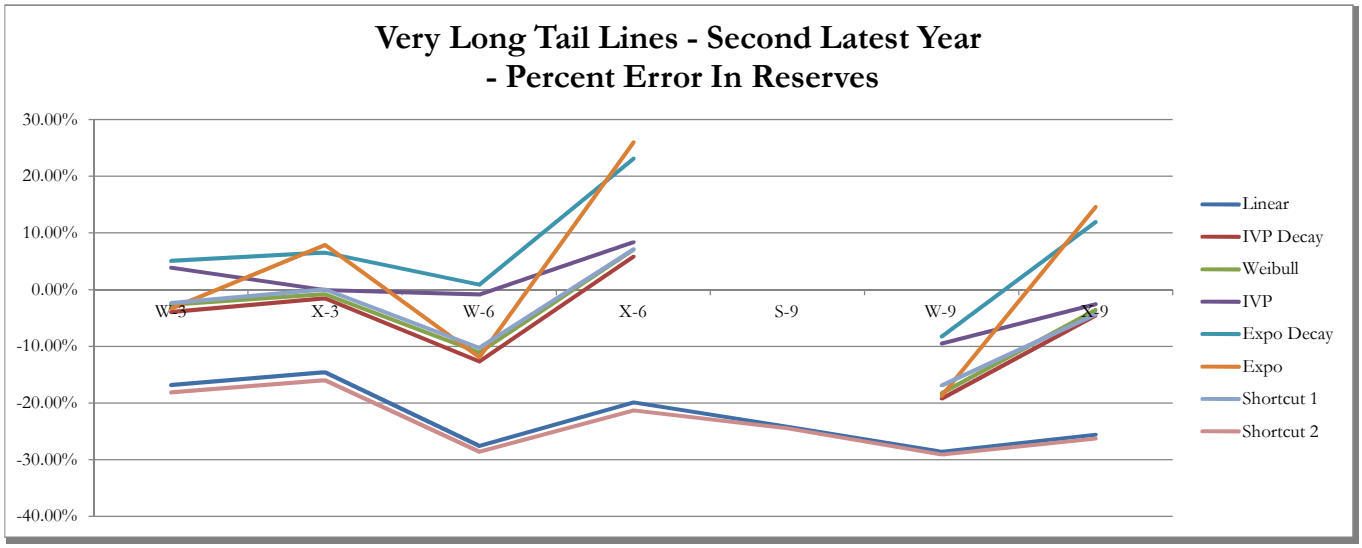


Very Long Tail Lines Paid Error

Appendix A
Sheet 17

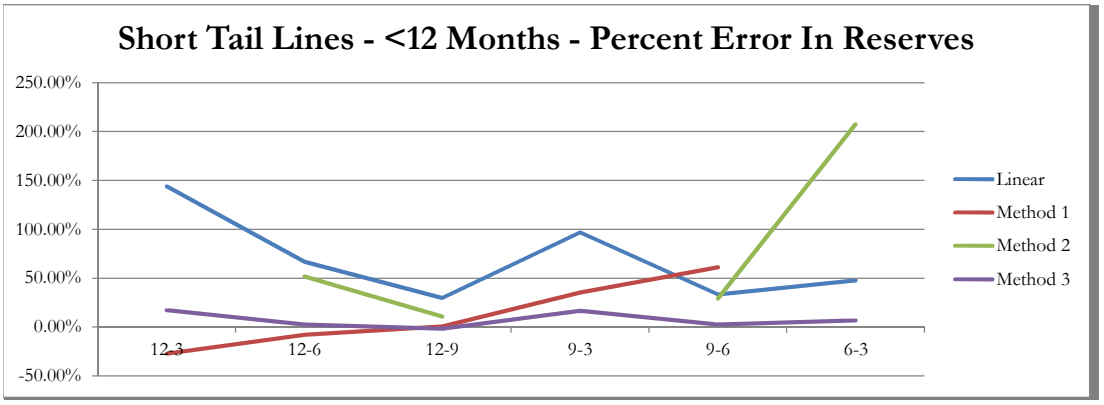


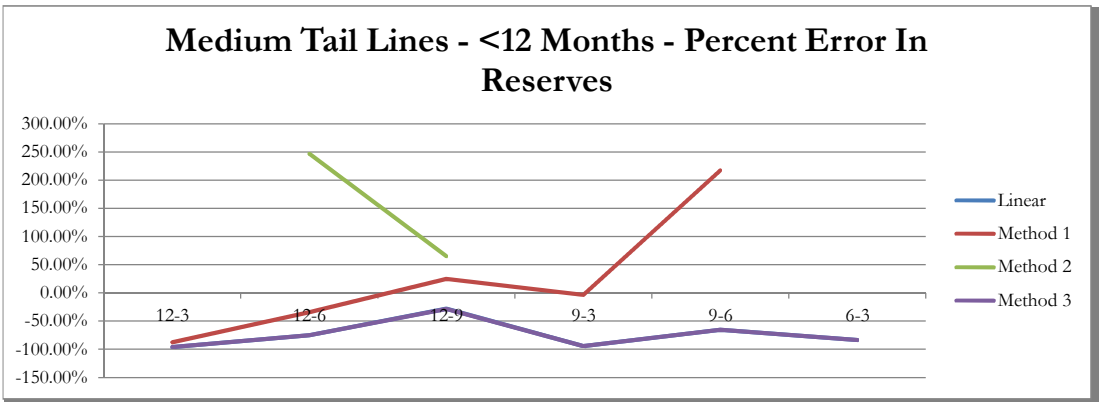


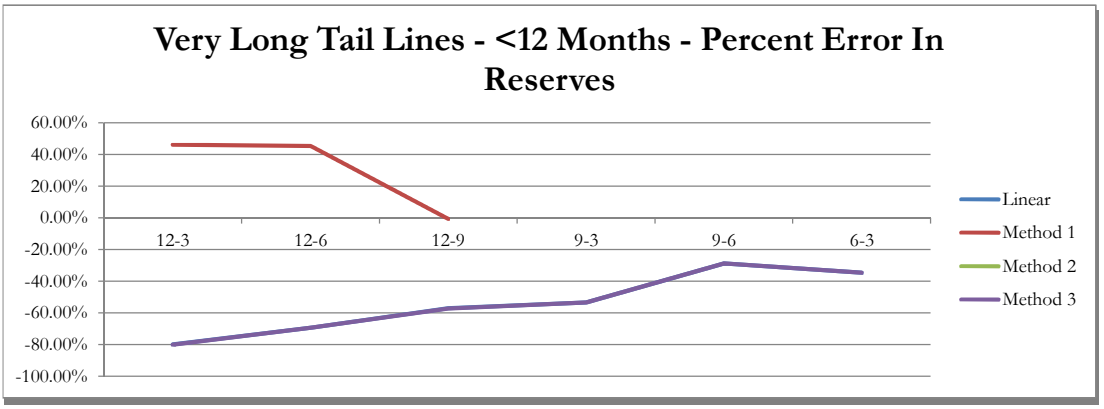


Paid Extrapolation Errors

Appendix A
Sheet 18

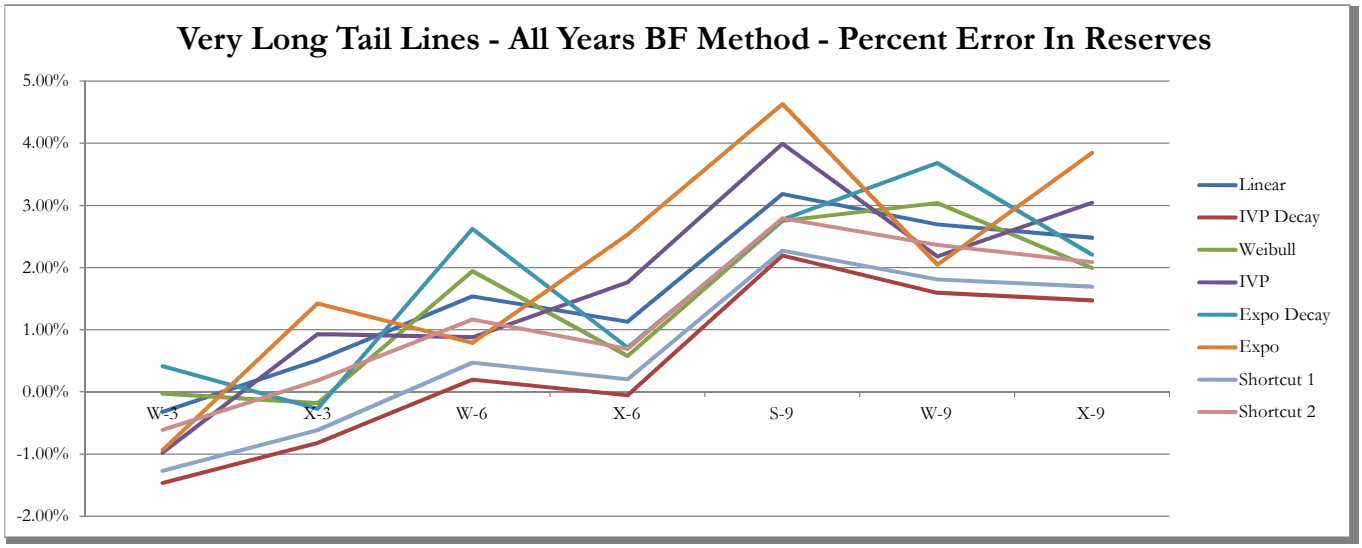


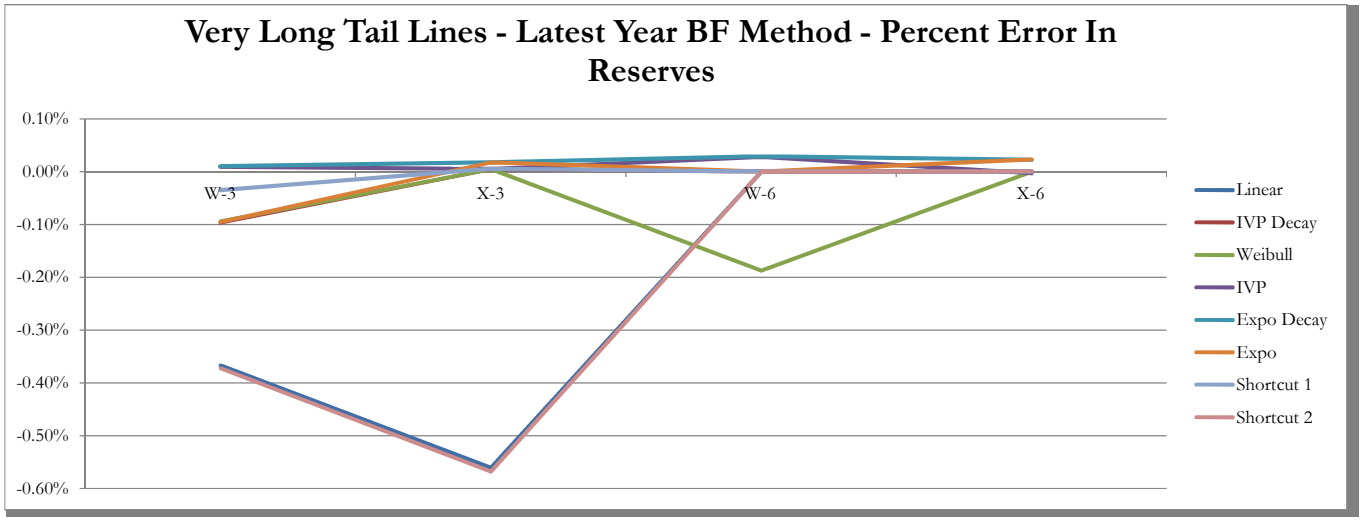


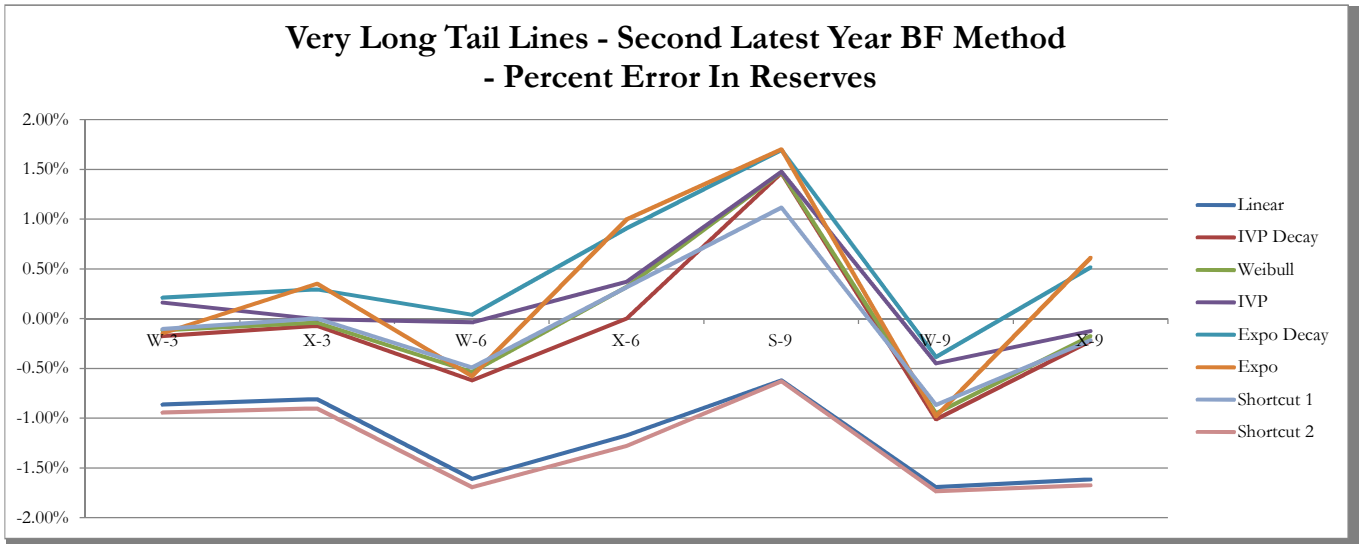


Very Long Tail Lines BF Error

Appendix A
Sheet 19

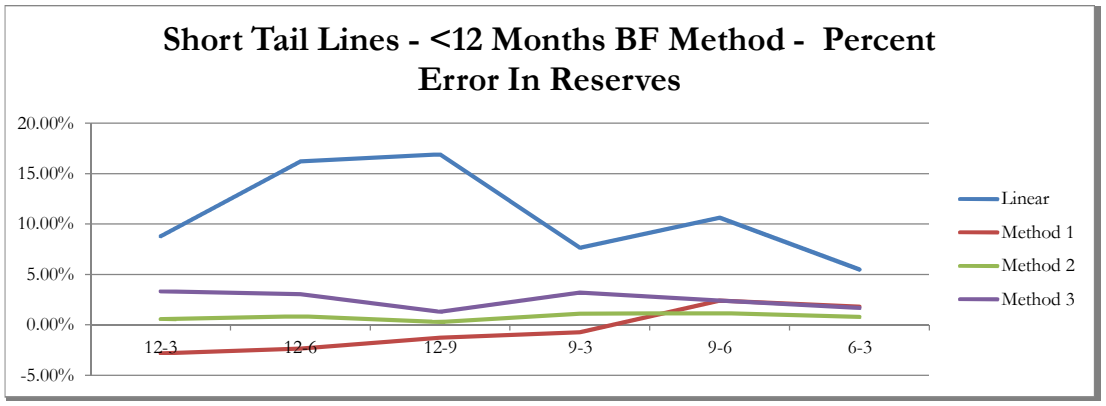


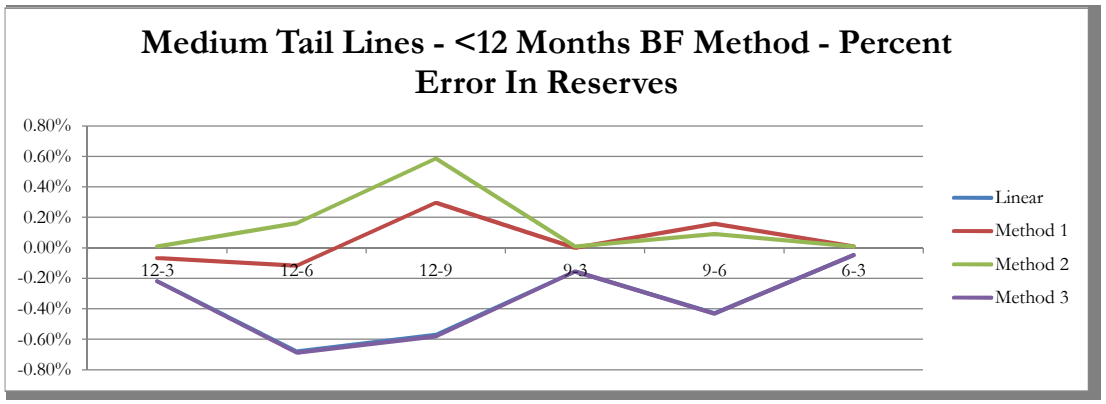


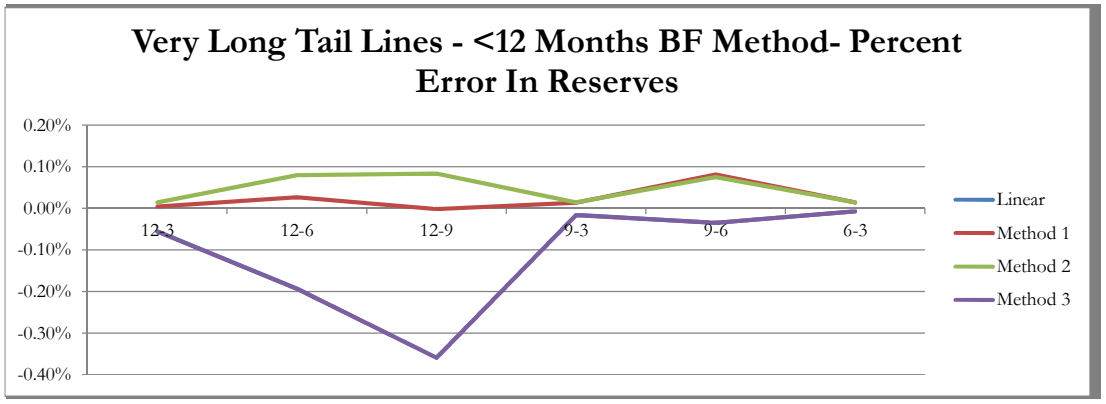


Paid BF Extrapolation Errors

Appendix A
Sheet 20







Appendix B
Sheet 1

			3Q 2014			4Q 2014			1Q 2015			2Q 2015		
Accident Year	Maturity in Months	Paid CDF 2nd Quarter	Maturity	Interpolated Paid CDF	Incremental Percent Paid	Maturity	Interpolated Paid CDF	Incremental Percent Paid	Maturity	Interpolated Paid CDF	Incremental Percent Paid	Maturity	Interpolated Paid CDF	Incremental Percent Paid
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
					1/(4) - 1/(2)			1/(7) - 1/(4)			1/(10) - 1/(7)			1/(13) - 1/(10)
2004	126	1.001	129	1.000	0.001	132	1.000	-	135	1.000	-	138	1.000	-
2005	114	1.003	117	1.002	0.001	120	1.002	0.001	123	1.001	0.000	126	1.001	0.000
2006	102	1.008	105	1.006	0.002	108	1.005	0.001	111	1.004	0.001	114	1.003	0.001
2007	90	1.018	93	1.015	0.003	96	1.012	0.003	99	1.010	0.002	102	1.008	0.002
2008	78	1.038	81	1.032	0.006	84	1.026	0.005	87	1.022	0.004	90	1.018	0.004
2009	66	1.064	69	1.057	0.007	72	1.050	0.006	75	1.044	0.005	78	1.038	0.005
2010	54	1.108	57	1.095	0.011	60	1.083	0.010	63	1.073	0.009	66	1.064	0.008
2011	42	1.191	45	1.165	0.019	48	1.143	0.017	51	1.124	0.015	54	1.108	0.013
2012	30	1.334	33	1.290	0.026	36	1.252	0.024	39	1.219	0.021	42	1.191	0.019
2013	18	2.068	21	1.780	0.078	24	1.580	0.071	27	1.438	0.063	30	1.334	0.054
2014	6	2.843	9	2.596	0.033	12	2.390	0.033	15	2.216	0.033	18	2.068	0.032
2014 Full Year	6	5.686	9	3.462	0.113		2.390	0.130		2.216	0.033		2.068	0.032

Appendix B
Sheet 2

(a) 4th quarter increase factor **50%**

	Percent of Year Paid in					Restated Percent				Restated Pattern			
Accident Year	3Q2014	4Q 2014	1Q 2015	2Q 2015	Total	3Q2014	4Q 2014	1Q 2015	2Q 2015	3Q2014	4Q 2014	1Q 2015	2Q 2015
	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)
2004													
2005	36.0%	27.4%	20.8%	15.8%	100.0%	29.2%	41.1%	16.9%	12.8%	1.002	1.002	1.001	1.001
2006	34.7%	27.2%	21.3%	16.7%	100.0%	28.2%	40.8%	17.3%	13.6%	1.007	1.005	1.004	1.003
2007	32.9%	27.0%	22.1%	18.1%	100.0%	26.8%	40.4%	18.0%	14.7%	1.015	1.011	1.009	1.008
2008	32.1%	26.8%	22.4%	18.7%	100.0%	26.2%	40.2%	18.3%	15.2%	1.033	1.025	1.021	1.018
2009	29.6%	26.3%	23.3%	20.7%	100.0%	24.3%	39.5%	19.2%	17.0%	1.058	1.048	1.043	1.038
2010	29.4%	26.3%	23.4%	20.9%	100.0%	24.2%	39.4%	19.3%	17.2%	1.097	1.080	1.072	1.064
2011	29.4%	26.3%	23.4%	20.8%	100.0%	24.2%	39.4%	19.2%	17.1%	1.170	1.137	1.121	1.108
2012	28.7%	26.1%	23.7%	21.5%	100.0%	23.6%	39.2%	19.5%	17.7%	1.297	1.241	1.214	1.191
2013	29.4%	26.7%	23.6%	20.3%	100.0%	24.1%	40.0%	19.3%	16.6%	1.826	1.529	1.418	1.334
2014	25.3%	25.2%	24.9%	24.5%	100.0%	21.1%	37.9%	20.7%	20.4%	2.635	2.329	2.189	2.068
2014 Full Year										3.513	2.329	2.189	2.068

Notes:

(15) (5)/ [(5)+(8)+(11)+(14)]
(16) (8)/ [(5)+(8)+(11)+(14)]
(17) (11)/ [(5)+(8)+(11)+(14)]
(18) (14)/ [(5)+(8)+(11)+(14)]
(20) (15)/ [(15)+(17)+(18)]*[1-(21)]
(21) (16)/ [1 + (a)]
(22) (17)/ [(15)+(17)+(18)]*[1-(21)]
(23) (17)/ [(15)+(17)+(18)]*[1-(21)]

(24) 1/[1/(2)+(20)*{1/Prior(2)-1/(2}]
(25) 1/[1/(2)+sum(20:21)*{1/Prior(2)-1/(2}]
(26) 1/[1/(2)+sum(20:22)*{1/Prior(2)-1/(2}]
(27) 1/[1/(2)+sum(20:23)*{1/Prior(2)-1/(2}]

Accident Year / Development Year Interactions

David R. Clark, FCAS, MAAA

Diana Rangelova, ACAS, MAAA

Abstract

While traditional actuarial reserving methods assume that development patterns are stable over time, changes are often observed in practice. This paper explores the reasons for these changes and surveys the most relevant literature on methods that address the changes in development patterns. Finally, the paper suggests possible research for further improvements in reserving techniques.

Keywords. Loss Reserving, Interaction Terms

1. INTRODUCTION

1.1 Research Context

Common reserving methods, such as Chain-Ladder and Bornhuetter-Ferguson, rely upon an assumption that loss development patterns are stable over time. That is, loss development patterns do not change from one accident year[1] to the next. In practice, however, reserving actuaries observe changes in these patterns and make adjustments in the use of their methods to account for the changes.

When the loss data is summarized in a triangular format, it can be analyzed from three directions: accident year (AY), development year (DY), and payment/calendar year (CY). Most reserving methodologies assume that the AY and DY directions are independent. However, many factors can create dependencies between the three directions and violate this assumption. In the actuarial literature, these dependencies are sometimes referred to as "CY effects", reflecting the fact that frequently they are caused by CY trends or shocks. A more general way to describe these effects is to say that there is some interaction between the AY and DY variables, or that there is some other confounding variable that we have not accounted for. The main danger from ignoring these changes is "omitted variable bias" in our estimated reserves.

Recently, this danger has been highlighted empirically through the test of common reserving methods on a sample of actual triangles. The Casualty Actuarial Society (CAS) has made available a database of loss triangles from Schedule P to test common reserving methods. These methods were applied mechanically, generally using all-year averages to select patterns. The results showed some

[1] The discussions and techniques presented in this paper can be easily applied to a policy/underwriting year triangle.

systematic biases in the estimates, confirming in many cases, that patterns were not unchanged over even a ten year period.

Practicing reserving actuaries have always been aware of this phenomenon, and would not naively apply an all-year weighted average without looking for changes in the pattern. Various practical approaches are used when changing patterns are observed. The simplest approach is to base the selected pattern only on the latest diagonals, ignoring the upper left portion of the triangle. This method is clearly not the optimal option, but it is an easy solution. Other practical techniques have been created, which generally try to adjust the historical data such that the triangle of adjusted losses will have consistent patterns by year and therefore allow the analyst to use more diagonals or even the whole triangle.

This call paper will survey the actuarial literature for the methods that address AY/DY interactions and will give a brief description of each of these techniques, including some numerical examples. The purpose, however, is to communicate only the major concepts. The relevant papers will be referenced so that the interested reader can find the specific calculations needed to implement the techniques. There are also more advanced statistical models which will be described in much less depth.

All of the methods presented have some limitations that will be discussed in this survey. A common theme is that the methods generally assume that there is a single cause for the changing development pattern, and that an adjustment to the triangle can be made that will make the patterns consistent over time "all else being equal." The difficulty is that often multiple types of changes have taken place over the experience history, and the practical methods may not satisfactorily handle changes from multiple causes.

1.2 Objective

The purpose of the present paper is to explore the reasons for the changes in development patterns, survey some of the relevant literature on methods that address the changes in development patterns, and suggest future research.

1.3 Outline

The remainder of the paper proceeds as follows.

Section 2 will discuss the basic reasons as to why loss development patterns are different from one year to the next.

Section 3 will provide some diagnostics for evaluating whether or not a development pattern is changing over time.

Section 4 will survey the actuarial literature for the most common methods to account for changing patterns.

Section 5 will present opportunities for future research for practical and advanced methods.

Section 6 will present our conclusions.

2. BACKGROUND

Many factors can cause the loss development patterns to change from one accident year to the next. They can be internal (e.g., shift in the mix of business, change in claim settlements procedures) or external (e.g., law changes, inflation) to the company. They can also occur alone or simultaneously, making the identification of the real cause of the change more challenging.

2.1 Internal Changes Impacting the Patterns

Internal changes impacting the loss development patterns often relate to changes in the company's business and processes that directly or indirectly impact the loss data.

The change in the mix of business, for example, can manifest itself as a change in the geographical distribution, frequency or severity level of the claims, the retention limits, the deductible levels and others. For reinsurance companies, a change in the mix of business can come from a change in virtually every clause of the reinsurance contract: program type (excess vs. quota share), quota share percentage, attachment points, excess retention and limits, special features (corridors, caps), coverage of expenses, statute of limitation, and others. The type of marketing (direct vs. broker) may cause a shift from regional accounts, that are dominant when direct marketing is used, to national accounts, which rely more on brokers. Consequently, the actuary may observe a change in attachment points, limits and reporting lags. Additionally, changes in underwriting guidance can shift the focus from a profit seeking portfolio to a growth strategy, from small to large risks, or simply to a new type of risk with different development characteristics.

Changes in a company's procedures are also a major source for pattern distortion. The change can be related to the way the initial case reserves are established or the way claims are settled. For example, the settlement of claims can be impacted by a desire to fight claims, a change in guideline on whether to prioritize large claims or small claims, or other factors that cause a speed up or a delay

in claim payments or reserve re-evaluation. A period of time with an understaffed claim department may create artificial changes in paid and reported loss development patterns. Expense related changes impacting the patterns can come from a simple change in the definition of allocated loss adjustment expenses, a shift from internal handling of claims to a Third Party Administrator (TPA), or a change in the TPA. This also creates opportunities for errors and delays in the claim processing.

Commutations can create one of the most significant pattern distortions for Schedule P loss triangles. In a typical commutation, the reinsurer transfers its current and future liability from particular ceded contracts back to the original insurer, along with an agreed upon payment. The reinsurer's loss triangles will no longer show any development for losses related to these commuted contracts. Any related reserves will be taken down and the final lump sum (or periodic payments) of the commutation "price" will be recorded as a paid loss. As a result, the ceding company will now start recording the reporting, payment, and development of these losses. Actuaries usually restate the historical loss triangles so that such transactions do not affect the development patterns. However, many industry studies and comparisons are done using Schedule P data, which is not restated for commutations. Thus, extra care must be used when applying reserving methods to Schedule P data.

Missing or incomplete loss data is a common issue for insurance and reinsurance companies. Whether due to a switch in data processing systems, a desire to start organizing the data differently (example: distinguish the medical and indemnity piece of a workers' compensation claim), or a significant delay in claim reporting, the missing or incomplete loss data compromise the reliance on historical patterns. In that case, actuaries usually exclude parts of the triangle from the analysis or try to find alternative methods to overcome this problem.

2.2 External Changes Impacting the Patterns

There are several external changes affecting the loss development patterns. One of them is related to changes in law and more specifically tort reforms. As discussed by Kerin and Israel (1998), most often, tort reforms limit the amount of damages that can be paid in total, restrict the conditions under which a damage is paid, modify the rule of evidence and change the litigation behavior. Their impact on loss payments and reserves is not easily predicted and it is also difficult to restate the historical data when significant changes occur. Examples of such reforms include no fault repeal in auto liability, caps in damage awards in medical malpractice, revised interpretation of coverage provisions and changes in workers compensation benefit laws.

Accident Year / Development Year Interactions

Another major external factor impacting the development pattern is the change in inflation. Payments are impacted by an increase in the cost of goods and services, medical costs, attorneys' fees and jury awards. Other economic and social influences may also distort the patterns. Examples include the increased workers compensation claim frequency after the 2007-2009 recession period and the reduced delays in claim reporting due to new technology.

Knowing the variety of factors that can create pattern distortions, the actuary's goal is 1) to explore the loss triangle data and identify if such distortions exist; 2) to identify what caused them; and 3) to find the appropriate reserving method to overcome these distortions. The following section provides a discussion of practical techniques that can help the actuary detect and analyze changes in loss development patterns.

3. LOSS DEVELOPMENT PATTERN DIAGNOSTICS

3.1 Examples of Practical Diagnostic Techniques

The first step in the identification of any change in the AY development pattern consists of an analysis of the triangular data. The analysis can start with a review of ratios of available loss data.

The changes in patterns can be detected directly in the loss development factor (LDF) triangle. A review of the incremental paid loss, reported loss or claim count triangles may also be helpful in identifying the effects of changes in business mix, missing data and others forces. The actuary can also look at ratio diagnostics. Cicci, Banerjee, and Jha (2011) and Friedland (2010) list the following examples of diagnostic tests:

- Paid loss to reported loss ratios
- Paid loss to on-level earned premium (or other on-leveled exposure measure)
- Reported loss to on-level earned premium (or other on-leveled exposure measure)
- Reported loss to reported counts (reported severity)
- Paid loss to closed with payment counts (paid severity)
- Case reserve to open counts (average case outstanding)
- Closed with payment counts to reported counts ratios
- Closed without payment counts to reported counts ratios

- Open counts to reported counts ratios

The ratio diagnostics are useful in identifying any of the pattern shifts discussed earlier. For example, the average case outstanding and paid loss to reported loss ratios could indicate changes in case reserve adequacy; changes in settlement rate could be revealed by any ratio involving paid losses and claim counts or the paid loss to reported loss ratio; other changes could be indicated by the closed to reported claim counts. As noted by Friedland (2010), when the diagnostic is a ratio, a signal for a change in the pattern can come from the numerator or from the denominator and it may not always be clear what is causing it. Also, a lack of a signal could be due to offsetting changes in the numerator and the denominator.

Here is an example of paid loss to reported loss ratios indicating a change in the most recent diagonals:

Table 1: Example for Ratio Diagnostics

Paid Loss to Reported Loss Ratios

AY	12	24	36	48	60
1	0.33	0.67	0.91	0.98	1.00
2	0.33	0.67	0.91	0.98	1.00
3	0.33	0.67	0.91	0.98	1.00
4	0.33	0.67	0.91	0.98	0.95
5	0.33	0.67	0.91	0.94	0.95
6	0.33	0.67	0.80	0.94	
7	0.33	0.60	0.80		
8	0.27	0.60			
9	0.27				

Paid Loss Triangle

AY	12	24	36	48	60
1	159	413	677	775	791
2	154	401	656	778	793
3	145	389	615	769	785
4	151	394	644	755	788
5	146	399	620	762	770
6	161	411	626	739	
7	158	412	556		
8	150	367			
9	113				

Reported Loss Triangle

AY	12	24	36	48	60
1	477	620	744	791	791
2	462	601	721	793	793
3	434	584	677	785	785
4	454	591	709	770	830
5	437	598	682	811	811
6	482	617	783	786	
7	473	687	695		
8	556	612			
9	420				

This example was constructed so that the two most recent diagonals show lower paid loss to reported loss ratios compared to prior diagonals. However, the reason for the shift is different for each diagonal. CY 8 was impacted by an increase in reported loss (i.e. increase in the denominator of the ratio diagnostic) that could be an indication of a case reserve strengthening. CY 9

experienced a decrease in payments (i.e. decrease in the numerator of the ratio diagnostic) that could be an indication of a slowdown in payments. In cases like this, a review of several ratio diagnostics can help isolate the effect of simultaneous changes and will provide more direction in identifying the real cause for the pattern instability.

3.2 Heatmaps

A practical tool for identifying patterns in any type of data is the heatmap, which is just a visual representation of the data, where the values are emphasized with colors. Its purpose is to reveal patterns or clusters that may not be visible without additional analysis. For example, a heatmap may be very helpful in the analysis of a large triangle with more than 20 accident and development periods, where changes in patterns may be difficult to spot through visual inspection. Heatmaps are convenient because they are easily created in an excel spreadsheet using conditional formatting.

The tables below provide examples of heatmaps. Let's take a look again at the paid and reported loss triangles from Table 1. Even without calculating the paid loss to reported loss ratio diagnostic, it is clear that both loss triangles experienced some changes. The paid triangle has a very light colored last diagonal indicating lower payments and the reported triangle has a bright colored diagonal for CY 8.

Table 2: Heatmaps of Paid and Reported Triangles

Paid Loss to Reported Loss Ratios

AY	12	24	36	48	60
1	0.33	0.67	0.91	0.98	1.00
2	0.33	0.67	0.91	0.98	1.00
3	0.33	0.67	0.91	0.98	1.00
4	0.33	0.67	0.91	0.98	0.95
5	0.33	0.67	0.91	0.94	0.95
6	0.33	0.67	0.80	0.94	
7	0.33	0.60	0.80		
8	0.27	0.60			
9	0.27				

Accident Year / Development Year Interactions

Paid Loss Triangle

AY	12	24	36	48	60
1	159	413	677	775	791
2	154	401	656	778	793
3	145	389	615	769	785
4	151	394	644	755	788
5	146	399	620	762	770
6	161	411	626	739	
7	158	412	556		
8	150	367			
9	113				

Reported Loss Triangle

AY	12	24	36	48	60
1	477	620	744	791	791
2	462	601	721	793	793
3	434	584	677	785	785
4	454	591	709	770	830
5	437	598	682	811	811
6	482	617	783	786	
7	473	687	695		
8	556	612			
9	420				

When starting an analysis, the actuary may not know in advance if there will be any data distortions. A heatmap can save time and effort by immediately focusing the actuary's attention to the problem area. Table 3 first shows a paid loss development factors triangle with changing patterns and then shows the heatmap of the same triangle. The heatmap immediately identifies that in CY 8, all AYs have larger payments when compared to other calendar years. This could be due to a speed up of payments or payments on larger number of claims that were reported with a delay. Also, the heatmap shows that the latest diagonal exhibits a much lower loss development.

Table 3: Heatmap of a Loss Development Triangle

Paid Age-to-Age Factors

AY	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120
1	2.209	1.416	1.140	1.090	1.049	1.038	1.021	1.056	1.011
2	2.050	1.313	1.180	1.095	1.058	1.040	1.061	1.014	
3	2.553	1.338	1.146	1.087	1.048	1.088	1.014		
4	2.159	1.326	1.158	1.084	1.101	1.019			
5	2.247	1.270	1.165	1.161	1.033				
6	2.395	1.311	1.375	1.025					
7	2.295	1.895	1.028						
8	4.517	1.031							
9	1.054								

Accident Year / Development Year Interactions

Heatmap of Paid Age-to-Age Factors

AY	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120
1	2.209	1.416	1.140	1.090	1.049	1.038	1.021	1.056	1.011
2	2.050	1.313	1.180	1.095	1.058	1.040	1.061	1.014	
3	2.553	1.338	1.146	1.087	1.048	1.088	1.014		
4	2.159	1.326	1.158	1.084	1.101	1.019			
5	2.247	1.270	1.165	1.161	1.033				
6	2.395	1.311	1.375	1.025					
7	2.295	1.895	1.028						
8	4.517	1.031							
9	1.054								

3.3 Limitations of the Diagnostics

These diagnostic tests are useful in identifying whether a problem exists in the triangle, either from changing patterns or due to missing data. Often the diagnostics cannot identify exactly what the problem is (as seen in the example with the paid loss to reported loss ratio). Some of the changes listed in Section 2 do not create sharp changes in the triangle, but rather gradual shifts over time. This makes it difficult for the analyst to hone in on the problem, or even to determine which dimension (e.g., accident year versus payment year) is involved.

For example, if our triangle is actually a combination of two types of risks – one with quick development and a second with slow development – and the mix is changing over time, then a changing development pattern will be observed. Our diagnostic tests will be unable to distinguish this mix problem from other possible causes such as, say, calendar year trend or changes in claim settlement practices.

Accident Year / Development Year Interactions

Table 4: Example of Diagnostic Limitations

Slow Developing Policies with Growing Volume
Loss Triangle

AY	12	24	36	48	60
1	100	200	250	275	290
2	300	600	750	825	
3	500	1000	1250		
4	700	1400			
5	900				

Age-to-Age Factors

AY	12-24	24-36	36-48	48-60
1	2.000	1.250	1.100	1.055
2	2.000	1.250	1.100	
3	2.000	1.250		
4	2.000			

Quick Developing Policies with Shrinking Volume
Loss Triangle

AY	12	24	36	48	60
1	900	1350	1395	1395	1395
2	700	1050	1085	1085	
3	500	750	775		
4	300	450			
5	100				

Age-to-Age Factors

AY	12-24	24-36	36-48	48-60
1	1.500	1.033	1.000	1.000
2	1.500	1.033	1.000	
3	1.500	1.033		
4	1.500			

All Policies Combined
Loss Triangle

AY	12	24	36	48	60
1	1000	1550	1645	1670	1685
2	1000	1650	1835	1910	
3	1000	1750	2025		
4	1000	1850			
5	1000				

Age-to-Age Factors

AY	12-24	24-36	36-48	48-60
1	1.550	1.061	1.015	1.009
2	1.650	1.112	1.041	
3	1.750	1.157		
4	1.850			

This is an example of Simpson's Paradox, as described in more detail in Stenmark and Wu (2004). The "paradox" is that the sub-portfolios each have patterns that are unchanging and perfectly stable over time, but the changing mix gives an appearance of a changing pattern for the combined business. This phenomenon occurs frequently in insurance applications because data are aggregated to produce more credible volumes, and that aggregation means that the data are no longer truly homogeneous; conversely, when data is broken out into smaller homogeneous pieces, it is no longer easy to see the signal hidden in the noise.

A practical example is US Workers' Compensation loss development. The development patterns are different for medical and indemnity coverages, with medical coverage generally having a longer development tail. Over time, the portion of losses in the medical coverage has been growing. Even if the patterns for medical and indemnity were each stable on their own, the combined triangle would, all else being equal, show a slowing development pattern.

The triangle may therefore show that something is changing, but at an aggregated level the actuary will be unable to identify the nature of that change. This is sometimes referred to as the problem of "lurking" or a "confounding" variable. The unidentified confounding variable is not explicit in the model and manifests as an AY/DY interaction.

3.4 Communication

Once the actuary has detected a change in the pattern, he or she needs to investigate what caused it. Knowing the source of the problem is important because it provides a better insight into what pattern to expect in the future. It tells us what data we can trust and what data we need to adjust (example: if the paid loss to reported loss ratio is distorted it is necessary to know whether it is the paid or the reported data that experienced a change). Finally, it helps the actuary decide which reserving method to use.

As we have seen the diagnostics may be misleading. Even in the most obvious case of distortion, the actuary needs to confirm his or her findings with other parties involved in the data processing who may be closer to the source of change. Berquist and Sherman (1977) and Friedland (2010) provide questions that can help the actuary investigate and confirm the change in data through communication with other departments of the (re)insurance firm. For example, the actuary can ask a claim executive if there have been any recent significant changes in the guidelines for setting and reviewing the unpaid case reserves. A question for the underwriters could focus on the shift of business by territory or by type of distribution (direct vs. brokerage distribution). These types of conversations can provide insights into the causes of the pattern distortion. More importantly, they

can lead to additional information that can help quantify the total impact of the pattern change.

The important note to keep in mind is that, even with the best intention to collaborate, the other party may not have noticed the change or may not be willing to recognize an event as the source for pattern distortion (e.g., a case reserve weakening may not be easy to admit to the auditing actuary). In a presentation at the CAS 2007 Casualty Loss Reserving Seminar, Richard Sherman cautioned the audience to "beware of quick, slick answers" that are designed to bias the investigation of the cause of pattern changes. He also raised awareness of the importance of carefully selecting parties who will be able to provide the actuary with the most valuable information. Benefits could be found in a conversation with the most knowledgeable party (for example the department executive) or with the less biased party (for example a middle level staff).

4. CURRENT METHODS TO ACCOUNT FOR CHANGING PATTERNS

The fact that patterns can change over time due to a variety of reasons is well-known. A number of practical approaches are used by reserving actuaries to account for these changes.

Some of these approaches consist solely of data rearrangements and no method changes. They rely on additional data that can eliminate or explain the changes in the patterns. Berquist and Sherman (1977) discuss two means of obtaining data that is relatively unaffected by a given problem:

1) Data substitution – for example: the use of earned exposure in place of claim count when count data is disrupted. Earlier, it was noted that net data patterns can be easily distorted by changes in the reinsurance structure. In this case, the actuary may use the data substitution technique and rely on gross data. This approach relies on the assumption that the substitute data is available.

2) Subdivision of data into homogeneous groups of exposures – for example: when there have been changes in the mix of business. The actuary must be careful however of the decrease in credibility associated with the data split.

The most common approach currently used by actuaries is to eliminate distorted data. For example, when the actuary observes changing age-to-age factors down the columns of a triangle, he or she will make use of the latest few diagonals and ignore the earlier factors in the upper left corner of the triangle. This may be considered a default "only the latest diagonals" (OLD) method. This

approach not only diminishes the statistical accuracy of LDF averages (they will be based on only a few points) but it will also affect the credibility of any reserve variability estimates. In other words, the actuary should be looking to use more data, not less.

The methods described in the five sections below improve on this in several ways. This survey of the actuarial literature will briefly describe methods for handling CY trends, changes in case reserve adequacy, changes in settlement rates, and missing data problems. Finally, statistical models will be discussed at a high level.

4.1 Calendar Year Trends

As noted above, a basic assumption of the Chain-Ladder method is that the columns of a development triangle are proportional to each other. Taylor (1977) notes that this assumption holds when "exogenous influences" such as monetary inflation and mix of business are relatively stable. But he also notes that:

> "It is crucial to the logic underlying the Chain-Ladder method that the 'exogenous influences' should not be too great. If this assumption does not hold, then the conclusion, that the columns of the run-off triangle are proportional, goes away too, and the Chain-Ladder method can give misleading results."

Taylor provides a "separation" method to isolate the calendar year effects from the development year effects. In order to apply this method, we need a development triangle of paid losses and an exposure base of claim counts by accident year. As Taylor states, getting a good estimate of ultimate counts by year can be "problematic" but we will assume here that it is available.

The separation method as outlined by Taylor requires that we distinguish the frequency and severity components within each accident year, so that changes along the diagonal can be assumed to be due to severity effects only. This requires that the triangle be adjusted such that each row represents the average severity rather than the aggregate loss dollars. To make this adjustment, we divide each row by an estimate of its ultimate frequency.

The accuracy of the separation method depends upon getting quality claim count information. For the example below, we will assume that all numbers have been adjusted to a common level, but in practice this assumption needs care.

Accident Year / Development Year Interactions

Table 5: Separation Model – Example

Cumulative Payments

AY	12	24	36	48	60
1	500	1,106	1,530	1,764	1,903
2	505	1,141	1,608	1,888	
3	530	1,230	1,790		
4	583	1,423			
5	700				

Age-to-Age Factors

AY	12-24	24-36	36-48	48-60
1	2.212	1.384	1.152	1.079
2	2.260	1.409	1.174	
3	2.320	1.455		
4	2.440			

Taylor gives a direct algebraic method for calculating a CY or payment year trend factor from this data. The method does not require any iterative optimization routines or special software, so it can be performed in a simple spreadsheet.

The implied trends by payment year are 1.0%, 5.0%, 10.0%, and 20.0%, which apply to incremental payments in the triangle. In this example, we have deliberately made sharply increasing trends so that the resulting increasing age-to-age factors in each column are obvious. If the payment year trend is constant, then no change in age-to-age factors would be observed.

Table 6: Separation Model – Example (cont.)

AY	Trend	CY Index
1		1.000
2	1%	1.010
3	5%	1.061
4	10%	1.167
5	20%	1.400

Index by Calendar Year

AY	Year 1	Year 2	Year 3	Year 4	Year 5
1	1.000	1.010	1.061	1.167	1.400
2	1.010	1.061	1.167	1.400	
3	1.061	1.167	1.400		
4	1.167	1.400			
5	1.400				

These CY Index factors are used to de-trend the incremental losses in the nominal triangle. The

de-trended incremental losses are then accumulated by accident year to produce an inflation-free triangle. In the idealized example, this inflation-free triangle produces age-to-age factors that are constant down each column.

Table 7: Separation Model – Example (cont.)

Cumulative Payments

AY	12	24	36	48	60
1	500	1100	1500	1700	1800
2	500	1100	1500	1700	
3	500	1100	1500		
4	500	1100			
5	500				

Age-to-Age Factors

AY	12-24	24-36	36-48	48-60
1	2.200	1.364	1.133	1.059
2	2.200	1.364	1.133	
3	2.200	1.364		
4	2.200			

Future losses, estimated by completing the lower right portion of the data, then need to be put onto a nominal basis using an assumption about the future inflation.

Taylor notes that this method gives a good estimate so long as the change in patterns is due to a payment year effect, which is "particularly appropriate when claim costs are dominated by high rates of inflation." He goes on to caution that there may be other causes of changing patterns that would not be appropriately addressed by this method: "It is not so appropriate in respect of influences such as changing mix of business within a risk group, which is related rather to policy year." As we noted earlier, it is not easy to diagnose from the data what is causing the patterns we see, so investigation beyond the triangle is needed.

This method does have limitations though. We need a reliable measurement of counts as well as dollars, and the reserve estimate is dependent upon our ability to forecast the CY trend index into the future. In addition, this method applies only to paid loss data, and is not directly applicable to case incurred losses. Even with these limitations, however, it is an improvement over the OLD method, because it uses the entire triangle and not only the latest diagonals.

The use of calendar year trends has been advanced in several papers in the actuarial literature. Butsic (1981) produced a similar model to that of Taylor, adding interest rate discounting in the reserve. Barnet and Zehnwirth (2000) show how a calendar year trend can be estimated in a log-

linear regression model. Gluck and Venter (2009) give a survey of the literature to 2009, especially with regard to more advanced statistical models.

4.2 Case Reserve Adequacy

When an actuary sees changes in reported losses, it is important to investigate what the real cause for these changes is. Given that the consistency of the reported incurred loss data depends not only on stable average case reserve, but also on stable claim reporting, and stable average payments, we can easily see that a pattern distortion may be due to changes in any (possibly multiple) of these three elements. Depending on what the real source of the disruption is, different data adjustments may be appropriate.

When faced with changes in case reserve adequacy, the actuary may be able to perform exact adjustments to the case reserves if they are set by formula (e.g., workers' compensation indemnity tabular reserves). In these cases, the system can re-evaluate the case reserves using current assumptions on mortality or interest rates and produce an "as if" triangle using the more recent assumptions.

In situations for which we are uncertain of the reasons for changes in case reserve adequacy, Berquist and Sherman (1977) provide a method for making an appropriate adjustment. That approach is nicely described by Duvall (1993), as follows:

> "Given a shift in reserving practices, the Berquist-Sherman adjustment for the shift begins by obtaining the rate of inflation in average closed claims. Next, the average reserve at the most recent valuation date is calculated for each year. These average reserves are trended back to earlier valuation dates at the estimated trend rate to obtain the average reserve at each age for each year in the experience period. The computed average reserves are then multiplied by the number of open claims at each age to get the estimated cost of open claims. Cumulative claim payments are then added to get an estimate of incurred losses on a basis that is consistent with current reserving practice."

Thorne's (1978) discussion of the Berquist and Sherman method points out the difficulty and actuarial judgment involved in the selection of the severity trend used to trend back the most recent average reserves.

One way to determine, or at least confirm, the severity trend selection is to use Duvall's (1992) regression technique. Duvall's model has two purposes: 1) to detect shifts and trends in the loss development factor parameters and, if a change is observed, 2) to provide an objective way to restate

the reported incurred losses for early valuations on a basis that is consistent with recent valuations. The first step in his model is to present the reported incurred loss as a function of the number of claims, the average claim cost and the loss development factor at each valuation date. Next, for each of these factors, Duvall specifies a regression function and estimates the parameters using the triangular data. He states:

> "The LDF function is central to the objective of this paper. Changes in reserving practices must be manifest in changes in the parameters of this function if they are to be detected. Therefore, it is important that the function be capable of providing an excellent fit to the observed development patterns."

The estimates from this regression model can be used as an objective way to determine a severity trend and restate the recent reported incurred losses to earlier valuations on a basis that is consistent with the current valuations. This approach can also be applied in cases where we have a change in settlement rates.

4.3 Changing Settlement Rates

Berquist and Sherman (1977) also present a method for reducing the impact of changes in settlement rates by adjusting the cumulative closed claim and paid loss triangles.

The method starts with a review of disposal rates. The disposal rate can be seen as a type of ratio diagnostic. It is defined as the cumulative closed claim counts for each accident year and maturity, divided by the ultimate claim counts. A change in the disposal rate pattern is an indication of a change in the rate of claim settlement. Next, a representative disposal rate pattern is selected (for example the most recent diagonal) and it is assumed to be valid for all accident years. The adjusted closed claim counts are obtained by multiplying the selected disposal rate by the ultimate claim counts. The method approximates the relationship between the paid losses and the closed claim counts, before any adjustments, with a function. It then uses this relationship to obtain the adjusted paid losses based on the adjusted closed claim triangle.

Thorne's (1978) comments on this technique are that "lack of recognition of the settlement patterns by size of loss can be an important source of error" and "it may be necessary to modify the technique to apply to size of loss categories adjusted for 'inflation' ". Exhibit I of his discussion paper provides an example of how a shift in claim settlement (from small to large claims) increases the error in the reserves estimates.

Accident Year / Development Year Interactions

Fleming and Mayer (1988) propose a variation of the Berquist-Sherman method where the adjustment is made not only to the paid losses but also to the outstanding losses. The procedure is described in pages 196 -199 and an example is given in Exhibit 5 of their paper.

As with the other methods described, this method can only be applied if reliable count data is available. This can be a challenge because counts can be compiled differently over time (e.g., the treatment of closed-without-pay claims). Counts can also be distorted by accident year changes. For example, a small increase in deductibles can greatly reduce claim counts, giving the appearance of a slow-down in settlements. Similarly in Workers' Compensation, a change in the states or industries covered can alter the mix of "medical only" versus "lost time" claim counts, giving a misleading impression of claims handling practices. These types of "confounding variable" need to be investigated before the methods are applied.

The change in settlement rates can be also addressed with a Bayesian model. More details of this technique will be provided in Section 4.5.4.

4.4 Incremental Development

Often, when data are missing for older accident years or when changes in definitions or mix of business have made it inappropriate to combine the data in a cumulative triangle for the purpose of the reserve estimation, the general practice is to "cut" the triangle and work only with accident years that are not distorted but contain data for all maturities. Throwing away the data is not an optimal solution. Instead, the actuary can make use of any non-distorted incremental data from old accident years.

The Sherman-Diss (2004) paper describes the Mueller Incremental Tail (MIT) method that can help achieve this goal. This method works for triangles that are missing values in the upper left corner, but have incremental amounts for the more mature years. The method consists of three steps:

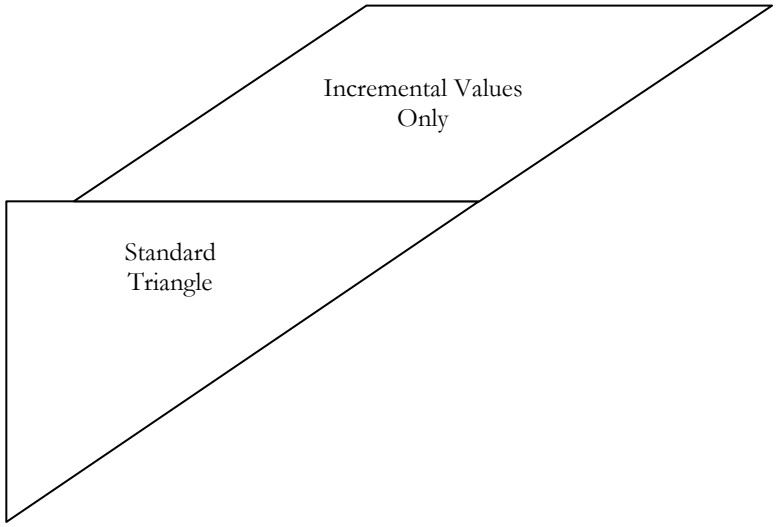
1) Calculation of incremental age-to-age factors for all available data. This is done by taking the ratio of incremental paid at age n+1 to incremental paid at age n

2) Calculation of an anchored decay factor representing the incremental payments made in year n relative to payments made in an anchor year. For example, it calculates the payments for years 16 to 37 relative to the incremental payments in year 15. The sum of the decay factors for years 16 to 37 can be viewed as a "cumulative decay factor" relative to year 15;

3) Calculation of a tail factor: The cumulative decay factor is then combined with a traditional

age-to-age factor for year 14 to 15, based on more recent data, to create a full cumulative loss tail factor.

Figure 1: Graphical representation of the MIT method

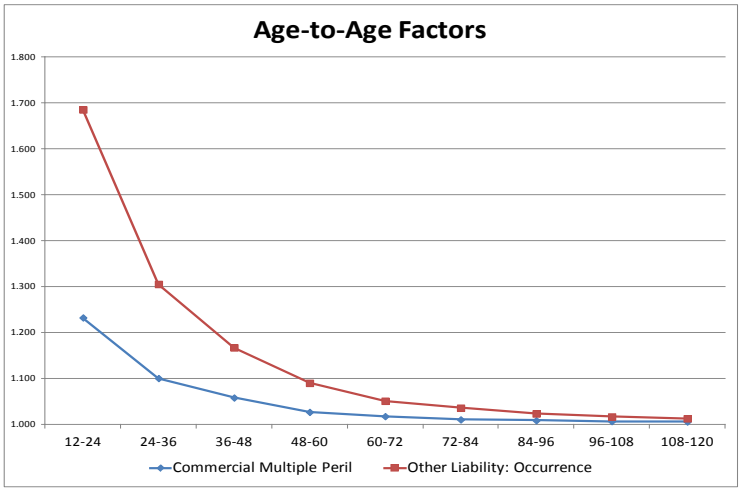


The MIT method was originally created for incremental payments on long-tailed Workers' Compensation losses in a database that did not include payments for early periods (the upper left triangle). In this case, the data was missing because it was not available to the analysts. However, the technique could also be applied if the early payments were missing because they were below a self-insured retention. This suggests that the incremental method may be useful when there is a changing mix of primary and excess business in a portfolio or when commutations are not excluded from the data.

An interesting application of this method could be made in situations where the mix of business is different by accident year. To illustrate this, we can look at two triangles from consolidated industry Schedule P. The loss triangles and the calculation of the "normal" and anchored loss development patterns are provided in Appendix I. The development patterns for Other Liability (occurrence) and Commercial Multiple Peril (CMP) are quite different. Other Liability includes monoline, ground-up losses along with some losses from excess and umbrella policies. CMP includes losses from property as well as from liability. As we would expect, a much larger percentage of total losses are reported within the first few years for CMP. This early loss reporting acts as a ballast to reduce the age-to-age factors.

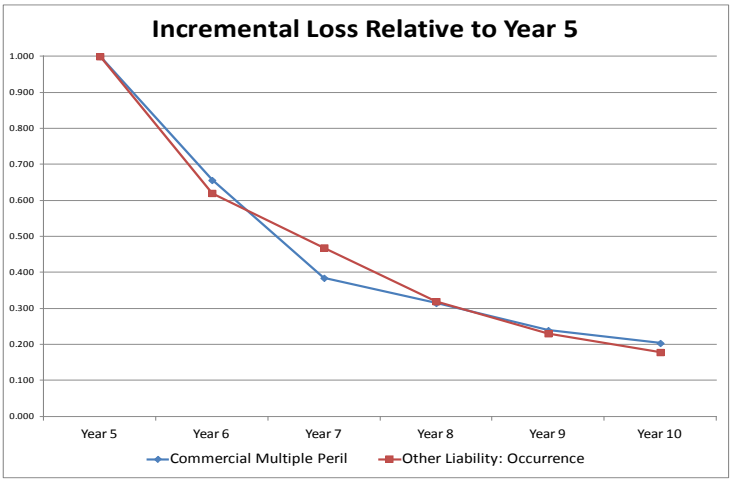
Accident Year / Development Year Interactions

Figure 2: CMP vs Other Liability Loss Development



However, if we use the Mueller Incremental Tail (MIT) concept, then instead of age-to-age factors on cumulative losses, we anchor the factors at a later age. Figure 3 shows the incremental reported amounts relative to the amount incurred in the fifth year. These patterns look much more similar, implying that the losses contributing to the "tail" may be similar in both triangles.

Figure 3: CMP vs Other Liability Relative Incremental Loss Development



This approach is not recommended as an alternative to segregating the data into homogeneous groups. However, it does suggest that the incremental method may be useful on triangles where the

changing mix of business cannot be fully identified in the historical triangle, or where the credibility of the data will be significantly impaired by the split of the data.

4.5 Statistical Models

The methods listed above have been designed so that the actuary can adjust the development triangle and make selections from the data without the need for special software. The transparency of assumptions and ease of calculation are clear advantages.

The methods do not make explicit assumptions about the variances or shape of the random variables that give rise to the observed data. For this reason, it is difficult to evaluate the results in terms of whether the final estimates are "best" (unbiased, minimum variance) estimates, or whether the adjustments are based on significant signals or could have been produced by random noise. Assumptions that are never made explicit are, by definition, untestable.

The main hurdle to implementation of a statistical model is the learning curve required to master the concepts and software. The statistical models listed below are roughly in order of difficulty in the learning curve required.

4.5.1 Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are a generalization of linear regression models that allow for much greater flexibility in the relationship between the explanatory variables and the response variable being forecast, and in the variance structure of that response variable. The recent text "Predictive Modeling Applications in Actuarial Science" includes a good description of GLM (Dean 2014), along with the connection to reserving (Taylor 2014).

The use of GLMs for reserving was first suggested by Wright (1990), but was given very clear exposition by Renshaw and Verrall (1998). The initial observation was that the traditional Chain-Ladder method was actually a GLM model, and therefore making that explicit allowed for statistical tests and variance calculations to be performed easily. The GLM framework also allows for the introduction of exposure measures, market cycles, and calendar year effects to be included. Taylor's "separation method" discussed above, as well as the MIT incremental method, are special cases in the same GLM.

Perhaps the greatest benefit of a GLM perspective is that interactions between the accident years and development years can be explicitly recognized and included in the model. Taylor (2014) gives a good introduction to the use of GLMs in reserving, including the introduction of interaction terms.

4.5.2 Hierarchical or "Mixed" Models

Generalized Linear Models can also be helpful to account for changes in the mix of business because they allow us to look at multiple triangles simultaneously. For example, two triangles may represent different businesses that both include general liability exposures; they may have different development patterns but share the same sensitivity to inflation changes. A GLM can estimate some parameters separately for each data set and some parameters which are common across data sets.

However, if the data is split into many triangles, then it may be impractical to estimate parameters for all of the components separately. This is where hierarchical models (also known as mixed or multi-level models) can be introduced.

A good example of the application of mixed models is given by Schmid (2012). He was looking at residual market triangles for Workers' Compensation. These pools are segregated by state and can have very different volumes by policy year as business shifts between voluntary and involuntary placement, resulting in several triangles with similar – but not identical – patterns. The hierarchical approach allows for separate parameters to be estimated for each state pool, but also controlled such that the parameters for any one state could not be too far apart from some overall average. This "total credibility" approach allows reserving for each pool, while also borrowing strength from the larger sample of triangles.

Guszcza (2008) introduces the use of hierarchical models for reserving, allowing individual accident year patterns to deviate from some overall average pattern.

4.5.3 Models Using Detailed Data

A more extreme case of segregating the reserving data is to go down to the individual claim level detail. Guszcza and Lommele (2006) describe the advantage of this approach because it would automatically capture the mix of coverages, types of losses, and changes in policy limits. They note that "A danger of using summarized loss triangles is that they can mask heterogeneous loss development patterns."

The danger that Guszcza and Lommele describe is another form of Simpson's Paradox, and is a result of using the highly aggregated data provided in the traditional development triangle.

This point has been made by England and Verrall (2002):

> "…it has to be borne in mind that traditional techniques were developed before the advent of desktop computers, using methods which could be evaluated using pencil and paper. With the continuing increase in computer power, it has to be questioned whether it

would not be better to examine individual claims rather than use aggregate data."

In many cases, however, the individual claim data may not be readily available. Reinsurers, for example, would not have access to the individual claim-level data from ceding companies. A compromise between complete aggregation and micro-level reserving might be a model that uses treaty-level data.

The use of individual claim level data also introduces the problem that late reported or "incurred but not yet reported" (IBNYR) claims must be modeled separately.

4.5.4 Bayesian Models

Some of the recent literature on statistical modeling in loss reserving has proposed the use of Bayesian models.

Bayesian models allow (in fact, require) the user to specify prior knowledge of development factors and variables influencing the development patterns. The prior knowledge takes the form of a distribution of model parameters that is revised as actual loss data is observed. Increased computer speed and the availability of Markov Chain Monte Carlo (MCMC) simulation techniques have made the models more accessible to actuaries.

The key advantage of these models is that even very complex non-linear interactions between accident year and development year dimensions can be evaluated. If the prior distributions are set meaningfully, the models can also work with relatively sparse data sets and still produce useful information. Meyers (2015) shows that a non-linear "growth" function can include a non-linear interaction term, which he termed the "Changing Settlement Rate" (CSR) model, and found that it was able to correct bias in some of the data sets he reviewed.

There are still challenges to making Bayesian models fully accessible to reserving actuaries. First, they require "prior knowledge" about patterns to be explicitly incorporated as multivariate distributions of parameters. These prior distributions are not trivial to create. Second, the MCMC algorithm needs to be calibrated and monitored by the user to ensure that the results have truly converged to approximate the posterior distribution.

5. OPPORTUNITIES FOR FUTURE RESEARCH

This call paper has been intended as a brief survey of existing literature on methods addressing changes in development patterns over time.

We have seen that most of the methods are limited in that they assume "all else being equal" from other effects. In other words, the techniques may not be reliable if more than one type of change is taking place simultaneously. If policy limits written are changing, or business mix is shifting from manufacturing to service industry risks, and at the same time case reserve adequacy is changing, then we have no available methods for correctly adjusting the data. In technical language, this is an example of a misspecified model and can lead to biased results.

The way forward is to recognize that all of the factors that cause patterns to change can be viewed as different types of interaction terms. Viewed in terms of a regression model, our explanatory variables are the accident year and development year indices; the traditional Chain-Ladder model assumes that these two variables act independently. Instead, we need to include models that allow for interactions between these two explanatory variables. The exact form of this interaction may be different based on the cause of the change (mix of business, CY inflation, reserving practices, etc.), but they all fall under this concept.

We have also seen that identification of the cause of changing patterns is problematic when only highly aggregated triangles are available. More data such as information about the mix of business may be needed to help understand how and why the accident year and development year dimensions are not independent. In most cases this is done judgmentally with few practical suggestions in the literature as to how it can be quantified objectively.

While "technical" models such as GLM or Bayesian MCMC have begun to move in this direction, they have yet to allow for full flexibility in the types of interactions or – more importantly – to provide a friendly user interface for the average reserving analyst.

Some concrete suggestions on moving this forward:

1) Identify the types of additional data needed for evaluating pattern changes, such as
    a) Historical policy limit and risk profiles
    b) Historical rate change indices
    c) Inflation and benefit change indices

2) Advance research on models that can look at multiple triangles, potentially down to the individual claims level; focus on practical implementation.

3) Create a library of the form of interactions appropriate for different factors, such as

   a) Calendar year trend is the simplest interaction term as in GLM

   b) Glenn Myers monograph on the speed-up is a good start on settlement patterns.

6. CONCLUSIONS

In this call paper we have seen that there are a number of reasons that development patterns can be different from one accident year to the next. These include calendar year trends, changing settlement patterns, changing case reserve adequacy, changing mixes of business, and others. All of these produce triangles in which the AY and DY dimensions are not independent, but instead show interactions. This violates a basic assumption of the Chain-Ladder method.

We have surveyed several practical methods for addressing these interactions. The methods can be as simple as ignoring portions of the triangle, or adjusting the historical data for known changes. These methods have proven useful to reserving actuaries because they are easy to implement, but also because they are tied to the reasons that patterns are changing, and therefore, help to give a more complete story for the reserve estimate.

However, most of these tools depend upon knowing *a priori* what adjustments need to be made to the data, and then restating the development triangles to current cost levels (or case reserve adequacy, or settlement rate) using a reliable measure of claim counts. Having reliable counts is necessary, but it is the assumption that we know the cause of the changing patterns that is most critical. If multiple changes are happening simultaneously – for example, a change in policy limits as well as a change in case reserve adequacy – then the methods will fail.

The long-term improvement in reserving models points us to the use of more data: including more detailed loss statistics, policy limit profiles, measures of exposure, and external indices such as cost inflation. Statistical modeling is the recommended framework for bringing in additional information.

For those building statistical models, the challenge is to make the models more accessible to the practicing actuary, including the flexibility to allow clear intervention points where the knowledgeable actuary can adjust the intermediate results when needed. Statistical models may be

better absorbed by practicing actuaries if they can easily incorporate adjustments such as changes in case reserve adequacy or claim closure rates.

Abbreviations and notations:

AY, accident year (row dimension of triangle)
BF, Bornhuetter-Ferguson method
CL, Chain-Ladder method
CY, calendar year
DY, development year (column dimension of triangle)
GLM, generalized linear models
GLMM, generalized linear mixed models
MIT, Mueller Incremental Tail method
OLD, only the latest diagonal(s) method
TPA, Third Party Administrator

Acknowledgment

The authors gratefully acknowledge the helpful reviews and feedback from Ira Robbin, Brian Archdeacon, Caroline Ferrara, Tho Ngo, Andy Kirtland, Kenneth Easlon, Chad Schlippert, and Arlie Proctor. All errors remain the responsibility of the authors.

Biographies of the Authors

David R. Clark is a senior actuary with Munich Reinsurance America Inc., working in the Actuarial Research and Modeling team. He is a Fellow of the CAS and a member of the American Academy of Actuaries. He received the 2003 Reserves Call Paper prize for the paper "LDF Curve-Fitting and Stochastic Reserving: A Maximum Likelihood Approach."

Diana Rangelova is a senior actuarial manager with Munich Reinsurance America Inc., working in Corporate Reserving. She is an Associate of the CAS and a Member of the American Academy of Actuaries. She is also a Member of the French Institute of Actuaries.

Accident Year / Development Year Interactions

Appendix I-A

Commercial Multiple Peril

Cumulative Incurred Loss+ALAE

	12	24	36	48	60	72	84	96	108	120
1974-2003					…					
2004	10,062,877	12,019,810	13,120,330	13,800,717	14,048,563	14,228,194	14,343,555	14,483,141	14,548,973	14,631,706
2005	10,807,279	13,426,225	14,338,169	15,034,751	15,342,112	15,537,527	15,597,310	15,688,265	15,790,959	
2006	9,497,881	11,602,993	12,651,289	13,302,004	13,668,841	13,806,668	13,938,202	14,051,374		
2007	10,595,875	12,728,205	13,930,135	14,601,754	14,888,062	15,094,464	15,193,434			
2008	14,050,047	16,969,555	18,106,007	18,766,853	19,210,347	19,408,707				
2009	11,339,648	13,883,191	14,991,666	16,232,739	16,558,188					
2010	12,302,948	14,937,969	16,189,574	16,944,556						
2011	15,602,014	18,354,523	19,759,192							
2012	13,342,603	16,316,017								
2013	11,939,724									

Age-to-Age Loss Development Factors[1]

	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120
All Year Weighted Avg	1.232	1.100	1.058	1.026	1.017	1.010	1.008	1.006	1.005

Incremental Triangle

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
1974-2003					…					
2004	10,062,877	1,956,933	1,100,520	680,387	247,846	179,631	115,361	139,586	65,832	82,733
2005	10,807,279	2,618,946	911,944	696,582	307,361	195,415	59,783	90,955	102,694	
2006	9,497,881	2,105,112	1,048,296	650,715	366,837	137,827	131,534	113,172		
2007	10,595,875	2,132,330	1,201,930	671,619	286,308	206,402	98,970			
2008	14,050,047	2,919,508	1,136,452	660,846	443,494	198,360				
2009	11,339,648	2,543,543	1,108,475	1,241,073	325,449					
2010	12,302,948	2,635,021	1,251,605	754,982						
2011	15,602,014	2,752,509	1,404,669							
2012	13,342,603	2,973,414								
2013	11,939,724									

Age-to-Age Loss Development Factors Anchored to Year 5

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
1974-2003					…					
2004					1.000	0.725	0.465	0.563	0.266	0.334
2005					1.000	0.636	0.195	0.296	0.334	
2006					1.000	0.376	0.359	0.309		
2007					1.000	0.721	0.346			
2008					1.000	0.447				
2009					1.000					

0.465 = Incremental Loss in Year 7 (115,361) / Incremental Loss in Year 5 (247,846)

Anchored Loss Development Factors[2]

	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
All Year Weighted Avg	1.000	0.656	0.384	0.314	0.239	0.203

1. The average age-to-age loss development factors were used in Figure 2
2. The anchored age-to-age loss development factors were used in Figure 3

Appendix I-B

Other Liability

Cumulative Incurred Loss+ALAE

	12	24	36	48	60	72	84	96	108	120
1974-2003					…					
2004	4,490,851	6,542,233	8,412,758	9,851,474	10,929,582	11,596,339	11,830,132	12,126,474	12,415,106	12,510,225
2005	4,314,808	6,802,700	8,537,640	10,029,125	10,959,584	11,312,724	11,735,144	12,152,153	12,429,060	
2006	4,442,035	7,156,625	9,360,538	10,885,689	11,757,797	12,400,293	13,002,514	13,236,483		
2007	4,555,135	7,646,174	9,975,072	11,537,284	12,772,506	13,788,485	14,272,563			
2008	4,068,513	7,126,421	9,490,827	11,206,561	12,664,704	13,367,708				
2009	4,096,903	7,034,898	9,119,711	10,963,616	12,109,594					
2010	3,752,463	6,649,357	9,213,646	10,933,500						
2011	3,670,262	6,637,029	9,098,742							
2012	3,571,801	6,543,045								
2013	3,584,497									

Age-to-Age Loss Development Factors[1]

	12-24	24-36	36-48	48-60	60-72	72-84	84-96	96-108	108-120
All Year Weighted Avg	1.686	1.305	1.167	1.090	1.051	1.035	1.023	1.016	1.012

Incremental Triangle

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
1974-2003					…					
2004	4,490,851	2,051,382	1,870,525	1,438,716	1,078,108	666,757	233,793	296,342	288,632	95,119
2005	4,314,808	2,487,892	1,734,940	1,491,485	930,459	353,140	422,420	417,009	276,907	
2006	4,442,035	2,714,590	2,203,913	1,525,151	872,108	642,496	602,221	233,969		
2007	4,555,135	3,091,039	2,328,898	1,562,212	1,235,222	1,015,979	484,078			
2008	4,068,513	3,057,908	2,364,406	1,715,734	1,458,143	703,004				
2009	4,096,903	2,937,995	2,084,813	1,843,905	1,145,978					
2010	3,752,463	2,896,894	2,564,289	1,719,854						
2011	3,670,262	2,966,767	2,461,713							
2012	3,571,801	2,971,244								
2013	3,584,497									

Age-to-Age Loss Development Factors Anchored to Year 5

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
1974-2003					…					
2004					1.000	0.618	0.217	0.275	0.268	0.088
2005					1.000	0.380	0.454	0.448	0.298	
2006					1.000	0.737	0.691	0.268		
2007					1.000	0.823	0.392			
2008					1.000	0.482				
2009					1.000					

0.217 = Incremental Loss in Year 7 (233,793) / Incremental Loss in Year 5 (1,078,108)

Anchored Loss Development Factors[2]

	Year 5	Year 6	Year 7	Year 8	Year 9	Year 10
All Year Weighted Avg	1.000	0.620	0.468	0.319	0.230	0.178

1. The average age-to-age loss development factors were used in Figure 2
2. The anchored age-to-age loss development factors were used in Figure 3

7. REFERENCES

[1] Antonio, Katrien; Yanwei Zhang, "Nonlinear Mixed Models" in Predictive Modeling Applications in Actuarial Science, Volume 1: Predictive Modeling Techniques. Cambridge University Press, **2014**.

[2] Barnett, Glen; Ben Zehnwirth, "Best Estimate for Reserves," *PCAS* **2000**, Vol. LXXXVII, 245-321.

[3] Berquist, James; and Richard Sherman, "Loss Reserve Adequacy Testing: A Comprehensive Systematic Approach," *PCAS* **1977**, Vol. LXVII, 123-184. Including discussion of paper: Thorne, J.O., PCAS LXV, **1978**, 10-33.

[4] Butsic, Robert, "The Effect of Inflation of Losses and Premium for Property-Liability Insurers" CAS Discussion Paper Program **1981**: May 58-102.

[5] Cicci Charles, Banerjee Debarshish, Jha Raunak, "Applying a Robust Actuarial Reserve Analysis to Long-Tailed General Insurance Coverage", Institute of Actuaries of India, 13GCA, **2011**

[6] Dean, Curtis Gary, "Generalized Linear Models" in Predictive Modeling Applications in Actuarial Science, Volume 1: Predictive Modeling Techniques. Cambridge University Press, **2014**.

[7] Duvall, Richard M. "Testing for Shifts in Reserve Adequacy," *PCAS* **1992**, Vol. LXXIX, 1-20.

[8] England, Peter and R.J. Verrall, "Stochastic claims reserving in general insurance," *British Actuarial Journal*; **2002**; 8:443-544.

[9] Friedland, Jacqueline, "Estimating Unpaid Claims Using Basic Techniques," CAS Study Note **2010**.

[10] Fleming, Kirk, Mayer, Jeffrey, "Adjusting Incurred Losses for Simultaneous Shifts in Payment patterns and case reserve adequacy Levels" CAS Discussion Paper Program, **1988**, 189-214.

[11] Gluck, Spencer M; Gary G. Venter; "Stochastic Trend Models in Casualty and Life Insurance" Enterprise Risk Management Symposium Monograph, **2009**; M-AS09-1, 1-39.

[12] Guszcza, James C. and Jan A. Lommele, "Loss Reserving Using Claim-Level Data" *CAS Forum*, Fall **2006**; 111–140.

[13] Guszcza, James C., "Hierarchical Growth Curve Models for Loss Reserving," *CAS Forum* **2008**:Fall, 146-173.

[14] Kandemirli, Özlen Kurt, "Turkey: Changing times in claims reserving. Reserving methods to deal with calendar year case reserving effects", *Insights,* Towers Watson, **2014**.

[15] Kerin, Allan A.; Israel, Jason, "The Analysis of the Effect of Tort Reform Legislation on Expected Liability Insurance Losses," *CAS Forum* **1998**:Winter, 153-192.

[16] Meyers, Glenn G., Peng Shi, "The Retrospective Testing of Stochastic Loss Reserving," *CAS Forum* **2011**:Summer, 1-37.

[17] Meyers, Glenn G., "Stochastic Loss Reserving Using Bayesian MCMC Models," *CAS Monograph Series*, Number 1, **2015**.

[18] Renshaw, AE, and R.J. Verrall, "A Stochastic Model Underlying the Chain-Ladder Technique," *British Actuarial Journal*; 4: IV; 903-923, **1998**.

[19] Schmid, Frank, "A Total Credibility Approach to Pool Reserving," *CAS Forum*, Summer Vol. 2 **2012**; 1-22.

[20] Sherman, Richard E., Gordon F. Diss, "Estimating the Workers' Compensation Tail," *CAS Forum* **2004**:Fall, 207-282.

[21] Stenmark, John A.; Cheng-Sheng Peter Wu, "Simpson's Paradox, Confounding Variables, and Insurance Ratemaking," *PCAS* **2004**, Vol. XCI, 133-198.

[22] Struzzieri, Paul J.; Paul R. Hussian, "Using Best Practices to Determine a Best Reserve Estimate," *CAS Forum* **1998**:Fall, 353-413.

[23] Taylor, Greg, "Claims Triangles/Loss Reserves" in Predictive Modeling Applications in Actuarial Science, Volume 1: Predictive Modeling Techniques. Cambridge University Press, **2014**.

[24] Taylor, Greg "Separation of Inflation and Other Effects from the Distribution of Non-Life Insurance Claim Delays," *ASTIN Bulletin*, **1977**,Vol.9:1-12, 219-230.

[25] Verbeek, H.G. "An Approach to the Analysis of Claims Experience in Motor Liability Excess of Loss Reassurance," *ASTIN Bulletin*, **1977**,Vol.9:1-12, 219-230.

[26] Wright, T.S. "A Stochastic Method for Claims Reserving in General Insurance," *Journal of the Institute of Actuaries* (J.I.A). 117, 677-731, **1990**.

The Actuary's Role in a Risk-Focused Statutory Examination

Alan M. Hines, FCAS

Abstract: This paper is being written for the benefit of company actuaries to help them prepare for their statutory financial examination and for consulting actuaries who assist state regulators with the examination of actuarial areas.

States have recently changed the way they perform statutory examinations. The National Association of Insurance Commissioners (NAIC) adopted a risk-focused examination approach as the accreditation standard for statutory examinations. One enhancement of the risk-focused approach is that regulators now leverage more work performed by independent auditors and evaluate company controls to gain comfort in areas that present less financial risk. This change allows regulators to spend more time testing areas with greater risk of material misstatement and assess prospective risk. As a result, many areas that involve the use of actuarial estimates are now getting more scrutiny.

By gaining a better understanding of how examiners assess risk, company actuaries will be better prepared for the examination and be more effective at demonstrating that company controls mitigate risk. This may result in a more efficient examination process by reducing the testing procedures required by the examination team. The information presented in this paper will prepare actuaries to expand their role assisting the examiner-in-charge (EIC) with all phases of the examination. An enhanced understanding of the risk-focused examination process will allow actuaries to assist with the risk assessment process, develop risk-focused testing plans for loss reserves, and add value in other actuarial areas of the examination.

1. INTRODUCTION

The risk-focused exam is now the NAIC standard for insurance company statutory financial examinations. Company[i] actuaries and actuaries on the examination team have seen their role in the examination process expand. The risk-focused examination goes beyond evaluating the adequacy of loss reserves and auditing the financial statement for the examination year. Regulators are spending more time during the examination evaluating company controls over the actuarial areas, considering operational risks, and determining whether there are prospective risks that threaten the future financial stability of the insurer.

Before developing a testing plan to evaluate loss reserves, the examination team evaluates all risks associated with the reserving process, beginning with the process to gather

[i] Company Actuary is being used in this paper to refer to the actuary providing the analysis company management relies on for making decisions on reserves, rate levels, and other areas of work commonly performed by actuaries.

and organize the claim data, and ending with the recording of management's best estimate. In addition, the examination actuary may collaborate with other members of the examination team to assess other areas of risk, including pricing and underwriting risk, concentration of exposure, reinsurance, and other activities that could impact financial results or insurer solvency. As a result, company actuaries working in pricing areas and involved in other enterprise risk management functions may be involved in the examination process.

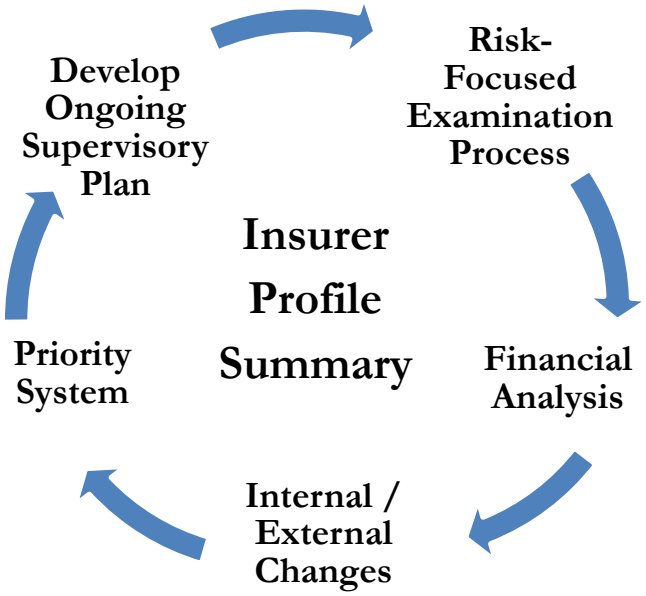
One of the expected benefits of risk-focused examinations was to create efficiencies in the examination process. Using a risk-focused approach, regulators evaluate the work being performed by the company, the company's auditors, and third party consultants to identify risk and evaluate the effectiveness of controls used to mitigate risk. A testing plan is usually developed to evaluate all areas deemed to have high inherent risk and areas where the company's documented controls and mitigation techniques are not effective at reducing risk to a low level. However, if the company can demonstrate that its controls are effective at mitigating the risk to a low residual level, no additional testing procedures may be required by the examination team. By gaining an understanding of how examiners assess risk, company actuaries will be better prepared for the examination and will know the type of information to provide to the examination team that could result in reduced testing procedures.

The sections that follow provide an overview of the NAIC risk-focused examination, with a concentration on areas of the exam where actuaries may be involved. The role of the actuary during a state exam will be addressed from two perspectives: the role of the examination actuary and the role of the company actuary. This paper provides an example of the risk-focused assessment and includes a sample template for documenting the risk assessment and testing plan. The paper will identify actuarial activities creating risk and the common mitigation strategies used by some companies that the examination team may evaluate during the risk assessment process. The paper will also provide a process for developing an efficient testing plan for loss and loss adjustment expense reserves using a risk-focused approach. Many sections of the paper conclude by addressing how the company actuary can better prepare for the statutory exam and providing suggestions to facilitate an efficient examination.

2. OVERVIEW OF RISK-FOCUSED STATUTORY EXAMINATIONS

While some states have been using a risk focused approach since 2007, as of January 1, 2010, the risk-focused surveillance and examination approach became the standard for NAIC accreditation. The main purpose of the surveillance process is to detect (a) financially troubled companies and (b) noncompliance with statutory requirements. The NAIC refers to the risk-focused regulatory process as a surveillance cycle since each element of the process feeds into other steps on a continuous basis. The surveillance cycle is described in detail in the National Association of Insurance Commissioners Financial Condition Examiners Handbook (the NAIC Handbook)[ii] and depicted in the graphic below.

Regulatory Surveillance Cycle



For each insurance company in its jurisdiction, state regulators create an insurer profile summary. Regulators use the insurer profile summary to develop a priority system and supervisory plan for financial solvency. Regulators analyze the company's quarterly statements and calculate key financial ratios to update the company's profile and priority score. Regulators also monitor significant changes in company management, changes in company operations, and reports from external sources. They use the company's priority score to determine how often the statutory exam will be performed. However, to maintain

[ii] 1976-2014 National Association of Insurance Commissioners, Financial Condition Examiners Handbook, 2014 Edition, pages 11-14. Future references of this publication will be denoted "NAIC Handbook, page xx".

NAIC accreditation, all companies under a state's jurisdiction must be examined at least once every three to five years. The information in the insurer profile summary is used by the EIC to develop an examination plan focused on the areas that represent the company's greatest risk. The examination findings are then used to update the supervisory plan.

The NAIC Handbook, page 183, has nine branded areas of risk that must be considered during the examination:

- Credit Risk
- Market Risk
- Pricing/Underwriting Risk
- Reserving Risk
- Liquidity Risk
- Operational Risk
- Legal Risk
- Strategic Risk
- Reputational Risk

Just as the role of the company actuary continues to expand across many operations of the typical property/casualty insurance company, the EIC is now asking examination actuaries to collaborate with other members of the examination team to assess the NAIC branded risks in several areas of the insurance company's operations[iii]. While the EIC is ultimately responsible for the design and execution of the examination, once the examination actuary develops a strong understating of the risk focused approach, the actuary may be asked to take a leadership role in the efforts to evaluate reserving risk and pricing/underwriting risk. In addition, on some examinations, the actuary plays a critical role in evaluating the company's reinsurance programs and assessing whether they effectively mitigate the company's liquidity risk. The role of the actuary and the risk considerations in these areas will be described below for each phase of the exam.

[iii] The EIC develops procedures specifically for the company being examined. Some of the procedures described in this paper may not be incorporated into an examination since the EIC may determine that a targeted examination of specific areas is most appropriate. It is the author's experience that the risk-focused examination approach has required the EIC to seek actuarial expertise to effectively assess the variety of risk that exists in insurance companies today.

Similar to an independent audit, an examination is organized into key functional areas of an insurance company's operation.[iv] For each of these functions, a risk assessment is performed to evaluate the nine types of risk. The NAIC Handbook, page 573, includes a template to document the risk assessment, listing all "risk activities"[v] and mitigation strategies, and documenting the testing plan. The examination team will usually develop a matrix for each functional area. To facilitate the sharing of information[vi] and ensure that the phases of an examination are documented in a consistent manner, many states use an electronic repository system called TeamMate to compile examination workpapers and track the progress of each phase of the exam. The NAIC Handbook identifies sub-activities for each of the key functional areas and lists common risks, best practice controls, and potential tests of the controls for the examination team to consider. The following is a list of key functional areas for property/casualty insurance companies[vii]; the bolded areas listed below are those in which actuaries are most commonly involved:

- Premium
- Claims
- **Reinsurance**
- **Reserves**
- Investments
- Taxes
- Expenses
- Other Liabilities and Surplus
- **Underwriting**

The NAIC Handbook, page 14, depicts the steps for the risk-focused examination. The examination process includes seven separate and distinct phases. Each phase is performed sequentially by the examination team and must be completed and approved by the EIC in the following order:

[iv] Auditors may refer to these as cycles or significant business processes.

[v] In this paper "risk activity" is used to describe the steps inherent in a business process that may result in a risk of material misstatement or other significant business risk.

[vi] Many times, regulators from multiple states will participate in a company's exam when a company has affiliates domiciled in other states. While the EIC for the lead state has the ultimate responsibility for the examination process, state regulators will collaborate to ensure all areas of risk important to their state are addressed during the exam.

[vii] See NAIC Handbook, page 299, for Reinsurance - Ceding Insurer; page 333, for Reserves; and page 421, for Underwriting.

- Phase 1 - Understand the company and identify key functional activities to be reviewed
- Phase 2 - Identify and assess inherent risk in activities
- Phase 3 - Identify and evaluate risk mitigation strategies/controls
- Phase 4 - Determine the residual risk
- Phase 5 - Establish/conduct examination procedures
- Phase 6 - Update prioritization and supervisory plan
- Phase 7 - Draft examination report and management letter based upon findings

The first two phases are considered the planning phases. During Phase 1, the examination team gains an understanding of the company's business and operational procedures through interviews with the company's management and "walk-throughs" of the company's operational processes. In Phase 2, the examination team reviews information gathered from Phase 1 and identifies the risk activities. All areas with significant risks are initially listed in the matrix and the examination team assesses the inherent risk related to those activities. During the Phase 3 procedures, the examination team identifies the company's controls and risk mitigation strategies and begins to evaluate the effectiveness of these controls at reducing risk. In Phase 4, the examination team determines a residual risk rating for each risk listed in Phase 2[viii]. The residual risk rating reflects both the inherent risk identified in Phase 2 and the degree to which the controls and risk mitigation strategies reduce the potential impact of these risks. In Phase 5, the examination team determines a testing plan commensurate with the residual risks assigned in Phase 4. In general, areas with high residual risk require more substantive testing, while abbreviated testing procedures may be used for areas with moderate residual risk. No additional testing procedures may be required for areas with low residual risk. The testing results are then used by the EIC to update the Insurer Profile Summary, the insurer's priority and supervisory plan in Phase 6. The final examination report and, if necessary, a comment letter to company management, are issued in Phase 7.

A more detailed description of the activities performed by the actuary in each phase of the exam follows.

[viii] Some EICs may eliminate risk activities with low residual risk from the risk assessment matrix if no examination testing procedures are deemed necessary.

2.1 Phase 1 – Understand the Company and Identify Key Functional Activities to be Reviewed

The examination team needs to have a complete understanding of the company's operations in order to identify risk activities and the company's risk mitigation strategies. The examination actuary's first step in understanding the company is to gather and evaluate relevant public information and review the findings from prior examinations. The examination actuary may want to review the company's Annual Report, the 10-K, and 10-Qs and search for relevant press releases. The EIC will usually provide the examination actuary with copies of the company's statutory financial statements, actuarial opinions and actuarial opinion summaries, and discuss the areas of significant risk from the prior exam.

A review of the company's statutory financial statements will allow the actuary to identify areas of risk. The actuary may want to note changes in premium volumes, loss ratios, and the one-year and two-year runoff statistics shown in the Five-Year Historical Data section of the annual statement. A preliminary review of Schedule P will allow the examination actuary to better understand the company's mix of business and determine if there has been a recent shift in the insured exposure. Loss ratios and reserve balances by accident year shown in Schedule P, Part 1 and the change in prior year estimates shown in Schedule P, Part 2 may provide the examination actuary with a basic understanding of the inherent reserving risk. A review of Schedule F, Part 3 will provide the examination actuary with a preliminary understanding of the amount and quality of the reinsurance placements. The purpose of the initial review is to develop a broad list of questions and issues that will be discussed during the company interviews.

The examination actuary may benefit by attending the examination's initial kick-off meeting, during which company management provides a high level overview of its operations and highlights changes, or issues that have emerged, since the prior exam. The examination team also uses this meeting to provide an overview of the scope and timing of the exam. A series of "C-suite" meetings are held with the company's Chief Executive Officer (CEO), Chief Financial Officer (CFO), Chief Risk Officer (CRO), Chief Information Officer (CIO), Appointed Actuary and other company leadership[ix] to allow the examination team to gain a better understanding of the company's operations and any significant business activities.

[ix] It should be noted that the corporate structure does not include all of the "Chief" officers designated in this list for some insurance companies. During the kick-off meeting the company will usually identify the person responsible for each of the designated areas. The examination team usually schedules interviews with the company's leadership in each of these areas.

During the C-suite meetings, the examination team begins to develop a basic understanding of how the company manages its business, its governance, and the controls management uses to mitigate risk. These meetings allow the examination team to gain a better understanding of the "tone from the top" related to the company's controls. The examination actuary may use the C-suite meetings to ask the CFO to explain how management establishes its best estimate for the recorded reserves and how management documents the rationale for reserves that differ from the actuarial central estimate. The examination actuary may use the meeting with the CRO or CEO to inquire how the company establishes its risk tolerances, evaluates its catastrophe exposure, and establishes retention levels for its reinsurance programs. Finally, the examination actuary may want to inquire how the company manages its underwriting and pricing activities. The responses from these inquiries will allow the examination team to organize the next level of meetings to gain a detailed understanding of the company's processes.

The actuary will usually work with the EIC and collaborate with other members of the exam team to coordinate meetings with the chief actuary, the actuary in charge of reserving, the actuary or executive in charge of pricing and product development, the head of the claims department, and the actuary or executive who develops and places the reinsurance program. While other members of the examination team usually attend meetings with the company actuaries, the examination actuary usually prepares an agenda and questions related to the actuarial aspects of the exam. Following these meetings, separate meetings are scheduled with the company auditor to evaluate the audit testing plans in the actuarial areas.

During this phase of the exam, the examination team may perform a "walk-though" of the company's processes. The findings from the walk-throughs can be used in Phase 2 to identify risk activities and evaluate inherent risk and can also be used in Phase 3 to evaluate the company's risk mitigation techniques. The "walk-through" is similar to the process used by auditors during the Sarbanes-Oxley[x] testing of internal controls, required for public companies. The examination team usually reviews the documentation supporting the company's internal controls and may be able to leverage the company's flow-charts and the Sarbanes-Oxley control matrices to identify risk activities and company controls.

Phase 1 documentation usually includes the company's responses to the examination team's questions, the agendas, and minutes of the meetings.

[x] The federal Sarbanes-Oxley Act of 2002 established corporate governance and risk management standards that required public companies to document internal controls.

2.1.1 Notes for the Company Actuary on Phase 1

Understanding the risk-focused exam procedures and the examination team's scope and objectives for the examination meeting will allow the company actuary to be better prepared. Since one of the examination team's objectives is to perform a risk assessment of the actuarial process, the company actuary may want to gather and prepare information regarding the company's procedures, oversight, controls, and other risk mitigation techniques inherent in the actuarial process. One of the goals for the company actuary, as it relates to the risk assessment process, is to demonstrate that the level of oversight for the process is commensurate with the inherent risk. If the company actuary is successful at demonstrating that the company controls are effective at mitigating risk, then less testing may be required by the examination team, resulting in a more efficient exam.

Using the reserving process as an example, the company actuary in charge of the reserves may want to demonstrate that data controls are performed at various stages of the review and quality controls are built into the actuarial analysis, allowing the actuary to easily identify material errors. Quality documentation of the actuarial review may allow the examination team to leverage more of the company's work. Additionally, showing evidence of a robust peer review of key actuarial judgments and a formal process to evaluate changes in prior estimates may reduce the examination team's assessment of residual risk, which can result in a more narrowly-focused testing plan. Prior to meeting with the examination team, the company actuary may want to review and update the documentation of the company's actuarial reserving process and ensure that the report supporting the actuarial opinion includes text that memorializes the key actuarial judgments and assumptions. The company actuary may want to schedule a preliminary meeting with the auditor to ensure that the auditor's actuaries are prepared to discuss all of their oversight activities and audit testing plan. The company actuary may supplement these discussions with the procedures the company uses to reconcile the company's estimates with those produced by the auditor. Showing the examination team how the two independent estimates have performed over time may be an effective way to convince the examination actuary to leverage more of the work performed by the auditor's actuary.

To be better prepared for the meetings, the company may request that the examination team provide an agenda and list of questions in advance of the meeting. The organization and quality of the information provided during these initial meetings influence the examination team's assessment of inherent risk and the effectiveness of the company's controls. Advanced preparation by the company actuary will usually result in a more

efficient and effective meeting, with the examination team gaining a better understanding of the company's controls.

2.2 Phase 2 – Identify and Assess the Inherent Risk

During Phase 1, the examination team gains an understanding of the company's operational procedures and begins to identify the activities that will be evaluated for inherent risk. For each of the key functional areas, the NAIC Handbook identifies major activities and common risks for the examination team to consider. As noted above, three of the key functional areas have a significant actuarial component. The reserving function and the underwriting/pricing function are directly tied to an NAIC branded risk, and the reinsurance function is tied to liquidity risk. This section will specifically address the actuarial aspects of the reserving process, the underwriting process and reinsurance. It should also be noted that all functional areas also include many other risk activities that are not considered "actuarial" in nature. For example, the examination of the reserving function may begin with the risk activities related to information systems and the process to accumulate the data used in the actuarial analysis and may end with the activities to record reserve changes or other financial reporting activities performed by the accounting department. The complete spectrum of risk activities included in the risk assessment matrix is not the focus of this paper. Even though the EIC may ask the examination actuary to collaborate with other members of the examination team on other areas, the focus of the discussion that follows is related to areas that are most often reviewed by the examination actuary.

As noted above, for each of the key functional areas, the NAIC Handbook lists general activities, common risks, and best practice controls to be considered for inclusion in the documentation matrix[xi]. However, because of the diversity in organizational structure among property/casualty insurers and unique nature of the risk activities that exist for each company, the documentation matrices used in practice are developed specifically for each company being examined. For smaller insurers, the general activities listed in the NAIC Handbook may be sufficient and the examination team may simply include the risks that are appropriate for the company being examined. However, a large, more complex insurance company may have unique processes that require a more detailed listing of risk activities or sub-activities to identify the risks associated with the process.

The organizational structure within the company being examined may necessitate multiple risk matrices for each key functional area. Some property/casualty companies use

[xi] See NAIC Handbook, page 299, for Reinsurance - Ceding Insurer; page 333, for Reserves; and page 421, for Underwriting.

different processes for various business units within the company. For example, the process used by a company for its personal lines exposure may differ from the process used for its commercial business. Some companies may also use different processes for subsidiary legal entities or branch offices, and others may have a separate and distinct process for unique types of businesses written in the company. If the inherent risk and controls to mitigate risks differ within the company, the examination team may consider performing separate risk assessments and documenting the results in separate matrices. Since the risk assessment may result in a different residual risk, the associated testing plan developed for each area may also differ.

2.2.1 Activities related to reserving risk

Many risk activities for the reserving process cross all lines of business (or reserving segments). Even if the company uses a best practice reserving process and has strong risk mitigating controls, the risk inherent in the exposure for some reserving segments (or for certain activities within the reserving process) may be sufficiently high that risk mitigation techniques will not effectively reduce the reserve risk to a low level. Different levels of inherent risk for the various reserving segments may create situations where certain activities and controls result in a high or moderate residual risk for one review segment but low residual risk for another. To develop a testing plan that is directly tied to the risk assessment process, it may be appropriate to perform a risk assessment at the reserve segment level. An example of a risk assessment performed at the reserve segment level is contained in Appendix B.

The first step of the risk assessment process is to develop the framework of activities to be considered in the risk assessment matrix. The reserving process may begin with the process to aggregate data for the actuarial review, but many times, the examination actuary's process begins with ensuring that the data used in the actuarial analysis is appropriate for estimating the unpaid claim liabilities. The system activities related to the claims and exposure data and the other detailed data quality controls are important elements in the reserving process, but these activities are usually evaluated by other members of the examination team. However, since risks associated with the underlying data may impact the actuary's risk assessment and testing plan for the reserve risk, the EIC may want to review the risk assessment and consider the testing results for the related claims systems prior to the examination actuary's providing a conclusion on the actuarial aspects of the reserving risk.

The following is a sample of risk activities associated with the reserving process and examples of risks that may be considered by the examination actuary.

- **Data aggregation and reconciliation** – The actuarial data is inaccurate, incomplete or otherwise inappropriate for estimating the unpaid claims.
- **Segmentations used in the actuarial reviews** – Improper segmentation of underlying actuarial data may inhibit the detection of loss trends, development patterns, or shifts in types of loss.
- **Environmental or operational changes impacting the actuarial analysis** – Changes in the company's policies, written exposure, claim processing, or environment are not adequately contemplated in the actuarial estimates.
- **Consideration of reinsurance** – Historical changes in the reinsurance program are not properly reflected in the estimation of net or ceded reserves.
- **Consideration of special policy provisions** – The actuarial estimates of unpaid claims do not adequately consider unique risks related to special policy provisions. Examples of special policy provisions include retrospective premium reserves, credit risk from large deductible policies, and long duration contracts that may require unearned premium reserve testing.
- **Actuarial methods and techniques used** – The company's actuarial reserving software does not include adequate or appropriate actuarial methods and techniques to evaluate the exposure. This assessment may include an evaluation of the company's reserving software and the system and spreadsheet controls related to the actuarial reserving process.
- **Quality controls and oversight –** Inadequate quality controls and oversight may result in material errors in the estimates, a bias in the actuarial assumptions, or unreasonable selections resulting in inadequate estimates of unpaid claims. Management's influence on the actuarial estimates may be evaluated in this area of the process.
- **Loss Adjustment Expense (LAE) considerations –** The actuarial process to estimate unpaid allocated loss adjustment expense (ALAE), or defense and cost containment (DCC) expense, does not adequately address changes in defense strategies or trends in legal defense costs. The process used by the company to estimate unpaid unallocated loss adjustment expense (ULAE), or adjusting and other payments (AOP), does not adequately reflect the company's average cost to settle claims or the expected duration of the remaining open claims. For some companies the reserving risks for loss and LAE are similar and a separate risk assessment may

not be required. However, differences in the company's process for estimating LAE reserves or unique risks related specifically to LAE may necessitate separate consideration.

- **Recording differences between Actuarial Central Estimate (ACE) and Management's Best Estimate (MBE) –** The recorded reserves based on the MBE are inadequate or have been selected in a manner that distorts reported earnings resulting in a material reputational risk for the company. The examination actuary may need to consider the risks and controls related to management's selection process if the recorded reserve differs from the company's actuarial central estimate.

2.2.2 Activities related to underwriting risk

The examination actuary's review for underwriting risk is usually focused on the actuarial ratemaking process, management's oversight of rate level changes, and the interaction between the ratemaking and actuarial reserving functions. The examination actuary may be asked to evaluate areas considered to have elevated underwriting risk. This may include segments of business with significant growth, newly emerging markets, segments with a high concentration of exposure, segments with perpetually high loss ratios, or lines with significant variability. Underwriting/pricing risk is more of an operational risk than a financial reporting risk. It is also more prospective in nature since the examination actuary is evaluating whether current or future policies may be written at inadequate rate levels, resulting in a future drain on surplus.

The following is a summary of some specific risks that may be evaluated by the examination actuary related to underwriting risk:

- Inappropriately selected ratemaking methods, resulting in inadequate rate levels;
- Inadequate actuarial expertise, impacting the quality and timeliness of rate adequacy reviews and rate filings;
- Significant growth in new markets, resulting in books of business with optimistic or inadequate pricing that may not be detected and corrected in a timely manner;
- Improper use of predictive modeling or other underwriting tools, leading to poor risk selection, adverse selection, and inadequate rate levels;
- Inadequate monitoring of rate levels and use of flexible pricing adjustments, leading to inadequate rate levels and underwriting deterioration; and

- Material unreconciled differences between the ultimate loss estimates derived for ratemaking and those estimated for reserving, resulting in inadequate rate levels and unfavorable underwriting results.

For well-established companies, many of the risks noted above may have a moderate to low inherent risk, which may be further mitigated by company controls. However, many well-established companies do acquire less successful companies or expand their operations to achieve growth objectives. The integration of new business or expansion plans may increase a company's underwriting risk. A walk-through of the company's product development and ratemaking process will allow the examination actuary to identify other risk activities that may require further review.

2.2.3 Activities related to liquidity risk and reinsurance

One of the more common areas the EIC asks the actuary to review for liquidity risk is the process used by the company to develop its company's reinsurance program. For property insurers, this may also involve evaluating the company's catastrophe exposure. However, under certain circumstances, the examination actuary may be asked by the EIC to evaluate company's payment patterns and perform cash flow testing to evaluate whether there is sufficient liquidity in invested assets.

To evaluate the company's reinsurance program, the examination team usually begins by gaining an understanding of the company's stated risk tolerances and procedures to establish the reinsurance program to mitigate fluctuation in the company's retained losses. The examination actuary usually reviews a history of the per-occurrence retention, the limits of coverage, and quality of placements for the company's reinsurance covers. The actuarial risk assessment may include reviewing variability studies supporting the company's per-occurrence retention or the exposure modeling used to evaluate the company's concentration of property risk. Some of the risks that the examination actuary may include in the assessment of the reinsurance program include:

- Inadequate actuarial expertise or system capabilities to perform exposure modeling and evaluate concentration of risk, leading to retained exposure that exceeds the company's stated risk tolerances.
- Inadequate process to monitor and manage new business writings, leading to excess exposure or inadequate rate levels due to increased reinsurance costs.
- Inadequate governance of risk tolerances by the company's Board or Enterprise Risk Management (ERM) committee, resulting in retained risk that exceeds rating agency risk tolerance levels.

- Inadequate management controls over the reinsurance program, leading to policy provisions or a design that does not effectively limit the company's exposure.
- Inadequate underwriting controls, resulting in the issuance of primary policies that do not meet retention levels or coverage limitations required by the company's reinsurance programs.

2.2.4 Determining inherent risk

Once the risk activities are identified, the examination team must assess the inherent risk for each risk. For most activities, the risk assessment is usually performed at the company level since the inherent risk defined in the NAIC Handbook relates to the frequency and magnitude of risk at the company level. If the risk assessment is performed in greater detail, the risk assessments for the individual segments must be aggregated to determine the company's total inherent risk for that activity.

In the example shown in Appendix B, the risk assessment is performed at the reserve review segmentation level. By approaching the risk assessment at this level of granularity, the resulting residual risk and recommended testing plan for each review segment may be tied directly to the risk assessment process. If risks are evaluated at the reserve segment level, both the magnitude of risk and the aggregation of risk with other segments need to be considered when assigning the inherent risk at the company-wide level. For some risks, such as random (independent) calculation errors, it would be less likely that multiple smaller errors would occur and aggregate to the magnitude required to be classified as a High Risk for the company. However, for highly correlated risks, such as errors in a reserving template impacting all review segments or a bias in actuarial judgment for a long tailed line of business, the aggregation of smaller risks in multiple reserve segments may result in a high risk for the company. Therefore, actuarial judgment must be applied when considering the appropriate magnitude of risk for each segment. If magnitude and aggregation are considered in assigning the inherent risk ratings for each segment, an averaging technique may be appropriate to determine the inherent risk for the company.

The NAIC Handbook, pages 185 – 188, includes a framework and a rating system for determining the three classifications of inherent risk. The NAIC Handbook suggests that a "High" inherent risk be assigned to risk activities that are large (in relation to the company's financial strength) and that could result in significant and harmful financial and/or reputational loss to the organization. A "Moderate" inherent risk is considered significant (moderate in size in relation to the company's financial strength) and the loss to the insurer could be absorbed in the normal course of the business. A "Low" inherent risk

results in an error that would have an insignificant negative impact on the insurer's financial strength and reputation.

The NAIC Handbook recommends the use of a frequency and severity approach to evaluate both the likelihood of an occurrence and the magnitude of the impact for each inherent risk. If the event being evaluated is likely to occur "most of the time", the risk is assigned a "High" frequency rating. If the event only rarely occurs, the risk is assigned a "Low" frequency rating. Events that will probably occur some of the time are assigned a "Moderate-High" rating and events that could occur some of the time are assigned a "Moderate-Low" rating. To develop the ratings, both qualitative and quantitative assessments are used, along with actuarial judgment.

The NAIC Handbook, pages 185-188, four classifications for the magnitude, or severity, of the impact:

- **Threatening -** The risk is classified as threatening if the event could result in an impact greater than 5% of surplus or material rating agency downgrade, or could otherwise give rise to financial solvency concerns.
- **Severe** - The risk is classified as severe if the event could result in an impact between 3% and 5% of surplus, have a serious impact on shareholder value and reputation with adverse publicity, or result in board and senior management attention.
- **Moderate** - The risk is classified as moderate if the event could result in an impact between 1% and 3% of surplus, have an impact on shareholder value and/or reputation, or result in senior and middle management attention.
- **Immaterial** - The risk is classified as immaterial if it results in an impact less than 1% of surplus, has no potential impact on shareholder value and/or the reputation of the company, and is expected to be addressed and resolved by the company's middle management.

The NAIC Handbook suggests the overall inherent risk assessment be determined by considering both the frequency and severity components as shown in the table on the following page:

Overall Inherent Risk Rating Scale

Likelihood of Occurrence	Magnitude of the Impact			
	Threatening	**Severe**	**Moderate**	**Immaterial**
High	High	High	High	Moderate
Moderate-High	High	High	Moderate	Moderate
Moderate-Low	High	Moderate	Moderate	Low
Low	Moderate	Moderate	Low	Low

Phase 2 is completed once the inherent risk assigned to the list of risks is approved by the EIC. The effectiveness of the company's risk mitigation techniques is evaluated in Phase 3.

2.3 Phase 3 – Identify and Evaluate Risk Mitigation Strategies

The examination actuary will learn about the company's risk mitigation techniques during the Phase 1 meetings, review of the company's documented Sarbanes-Oxley or Model Audit Rule (MAR)[xii] controls, and walk-throughs of the processes being evaluated. However, while many company actuaries naturally incorporate risk mitigation techniques in their processes, they may not think about the various quality control checks and balances as "risk mitigation strategies." The following sections are intended to identify some of the risk mitigation techniques commonly used by companies.

2.3.1 Reserving risk mitigation techniques

The loss and LAE reserve estimates are inherently a high risk area for most property/casualty insurance companies. However, most companies have a number of controls and risk mitigation strategies imbedded in their actuarial reserving process. These include controls that are built into the actuarial analyses (or models) underlying the estimates of the unpaid claims, as well as management controls over the change in estimates. The company's auditor (or a third party actuary) may also produce independent estimates that serve as a control over the reserve estimates.

[xii] The Model Audit Rule is the common name for the NAIC Annual Financial Reporting Model Regulation (#205). MAR requires large non-public insurance companies to document their assessment of internal controls. For smaller companies, professional auditing standards, established by the American Institute of Certified Public Accountants (AICPA) requires the auditor to document and review the company controls.

Potential risk mitigation strategies that may be used by companies include:

- The existence of data controls and reconciliations performed before and after the actuarial review to ensure the data provided for actuarial analysis reconciles to the financial statements.
- The use of procedures to verify that prior valuations of claims data have not changed.
- Robust discussions between the reserving actuary and management, the claims department, and key personnel in other operational areas to identify potential changes in business and other industry trends to be incorporated into the reserving process.
- An adequate team of actuarial experts assigned to develop the actuarial central estimate and range of reasonable estimates of unpaid claims.
- The use of a protected loss reserving system that includes multiple actuarial techniques and the application of appropriate methods to evaluate the exposure.
- Adequate actuarial oversight of the methods and assumptions, with documented peer reviews.
- A formal process to monitor and respond to changes in prior estimates.
- Detailed reconciliations and analysis of differences between company estimates and those developed by the independent auditor's actuary or other third party.
- The inclusion of claim diagnostics and other statistical controls to evaluate environmental or operational changes that may impact the actuarial estimates of reserves.
- Proper procedures in place to estimate ceded reserves.
- Documentation of management's best estimate for the recorded reserves, with sufficient rationale for differences with the appointed actuary's central estimates.
- The existence of a functional reserve committee that meets regularly and documents meeting minutes.
- A well-written actuarial opinion identifying risk of material adverse deviation and a complete actuarial report with text memorializing the key assumptions inherent in the estimates.

- Sufficient interaction between the appointed actuary and the Board or Audit Committee.

These and other mitigation techniques may effectively reduce the risk inherent in developing estimates of unpaid claims. However, just as the inherent risk may differ by reserving segment, so may the effectiveness of risk mitigation strategies at reducing the risk. Therefore, once the mitigation strategies are identified, the actuary needs to determine if the effectiveness of the mitigation technique can be evaluated at the company level or whether it may be more appropriate to evaluate the effectiveness of the controls for each reserving segment.

As noted above, the example provided in Appendix B was prepared at the reserve segment level, and the effectiveness of the risk mitigation techniques was evaluated for each segment individually. This process allows the examination actuary to develop a testing plan commensurate with the residual risk for each reserving segment. However, to determine the overall effectiveness of controls, an aggregation of the results by segment is needed to complete the risk assessment at the company level. Similar to assessing the inherent risk, the effectiveness of the control may be assigned at the reserving segment level in consideration of the aggregation technique to be applied. Once completed, the aggregation of the results for each segment is used to determine the overall rating documented in the company's risk assessment matrix.

2.3.2 Underwriting risk mitigation techniques

Many of the mitigation techniques listed above for reserving risk may also be considered mitigating controls for underwriting risk. Some of the common mitigation techniques for underwriting risk are:

- An adequate number of experienced actuaries overseeing the rate review process.
- The involvement of actuaries in product development and evaluating the costs of coverage changes.
- The existence of a robust planning process and comparison of plan to actual results.
- A process to reconcile differences between projected (budgeted) premiums and actual premium.
- A process to monitor rate level changes, flexible pricing changes, and the use of pricing tiers.
- A process to determine profitability by line, branch office, agency, and geographical region.

- A process for integrating newly acquired businesses and entry into new markets.
- A process to compare company loss costs and rate levels with industry benchmarks or key competitor rates.
- A well-defined and documented process to develop and review underwriting models.

For many well-established companies, underwriting risks have a low inherent risk and the assessment of the effectiveness of the controls may be performed on a company-wide basis. However, variability in operating results and change in business operations may elevate the inherent risk. Larger companies may have separate personal and commercial units, and very large companies may have a regional organizational structure. During Phase 1, the examination team may want to perform a walk-through for each unit to determine if the underwriting process is more appropriately evaluated on some basis other than at the overall company level.

2.3.3 Liquidity risk mitigation techniques

The most common liquidity-related risk mitigation technique evaluated by the examination actuary is the design and placement of the reinsurance program. Therefore, the examination actuary may need to understand how the company's risk tolerances were established and how those risk tolerances compare to targets established by rating agencies. The examination actuary may also be asked by the EIC to evaluate whether the reinsurance program is designed to meet the specific thresholds established by the Risk Based Capital requirements.

For companies with a large property insurance exposure, the EIC may ask the examination actuary to review the results of the company's catastrophe modeling and evaluate how the company manages its concentration of risk. The examination actuary may want to compare the catastrophe model results reported to the rating agencies to the results produced by the reinsurer in underwriting the exposure. The type and quality of the model used and the abilities of the company actuaries to evaluate the exposure are important elements in the risk assessment. If there has been a recent change in the company's per risk retention, the examination actuary may also want to evaluate any actuarial variability studies performed to evaluate the change in retained risk. Common liquidity risk mitigation techniques used by companies include:

- An adequate amount of actuarial expertise involved in the design and development of the reinsurance program.
- An active program to measure and monitor concentration of risk.

- The use of appropriate catastrophe models and documentation of the company's catastrophe results.
- An annual presentation to the company's Board, ERM committee, or other governing committee.
- Documentation of the company's historical reinsurance placements with quality reinsurers.

2.3.4 Determining the effectiveness of the control

Phase 3 requires an assessment of the effectiveness of the mitigation strategy for each risk. The NAIC Handbook considers a risk control "strong" when it is deemed to be effective at reducing the assessed risk, "moderate" when it is only partially effective at reducing risk or will reduce the risk some of the time, and "weak" when there are no risk mitigation procedures in place or if there is material weakness identified during the controls testing. Under certain circumstances, a weak risk control may actually increase the risk for the activity, and the examiner may revise the rating of the inherent risk and recalculate the residual risk.

2.3.5 Notes for the company actuary on risk mitigation strategies

The effectiveness of the company controls impacts the amount of testing to be performed by the examination team. If the company actuary is able to show the examination team there are strong risk mitigation techniques in place and provide evidence to demonstrate that these controls are effectively reducing the company's risk, less testing may be required by the examination team. However, it may not be adequate for company actuaries, or other staff, to show that they perform risk mitigation activities on an informal or periodic basis. To be considered an effective control, many examination teams request to see documentation of the process and evidence to demonstrate that the process is consistently executed by the company. The company actuary will benefit by developing a strong understanding of the activities considered "risk mitigation techniques" and documenting these processes as part of the company's internal controls. Maintaining organized files and documentation of the risk mitigation procedures will facilitate the risk assessment process and may ultimately result in fewer testing procedures.

One of the controls the examination actuary may consider is the company's peer review process and the oversight of the actuarial estimates. When the company's peer reviews are not adequately documented, or there is not adequate evidence to validate that the peer reviewer has actually performed a robust review of the key actuarial assumptions and judgments, the examination team may not be able to place strong reliance on this control.

Some company actuaries maintain separate peer review files where the second reviewer adds comments, questions, and suggestions. These files may demonstrate that there has been a thorough peer review and robust professional discussion about the underlying assumptions used in the final estimates. This documentation also provides the examination actuary with additional insight on the final selections.

Another control that may be considered by the examination actuary is the quality of the auditor's actuarial review and the consistency between the auditor's estimates and the company's actuarial estimates. If the company actuary does not maintain a history of the auditor's estimates in their files or does not maintain documentation of their assessment of the difference in estimates, the examination team may need to perform additional procedures to evaluate both sets of estimates. Company actuaries that maintain a history of how the company estimates compared to the auditor estimates, or other independent actuarial estimates, are able to more efficiently identify and address the difference in assumptions that produced the estimates. It is also effective when the company actuary is able to show the examination team how the company's prior actuarial estimates have run off compared to those selected by the auditor's actuary and discuss the rationale for changes in prior reserve estimates. If the company actuary is able to provide this level of detail, the examination actuary may be able to place a greater reliance on the auditor's independent estimates as an effective control.

2.4 Phase 4 – Residual Risk Assessment

The residual risk for each identified risk activity is determined in Phase 4. The residual risk is based on both the inherent risk assigned in Phase 2 and the effectiveness of the controls assigned in Phase 3. The NAIC Handbook, page 198, includes the table shown on the following page to depict the process used to determine residual risk. The examining actuary may want to review the resulting residual risk assigned to each activity and apply sound actuarial judgment to reconsider the inherent risk and effectiveness of the risk mitigation strategy if the residual risk is not appropriate. For areas in which the company has weak controls or testing has identified a material weakness in the execution of the company controls, the inherent risk may be elevated to reflect the increased residual risk.

Calculating the Residual Risk

Inherent Risk Assessment	Control Risk Assessment		
	Strong Control	**Moderate Control**	**Weak Control**
High	Moderate to High	Moderate to High	High
Moderate	Low to Moderate	Moderate	Moderate*
Low	Low	Low	Low*

* The inherent risk may be reassessed in light of the control weakness

The risk assessment is documented and the resulting residual risk is approved by the EIC. If the examination actuary issues a memo to describe the risk assessment process, it is usually referenced in Columns 3b and Column 4b of the NAIC Risk Assessment Matrix shown in Appendix A.

2.5 Phase 5 – Establish and Conduct Examination Testing Procedures

Under a risk-focused approach, the examination testing plan is developed based on the level of residual risk. No additional testing procedures may be required for areas with low residual risk, but the EIC may ask the examination actuary to perform additional analytical procedures to document the risk assessment. Testing procedures are required for moderate and high residual risks unless the rationale is documented and approved by the EIC. More robust independent testing procedures are required for areas with high residual risk. For areas with moderate residual risk, the examination team may leverage more of the testing procedures performed by the company or the company's auditors in its testing plan.

For most insurance companies, the reserving risk poses the greatest risk of material financial misstatement. Even if the company uses appropriate actuarial estimation techniques and uses best practice mitigation techniques, it is unlikely the risk assessment will result in a low residual reserving risk. As a result, a Phase 5 testing plan is usually required to be developed and approved by the EIC.

Tying the testing plan back to the NAIC Risk Assessment Matrix is sometimes a challenge. Many risks in the reserving Risk Assessment Matrix are related to company procedures for developing reserve estimates. While some testing procedures may be designed to evaluate specific risks[xiii], many reserving risks are interrelated. Some examination

[xiii] For example, data reconciliations can be performed and data testing procedures may be designed to evaluate the accuracy and completeness of the data used in the actuarial analysis.

teams simply default to performing an actuarial analysis to evaluate the reasonableness of the recorded reserves. However, performing independent testing to validate that the company's recorded reserves are reasonable does not necessarily provide insight on the appropriateness of company's reserving process or the effectiveness of the company's controls. The examination actuary may want to consider testing procedures to evaluate the company's processes and controls in order to leverage the company's or the auditor's actuarial estimates.

Even though a testing plan will likely be required to evaluate the company's reserving risk, it is not readily apparent how to develop an efficient testing plan. If a risk assessment is performed at the reserve segment level, it will likely show that the estimates have different residual risk. This risk assessment will allow the examination actuary to develop a testing plan for each reserving segment that is commensurate with the residual risk.

For reserve segments with lower residual risk, diagnostic statistics may allow the examination actuary to determine that the company's reserve balances are not materially misstated. The examination actuary may compare the company's estimates of the unpaid claims to the auditor's actuarial estimates and conclude the reserves are reasonable based on the proximity of the company's and auditor's current estimates, the consistency of the estimates over time, and the runoff of prior year estimates. The examination actuary may also calculate and evaluate other diagnostic statistics[xiv] using Schedule P data to reach a similar conclusion.

For moderate risk segments, the testing approach may include a methods and assumptions review[xv] of the actuarial analysis supporting the company's estimates or the analysis performed by the auditor. A methods and assumptions review of an actuarial analysis can take many forms[xvi]. When documenting the peer review approach used in the exam, the author sometimes finds it useful to differentiate between a methods and assumptions review and a technical peer review where the differences in actuarial judgments are quantified. For the former, the reviewing actuary generally reviews the work papers, methods and key assumptions in the analysis. If the analysis is deemed reasonable, the actuary adopts the reserve estimates as the examination estimate with little modification. For the latter, a more robust peer review is performed, and differences in actuarial judgment are

[xiv] Various ratios may be compared to industry benchmark ratios and the company's ratios from prior statement years. Accident year ratios that may be considered include IBNR to case reserve ratios, the implied paid and reported development factors, various reserve to premium ratios, and ultimate LAE to loss ratios.
[xv] A methods and assumptions review is sometimes referred to as a peer review.
[xvi] The reader may want to review Balester, Jennifer Lynn and Kirschner, Gerald S. Casualty Actuarial Society Forum Casualty Actuarial Society - Arlington, Virginia 2013: Fall, Vol. 1 1-30 Structured Tools to Help Organize One's Thinking When Performing or Reviewing a Reserve Analysis

quantified. For this procedure, the selected ultimate losses and key parameters of the company's actuarial analysis[xvii] are entered into a spreadsheet allowing the reviewing actuary to independently select his parameters and quantify the difference in actuarial estimates.

Other testing procedures may also be appropriate for segments with moderate residual risk. These may include: tests to evaluate a specific aspect of the estimate, supplemental tests not included in the company's procedures, and re-performing selected actuarial methods to validate the results.

If these abbreviated testing procedures indicate there is elevated risk of a significant difference in estimates, the examination actuary may need to perform additional testing procedures or revert to developing independent estimates in order to quantify the amount of the potential misstatement. For segments with high residual risk, the examination actuary may need to develop independent estimates to efficiently evaluate the reasonableness of the company's reserve.

When using a detailed risk assessment to develop a testing plan, the examination actuary may need to consider the aggregation of many small to moderate differences in reserve estimates that may result in a material misstatement. For some companies, it may be appropriate to independently test a sample of low and moderate risk lines to evaluate if there is a bias in the company's estimates. Similarly, for some companies, it may be appropriate to include a mix of testing procedures for segments with a high residual risk rating.

When a mix of testing procedures is used, it is usually helpful for the examination actuary to develop a summary showing the distribution of the reserves by testing procedure as shown in the table below. This will allow the EIC to efficiently evaluate the mix of testing procedures by level of residual risk.

[xvii] The key parameters will depend on the analysis, but loss development factors, expected losses, and weights assigned to each of the methods are elements that may be included.

Residual Risk Assessment vs Review Testing Approach

Carried Reserves in $000

	Residual Risk Assessment						
Review Approach	**Low**		**Moderate**		**High**		**Total**
Independent Estimates	1,222	**1%**	6,667	**6%**	53,333	**48%**	**55%**
Technical Peer Review or Supplemental Procedures	1,100	**1%**	10,000	**9%**	8,889	**8%**	**18%**
Peer Review	1,111	**1%**	10,889	**10%**	4,441	**4%**	**15%**
Diagnostic/Non-Review	10,000	**9%**	3,556	**3%**	333	**0%**	**13%**
Total	**13,433**	**12%**	**31,111**	**28%**	**66,997**	**60%**	**100%**

The example provided in Appendix B demonstrates how a risk assessment process can be structured to evaluate the residual risk for each review segment and used to develop an efficient testing plan.

To complete the Phase 5 testing procedures for reserving risk, the actuary will be required to evaluate whether recorded reserves are reasonable or quantify the resulting differences for the EIC. Examination testing procedures and results are usually documented in an actuarial report, consistent with actuarial standards of practice.

2.5.1 Testing procedures for underwriting/pricing risk and liquidity risk

The level of actuarial involvement in testing the underwriting risk and liquidity risk varies significantly by exam. To evaluate the underwriting risk, the EIC may ask the examination actuary to perform testing procedures to validate the quality of the company's rate reviews or underwriting models. To evaluate liquidity risk, the testing plan may include an actuarial review of the company's catastrophe model or an analysis to ensure that the reinsurance program is designed to meet the company's stated risk tolerances. Detailed descriptions of the testing procedures used by the examination actuary in these areas are beyond the scope of this paper.

2.6 Phase 6 and Phase 7

The examination results are used by the EIC in Phase 6 to update the Insurer Profile Summary and prioritization plan. Once the examination actuary's report is approved, the EIC may schedule meetings to discuss the examination results with the company actuaries. The actuary's examination findings will be incorporated into the EIC's final examination report in Phase 7. Any significant findings in the examination actuary's report related to the company's actuarial process or the company's risk mitigation strategies may be addressed in the EIC's final report or the management letter issued by the EIC at the conclusion of the exam.

3. CONCLUSION

Developing a detailed understanding of the risk-focused examination process will allow the company actuary to facilitate a more efficient examination of the company's actuarial processes, and allow examination actuaries to add value to the EIC in more phases and operational areas of the exam. The risk-focus examination encourages examination actuaries to develop targeted testing plans and concentrate their efforts in the areas that represent the greatest risk for the company.

Appendix – Example Risk Assessment

This is an example risk assessment for Sample Company's reserving risk. The NAIC Risk Assessment Matrix is shown in Appendix A. A sample of actuarial risk activities begins in activity 3.1 and ends with activity 5.1. The risk assessment for Sample Insurance Company was performed on a reserve segment basis and is shown in Appendix B.

Sample Company reviewed their reserves by legal entity and separately for commercial and personal lines. To account for these differences, the detailed risk assessment was performed at the company and reserving segmentation level (Appendix B, Sheets 2-3). Due to space limitations, only a selection of segments is actually shown in the exhibits. In Appendix B, Sheet 1, a weighted average of the risk assessments for each reserve segment was used along with actuarial judgment to aggregate the risk assessments to a company level. The aggregated company risk ratings in this summary are used to complete the reserving risk matrix shown in Appendix A. However, there is not a one-to-one correspondence between the detailed risk assessment performed on a reserving segment basis in Appendix B and the risks for the actuarial reserving process shown in Appendix A. The assessments shown in Appendix A may also include a review of procedures and controls that would be referenced in columns 3B and 4B. Due to space limitations, the text and reference to key documents in the database would be entered in the Reserving Risk Matrix for Phase 3 through Phase 7 are not shown. By considering the residual risk at the review segment level, a testing plan can be selected for each reserve review segment based on the risk characteristics and effectiveness of the mitigation techniques for that specific segment. The testing method is shown in Appendix B, Sheets 2-3.

The factors considered in the detailed risk assessment included the following:

- **Quality of the Company's Actuarial Reserve Analyses –** The quality and completeness of the actuarial review process and the documentation supporting the estimates was considered. The types of methods used and the specific diagnostics evaluated in the actuarial reserving package were considered, including: settlement rates, case reserve adequacy, frequency, severity, runoff of prior estimates, and other supporting analysis to support the estimates of unpaid claims.
- **Management's Differences –** The variances between the actuarial central estimate (ACE) and management's best estimate (MBE), which is the basis for the held reserves, were considered in the risk assessment. The larger the variation, the higher the assessed risk.
- **Results of Auditor's Reserve Analyses –** The type of review performed by the auditor and comparison of the auditor's estimates with the company's estimates were considered. Lines of business or review segments where the auditor showed larger

variances to the company's actuarial central estimate were assigned higher risk. Lines of business where the auditor did not test the reserves may also have elevated the risk assessment, considering other factors.

- **Prior Results –** The historical change in ultimate losses from prior years was used as an indication of the inherent risk in the estimate.
- **Inherent Risk of Particular Line of Business/Segment –** The unpaid claims for some segments are inherently difficult to estimate. Longer-tailed casualty lines, lines with large concentrations of reserves and/or lines of business where the loss development patterns or loss ratios demonstrate significant variability were considered higher risk.

SAMPLE INSURANCE COMPANY RISK ASSESSMENT MATRIX

Calculated Cell Do Not Enter Data	Enter Data / Information using Drop Down Boxes or Message Prompts	1a	Key Activity	P&C Reserving

1b – Overall Risk Statement	The risk that reserve accounts are not properly reported, misstated, or improperly valued.
1c – Analytical Assessment:	Refer to analytical procedures performed as part of phase 1 for further information.

Phase One		Phase Two						Phase Three						Phase Four			Phase Five	Phase Six	Phase Seven
1d		2a	2b	2c	2d	2e	2f	3a				3b	3c	4a	4b	4c	5	6	7
		Risk Identification			Inherent Risk Assessment			Risk Mitigation Strategy/Control Assessment						Residual Risk Assessment					
Sub-activities	Identified Risk Number	Identified Risks	Branded Risk(s)	Exam Assertion(s)	Likelihood	Impact	Overall Inherent Risk Assessment	Risk Mitigation Strategy	Frequency	Samples Tested	Obtained from	Evidence & Document Testing Controls	Overall Risk Mitigation Strategy Assessment	Calculated Residual Risk	Judgmental Residual Risk	Overall Residual Risk Assessment	Examination Procedures / Findings	Prioritization Results Supervisory Plan	Report Findings & Management Letter Comments
Risks Other than Financial Reporting																			
	1.1	The Company Board of Directors are not involved in establishing and/or reviewing the insurer's overall reserving policy and methodology.	ST, RV		Moderate-Low	Severe	Moderate												
	1.2	The Company is not following the reserving policy and methodology that has been adopted and reviewed by the Board of Directors.	OP, RV, ST		Moderate-Low	Severe	Moderate												
Financial Reporting Risks																			
Accumulation of Data for Reserving	2.1	Claims data maintained by the Company is not complete, accurate (including line of business classification) and properly cut off.	OP, RV	CO, AC	Moderate-Low	Moderate	Moderate												
Accumulation of Data for Reserving	2.2	The claims data utilized by the actuary to estimate reserves does not correspond to the data in the Company's claims system and to the data in the insurer's accounting records.	OP, RV	CO, AC	Moderate-Low	Moderate	Moderate												
Accumulation of Data for Reserving	2.3	Loss adjustment expense data is not properly classified as defense and cost containment (DCC) or adjusting and other (AO).	OP	AC	Low	Moderate	Low												
Reserving Assumptions and Methodologies	3.1	The methodologies used by the insurer to estimate loss and LAE reserves are not performed using standard actuarial techniques or are not appropriate for the exposure.	RV	VA, AC, PD	Moderate-High	Severe	High					**See note below**	Strong Risk Controls	Moderate or High	Moderate-High				
Reserving Assumptions and Methodologies	3.2	Changes in the legal environment or changes in the insurer's underwriting, case reserving, or claims handling processes are not appropriately considered within the insurer's reserving assumptions and methodologies.	OP, RV, ST	VA, PD, AC	Moderate-Low	Moderate	Moderate						Moderate Risk Controls	Moderate	Moderate				
Performance of Reserve Calculations	4.1	The company does not use year end data to estimate its reserves. Errors may occur when the actuarial estimates are rolled forward to adjust to the reporting date reserves. The actuary does not reconcile data used in the loss development analysis with the financial statements.	RV	AC, VA, CO	Moderate-Low	Moderate	Moderate						Moderate Risk Controls	Moderate	Moderate-High				
Performance of Reserve Calculations	4.2	The actuarial calculations are not accurate or the actuarial assumptions and judgements are not appropriate, or selected estimates are not reasonable.	OP, RV	AC, VA, PD	Moderate-High	Severe	High						Moderate Risk Controls	Moderate or High	High				

Note: Column 3b includes references to the actuarial risk assessment memo, the analysis shown in Appendix B, and other company documents reviewed to assess the specific risks and controls for each row.

Phase One		Phase Two						Phase Three						Phase Four			Phase Five	Phase Six	Phase Seven
1d		2a	2b	2c	2d	2e	2f	3a				3b	3c	4a	4b	4c	5	6	7
		Risk Identification			Inherent Risk Assessment			Risk Mitigation Strategy/Control Assessment						Residual Risk Assessment					
Sub-activities	Identified Risk Number	Identified Risks	Branded Risk(s)	Exam Assertion(s)	Likelihood	Impact	Overall Inherent Risk Assessment	Risk Mitigation Strategy	Frequency	Samples Tested	Obtained from	Evidence & Document Testing Controls	Overall Risk Mitigation Strategy Assessment	Calculated Residual Risk	Judgmental Residual Risk	Overall Residual Risk Assessment	Examination Procedures / Findings	Prioritization Results Supervisory Plan	Report Findings & Management Letter Comments
Performance of Reserve Calculations	4.3	The computation of ceded reinsurance credits within loss and LAE reserves for internal and external reinsurance programs are not performed correctly or are not reasonable.	CR, RV	AC, VA	Moderate-Low	Moderate	Moderate					**See note below**	Moderate Risk Controls	Moderate	Moderate				
Performance of Reserve Calculations	4.4	The defense and cost containment loss adjustment expense (DCC or ALAE) estimates for direct business are not estimated using standard actuarial techniques, are not performed correctly, or the selected estimates are not reasonable.	CR, RV	AC, VA, CO	Moderate-High	Severe	High						Moderate Risk Controls	Moderate or High	Moderate-High				
Performance of Reserve Calculations	4.5	The unallocated loss adjustment expense (AOE, or ULAE) estimates are not estimated using standard actuarial techniques, are not performed correctly, or the selected estimates are not reasonable.	CR, RV	AC, VA, CO	Moderate-High	Moderate	Moderate						Strong Risk Controls	Low or Moderate	Moderate				
Performance of Reserve Calculations	4.6	New business may result in a development patterns that are different from the historical data. The actuarial methods and assumptions used to estimate reserves may contain a bias resulting in a build-up of differences across many years. Significant growth or expansions into new areas make it difficult to estimate the initial loss reserves.	OP, RV	VA	Moderate-High	Severe	High						Moderate Risk Controls	Moderate or High	Moderate-High				
Recording and reporting of loss reserves	5.1	Management books reserves that are materially different than the actuary's best estimate.	OP, ST, LG	VA, PD	Moderate-High	Moderate	Moderate						Moderate Risk Controls	Moderate	Moderate				
Recording and reporting of loss reserves	5.2	Loss reserves and loss adjustment expenses are not properly distributed and recorded amongst insurers in the reinsurance pooling arrangement.	OP	OB/OW, AC, CM	Moderate-High	Moderate	Moderate						Strong Risk Controls	Low or Moderate	Low				
Recording and reporting of loss reserves	5.3	Unauthorized changes could be made to adjuster limit's within the system allowing unauthorized changes in case basis claim reserves.	OP	AC, VA	Moderate-Low	Moderate	Moderate						Strong Risk Controls	Low or Moderate	Low				
Recording and reporting of loss reserves	5.4	Reserves are not properly monitored within management expectations.	OP, ST, LG	VA, PD	Moderate-Low	Moderate	Moderate						Moderate Risk Controls	Moderate	Moderate				

Note: Column 3b includes references to the actuarial risk assessment memo, the analysis shown in Appendix B, and other company documents reviewed to assess the specific risks and controls for each row.
Highlighted risks are documented by the examination actuary. Other items completed by other members of the examination team after collaboration with examination actuary.

Appendix B

Sheet 1

Sample Insurance Company Aggregated Risk Assessment

		Reserve Risk Assessment				
		Risk 5.1	Includes Risks 3.1, 3.2, 4.2, 4.3, 4.4, 4.5		Risk 4.6	
		MBE Difference	Auditor vs ACE Difference	Inherent Variability	Runoff Risk	Overall Residual Risk
1	**Loss**	**Moderate**	**Moderate**	**Moderate**	**Moderate-High**	**Moderate-High (1)**
2	Personal Company 1	High	Moderate	Moderate	High	Moderate-High
3	Personal Company 2	Low	Moderate-High	Moderate-High	Moderate-High	Moderate-High
4	Commercial Company 1	Moderate	Moderate	Moderate	Moderate	Moderate-High
5	**ALAE**	**Moderate**	**Moderate**	**Moderate**	**Moderate-High**	**Moderate-High (2)**
6	Personal Company 1	Moderate	Moderate	Moderate	Moderate-High	Moderate-High
7	Personal Company 2	Low	Moderate-High	Moderate-High	Moderate-High	Moderate-High
8	Commercial Company 1	Moderate	Moderate	Moderate	Moderate-Low	Moderate-High
9	**ULAE**	**Low**	**Moderate**	**Moderate**	**Low**	**Moderate (3)**
10	Personal Company 1	Low-Moderate	Moderate	Moderate	Low	Moderate
11	Personal Company 2	Low	Moderate	Moderate	Low	Moderate
12	Commercial Company 1	Low	Moderate	Moderate	Low	Moderate
13	**Assumed**	**Low**	**Low**	**Low**	**Low**	**Low**
14	Personal Company 1	Low	Low	Low	Low	Low
15	Personal Company 2	Low	Low	Low	Low	Low
16	Commercial Company 1					
17						
18	**Ceded**	**Moderate**	**Moderate**	**Moderate**	**Moderate**	**Moderate (4)**
19	Personal Company 1	Moderate	Moderate	Moderate	Moderate	Moderate
20	Personal Company 2	Low	Low	Low	Low	Low
21	Commercial Company 1	Low	Low	Low	Low	Low
22						
21	**All Segments**	**Moderate (5)**	**Moderate (7)**	**Moderate (7)**	**Moderate (6)**	**Moderate**

The ratings on Sheets 2-3 are based on our review of the inherent risk and the effectiveness of the company controls applied to a reserve segment basis.
The ratings by reserving segment are aggregated to the company level on Sheet 1, using a weighted average with reserve balances as weights.
The average residual risk represents a composite of many reserving risks identified in the Reserving Risk Matrix - Appendix A

(1) This represents a composite residual risk for the direct loss reserves. This rating is used to evaluate risks 3.1, 3.2, and 4.2 in Appendix A.
(2) This represents a composite residual risk for ALAE reserves. This rating is used to evaluate risks 4.4 in Appendix A.
(3) This represents a composite residual risk for ULAE reserves. This rating is used to evaluate risks 4.5 in Appendix A.
(4) This represents a composite residual risk for ceded reserves. This rating is used to evaluate risks 4.3 in Appendix A.
(5) This represents a composite residual risk for MBE-ACE differences. This rating is used to evaluate risks 5.1 in Appendix A.
(6) This represents a composite residual risk for actuarial bias and reserve runoff. This rating is used to evaluate risks 4.6 in Appendix A.
(7) This is a composite residual risk for all loss and LAE reserves. This rating is adjusted if the aggregation of small differences in estimates have increased the risk.

Appendix B
Sheet 2

Sample Insurance Company Company 1: Personal Lines - Risk Assessment on an Actuarial Review Segment Basis

	Estimates as of 12/31/2013				Auditor Estimates as of 12/31/2013				Runoff		Examination Team Risk Assessment					Phase Five
	Company					$ Difference	% Difference	% Difference	Change in	R = Auditor	Risk 5.1	Risks 3.1, 3.2, 4.2, 4.3, 4.4, 4.5		Risk 4.6		Test
Line of Business	MBE	ACE	Difference $	Difference %	Select	to Booked	to Booked	to ACE	Estimate	Reviewed	MBE Difference	Auditor vs ACE Difference	Inherent Variability	Runoff Risk	Overall Residual Risk	Plan
DIRECT																
TOTAL LOSS	**431,630**	**458,666**	**(27,036)**	**-5.9%**	**472,321**	**(40,691)**	**-8.6%**	**-2.9%**			**High**	**Moderate**	**Moderate**	**High**	**Moderate-High**	
PPA BI/UM Liability	305,000	330,000	(25,000)	-7.6%	342,100	(37,100)	-10.8%	-3.5%	**16,301**	**R**	High	Moderate	Moderate	High	High	Ind
PPA Prop. Damage Liability	21,500	21,636	(136)	-0.6%	22,450	(950)	-4.2%	-3.6%		**R**	Low	Moderate	Low	Low-Moderate	Moderate	Peer
PPA PIP/NF Liability	15,800	16,800	(1,000)	-6.0%	17,300	(1,500)	-8.7%	-2.9%		**R**	High	Moderate	Moderate	High	Moderate	Ind
Homeowners	38,110	38,110	(0)	0.0%	37,948	162	0.4%	0.4%	**953**	**R**	Low	Low	Low-Moderate	Low-Moderate	Low-Moderate	M&A
Umbrella	12,950	12,950	0	0.0%	13,353	(403)	-3.0%	-3.0%		**R**	Low	Moderate	Moderate	Low-Moderate	Low-Moderate	Peer
Dwelling Fire	2,450	2,450	0	0.0%	2,450	0	0.0%	0.0%			Low	Low	Low	Low-Moderate	Low	Diag
Inland Marine	420	420	0	0.0%	420	0	0.0%	0.0%			Low	Low	Low	Low-Moderate	Low	Diag
TOTAL ALAE	**42,382**	**39,855**	**2,527**	**6.3%**	**42,862**	**(480)**	**-1.1%**	**-7.0%**			**Moderate**	**Moderate**	**Moderate**	**Moderate-High**	**Moderate-High**	
PPA Liability	33,289	30,844	2,445	7.9%	33,755	(466)	-1.4%	-8.6%	**4,000**	**R**	Moderate	Moderate	Moderate	High	Moderate-High	Ind
PPA Physical Damage	333	333	0	0.0%	333	0	0.0%	0.0%			Low	Low	Low	Low	Low	Diag
Homeowners	6,800	6,713	87	1.3%	6,809	(9)	-0.1%	-1.4%	**844**	**R**	Low	Low-Moderate	Low-Moderate	Moderate	Low-Moderate	Peer
Umbrella	1,500	1,512	(12)	-0.8%	1,512	(12)	-0.8%	0.0%			High	Low	Low	Low	Low	Diag
Dwelling Fire	450	445	5	1.1%	445	5	1.1%	0.0%			Low	Low	Low	Low	Low	Diag
Inland Marine	10	8	2	24.8%	8	2	24.8%	0.0%			Low	Low	Low	Low	Low	Diag
ULAE	**43,000**	**44,000**	**(1,000)**	**-2.3%**	**43,000**	**0**	**0.0%**	**2.3%**		Auditor Accepted Booked!	**Low-Moderate**	**Moderate**	**Moderate**	**Low**	**Moderate**	Ind
TOTAL DIRECT LOSS & LAE	517,012	542,521	(25,509)	-4.7%	558,183	(41,171)	-7.4%	-2.8%								
TOTAL ASSUMED											**Low**	**Low**	**Low**	**Low**	**Low**	
FAIR Plan/Other Pools	6,428	6,428	0	0.0%	6,428	0	0.0%	0.0%			Low	Low	Low	Low	Low	Diag
TOTAL DIRECT & ASSUMED	**523,440**	**548,948**	**(25,509)**	**-4.6%**	**564,611**	**(41,171)**	**-7.3%**	**-2.8%**								
CEDED											**Moderate**	**Moderate**	**Moderate**	**Moderate**	**Moderate**	
Auto Liability	29,125	26,200	(2,925)	-11.2%	29,125	0	0.0%	10.0%		Auditor	Moderate	Moderate	Moderate	Moderate	Moderate	Ind
Homeowners	17,426	16,250	(1,176)	-7.2%	17,426	0	0.0%	6.8%		Accepted	Moderate	Moderate	Moderate	Moderate	Moderate	Ind
Fair Plan	2,800	2,800	0	0.0%	2,800	0	0.0%	0.0%		Booked!	Low	Moderate	Low	Low	Low	Diag
TOTAL CEDED LOSS & LAE	**49,351**	**45,250**	**(4,101)**	**-9.1%**	**49,351**	**0**	**0.0%**	**8.3%**								
TOTAL NET LOSS & LAE	**474,089**	**503,699**	**(29,610)**	**-5.9%**	**515,261**	**(41,171)**	**-8.0%**	**-2.2%**								

Sample Insurance Company Company 2: Personal Lines - Risk Assessment on an Actuarial Review Segment Basis Appendix B

Note * Detail for only two segments of business are shown in the example. Loss, ALAE and ULAE Total represent sum of all segements. Sheet 3

	Estimates as of 12/31/2013				Auditor Estimates as of 12/31/2013				Runoff		Examination Team Risk Assessment					Phase
	Company					$ Difference	% Difference	% Difference	Change in	R = Auditor	Risk 5.1	Risks 3.1, 3.2, 4.2, 4.3, 4.4, 4.5		Risk 4.6		Five
Line of Business	MBE	ACE	Difference $	Difference %	Select	to Booked	to Booked	to ACE	Estimate	Reviewed	MBE Difference	Auditor vs ACE Difference	Inherent Variability	Runoff Risk	Overall Residual Risk	Test Plan
DIRECT																
Loss	**113,875**	**114,750**	**(875)**	**-0.8%**	**121,345**	**(7,470)**	**-6.2%**	**-5.4%**			**Low**	**Moderate-High**	**Moderate-High**	**Moderate-High**	**Moderate-High**	
PPA BI	66,000	66,400	(400)	-0.6%	72,242	(6,242)	-8.6%	-8.1%	**1,406**	**R**	Low	High	High	High	High	Ind
PPA PIP	5,000	5,100	(100)	-2.0%	5,853	(853)	-14.6%	-12.9%	**1,600**	**R**	Low	High	Moderate	High	Low	All
*																
ALAE	**26,000**	**26,005**	**(5)**	**0.0%**	**27,972**	**(1,972)**	**-7.0%**	**-7.0%**			**Low**	**Moderate-High**	**Moderate-High**	**Moderate-High**	**Moderate-High**	
PPA BI	14,500	14,650	(150)	-1.0%	16,115	(1,615)	-10.0%	-9.1%	**1,743**	**R**	Low	Moderate-High	High	High	High	Ind
PPA PIP	4,250	4,130	120	2.9%	4,632	(382)	-8.3%	-10.8%	**1,030**	**R**	High	Moderate-High	High	High	Moderate	All
*																
Total Loss & ALAE	**139,875**	**140,755**	**(880)**	**-0.6%**	**149,317**	**(9,442)**	**-6.3%**	**-5.7%**								
ULAE	**9,000**	**9,000**	**0**	**0.0%**	**9,450**	**(450)**	**-4.8%**	**-4.8%**			**Low**	**Moderate**	**Moderate**	**Low**	**Moderate**	Ind
*																
TOTAL DIRECT Loss & LAE	**148,875**	**149,755**	**(880)**	**-0.6%**	**158,767**	**(9,892)**	**-6.2%**	**-5.7%**								
ASSUMED	8,950	8,950	0	0	8,950	0	0.0%	0.0%			Low	Low	Low	Low	Low	Diag
TOTAL DIRECT & ASSUMED	**166,775**	**167,655**	**(880)**	**-0.5%**	**176,667**	**(9,892)**	**-5.6%**	**-5.1%**								
TOTAL CEDED	**575**	**575**	0	0.0%	575	0	0.0%	0.0%			Low	Low	Low	Low	Low	Diag
TOTAL NET	**166,200**	**167,080**	**(880)**	**-0.5%**	**176,092**	**(9,892)**	**-5.6%**	**-5.1%**								

Sample Insurance Company Company 1: Commercial Lines - Risk Assessment on an Actuarial Review Segment Basis Appendix B

Note * Detail for only two segments of business are shown in the example. Loss, ALAE and ULAE Total represent sum of all segements. Sheet 4

	Estimates as of 12/31/2013				Auditor Estimates as of 12/31/2013				Runoff		Examination Team Risk Assessment					Phase
	Company					$ Difference	% Difference	% Difference	Change in	R = Auditor	Risk 5.1	Risks 3.1, 3.2, 4.2, 4.3, 4.4, 4.5		Risk 4.6		Five
Line of Business	MBE	ACE	Difference $	Difference %	Select	to Booked	to Booked	to ACE	Estimate	Reviewed	MBE Difference	Auditor vs ACE Difference	Inherent Variability	Runoff Risk	Overall Residual Risk	Test Plan
DIRECT																
Loss	**661,438**	**674,194**	**(12,756)**	**-1.9%**	**681,626**	**(20,188)**	**-3.0%**	**-1.1%**			**Moderate**	**Moderate**	**Moderate**	**Moderate**	**Moderate-High**	
CMP	291,724	291,724	0	0.0%	292,099	(376)	-0.1%	-0.1%	**(2,169)**	**R**	Low	Low	Moderate	Low	Low- Moderate	Peer
Commercial Auto Liability	101,761	101,761	0	0.0%	105,268	(3,507)	-3.3%	-3.3%	**(500)**	**R**	Low	Moderate	Moderate	Low	Moderate	Peer
*																
ALAE	**83,806**	**85,899**	**(2,092)**	**-2.4%**	**89,614**	**(5,808)**	**-6.5%**	**-4.1%**			**Moderate**	**Moderate**	**Moderate**	**Moderate-Low**	**Moderate-High**	
CMP	32,090	32,090	0	0.0%	32,987	(897)	-2.7%	-2.7%	**871**	**R**	Low	Moderate	Moderate	Moderate	Moderate	Ind
Commercial Auto Liability	9,667	9,667	0	0.0%	9,905	(238)	-2.4%	-2.4%	**(40)**	**R**	Low	Moderate	Moderate	Low	Low- Moderate	Peer
*																
Total Loss & ALAE	**745,245**	**760,093**	**(14,848)**	**-2.0%**	**771,240**	**(25,995)**	**-3.4%**	**-1.4%**								
ULAE	**79,373**	**80,903**	**(1,531)**	**-1.9%**	**80,903**	**(1,531)**	**-1.9%**	**0.0%**			**Low**	**Moderate**	**Moderate**	**Low**	**Moderate**	Ind
*																
TOTAL DIRECT Loss & LAE	**824,617**	**840,996**	**(16,379)**	**-1.9%**	**852,143**	**(27,526)**	**-3.2%**	**-1.3%**								
TOTAL CEDED	5,546	5,546	0	0.0%	5,546	0	0.0%	0.0%			Low	Low	Low	Low	Low	Diag
TOTAL NET	**819,071**	**835,450**	**(16,379)**	**-2.0%**	**846,597**	**(27,526)**	**-3.3%**	**-1.3%**								

Premium Deficiency Reserve Evaluation for Mortgage Insurers

David Kaye, FCAS, MAAA

Abstract
This paper will provide practical guidance for the actuary evaluating premium deficiency reserves for mortgage insurers. The paper includes a brief discussion of the premium deficiency accounting considerations for mortgage insurance, and introduces a practical deterministic approach for evaluating whether a premium deficiency reserve is necessary for mortgage insurers.

Keywords. Mortgage insurance; premium deficiency reserve; PDR.

1. INTRODUCTION

Beginning in 2007 and continuing for several years, the mortgage insurance (MI) industry experienced significant increases in losses, driven by the deterioration of several interrelated macroeconomic factors, principally, negative home price appreciation (HPA) and elevated unemployment levels. Given the prolonged period of elevated MI losses, mortgage insurers, their auditors and insurance regulators placed greater emphasis on the importance of evaluating whether a premium deficiency reserve (PDR) should be recorded on mortgage insurers' balance sheets.

The meaningful differences that exist between MI and short duration insurance products (e.g., a workers' compensation policy) require a specialized framework when evaluating MI PDR.

1.1 Objective

This objective of this paper is to provide the practicing actuary with:

- Sufficient background on the MI accounting requirements to understand the evaluation of premium deficiency reserve; and
- A simple deterministic framework for evaluating MI PDR.

This paper uses a simulated data set to familiarize the practicing actuary with the MI loss and premium process and provides a basic deterministic framework for analyzing PDR for MI companies. It should be noted, however, that there are various macroeconomic factors (principally, home price appreciation, unemployment and interest rates) that have a significant impact on MI claim and premium experience; these factors are not explicitly addressed in this paper but should be considered when evaluating PDR. These factors result in claim and premium processes that are

more complex than the simulated data utilized to demonstrate the methodology presented in this paper.

1.2 Outline

The remainder of the paper proceeds as follows: Section 2 provides MI background including a brief discussion regarding MI accounting framework with particular emphasis on PDR requirements and a description of common terminology used throughout this paper[1]; Section 3 provides a deterministic framework for evaluating MI PDR and discusses limitations and potential enhancements of the model presented herein.

2. BACKGROUND

2.1 Background on Mortgage Insurance

MI policies differ from typical short duration products familiar to most P&C actuaries in one key way: mortgage insurance policies have effectively unlimited terms and generate premium and losses for many years (in contrast to typical short duration contracts which generally remain in effect for one year or less). The implication for PDR for MI exposure is that the practitioner must project future premium and losses for loans originated on or before the evaluation date for many years into the future. Over the projection period, however, macroeconomic factors that influence claims and policy persistency can change significantly and result in significant deviation between historical performance and performance over the projection period.

Additional key features of MI policies include the following[2]:

- MI policies are issued at the time that the mortgage is issued and can either be paid by the borrower (most common) or lender (less common).

[1] This paper presumes a level of familiarity with MI. For a more detailed primer on MI, see reference [1].
[2] Reprinted from [1].

- Premiums are paid on either a monthly (most common) or single up-front (less common) basis. The premium associated with monthly pay policies is typically paid as part of the monthly mortgage payment.

- The collected monthly premiums are generally recognized as income in the period in which they are collected (that is, the monthly premiums are written and earned at the same time) meaning that there is typically a very small (or no) unearned premium reserve associated with monthly paid MI policies. There is an unearned premium reserve associated with single up-front premium policies, which is amortized over the life of the MI contract as losses associated with the contract is expected to emerge; however monthly pay policies are more common than up front policies.

- MI coverage is typically expressed as a percentage of a loan's unpaid principal balance ("UPB"). These coverage percentages vary from loan to loan, but a typical average coverage percentage is around 25%.

- MI policies provide lenders coverage for a portion of the UPB stipulated in the contract (generally around 25%). In addition, the MI policy generally reimburses the coverage beneficiary for lost interest payments and certain foreclosure-related expenses.

- Unlike typical Property and Casualty insurance policies, which are generally in force for one year and have defined termination dates, MI policies often generate premiums and losses for a number of years and there is uncertainty with regard to how long each policy will remain in force. The MI policy holder may exit the insured population for a number of reasons, including defaulting on the mortgage (i.e., becoming a claim), refinancing the loan, or paying down the principal on the loan to the point that the loan no longer requires MI.

- MI losses are highly correlated with macroeconomic factors such as home price appreciation and unemployment. As was highly evident in 2007-2011, MI company results were adversely affected by a steep drop in home prices followed by rising levels of unemployment. Not surprisingly, the states with the sharpest decreases in home prices – CA, FL and NV – were significant drivers of adverse loss experience for the MI industry.

- As explained further below, MI loss reserves are recorded at the time when a borrower is "delinquent" in paying their mortgage. This results in an unusual accounting construct where the timing of premium earning and loss accrual are not matched. In other words, premium revenue from MI policies is recognized (i.e., earned) prior to the associated losses being recognized.

- For a cohort of monthly paid MI policies issued during a year, the premium revenue generated by the policies is the greatest during the first year and then decreases over the next ten years as policies exit the population. Those that exit the population through delinquency, thereby giving rise to the recording of MI loss reserves tend to rise through third or fourth year after loan origination. After peaking, incremental losses tend to decrease as policies continue to exit the population.

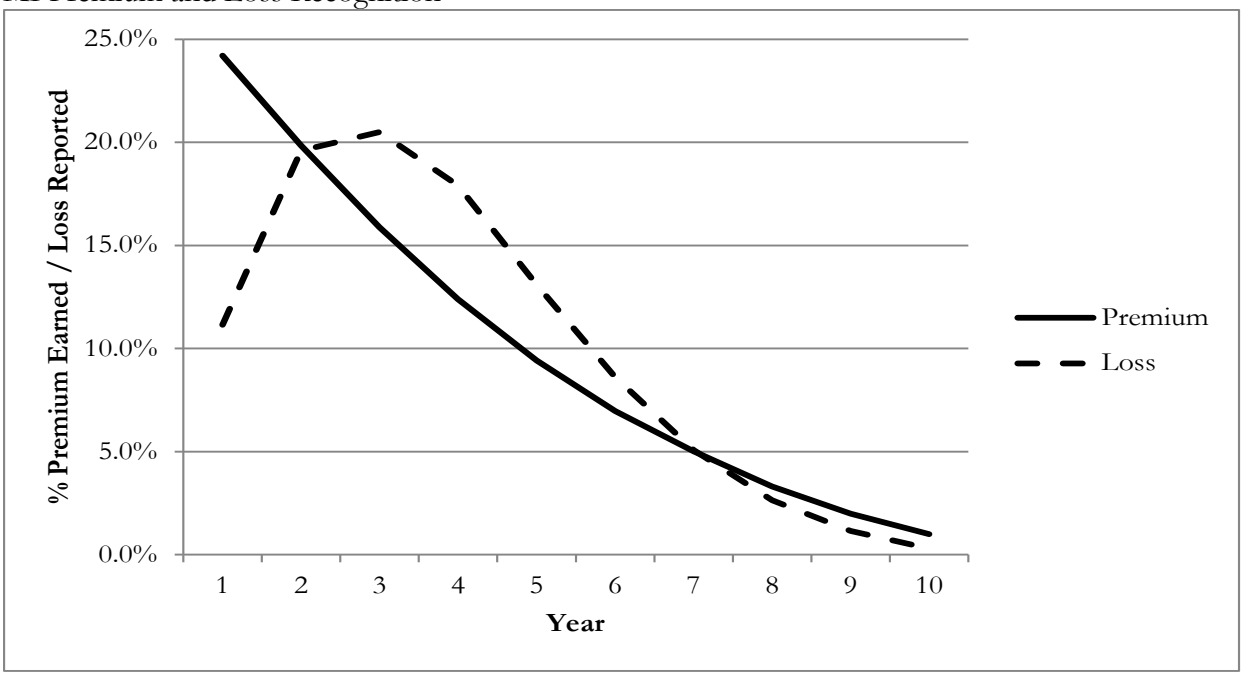
2.2 Accounting for Mortgage Insurance Losses

The accounting framework for MI results in a departure of one of the principal objectives of accounting: revenue and expense matching. For typical, single-year (or shorter duration) P&C insurance products, both revenue (premium) and expense (claim costs) are recognized uniformly through the period the policy is effective[3]. For the majority of MI policies written in the U.S., premium is earned on a monthly basis, while losses on MI policies are not recognized until the borrower stops paying the monthly mortgage payment and the lender or loan servicer notifies the MI company that the borrower is delinquent. The result of this accounting framework is that premiums are highest during the first year following the loan origination for a cohort of policies

[3] This description is generally accurate, although there are exceptions such as property catastrophe cover where premium and loss might not be recognized uniformly through the policy period.

while losses generally rise through the first few years after origination, peak around the third or fourth year and then decline over time as loans from the cohort exit the population. Graph 1 demonstrates this relationship for a cohort of loans written during the same calendar year[4]:

Graph 1
MI Premium and Loss Recognition



While MI companies follow the general framework described above for general premium and loss accounting, Statement of Statutory Accounting Principle (SSAP) 58 provides statutory accounting guidance for PDR[5]. SSAP 58, paragraph 23 states:

> When the anticipated losses, loss adjustment expenses, commissions and other acquisition costs, and maintenance costs exceed the recorded unearned premium reserve, contingency reserve, and the estimated future renewal premium on existing policies, a premium deficiency reserve shall be recognized by recording an additional liability for the deficiency with a corresponding charge to operations. Commissions and other acquisition costs need not be considered in the premium deficiency analysis to the extent they have not been

[4] In this paper, we refer to years in which loans are underwritten as "book year".

[5] GAAP accounting guidance does not specifically address MI, therefore, MI companies typically utilize statutory accounting guidance in preparing the GAAP accounting statements. A primary difference between statutory and GAAP accounting statements is the existence of a contingency reserve required by statutory accounting guidance, but not allowed under GAAP.

expensed. If an insurer utilizes anticipated investment income as a factor in the premium deficiency calculation, disclosure of such shall be made in the financial statements[6].

A key point to note is that a PDR is only required when the sum of projected future losses and expenses (outflows) exceeds the sum of projected future premiums (inflows), unearned premium reserves, and contingency reserves; nothing is recorded in the event that the opposite of a PDR (a "premium sufficiency") is estimated. In practice, when estimating PDR, MI companies project cash payments related to all loans in the portfolio as of the measurement date inclusive of cash payments related to loans that are delinquent (and therefore included in the recorded loss reserves) at the evaluation date (facilitating discounting of the projected cash flows). The recorded loss reserves are deducted from the projected cash flows to avoid double counting the outflows related to loans making up the recorded loss reserves at the financial statement date.

2.3 Terminology and Organization of Data

Before providing a framework for estimating MI PDR, it is important to introduce several additional terms as well as to lay out the key characteristics used to organize the data.

2.3.1 Terminology

Although the terminology below is not necessarily universal, it is used throughout the remainder of this paper.

- Book year (and book half year): The year (or half year) in which a cohort of MI policies is issued. For example, MI policies written during 2014 will be referred to as "book year 2014 policies". In Section 3, we organize the data by book half year with the format 20XX-1 representing loans originated during the first six months of 20XX and 20XX-2 representing loans originated during the second six months of 20XX[7].
- Risk in force ("RIF"): The exposure to loss faced by MI companies. The RIF is calculated by multiplying the MI's coverage percentage by the loan's UPB. In addition to the coverage percentage multiplied by the UPB, the MI company may also be required to pay lost interest

[6] MI companies typically recognize investment income by discounting projected cash flows at an appropriate discount rate. The selection of an appropriate discount rate is beyond the scope of this paper; for illustrative purposes, a 2% discount rate has been used in the calculations shown in the Appendix.

[7] Although we use book half year in the Appendix, we note that the actuary could consider book year or book quarter.

and certain foreclosure expenses; for this reason, the ratio of claim payments to RIF may be greater than 100%.

- Policies in force ("PIF"): The total number of policies in force as of a particular evaluation date.
- Policy persistency (persistency): The portion of policies that remain in force from one period of time to the next.
- Submitted claim: A delinquent loan where the borrower has not made mortgage payments, the lending institution has foreclosed on the subject property, and a claim has been submitted to the MI company. Within this paper we will consider both claim payments (dollars) and claim counts.
- Outstanding delinquency: A loan reported to the MI company when the borrower has fallen two mortgage payments behind (note, there is some variance about when a loan is identified as delinquent in the MI industry, here we are assuming the MI company has set the definition as the borrower being behind two or more payments).

2.3.2 Data organization

For purposes of the method described in Section 3, the data will be organized by book half year with semi-annual evaluations. Further segmentation of the data in the Appendix is not addressed; however, in practice, the actuary should consider how to best segment the data for use in the model described in Section 3[8]. There are several items that an actuary might consider when determining appropriate segmentation for the data within the PDR analysis as described below:

- Unemployment: Varying levels of unemployment can have significant impacts on MI claim activity. States or regions with higher levels of unemployment are more likely to also experience elevated claim activity which might have a meaningful impact on the PDR analysis.
- Credit worthiness of borrowers: The credit worthiness of borrowers can be a significant predictor in determining borrower behavior. FICO score[9] or distinguishing between Prime and Subprime loans in developing estimates can result in better data stratification.

[8] Actuarial literature contains a number of papers written about effective data segmentation. For example, see [2].
[9] FICO is a common credit scoring mechanism developed originally by the Fair Isaac Corporation. The FICO score is a numerical representation of the credit worthiness of a borrower.

- Home price appreciation or depreciation: Different states or regions might have different levels of home price appreciation that can impact both claim behavior and policy persistency.

During the 2008 housing market downturn, MI companies observed elevated MI claim submissions from states that had significant increases in home prices prior to the housing market downturn followed by significantly elevated unemployment levels resulting from the subsequent recession. For this reason, during the last market downturn, some MI companies chose to separately analyze California, Florida, and Nevada; these "sand states" were particularly hard hit by the combination of a significant housing market collapse and elevated unemployment and displayed similar, elevated claim characteristics.

Accounting principles require that the need for a PDR be evaluated at the level at which the MI company manages its insurance portfolio; the actuarial evaluation of PDR may or may not coincide with the level at which the MI company manages the insurance portfolio so the actuarial analysis may need to be aggregated in order to align with the Company's required PDR segmentation.

3. DETERMINISTIC FRAMEWORK FOR MI PDR

As described in the previous section we will estimate PDR by comparing:

a) the net present value of expected future losses and policy maintenance expenses for the MI company's business in force and

b) the net present value of expected future premiums, existing unearned premium reserves, unpaid claim reserves, DAC (if any), and contingency reserves (if any).

If, in our evaluation a) exceeds b), then the MI company should record a PDR.

We note that the focus of this paper is on the estimation of the present value of cash flows related to future premiums, policy maintenance expenses and losses. The financial statement items (unearned premium reserves, unpaid claim reserves and contingency reserves) are presumed to be known by the actuary at the financial statement date and their estimation is beyond the scope of this paper.

3.1 Estimating the Cash Flows for PDR

In developing the cash flows used to assess whether a PDR is necessary, we will separately estimate premiums (adding a provision for policy maintenance expenses) and losses and then

consider the financial statement items to determine whether a PDR is necessary. Sections 3.1.1 through 3.1.3 below provide a description of the PDR calculations and Section 3.1.4 provides a highly simplified sample PDR calculation. The Appendix includes a more realistic example based on simulated MI data (the sections below provide references to the calculations detailed in the Appendix).

3.1.1 Estimating future premiums

To estimate the future premiums, we will organize historical PIF data into a triangular data format familiar to P&C actuaries, with the rows representing book half year exposure periods and the columns representing semi-annual evaluation periods. The PIF data gathered to create the data triangle represents the remaining PIF at the end of each evaluation date, therefore, the PIF data decreases over time as the MI portfolio unwinds.

Estimating the future premiums as described in this paper is a three step process (the calculations are detailed in Appendix, Exhibit 3):

1. First, we evaluate the decline in PIF over time as the MI portfolio unwinds by estimating PIF persistency ("persistency"). The persistency is developed by calculating ratios of PIF at each evaluation period, i+1, divided by the PIF at the preceding evaluation period, i. The triangle of PIF is completed by selecting a persistency factor for each evaluation period and then applying the selected decay factor at each period to the PIF observed (or projected) at the end of the prior evaluation period. Performing these calculations allows the actuary to estimate the PIF for each exposure period at each future evaluation period.

   As an example, Appendix Exhibit 3, Table 1 indicates that at the end of December 2015, there are 6,923 policies in force for loans written during the second half (July 1 – December 31) of 2015 (the 2015-2 cohort). By using historical relationships of PIF decay (see Appendix Exhibit 3, Table 2), we estimate that 90.7% of policies will remain in force at the end of the next evaluation period. Therefore, we project 6,283 (=6,923 x 90.7%) policies in force at June 30, 2016 from the 2015-2 cohort. Proceeding in this manner for all projection periods, we can project PIF for each future period as shown in Appendix Exhibit 3, Table 4.

2. Based on the projected PIF, we calculate the average PIF for each future evaluation period as shown in Appendix Exhibit 3, Table 5. Note, that we have shifted the triangle

in Appendix Exhibit 3, Table 5 so that each column represents a future calendar half year; shifting the projections in this manner facilitates discounting of the projected premium cash flows (related to this point, below Appendix Exhibit 3, Table 5, discount factors are calculated assuming on average, premiums are collected in the middle of each projection period[10]).

3. Using the projected average PIF, we calculate the projected future premium by multiplying the average monthly premium for each book half year by the average PIF over each projection period and multiply the result by 6 (the premiums are monthly average premiums while each projection period represents 6 months of exposure). The calculations and results are shown in Appendix Exhibit 3, Table 6.

   Note that Appendix Exhibit 3, Table 6 utilizes the discount factors calculated in Appendix Exhibit 3, Table 5 to determine the discounted projected future premium necessary for the PDR calculation. For illustrative purposes, the cash flows are discounted using a 2% discount rate assumption.

3.1.2 Estimating future claims and loss adjustment expenses

In order to estimate the future claim payments, we begin with a triangle of paid claim counts, which we will utilize to perform two standard actuarial methodologies – a traditional "chain ladder" development method (referred to in this paper and exhibits as Claim Development Method, or CDM and a Bornhuetter-Ferguson method, or BFM). The future claim estimates are outlined in Exhibit 2 of the Appendix.

The CDM should be recognizable by P&C actuaries. In preparing our estimates, we calculate a triangle of claim count development factors (Appendix Exhibit 2, Table 2), compute average development factors, select claim development factors and apply the calculated cumulative development factors to latest evaluation of the claim count triangle in order to develop an estimate of ultimate claim counts[11]. The claim count estimate represents the projected ultimate claims for each book half year.

The CDM results are shown in Appendix Exhibit 2, Table 3, Column 4. We utilize the CDM

[10] Given that the policies are declining over the future periods, assuming the middle of each projection period results in slightly higher discount than what would be calculated using a more refined assumption (i.e., factoring in the declining portfolio). The mid-period assumption is utilized here for simplicity, although enhancements to the calculation could be made if the actuary chooses to do so.

[11] See [3] for a more detailed discussion of the CDM and BFM

results to provide guidance in developing expected loss estimates for use in the BFM by first dividing the CDM results by the number of loans originated in each half year and then selecting an expected claim count per loan count rate. The expected count per loan count and the claim development pattern underlying the CDM to develop a BFM as in Appendix Exhibit 2, Table 3, Column 7.

We select paid claims based on the results of the CDM and the BFM and then multiply the estimate by the expected average claims size. The expected average claim size, in turn, is calculated by dividing the claim dollars paid to date by the count of claims paid to date. In general, the claim severity tends to be more stable and predictable as claim payments are closely aligned with RIF and RIF tends to be stable over time, although the actuary should pay attention to observed or expected shifts in future claim severity relative to recent historical severity experience[12].

The estimate of ultimate claim costs is calculated by multiplying the ultimate claim count estimate by the projected average claim size; subtracting the claims paid as of the evaluation date results in the forecasted unpaid claims. The forecasted unpaid claims are discounted using the development factors underlying the CDM and a LAE factor is added to represent the total projected discounted loss and LAE cash flows.

3.1.3 PDR estimates

We use the discounted premium and loss cash flow projections to determine whether a PDR is necessary at the evaluation date. A PDR is recorded if the estimated net discounted loss and LAE) exceeds discounted premiums net of policy maintenance expenses plus financial statement items related to premiums and losses (unearned premium reserves, and recorded loss and LAE reserves)[13]. If the cash flows are negative and the absolute value of the cash flows is greater than the financial statement items related to premiums and losses, then we record a PDR (nothing is recorded if the conditions are not met). In the Appendix, Exhibit 1 displays the determination of whether a PDR is necessary at the evaluation date (in the example in the Appendix, no PDR is necessary).

[12] For mortgage insurance, separate estimation of frequency and severity is often preferable since severity tends to be closely tied to RIF and is therefore generally more stable and easier to estimate than frequency.
[13] On a statutory basis, we also include contingency reserves.

3.1.4 A Simplified PDR Example

The Appendix to this paper contains a detailed sample calculation showing the PDR estimation framework described above, however a simplified sample is presented in this section to facilitate understanding of the process described herein.

The following data and assumptions are provided for this simple example (note, in this example, we are determining whether a PDR is necessary for a single cohort of policies written during a single calendar period, which is not consistent with actual practice where we would determine whether a PDR is necessary across a portfolio of MI policies):

A. During 2015, 1,100 MI policies are written. At year-end 2015 (the financial statement date), 1,000 of the 1,100 policies remain.
B. All of the policies are monthly pay premium and there is no unearned premium reserve related to single-pay policies.
C. We expect 250 policies to exit the population during each subsequent calendar year (i.e., 750 policies remain at year-end 2016, 500 policies remain at year-end 2017, etc.) until all policies exit the population by year-end 2019.
D. During 2015, the average monthly policy premium was $100 / policy / month.
E. The contingency reserve recorded at year-end 2015 for the loans in the cohort is $300,000.
F. At year-end 2015, there are 50 delinquent loans with average RIF on the loans of $40,000 and a recorded loss and LAE reserve of $400,000. The projected average payment date for the loss reserves is June 30, 2016. Further, we assume that $40,000 is a reasonable estimated severity for claims paid during subsequent calendar years.
G. Using historical claim data, we have projected 10 claims to be paid during 2017, 25 in 2018, 20 in 2019 and 5 in 2020. We assume the average payment date is June 30 for the paid claims.
H. Policy maintenance expenses are assumed to be 3% of the forecasted premiums and loss adjustment expenses are assumed to be 5% of the forecasted paid losses. Both items (policy maintenance expenses and loss adjustment expenses) are assumed to be expenses in the period in which the premiums and losses are paid.
I. The illustrative discount rate selected for the example is 1.5% / annum.

Table 1 outlines the methodology used to develop the projected and discounted premium flows over the projection period. Note that the notes referenced in Table 1 reference data provided in the assumptions and data list directly above.

Table 1: Projected and discounted premiums								
		Note / formula	12/31/15	12/31/16	12/31/17	12/31/18	12/31/19	Total
(1)	PIF	Given in (A) & (C)	1000	750	500	250	0	N/a
(2)	Average PIF	Average of CY values in (1)		875	625	375	125	N/a
(3)	Premium / policy / month	Given in (D)		100	100	100	100	N/a
(4)	Projected annual premium	(2)x(3)x12.0		1,050,000	750,000	450,000	150,000	2,400,000
(5)	Discount years			0.5	1.5	2.5	3.5	
(6)	Discount factor	1.0 /1.015 ^ (5)		0.993	0.978	0.963	0.949	
(7)	Discounted premium	(4)x(6)		1,042,213	733,436	433,558	142,384	2,351,590

Note that the premium presented in item D is presented as an average per policy per month, therefore in Table 1, item 4, the calculation is multiplied by 12 to represent a full year of premium for the average number of policies in force in each calendar year. The resulting discounted premium shown in the total column of line 7 of $2,351,590 is used in Table 3 below to determine if a PDR is needed at year-end 2015.

Table 2 outlines the methodology used to develop the projected and discounted loss cash flows over the projection period.

Table 2: Projected and discounted paid losses

		Note / formula	6/30/16	6/30/17	6/30/18	6/30/19	6/30/20	Total
(1)	Paid claim counts	Given in (G)	N/a	10	25	20	5	60
(2)	Average size per claim	Given in (F)	40,000	40,000	40,000	40,000	40,000	N/a
(3)	Paid claim dollars	2016 from (F), other (1)x(2)	400,000	400,000	1,000,000	800,000	200,000	2,800,000
(4)	Discount years		0.5	1.5	2.5	3.5	4.5	
(5)	Discount factor	1.0 / 1.015 ^ (4)	0.993	0.978	0.963	0.949	0.935	
(6)	Discounted paid claim dollars	(4)x(6)	397,033	391,166	963,463	759,379	187,039	2,698,081

The resulting discounted paid claim dollars shown in the total column of line 6 of $2,698,081 is used in Table 3 below to determine if a PDR is needed at year-end 2015.

Table 3 uses the premium and paid claim amounts from Tables 1 and 2 to along with other amounts from the data and assumptions presented above to determine whether a PDR is necessary at year-end 2015.

Table 3: PDR Calculation

		Note / formula	Statutory Basis	GAAP Basis
(1)	Discounted premium	Table 1, Line (7)	2,351,590	2,351,590
(2)	Policy maintenance costs	3% x (1), 3% given in (H)	70,548	70,548
(3)	Discounted paid claim dollars	Table 2, Line (6)	2,698,081	2,698,081
(4)	Loss adjustment expense	5% x (3), 5% given in (H)	134,904	134,904
(5)	**Net cash flows**	**(1)-(2)-(3)-(4)**	**(551,942)**	**(551,942)**
(6)	Contingency reserve	Given in (E)	300,000	-
(7)	Recorded loss & LAE reserve	Given in (F)	400,000	400,000
(8)	Unearned premium reserve	Given in (B)	-	-
(9)	**Total - Financial Statement Items**	**(6)+(7)+(8)**	**700,000**	**400,000**
(10)	Net Cash Flows Plus Financial Statement Items	(5)+(9)	148,058	(151,942)
(11)	**Premium deficiency reserve**	**ABS{Min(0,(10)}**	**-**	**151,942**

Table 3 presents the PDR calculation using U.S. GAAP and U.S. statutory accounting principles. Because contingency reserves are not permissible under U.S. GAAP, the calculations outlined in Table 3 indicate a PDR of $151,942 on a GAAP basis, but no PDR on a statutory basis.

3.2 Benefits and Limitations of the Methodology Described in Section 3.1

The triangular methods for estimating premium and claim cash flows should have an appeal for actuaries since the triangular arrangement of the data is familiar to all actuaries and the mechanics of the model are intuitive and straightforward. The methodology is also appealing because it is more straightforward to describe to a non-actuarial audience than methods that require an understanding of statistical concepts (e.g., regression). Statistical methods are often referred to by non-technical audiences as "black box" methods because the inputs and outputs of the model are easy to describe, but the actual model mechanics are difficult to describe; the deterministic model described in this paper does not have this limitation.

In addition to being straightforward to describe to a non-actuarial audience, the organization of the data as outlined above and in the Appendix allows the actuary and management to estimate and monitor loss ratios for book years through time. Monitoring current and historical book year loss ratios can give actuaries and management insight on changes in underwriting, claims experience or portfolio persistency that drives the profitability of the MI company's insurance portfolio.

The key limitations of the deterministic framework are:

- Using aggregate data does not allow the actuary to explicitly model the factors that are most correlated with persistency and claim behavior. For example, persistency is highly correlated with interest rates; if interest rates fluctuate significantly during the historical experience period, but are not expected to fluctuate over the projection period, then the historical experience may not be representative of future performance. Conversely, regression models can be developed that utilize interest rates as an explanatory variable, which allows the actuary to quickly develop alternative estimates assuming different future interest rate paths.
- Related to the first point, the deterministic model does not allow for explicit sensitivity testing of the results to changes in macroeconomic factors. For example, if the MI company is concerned about the effect of an increase in unemployment on the Company's results, the effect cannot be explicitly incorporated into the framework; such modeling may be required by Government Sponsored Entities (GSE's)'s to determine the MI company's capital requirements.

Statistical regression models have the distinct advantage over the method described in this paper in that they allow for direct modeling of premium and losses in different macro-economic environments. However, the deterministic methods utilized in this paper could be enhanced by

looking at the impact of historical macroeconomic "shock" events on MI claim and persistency and using those historical relationships to calculate "stressed" scenarios of future performance.

4. CONCLUSIONS

The actuary who prepares premium and loss forecasts for MI companies must understand the unique MI accounting framework, including the evaluation of whether a PDR is required. Although the accounting for MI differs from traditional P&C insurance products, deterministic triangular methods commonly used to develop estimates for P&C products can help actuaries project delinquent loan behavior. After the actuary has a strong grasp of MI data, the accounting model and persistency and claim behavior, more complex regression or generalized linear model procedures can be utilized to further enhance MI premium and loss forecasts.

Acknowledgment

The author would like to thank Lynne Bloom, Bill Lakins, Tim Landick, and Miranda Ma for their valuable comments, suggestions and edits to this paper.

Supplementary Material

The Appendix follows this paper is also available electronically on the CAS website accompanying this paper. The dataset provided within the Appendix was simulated using constraints generally consistent with the author's knowledge of MI premium and claim.

5. REFERENCES

[1] Kaye, David, "Estimating Unpaid Claim Liabilities for Mortgage Insurance", *Casualty Actuarial Society E-Forum*, Fall 2013.
[2] Lakins, William, "Efficient Estimators Through Data Segmentation", *CAS Forum,* Fall 1998.
[3] Friedland, Jacqueline, "Estimating Unpaid Claims Using Basic Techniques", *Casualty Actuarial Society Practice Note*, July 30, 2010.

Abbreviations and notations

BFM, Bornhuetter-Ferguson method
CDM, claim development method
LAE, loss adjustment expenses
LTV, loan to value ratio
MI, mortgage insurance
PDR, premium deficiency reserve
PIF, policies in force
RIF, risk in force
UPB, unpaid principal balance

Biography of the Author

David Kaye is Director at PwC in Philadelphia, PA. He has a B.S. in Mathematics and a B.S. in Statistics from the Pennsylvania State University. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. David participates on the CAS Committee on Professionalism Education.

Cash Flows			
(1)	Discounted Premium Net of Policy Maintenance Flows		115,202,033
(2)	Discounted Loss & LAE Flows		92,714,955
(3)	Net Cash Flows	(1)-(2)	22,487,077
Financial Statement Items			
(4)	Recorded Loss and LAE Reserves		42,153,568
(5)	Unearned Premium Reserve		7,515,352
(6)	Statutory Contingency Reserve		111,251,356
(7)	Total - Financial Statement Items	Sum (4) - (6)	160,920,276
(8)	Net Cash Flows Plus Financial Statement Items	(3)+(7)	183,407,353
(9)	Premium Deficiency Reserve	Abs{Min[0,(8)]}	-

Table 1: Cumulative Paid Claim Count Data

Book Half Year	1	2	3	4	5	6	7	8	9	10	11
2004-1	1	8	15	22	29	36	43	50	80	125	147
2004-2	1	8	15	22	29	36	43	50	78	120	141
2005-1	4	11	18	25	32	39	53	60	81	123	137
2005-2	4	11	18	25	32	39	54	61	98	120	135
2006-1	3	10	17	24	31	38	45	52	79	106	127
2006-2	3	10	17	24	31	38	45	59	80	101	122
2007-1	5	12	19	26	33	40	54	61	82	117	138
2007-2	5	12	19	26	33	40	55	70	100	145	175
2008-1	1	8	15	22	29	36	51	66	88	133	163
2008-2	1	8	15	22	29	36	43	50	86	108	130
2009-1	3	11	19	27	35	43	51	66	89	119	149
2009-2	3	10	17	24	31	38	52	59	95	138	152
2010-1	4	11	18	25	32	39	54	69	106	136	166
2010-2	5	12	19	26	33	40	47	61	95	116	130
2011-1	5	12	19	26	33	40	55	62	99	143	
2011-2	5	12	19	26	33	40	55	70	100		
2012-1	1	8	15	22	29	36	50	64			
2012-2	3	10	17	24	31	38	45				
2013-1	5	12	19	26	33	40					
2013-2	4	11	18	25	32						
2014-1	3	10	17	24							
2014-2	2	9	16								
2015-1	1	8									
2015-2	2										

12	13	14	15	16	17	18	19	20	21	22	23	24
162	177	184	191	198	205	212	219	224	227	228	229	229
162	176	190	197	204	211	218	225	230	233	234	235	
165	172	179	186	193	200	207	214	219	222	223		
157	164	171	178	185	192	199	206	211	214			
154	168	175	182	189	196	203	210	215				
143	157	171	178	185	192	199	206					
152	159	173	180	187	194	201						
197	204	219	226	233	240							
178	185	200	207	214								
145	152	167	174									
164	179	187										
181	195											
188												

Table 2: Paid Claim Count Development Factors

Book Half Year	2/1	3/2	4/3	5/4	6/5	7/6	8/7	9/8	10/9	11/10	12/11
2004-1	8.000	1.875	1.467	1.318	1.241	1.194	1.163	1.600	1.563	1.176	1.102
2004-2	8.000	1.875	1.467	1.318	1.241	1.194	1.163	1.560	1.538	1.175	1.149
2005-1	2.750	1.636	1.389	1.280	1.219	1.359	1.132	1.350	1.519	1.114	1.204
2005-2	2.750	1.636	1.389	1.280	1.219	1.385	1.130	1.607	1.224	1.125	1.163
2006-1	3.333	1.700	1.412	1.292	1.226	1.184	1.156	1.519	1.342	1.198	1.213
2006-2	3.333	1.700	1.412	1.292	1.226	1.184	1.311	1.356	1.263	1.208	1.172
2007-1	2.400	1.583	1.368	1.269	1.212	1.350	1.130	1.344	1.427	1.179	1.101
2007-2	2.400	1.583	1.368	1.269	1.212	1.375	1.273	1.429	1.450	1.207	1.126
2008-1	8.000	1.875	1.467	1.318	1.241	1.417	1.294	1.333	1.511	1.226	1.092
2008-2	8.000	1.875	1.467	1.318	1.241	1.194	1.163	1.720	1.256	1.204	1.115
2009-1	3.667	1.727	1.421	1.296	1.229	1.186	1.294	1.348	1.337	1.252	1.101
2009-2	3.333	1.700	1.412	1.292	1.226	1.368	1.135	1.610	1.453	1.101	1.191
2010-1	2.750	1.636	1.389	1.280	1.219	1.385	1.278	1.536	1.283	1.221	1.133
2010-2	2.400	1.583	1.368	1.269	1.212	1.175	1.298	1.557	1.221	1.121	
2011-1	2.400	1.583	1.368	1.269	1.212	1.375	1.127	1.597	1.444		
2011-2	2.400	1.583	1.368	1.269	1.212	1.375	1.273	1.429			
2012-1	8.000	1.875	1.467	1.318	1.241	1.389	1.280				
2012-2	3.333	1.700	1.412	1.292	1.226	1.184					
2013-1	2.400	1.583	1.368	1.269	1.212						
2013-2	2.750	1.636	1.389	1.280							
2014-1	3.333	1.700	1.412								
2014-2	4.500	1.778									
2015-1	8.000										
Sel. Claim Development Factor	**4.271**	**1.701**	**1.409**	**1.289**	**1.225**	**1.293**	**1.212**	**1.493**	**1.389**	**1.179**	**1.143**
Cumulative DF	100.801	23.601	13.874	9.850	7.639	6.237	4.824	3.981	2.666	1.920	1.628

13/12	14/13	15/14	16/15	17/16	18/17	19/18	20/19	21/20	22/21	23/22	24/23
1.093	1.040	1.038	1.037	1.035	1.034	1.033	1.023	1.013	1.004	1.004	1.000
1.086	1.080	1.037	1.036	1.034	1.033	1.032	1.022	1.013	1.004	1.004	
1.042	1.041	1.039	1.038	1.036	1.035	1.034	1.023	1.014	1.005		
1.045	1.043	1.041	1.039	1.038	1.036	1.035	1.024	1.014			
1.091	1.042	1.040	1.038	1.037	1.036	1.034	1.024				
1.098	1.089	1.041	1.039	1.038	1.036	1.035					
1.046	1.088	1.040	1.039	1.037	1.036						
1.036	1.074	1.032	1.031	1.030							
1.039	1.081	1.035	1.034								
1.048	1.099	1.042									
1.091	1.045										
1.077											
											Factor to Ult
1.066	**1.065**	**1.039**	**1.037**	**1.036**	**1.035**	**1.034**	**1.023**	**1.014**	**1.004**	**1.004**	**1.002**	**1.002**
1.424	1.336	1.254	1.207	1.165	1.124	1.086	1.050	1.027	1.013	1.008	1.004	1.002

Table 3: Claim Estimates

	(1) Original Loan Population	(2) Total Paid Claims	(3) Development Factor	(4) CDM Estimated Claims (2)x(3)	(5) CDM Estimate / Loan Count (4)x(1)	(6) Expected Claims / Loan Count	(7) BFM Estimated Claims (A)	(8) Selected Estimated Claims	(9) Average Claim Size	(10) Estimated Ultimate Claims (8)x(9)	(11) Claims Paid A.o. 12/31/15	(12) Forecasted Unpaid Claims (10)-(11)	(13) Discounted Unpaid 2%	(14) LAE Load 5%	(15) Discounted Loss & LAE (13)x(14)
2004-1	7,483	229	1.002	229	3.1%	3.3%	229	229	38,952	8,937,848	8,920,008	17,840	17,752	1.05	18,640
2004-2	6,970	235	1.004	236	3.4%	3.3%	236	236	38,996	9,200,753	9,164,060	36,693	36,332	1.05	38,148
2005-1	6,923	223	1.008	225	3.2%	3.3%	225	225	40,050	9,005,735	8,931,150	74,585	73,690	1.05	77,374
2005-2	7,429	214	1.013	217	2.9%	3.3%	217	217	39,764	8,618,319	8,509,496	108,823	107,086	1.05	112,441
2006-1	6,845	215	1.027	221	3.2%	3.3%	221	221	38,862	8,577,167	8,355,330	221,837	218,499	1.05	229,424
2006-2	6,902	206	1.050	216	3.1%	3.3%	217	216	38,758	8,387,099	7,984,148	402,951	396,662	1.05	416,495
2007-1	7,095	201	1.086	218	3.1%	3.3%	220	218	39,146	8,546,174	7,868,346	677,828	666,119	1.05	699,425
2007-2	7,490	240	1.124	270	3.6%	3.3%	267	270	38,980	10,519,708	9,355,200	1,164,508	1,140,386	1.05	1,197,405
2008-1	7,498	214	1.165	249	3.3%	3.3%	249	249	39,647	9,881,804	8,484,458	1,397,346	1,362,634	1.05	1,430,765
2008-2	7,271	174	1.207	210	2.9%	3.3%	215	210	39,812	8,364,555	6,927,288	1,437,267	1,395,264	1.05	1,465,027
2009-1	7,500	195	1.254	245	3.3%	3.3%	245	245	39,564	9,674,521	7,714,980	1,959,541	1,893,593	1.05	1,988,272
2009-2	7,147	195	1.336	261	3.6%	3.3%	254	261	39,957	10,409,576	7,791,615	2,617,961	2,524,411	1.05	2,650,631
2010-1	7,405	188	1.424	268	3.6%	3.3%	261	268	39,218	10,501,088	7,372,984	3,128,104	3,006,236	1.05	3,156,548
2010-2	6,853	130	1.628	212	3.1%	3.3%	217	212	40,004	8,467,601	5,200,520	3,267,081	3,141,251	1.05	3,298,313
2011-1	7,359	143	1.920	275	3.7%	3.3%	259	275	38,868	10,670,000	5,558,124	5,111,876	4,909,406	1.05	5,154,876
2011-2	7,497	100	2.666	267	3.6%	3.3%	255	267	40,026	10,670,538	4,002,600	6,667,938	6,409,401	1.05	6,729,872
2012-1	6,869	64	3.981	255	3.7%	3.3%	234	255	39,138	9,972,839	2,504,832	7,468,007	7,161,251	1.05	7,519,314
2012-2	7,372	45	4.824	217	2.9%	3.3%	238	217	39,173	8,503,687	1,762,785	6,740,902	6,417,332	1.05	6,738,199
2013-1	7,244	40	6.237	249	3.4%	3.3%	241	241	39,931	9,616,498	1,597,240	8,019,258	7,582,637	1.05	7,961,769
2013-2	7,466	32	7.639	244	3.3%	3.3%	246	246	38,786	9,550,314	1,241,152	8,309,162	7,795,886	1.05	8,185,680
2014-1	7,274	24	9.850	236	3.2%	3.3%	240	240	40,049	9,602,935	961,176	8,641,759	8,046,733	1.05	8,449,069
2014-2	6,973	16	13.874	222	3.2%	3.3%	230	230	39,879	9,157,420	638,064	8,519,356	7,874,361	1.05	8,268,079
2015-1	7,027	8	23.601	189	2.7%	3.3%	230	230	38,906	8,955,273	311,248	8,644,025	7,932,295	1.05	8,328,910
2015-2	6,933	2	100.801	202	2.9%	3.3%	229	229	39,652	9,065,762	79,304	8,986,458	8,190,743	1.05	8,600,280
Total	172,825	3,333		5,632	3.3%		5,675	5,704		224,857,215	131,236,108	93,621,107	88,299,958		92,714,955
2004-1 - 2012-2	129,908	3,211		4,289	3.3%										

Notes
(A) (1)x(6)x[1.0-1.0/(3)]+(2)

Table 1: Policies in Force (PIF) Data Evaulation

Book Half Year	1	2	3	4	5	6	7	8	9	10	11
2004-1	7,474	6,951	6,256	5,568	4,956	4,213	3,539	2,831	2,406	2,045	1,575
2004-2	6,965	6,269	5,956	5,360	4,878	4,293	3,692	3,212	2,730	2,321	1,810
2005-1	6,916	6,432	5,724	5,094	4,636	3,941	3,507	3,016	2,413	1,955	1,564
2005-2	7,420	6,678	6,077	5,530	5,088	4,477	3,805	3,158	2,716	2,173	1,717
2006-1	6,839	6,497	5,847	5,262	4,683	4,262	3,708	3,078	2,524	2,070	1,697
2006-2	6,892	6,065	5,459	4,804	4,468	3,842	3,419	2,838	2,242	1,794	1,417
2007-1	7,089	6,238	5,552	5,052	4,547	3,910	3,363	2,690	2,233	1,742	1,446
2007-2	7,485	6,961	6,335	5,575	5,073	4,515	3,838	3,186	2,676	2,221	1,843
2008-1	7,492	6,818	6,477	6,088	5,297	4,555	3,963	3,369	2,864	2,263	1,788
2008-2	7,266	6,467	5,691	5,008	4,357	3,834	3,336	2,669	2,189	1,751	1,383
2009-1	7,490	7,041	6,689	5,953	5,417	4,659	3,820	3,094	2,599	2,183	1,768
2009-2	7,141	6,427	5,977	5,618	4,944	4,499	3,869	3,134	2,539	2,057	1,584
2010-1	7,399	6,585	5,861	5,392	4,637	4,034	3,469	2,845	2,333	1,936	1,588
2010-2	6,847	6,368	6,050	5,385	4,631	3,890	3,190	2,743	2,359	1,982	1,645
2011-1	7,350	6,542	5,888	5,417	5,038	4,282	3,597	3,021	2,507	2,006	
2011-2	7,488	6,814	6,405	6,021	5,359	4,823	4,244	3,692	3,175		
2012-1	6,864	6,384	5,873	5,521	4,969	4,224	3,675	3,050			
2012-2	7,365	6,629	6,298	5,857	5,388	4,849	4,267				
2013-1	7,239	6,732	6,261	5,510	5,069	4,309					
2013-2	7,457	6,860	6,174	5,433	4,672						
2014-1	7,269	6,469	5,887	5,475							
2014-2	6,968	6,132	5,825								
2015-1	7,021	6,600									
2015-2	6,923										

Table 1 (cont'd): Policies in Force (PIF) Data

12	13	14	15	16	17	18	19	20	21	22	23	24
1,197	886	611	403	258	157	88	53	28	14	7	3	1
1,412	1,031	722	484	329	197	118	65	34	16	8	4	
1,189	856	608	413	285	182	111	59	32	15	7		
1,322	1,018	743	505	313	194	109	61	35	19			
1,290	942	659	468	290	174	110	58	31				
1,091	851	604	417	271	173	97	51					
1,070	781	547	356	235	143	82						
1,456	1,092	819	549	346	225							
1,413	1,102	815	554	377								
1,106	852	622	429									
1,326	1,008	685										
1,236	902											
1,255												

Table 2: Incremental PIF Decay
Evaulation

	1	2	3	4	5	6	7	8	9	10	11
2004-1	100.0%	93.0%	90.0%	89.0%	89.0%	85.0%	84.0%	80.0%	85.0%	85.0%	77.0%
2004-2	100.0%	90.0%	95.0%	90.0%	91.0%	88.0%	86.0%	87.0%	85.0%	85.0%	78.0%
2005-1	100.0%	93.0%	89.0%	89.0%	91.0%	85.0%	89.0%	86.0%	80.0%	81.0%	80.0%
2005-2	100.0%	90.0%	91.0%	91.0%	92.0%	88.0%	85.0%	83.0%	86.0%	80.0%	79.0%
2006-1	100.0%	95.0%	90.0%	90.0%	89.0%	91.0%	87.0%	83.0%	82.0%	82.0%	82.0%
2006-2	100.0%	88.0%	90.0%	88.0%	93.0%	86.0%	89.0%	83.0%	79.0%	80.0%	79.0%
2007-1	100.0%	88.0%	89.0%	91.0%	90.0%	86.0%	86.0%	80.0%	83.0%	78.0%	83.0%
2007-2	100.0%	93.0%	91.0%	88.0%	91.0%	89.0%	85.0%	83.0%	84.0%	83.0%	83.0%
2008-1	100.0%	91.0%	95.0%	94.0%	87.0%	86.0%	87.0%	85.0%	85.0%	79.0%	79.0%
2008-2	100.0%	89.0%	88.0%	88.0%	87.0%	88.0%	87.0%	80.0%	82.0%	80.0%	79.0%
2009-1	100.0%	94.0%	95.0%	89.0%	91.0%	86.0%	82.0%	81.0%	84.0%	84.0%	81.0%
2009-2	100.0%	90.0%	93.0%	94.0%	88.0%	91.0%	86.0%	81.0%	81.0%	81.0%	77.0%
2010-1	100.0%	89.0%	89.0%	92.0%	86.0%	87.0%	86.0%	82.0%	82.0%	83.0%	82.0%
2010-2	100.0%	93.0%	95.0%	89.0%	86.0%	84.0%	82.0%	86.0%	86.0%	84.0%	83.0%
2011-1	100.0%	89.0%	90.0%	92.0%	93.0%	85.0%	84.0%	84.0%	83.0%	80.0%	
2011-2	100.0%	91.0%	94.0%	94.0%	89.0%	90.0%	88.0%	87.0%	86.0%		
2012-1	100.0%	93.0%	92.0%	94.0%	90.0%	85.0%	87.0%	83.0%			
2012-2	100.0%	90.0%	95.0%	93.0%	92.0%	90.0%	88.0%				
2013-1	100.0%	93.0%	93.0%	88.0%	92.0%	85.0%					
2013-2	100.0%	92.0%	90.0%	88.0%	86.0%						
2014-1	100.0%	89.0%	91.0%	93.0%							
2014-2	100.0%	88.0%	95.0%								
2015-1	100.0%	94.0%									
2015-2	100.0%										
Selection	**100.0%**	**90.7%**	**92.3%**	**90.5%**	**90.0%**	**87.5%**	**86.7%**	**85.0%**	**84.2%**	**82.0%**	**80.8%**

Table 2 (cont'd): Incremental PIF Decay

12	13	14	15	16	17	18	19	20	21	22	23	24
76.0%	74.0%	69.0%	66.0%	64.0%	60.9%	56.1%	60.2%	52.8%	50.0%	50.0%	42.9%	33.3%
78.0%	73.0%	70.0%	67.0%	68.0%	59.9%	59.9%	55.1%	52.3%	47.1%	50.0%	50.0%	
76.0%	72.0%	71.0%	67.9%	69.0%	63.9%	61.0%	53.2%	54.2%	46.9%	46.7%		
77.0%	77.0%	73.0%	68.0%	62.0%	62.0%	56.2%	56.0%	57.4%	54.3%			
76.0%	73.0%	70.0%	71.0%	62.0%	60.0%	63.2%	52.7%	53.4%				
77.0%	78.0%	71.0%	69.0%	65.0%	63.8%	56.1%	52.6%					
74.0%	73.0%	70.0%	65.1%	66.0%	60.9%	57.3%						
79.0%	75.0%	75.0%	67.0%	63.0%	65.0%							
79.0%	78.0%	74.0%	68.0%	68.1%								
80.0%	77.0%	73.0%	69.0%									
75.0%	76.0%	68.0%										
78.0%	73.0%											
79.0%												
78.0%	**76.0%**	**72.5%**	**67.3%**	**65.5%**	**62.4%**	**58.2%**	**53.6%**	**54.3%**	**49.6%**	**48.9%**	**46.4%**	**33.3%**

Table 3: Actual and Projected Loan Persistency Factors

	Evaulation										
	1	2	3	4	5	6	7	8	9	10	11
2004-1	100.0%	93.0%	90.0%	89.0%	89.0%	85.0%	84.0%	80.0%	85.0%	85.0%	77.0%
2004-2	100.0%	90.0%	95.0%	90.0%	91.0%	88.0%	86.0%	87.0%	85.0%	85.0%	78.0%
2005-1	100.0%	93.0%	89.0%	89.0%	91.0%	85.0%	89.0%	86.0%	80.0%	81.0%	80.0%
2005-2	100.0%	90.0%	91.0%	91.0%	92.0%	88.0%	85.0%	83.0%	86.0%	80.0%	79.0%
2006-1	100.0%	95.0%	90.0%	90.0%	89.0%	91.0%	87.0%	83.0%	82.0%	82.0%	82.0%
2006-2	100.0%	88.0%	90.0%	88.0%	93.0%	86.0%	89.0%	83.0%	79.0%	80.0%	79.0%
2007-1	100.0%	88.0%	89.0%	91.0%	90.0%	86.0%	86.0%	80.0%	83.0%	78.0%	83.0%
2007-2	100.0%	93.0%	91.0%	88.0%	91.0%	89.0%	85.0%	83.0%	84.0%	83.0%	83.0%
2008-1	100.0%	91.0%	95.0%	94.0%	87.0%	86.0%	87.0%	85.0%	85.0%	79.0%	79.0%
2008-2	100.0%	89.0%	88.0%	88.0%	87.0%	88.0%	87.0%	80.0%	82.0%	80.0%	79.0%
2009-1	100.0%	94.0%	95.0%	89.0%	91.0%	86.0%	82.0%	81.0%	84.0%	84.0%	81.0%
2009-2	100.0%	90.0%	93.0%	94.0%	88.0%	91.0%	86.0%	81.0%	81.0%	81.0%	77.0%
2010-1	100.0%	89.0%	89.0%	92.0%	86.0%	87.0%	86.0%	82.0%	82.0%	83.0%	82.0%
2010-2	100.0%	93.0%	95.0%	89.0%	86.0%	84.0%	82.0%	86.0%	86.0%	84.0%	83.0%
2011-1	100.0%	89.0%	90.0%	92.0%	93.0%	85.0%	84.0%	84.0%	83.0%	80.0%	80.8%
2011-2	100.0%	91.0%	94.0%	94.0%	89.0%	90.0%	88.0%	87.0%	86.0%	82.0%	80.8%
2012-1	100.0%	93.0%	92.0%	94.0%	90.0%	85.0%	87.0%	83.0%	84.2%	82.0%	80.8%
2012-2	100.0%	90.0%	95.0%	93.0%	92.0%	90.0%	88.0%	85.0%	84.2%	82.0%	80.8%
2013-1	100.0%	93.0%	93.0%	88.0%	92.0%	85.0%	86.7%	85.0%	84.2%	82.0%	80.8%
2013-2	100.0%	92.0%	90.0%	88.0%	86.0%	87.5%	86.7%	85.0%	84.2%	82.0%	80.8%
2014-1	100.0%	89.0%	91.0%	93.0%	90.0%	87.5%	86.7%	85.0%	84.2%	82.0%	80.8%
2014-2	100.0%	88.0%	95.0%	90.5%	90.0%	87.5%	86.7%	85.0%	84.2%	82.0%	80.8%
2015-1	100.0%	94.0%	92.3%	90.5%	90.0%	87.5%	86.7%	85.0%	84.2%	82.0%	80.8%
2015-2	100.0%	90.7%	92.3%	90.5%	90.0%	87.5%	86.7%	85.0%	84.2%	82.0%	80.8%

Table 3 (cont'd): Actual and Projected Loan Persistency Factors

12	13	14	15	16	17	18	19	20	21	22	23	24
76.0%	74.0%	69.0%	66.0%	64.0%	60.9%	56.1%	60.2%	52.8%	50.0%	50.0%	42.9%	33.3%
78.0%	73.0%	70.0%	67.0%	68.0%	59.9%	59.9%	55.1%	52.3%	47.1%	50.0%	50.0%	33.3%
76.0%	72.0%	71.0%	67.9%	69.0%	63.9%	61.0%	53.2%	54.2%	46.9%	46.7%	46.4%	33.3%
77.0%	77.0%	73.0%	68.0%	62.0%	62.0%	56.2%	56.0%	57.4%	54.3%	48.9%	46.4%	33.3%
76.0%	73.0%	70.0%	71.0%	62.0%	60.0%	63.2%	52.7%	53.4%	49.6%	48.9%	46.4%	33.3%
77.0%	78.0%	71.0%	69.0%	65.0%	63.8%	56.1%	52.6%	54.3%	49.6%	48.9%	46.4%	33.3%
74.0%	73.0%	70.0%	65.1%	66.0%	60.9%	57.3%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
79.0%	75.0%	75.0%	67.0%	63.0%	65.0%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
79.0%	78.0%	74.0%	68.0%	68.1%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
80.0%	77.0%	73.0%	69.0%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
75.0%	76.0%	68.0%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	73.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
79.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%
78.0%	76.0%	72.5%	67.3%	65.5%	62.4%	58.2%	53.6%	54.3%	49.6%	48.9%	46.4%	33.3%

Table 4: Actual and Projected PIF
Evaulation

Book Half Year	1	2	3	4	5	6	7	8	9	10	11
2004-1	7,474	6,951	6,256	5,568	4,956	4,213	3,539	2,831	2,406	2,045	1,575
2004-2	6,965	6,269	5,956	5,360	4,878	4,293	3,692	3,212	2,730	2,321	1,810
2005-1	6,916	6,432	5,724	5,094	4,636	3,941	3,507	3,016	2,413	1,955	1,564
2005-2	7,420	6,678	6,077	5,530	5,088	4,477	3,805	3,158	2,716	2,173	1,717
2006-1	6,839	6,497	5,847	5,262	4,683	4,262	3,708	3,078	2,524	2,070	1,697
2006-2	6,892	6,065	5,459	4,804	4,468	3,842	3,419	2,838	2,242	1,794	1,417
2007-1	7,089	6,238	5,552	5,052	4,547	3,910	3,363	2,690	2,233	1,742	1,446
2007-2	7,485	6,961	6,335	5,575	5,073	4,515	3,838	3,186	2,676	2,221	1,843
2008-1	7,492	6,818	6,477	6,088	5,297	4,555	3,963	3,369	2,864	2,263	1,788
2008-2	7,266	6,467	5,691	5,008	4,357	3,834	3,336	2,669	2,189	1,751	1,383
2009-1	7,490	7,041	6,689	5,953	5,417	4,659	3,820	3,094	2,599	2,183	1,768
2009-2	7,141	6,427	5,977	5,618	4,944	4,499	3,869	3,134	2,539	2,057	1,584
2010-1	7,399	6,585	5,861	5,392	4,637	4,034	3,469	2,845	2,333	1,936	1,588
2010-2	6,847	6,368	6,050	5,385	4,631	3,890	3,190	2,743	2,359	1,982	1,645
2011-1	7,350	6,542	5,888	5,417	5,038	4,282	3,597	3,021	2,507	2,006	1,620
2011-2	7,488	6,814	6,405	6,021	5,359	4,823	4,244	3,692	3,175	2,604	2,103
2012-1	6,864	6,384	5,873	5,521	4,969	4,224	3,675	3,050	2,570	2,107	1,702
2012-2	7,365	6,629	6,298	5,857	5,388	4,849	4,267	3,627	3,055	2,506	2,023
2013-1	7,239	6,732	6,261	5,510	5,069	4,309	3,738	3,177	2,676	2,195	1,773
2013-2	7,457	6,860	6,174	5,433	4,672	4,088	3,546	3,014	2,539	2,082	1,682
2014-1	7,269	6,469	5,887	5,475	4,927	4,311	3,740	3,179	2,678	2,196	1,774
2014-2	6,968	6,132	5,825	5,272	4,744	4,151	3,601	3,061	2,579	2,115	1,708
2015-1	7,021	6,600	6,089	5,510	4,959	4,339	3,764	3,199	2,695	2,210	1,785
2015-2	6,923	6,283	5,796	5,245	4,720	4,130	3,583	3,045	2,566	2,104	1,699

Table 4 (cont'd): Actual and Projected PIF

12	13	14	15	16	17	18	19	20	21	22	23	24
1,197	886	611	403	258	157	88	53	28	14	7	3	1
1,412	1,031	722	484	329	197	118	65	34	16	8	4	1
1,189	856	608	413	285	182	111	59	32	15	7	3	1
1,322	1,018	743	505	313	194	109	61	35	19	9	4	1
1,290	942	659	468	290	174	110	58	31	15	8	3	1
1,091	851	604	417	271	173	97	51	28	14	7	3	1
1,070	781	547	356	235	143	82	44	24	12	6	3	1
1,456	1,092	819	549	346	225	131	70	38	19	9	4	1
1,413	1,102	815	554	377	235	137	73	40	20	10	4	1
1,106	852	622	429	281	175	102	55	30	15	7	3	1
1,326	1,008	685	461	302	188	110	59	32	16	8	4	1
1,236	902	654	440	288	180	105	56	30	15	7	3	1
1,255	954	691	465	305	190	111	59	32	16	8	4	1
1,283	975	707	476	312	194	113	61	33	16	8	4	1
1,264	960	696	468	307	192	111	60	32	16	8	4	1
1,640	1,247	904	608	398	249	145	78	42	21	10	5	2
1,327	1,009	731	492	322	201	117	63	34	17	8	4	1
1,578	1,200	869	585	383	239	139	75	41	20	10	5	2
1,383	1,051	762	512	336	210	122	65	36	18	9	4	1
1,312	997	723	486	318	199	116	62	34	17	8	4	1
1,383	1,052	762	513	336	210	122	65	36	18	9	4	1
1,332	1,012	734	494	323	202	118	63	34	17	8	4	1
1,392	1,058	767	516	338	211	123	66	36	18	9	4	1
1,325	1,007	730	491	322	201	117	63	34	17	8	4	1

Table 5: Average Projected PIF

	Evaulation										
Book Half Year	1	2	3	4	5	6	7	8	9	10	11
2004-1	1										
2004-2	3	1									
2005-1	5	2	1								
2005-2	14	7	3	1							
2006-1	23	11	5	2	1						
2006-2	39	21	10	5	2	1					
2007-1	63	34	18	9	4	2	1				
2007-2	178	101	54	29	14	7	3	1			
2008-1	306	186	105	57	30	15	7	3	1		
2008-2	355	228	139	78	42	22	11	5	2	1	
2009-1	573	381	245	149	84	45	24	12	6	2	1
2009-2	778	547	364	234	142	80	43	23	11	5	2
2010-1	1,104	823	578	385	247	150	85	46	24	12	6
2010-2	1,464	1,129	841	591	394	253	154	87	47	25	12
2011-1	1,813	1,442	1,112	828	582	388	249	152	86	46	24
2011-2	2,889	2,353	1,871	1,443	1,075	756	503	323	197	111	60
2012-1	2,810	2,338	1,904	1,515	1,168	870	612	407	262	159	90
2012-2	3,947	3,341	2,780	2,264	1,801	1,389	1,035	727	484	311	189
2013-1	4,024	3,458	2,927	2,436	1,984	1,578	1,217	906	637	424	273
2013-2	4,380	3,817	3,280	2,777	2,311	1,882	1,497	1,154	860	604	402
2014-1	5,201	4,619	4,026	3,459	2,928	2,437	1,985	1,579	1,218	907	637
2014-2	5,548	5,008	4,448	3,876	3,331	2,820	2,347	1,911	1,520	1,172	873
2015-1	6,344	5,799	5,235	4,649	4,052	3,482	2,947	2,453	1,998	1,589	1,225
2015-2	6,603	6,039	5,520	4,983	4,425	3,857	3,314	2,805	2,335	1,902	1,512
	0.25	0.75	1.25	1.75	2.25	2.75	3.25	3.75	4.25	4.75	5.25
Discount Factor	0.995	0.985	0.976	0.966	0.956	0.947	0.938	0.928	0.919	0.910	0.901

Table 5 (cont'd): Average Projected PIF

12	13	14	15	16	17	18	19	20	21	22	23	24
1												
2	1											
6	2	1										
12	6	2	1									
32	16	7	3	2								
48	26	13	6	3	1							
107	58	30	15	7	3	2						
166	94	50	27	13	6	3	1					
259	157	89	48	25	12	6	3	1				
424	273	166	94	50	27	13	6	3	1			
614	409	263	160	90	49	26	13	6	3	1		
913	641	427	275	167	94	51	27	13	6	3	1	
1,166	869	611	406	261	159	90	48	25	13	6	3	1
5.75	6.25	6.75	7.25	7.75	8.25	8.75	9.25	9.75	10.25	10.75	11.25	11.75
0.892	0.884	0.875	0.866	0.858	0.849	0.841	0.833	0.824	0.816	0.808	0.800	0.792

Table 6: Premium Estimate

(1) Average Monthly Policy Premium	(2) Premium Scaling Factor	(3) Projected Future Premium	(4) Premium A.o. 12/31/15	(5) Total Premium (3)+(4)	(6) Discounted Projected Future Premium	(7) Policy Maintenance Expense Ratio	(8) Discounted Projected Prem. Net of Expenses (6)x[1.0-(7)]
75	6	450	23,446,397	23,446,847	448	5%	425
75	6	1,807	23,538,827	23,540,634	1,792	5%	1,703
76	6	3,795	23,607,461	23,611,256	3,757	5%	3,569
76	6	11,492	23,694,210	23,705,701	11,360	5%	10,792
76	6	19,897	23,756,921	23,776,818	19,648	5%	18,665
76	6	35,928	23,859,775	23,895,703	35,440	5%	33,668
77	6	60,088	23,931,198	23,991,285	59,220	5%	56,259
77	6	178,270	23,860,998	24,039,268	175,543	5%	166,766
77	6	328,722	23,782,663	24,111,386	323,347	5%	307,180
78	6	411,422	23,820,521	24,231,943	404,192	5%	383,982
78	6	709,735	23,570,671	24,280,407	696,350	5%	661,533
78	6	1,045,062	23,356,747	24,401,809	1,023,843	5%	972,651
78	6	1,627,253	22,847,761	24,475,014	1,591,486	5%	1,511,912
79	6	2,363,358	22,234,031	24,597,389	2,307,077	5%	2,191,723
79	6	3,195,937	21,499,842	24,695,779	3,113,575	5%	2,957,896
79	6	5,539,850	19,254,711	24,794,562	5,385,751	5%	5,116,464
80	6	5,849,567	19,068,968	24,918,535	5,674,436	5%	5,390,715
80	6	8,869,340	16,123,950	24,993,290	8,584,457	5%	8,155,234
80	6	9,719,246	15,324,030	25,043,277	9,385,388	5%	8,916,119
80	6	11,348,862	13,744,501	25,093,363	10,933,164	5%	10,386,506
81	6	14,531,726	10,662,011	25,193,737	13,965,083	5%	13,266,829
81	6	16,708,141	8,535,983	25,244,124	16,015,431	5%	15,214,659
81	6	20,599,330	4,720,526	25,319,857	19,692,463	5%	18,707,840
81	6	22,931,142	2,515,314	25,446,456	21,862,046	5%	20,768,944
		126,090,420	**460,758,018**	**586,848,438**	**121,265,297**		**115,202,033**

Reserving Styles —

Are Actuaries In-Sync with their Stakeholders?

Mark Littmann, FCAS, MAAA

Abstract

Motivation. Reserving actuaries are constantly faced with forming estimates that inherently reflect consideration of data and information that spans from initial expectations to actual claims experience. The actuaries and their stakeholders (e.g., members of management) may implicitly or explicitly apply different perspectives on the relative merits of projections based on actual experience or initial expectations, or projections that reflect a blending of the two. As an actuary associated with an audit firm, Mr. Littmann encounters these situations frequently, primarily in a reserving context. Apparently subtle differences in perspectives among actuaries and among various stakeholders when actual experience diverges from expectations (generating divergent projections of unpaid claim estimates) can generate substantial dialogue. The paper presents an exploration of historical progressions of recognizing accident year losses, casts light on certain implications of common actuarial methods, and provides insight on the notion of a reserving cycle akin to an underwriting cycle. The investigation provides a framework for dialogue among stakeholders to the reserving process, as well as identifies areas where actuaries may be able to enhance the technical aspects of, and their communications from, their work processes.

Method. The paper provides examples of the historical progression of accident year loss ratios booked by the industry in aggregate and for a sample of companies. A model is presented to demonstrate the extent to which a combination of cyclical accident year loss ratios and alternate views from stakeholders on their 'best estimates' to be adopted at a point in time can create differences in the estimates of unpaid claims liabilities.

Results. The outcomes are a framework for expressing views on responsiveness to the emerging claims data in relation to initial expectations, as well as illustrations that provide actuaries with insights on the implications of differing views on loss picks. The paper identifies matters for actuaries to discuss among themselves and with their stakeholders. Discussions around these concepts and implications in advance of the periodic reserves meetings may help the meetings go more smoothly.

Conclusions. Apparently small differences in styles for making loss picks from among projections that span from initial expectations to extrapolations from actual data can yield noticeable differences in reserve estimates. Differences in selection approach between stakeholders do matter and create the need for discussion, transparency and documentation.

Keywords. Reserving Methods. Management Best Estimate. Reserve Variability. Credibility.

Disclaimer. Beginning in Section 4, the paper includes commentary, tables, and charts that illustrate a scenario where management's loss picks (for ultimate losses and the associated reserves) are based on the paid Bornhuetter-Ferguson (BF) method and an actuary's loss picks are based on the reported BF method. Under no circumstance should the scenario (or anything else in the paper) be construed as indicative of the author's nor his employer's view on any insurance company management or actuary, nor the author's or his employer's view on any preferred actuarial projection method(s) as the basis for loss picks or booked amounts.

---

1. Background

The Casualty Actuarial Society's (CAS) literature and seminar archives include papers and presentations that analyze the performance of loss reserves established by insurance companies in terms of how original provisions have fared against the subsequent experience. Various descriptions and potential explanations have been offered for an apparent cyclical pattern to reserve adequacy, akin to the commonly regarded cycle of pricing adequacy. Certain approaches, frequently involving statistical metrics, for testing the performance of various actuarial techniques have been described, with an apparent purpose to enhance the technical strength of the actuarial estimates.

This paper takes a different perspective on the matter. To set the stage for this, I recall the CAS Centennial Celebration in New York in November 2014, at which a luncheon speaker offered the audience a simple challenge. If someone tosses a coin 12 times and 3 heads result, what is the probability of a head on the next toss? Of course, we actuaries have been trained to avoid falling into the trap of responding quickly with 25%, since we treat the 12 observations as a random sample from a population of possible outcomes where we believe that the probability of a head on any toss is 50%. Therefore, we ignore the actual experience and give full consideration to our expectation based on external information. But, if we were informed that the coin-flipper was a con-artist, which introduced the possibility that the coin was biased, then that supplemental information might influence how we respond to the 12 observations and consequently our view on the likelihood of a head on the next toss.

The example illustrates the dilemma that actuaries and management face when confronted with claims data and various actuarial projections of ultimate losses and the corresponding reserves. For medium to long tail lines, initial expectations of ultimate losses are often closely aligned with expectations based on pricing. The dilemma is to know when, and to what extent, to migrate from the original expectation to the experience-based projections. Stated another way, the dilemma is how to choose an ultimate loss estimate based on a collection of projections from different methods applied to alternate data sets and which reflect certain judgments for key parameters, including initial expected losses, development factors, and assessments on the effects of internal operational changes or external environmental conditions.

As multiple personnel are often involved in the analysis of unpaid claims estimates and in forming a view as to the level of reserves to be recorded in an entity's financial statements, differences in the perspectives of these personnel on the relative merit of alternate projections can drive differences in views as to the relative adequacy of the booked reserves.

2. Historical performance of ultimate loss estimates

Publicly-available Schedule P data were obtained and analyzed to assess the progression of accident year booked ultimate loss ratio estimates from the 12-month valuation to subsequent valuations, particularly for medium- to long-tail lines. For short-tail lines, where a substantial portion of ultimate losses are generally paid by the end of the accident period, there is generally lesser variation in the booked loss ratio from 12-months to subsequent valuations. For the longer-tail lines, insurance company management often books ultimate loss ratios at 12 months that are characterized as being "in line with pricing expectations." Hindsight often demonstrates that the ultimate losses are higher or lower than the amounts booked at 12 months, consistent with the historical phenomenon of the cyclical nature of pricing adequacy over time.

Table 1 shows accident year ultimate loss ratios at 12 months and at 72 months for the P&C insurance industry for four lines of business.[1]

Table 1
Comparison of Accident Year Loss Ratios at 12- and 72-months Maturity
Property/Casualty Insurance Industry

		2005	2006	2007	2008	2009
PAL	at 12 months	67%	66%	69%	69%	73%
	at 72 months	63%	63%	67%	67%	70%
	Ratio	0.94	0.96	0.97	0.97	0.97
CAL	at 12 months	61%	62%	62%	62%	63%
	at 72 months	58%	58%	61%	61%	60%
	Ratio	0.95	0.94	0.97	0.98	0.96
CMP	at 12 months	61%	53%	55%	69%	60%
	at 72 months	56%	47%	50%	65%	60%
	Ratio	0.92	0.90	0.92	0.94	0.99
GL-Occ	at 12 months	66%	64%	66%	67%	69%
	at 72 months	55%	54%	60%	61%	61%
	Ratio	0.84	0.85	0.91	0.92	0.89

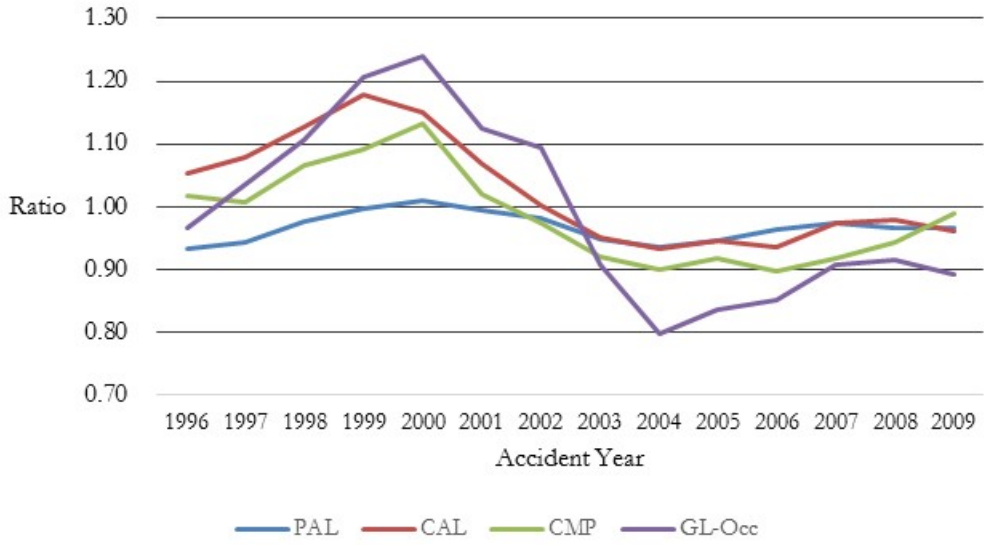
PAL = Private Passenger Auto Liability
CAL = Commercial Automobile Liability
CMP = Commercial Multi-Peril
GL-Occ = General Liability – Occurrence
Source: SNL Financial website. P&C Industry Composite.

[1] Throughout this paper, amounts are shown in various tables and charts. The actual amounts contain more digits than are displayed, and therefore, some apparent arithmetic may be influenced by rounding.

The booked ultimate loss ratios demonstrate varying degrees of change from the 12 month valuation to the 72 month valuation. The magnitude of change appears smallest for the automobile lines, with changes a bit larger for CMP, with still larger changes for GL-Occurrence. For these accident years, we also note that the changes are favorable, as the booked loss ratios at 72 months are less than those booked at 12 months.

Comparable data as shown in Table 1 are provided in Appendix A for a longer experience period, spanning accident years 1996 to 2009. Over the 14-year period, initial booked loss ratios deviated upward and downward with subsequent valuations. Chart 1 shows the ratios of the 1996 to 2009 accident year booked loss ratios at the 72-month valuation, in comparison to the loss ratio booked at the 12-month valuation.

Chart 1
Ratios of Accident Year Booked Loss Ratio at 72-months
Compared to Booked Loss Ratio at 12-months
US P&C Insurance Industry



For Personal Auto Liability (PAL), the ratios were in the range from 0.93 to 1.01 over the 14 accident years, with an average ratio of 0.97 (favorable 3%). In contrast, the booked loss ratios for General Liability – Occurrence at 72-months, on average, were within 1% of the loss ratios booked at 12-months. On an accident year by accident year basis, however, individual years' ratios were as low as 0.80 and as high as 1.24.

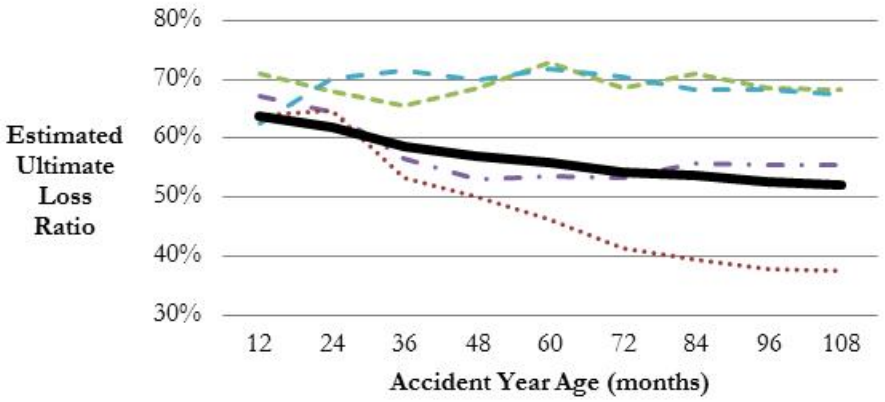
A particular focus area for this paper is assessing the progression of loss ratios from an initial valuation to subsequent valuations on the path toward "true" (and final) ultimate. Charts 2a and 2b show the progression for CMP and GL-Occurrence, respectively, for the 2006 accident year, from 12 months through the 72 month valuation, and continuing to the 108 month valuation at year-end

2014 reporting. The data are shown for the P&C insurance industry (bold/black line) and for four companies/groups from among the Top 10 based on market share for each line.

Chart 2a
Progression of Booked Ultimate Loss Ratios
CMP - Accident Year 2006
P&C Industry (bold/black) & 4 Top-10 Companies

For Commercial Multi-Peril for accident year 2006, the industry booked loss ratio at 12 months was 53%, and the booked loss ratio appeared to stabilize at the 72 month valuation at 47%. Thus, with hindsight, the initial booked loss ratio decreased by 10% over subsequent valuations. For the four companies in the sample from the Top 10, initial booked loss ratios decreased by 7% to 19%.

Chart 2b
Progression of Booked Ultimate Loss Ratios
General Liability - Occurrence -- Accident Year 2006
P&C Industry (bold/black) & 4 Top-10 Companies



For General Liability – Occurrence for accident year 2006, the industry booked loss ratio at 12 months (at year-end 2006) was 64%. The booked loss ratio decreased to 54% at the 72 month valuation, with further decreases to 52% at the 108 month valuation (at year-end 2014). With hindsight, the initial booked loss ratio decreased by 18% over subsequent valuations. For the four companies in the sample from the Top 10, one company's initial loss ratio decreased by about 40%, while another's increased by about 10%.

When the ultimate loss ratio is sufficiently different than the estimate at 12 months, there appears to be a tendency for the magnitude of the change to be related to the length of the paid/reported loss emergence pattern. Thus, it is not surprising that larger changes from initial booked loss ratios are observed for GL-Occurrence than for CMP, and, that the booked loss ratios for GL-Occurrence continue to evolve at valuations beyond 72 months, while CMP's loss ratio appears to have stabilized by that valuation.

Along the path from an accident year aging from 12 months to 72 months (or beyond), at what point was there sufficient claims data or other indicators that the ultimate estimates made at 12 months would not hold up? Stated another way, why didn't the industry (or individual companies) get it "right" sooner? If the early claims experience deviated from initial expectations, why didn't booked loss ratios demonstrate a greater response to the data?

In this paper, I explore the notion that, along the path of an accident year aging, different stakeholders to the reserving process take different positions on the degree of responsiveness to the emerging data, as evidenced by differing bases for ultimate loss estimates and the corresponding

reserves. What if actuaries' estimates respond more quickly to the emerging claims experience than management in the formation of the best estimate? In the next section, the nature and key features of common actuarial projection methods are identified and described.

3. Features of Actuarial Projection Models

Actuarial analysis of unpaid claims estimates is often performed utilizing multiple methods, which can be applied to various types of data. The table below identifies four common actuarial methods, and types of claims data to which the methods can be applied.

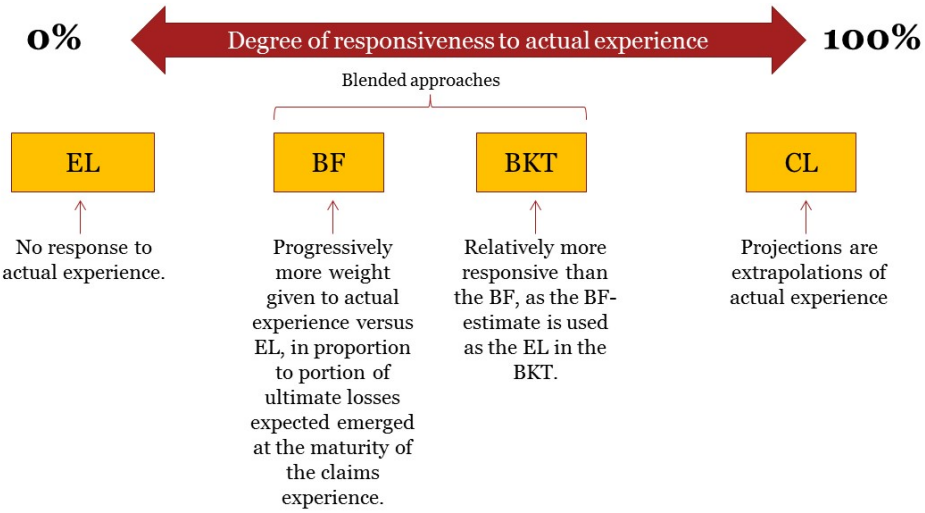
Methods	**Types of Data**
Expected Loss (EL)	Paid losses
Bornhuetter-Ferguson (BF)	Reported losses (payments plus case reserves)
Benktander (BKT)	Claim counts
Chain Ladder (CL) (also known as loss development)	

The key parameters of the methods require that judgments be made for each parameter in the application of the methods to a particular set of data. The table below identifies the parameters for each of the methods.

	Methods			
Parameter	**EL**	**BF**	**BKT**	**CL**
Initial expected losses	●	●	●	
Loss development factors (including a tail factor)		●	●	●

By their nature, the four actuarial methods have varying degrees of responsiveness to the actual claims experience. Figure 1 provides a comparison.

Figure 1
Comparative Responsiveness of Actuarial Methods to Actual Experience



In a rare situation where the actual claims experience for an accident period emerges in line with expectations (based on initial expected losses and the expected emergence pattern), all methods will generate the same (and accurate) projected ultimate losses, and there is no divergence among the methods' projections.

Actual claims experience inevitably deviates, to some degree, from expectations, whether in the level of ultimate losses once all claims are reported, settled, and closed, or in the pattern by which the losses emerge, or both. When actual experience deviates (whether favorably or adversely) from expectations, the projections from various methods will diverge, due to the different degree of responsiveness of each method to the actual loss experience. Table 2 shows illustrated BF- and CL-projections that reflect an initial expected loss of 100, a true ultimate of 92, and actual emerged losses being less than expectations at each valuation date, but demonstrating inconsistent deviations to expected amounts. (The assumed loss reporting pattern is shown in Appendix C.)

Table 2
Illustration of BF and CL Projections
when Actual Experience emerges Inconsistently Less than Expected

	Accident Period Age						
	1	2	3	4	5	6	7
Expected	35	55	70	85	90	95	100
Actual	35	52	61	71	78	89	92
% deviation	-1%	-6%	-13%	-16%	-13%	-7%	-8%
BF-estimate	100	97	91	86	88	94	92
CL-estimate	99	94	87	84	87	93	92

In Table 2[2], the actual reported losses at the 1st valuation are with 1% of expectations, such that the BF and CL projections are closely aligned with the initial expected ultimate. By the 4th valuation, the extent of the divergence in cumulative actual versus expected reported losses increased to 16%, thereby decreasing the CL projection to 84. The BF-projection has a tempered response to the actual experience, with an estimate of 86 at the 4th valuation. As the actual experience settles to an 8% favorable deviation at the 7th valuation, the BF and CL estimates are the same and converge at the true ultimate of 92.

Of course, in a scenario where actual loss emergence is greater than expectations, the relative positions of the projections would be reversed, with the CL projection becoming larger than the initial expected ultimate, with the BF method yielding a projection higher than the initial expected amount, but less than the CL projection.

For the remainder of this paper, the emphasis is on exploring implications of divergence of methods projections in terms of responsiveness to actual emerged claims experience, with an assumption that the pattern of actual emergence is in line with expectations, although perhaps on a path to a level of ultimate losses that differs from initial expectations. Therefore, the following examples reflect a consistency in the actual and expected pattern of emergence. Using the same assumptions underlying Table 2 above, Table 3 shows a scenario where actual experience deviates from expected experience consistently over the valuations.

Table3
Illustration of BF and CL Projections
when Actual Experience emerges Consistently Less Expected

	Accident Period Age						
	1	2	3	4	5	6	7
Expected	35	55	70	85	90	95	100
Actual	32	51	64	78	83	87	92
% deviation	-8%	-8%	-8%	-8%	-8%	-8%	-8%
BF-estimate	97	96	94	93	93	92	92
CL-estimate	92	92	92	92	92	92	92

[2] In the example, the BKT projection is deliberately not shown, for ease of presentation. The BKT projection is more responsive to the emerged claims experience than the BF, since its algorithm effectively re-cycles the BF projected loss as the input for another BF projection. Thus, the BKT projection generally falls between the BF and CL projections.

Cumulative actual reported losses emerge in the expected pattern, albeit 8% less than expected at each valuation date. The CL projection is consistent at 92 for all valuation dates, since the actual emergence pattern is in line with the expected pattern. The BF projection at the 1st valuation is slightly less than the initial expected amount, and it decreases progressively at successive valuations by the difference in actual versus expected emerged losses.

Appendix B contains an exhibit that provides details of the computations included in Table 3, including additional calculations for the BKT projection. With the spectrum of responsiveness to emerged data as illustrated in Figure 1 above in mind, the response of the BF is equivalent to the reciprocal of the loss development factor to ultimate (that is, the expected loss emergence percentage). In this example, the BF response at the 2nd valuation is 55%. The responsiveness of the BKT projection is dependent on both the expected emergence percentage and the degree to which actual experience diverges from expectations; in Appendix B, the BKT response at the 2nd valuation in this example is 80%.

Additional projections could be illustrated if the methods are applied to multiple types of data, for instance, paid losses and reported losses. This increases the potential divergence among the projections and illustrates another (implicit or explicit) judgment that actuaries and management must make in order to form a view on an actuarial central estimate and management's best-estimate for financial reporting.

The reader may wish to re-visit the charts shown in Section 2 with the progression of booked ultimate loss estimates for the industry and four companies. The progressions tend to follow a deliberate migration from initial expectations of ultimate loss at 12 months toward the value accrued by the 72-month to 108-month valuations. Nevertheless, neither I, nor the reader, can infer definitively whether the progressions followed an explicit, intentional path, (for instance, a reported-BF path) or reflected a changing mix of considerations over time.

4. When Styles Diverge (not just the Projections)

The implications around differing degrees of responsiveness to emerged claims data become apparent in the internal and joint discussions among insurance company actuaries and management, their external actuarial consultants, and the external audit firm's actuarial specialists that support the audit of the company's financial statements. To illustrate:

- Company management may form a view that it takes a while for the actual claims experience and the related projections to be sufficiently credible for management to deviate from initial expectations of ultimate loss for a particular accident period.

- A company's actuary may form a view that a staged approach to selecting ultimate losses is appropriate. For example, for the initial and second valuations, the EL method may be chosen (absent any individual large claims or losses arising from catastrophe events). For the third and subsequent valuations, the actuary may choose a BF estimate, and then shift toward a BKT- or CL-based estimate at valuations nearing the expected completion of the emergence pattern.

- An external consulting actuary (and/or the actuary supporting the external audit firm) may form a view that the ultimate losses for an accident period's initial period-end valuation are best represented by initial expected losses, but then may shift to a BF- or BKT-based estimate for subsequent valuations.

There are differing manners by which the parties may express their views as to the basis for the chosen estimate. These could be based strictly on the passage of time, the magnitude of the development factor, or the type of data.

It can be quite plausible and reasonable that management forms a view for best-estimate ultimate losses and the associated reserves that are different than the actuarial indication. Management may have valid and supportable rationale, considering features of the company's business and operations, as well as external trends and conditions, which management believes have not been fully incorporated within the actuarially-determined projections. For instance, for a portfolio that is exposed to individual large, late-reported claims, for which there has been an extended period of relatively benign claims experience, management may form a best-estimate that is greater than an actuarial indication that reflects a stated or unstated degree of response to the benign historical development experience.

Differences in judgments for forming a view on ultimate losses do not fall solely between actuaries and personnel from other backgrounds and functional roles. Indeed, differences in estimates arise among multiple actuaries involved in the analysis of unpaid claim liabilities for a particular business segment, legal entity, or an insurance company group.

Differences in how actuaries (whether company or external) and management pick ultimate loss estimates will generate differences in estimated unpaid claims liabilities. The illustrations above have shown the relative progression of projections for a single accident period over its successive valuations. Using the same set of assumptions above (where actual emerged losses deviate consistently and favorably from expectations), with initial expected losses of 100 and ultimate losses of 92, Chart 3 shows the progression of ultimate loss projections from the expected loss, paid BF, reported BF, and chain ladder methods.

Chart 3
Comparative Projections of Ultimate Losses

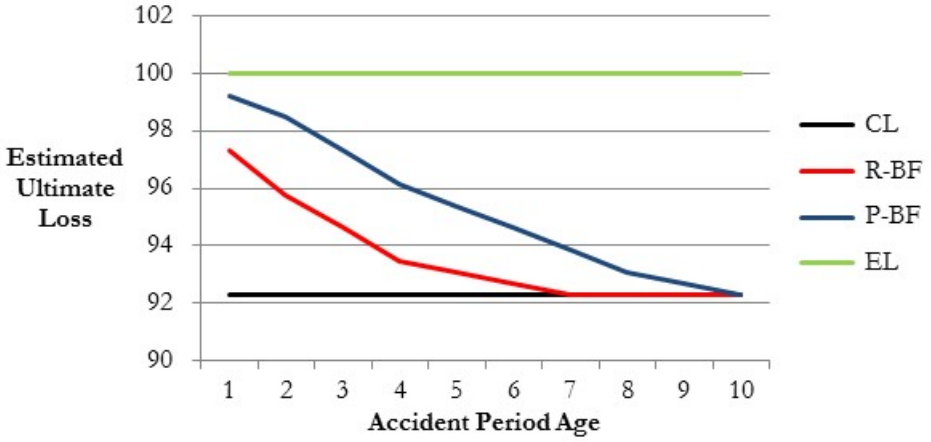


Table 4 shows the array of estimates of ultimate loss by method (as shown in Chart 3), as well as the cumulative payments at each age.

Table 4
Comparative Projection of Ultimate Losses

	Estimated Ultimates				
Age	EL	CL	P-BF	R-BF	**Paid**
1	100	92	99	97	9
2	100	92	98	96	18
3	100	92	97	95	32
4	100	92	96	93	46
5	100	92	95	93	55
6	100	92	95	93	65
7	100	92	94	92	74
8	100	92	93	92	83
9	100	92	93	92	88
10	100	92	92	92	92

The CL-projection is consistently $92 over the valuations, as the claims experience, although less than expectations, is following the expected loss emergence pattern. The EL has a 0% response to the emerging data, maintaining the estimate at $100 over time. The paid and reported BF projections reflect a blending of the CL and EL estimates. Table 5 shows the corresponding progressions of estimates of unpaid claims arising from the methods.

Table 5
Comparative Projections of Unpaid Claims

Age	EL	CL	P-BF	R-BF
1	91	83	90	88
2	82	74	80	77
3	68	60	65	62
4	54	46	50	47
5	45	37	40	38
6	35	28	30	28
7	26	18	20	18
8	17	9	10	9
9	12	5	5	5
10	8	-	-	-

In this example, at the 1st valuation, the $8 difference between the CL and EL estimates of unpaid claims liabilities represents 9% of the CL estimate ($83). By the 5th valuation, the $8 difference between the EL and the CL estimates represents 21% of the CL-estimate of unpaid claims ($37) for the accident period. Maintaining the initial expected losses as the estimated ultimate at the 10th valuation yields an unpaid claim estimate of $8, even though the expected payment pattern suggests that no further payments are expected. At some point along the way from accident year inception, to initial period-end valuation, and to final settlement of all attendant claims, stakeholders need to move off the initial expected loss estimate and respond to the actual claims experience. But when? And to what?

Extending the investigation to the recognition of the accident year incurred losses in a calendar year income statement of an insurance company, Table 6 shows the progression of ultimate loss estimates based on the CL and the paid and reported BF methods, along with the calendar year recognition.

Table 6
Recognition of Accident Year Losses

	Accident Year Ultimates				Calendar Year Incurred Losses		
Age	CL	P-BF	R-BF	Cal Yr	CL	P-BF	R-BF
1	92	99	97	1	92	99	97
2	92	98	96	2	-	(1)	(2)
3	92	97	95	3	-	(1)	(1)
4	92	96	93	4	-	(1)	(1)
5	92	95	93	5	-	(1)	(0)
6	92	95	93	6	-	(1)	(0)
7	92	94	92	7	-	(1)	(0)
8	92	93	92	8	-	(1)	-
9	92	93	92	9	-	(0)	-
10	92	92	92	10	-	(0)	-
				Sum	92	92	92

As illustrated, the CL estimate of ultimate losses for the accident year is accurate at the 1st valuation, and so the recognition of incurred losses is fully contained to the corresponding calendar year. For the reported BF projection, which reflects a blending of initial expectations ($100) and actual reported emergence over time, the initial recognition is $97. Subsequent calendar year results reflect favorable development, in total of $(5) for the reported BF, until the true ultimate of $92 is recognized by the 7th year on a reported basis. The recognition of the true ultimate losses from the paid BF approach is slower, with $99 recognized in the 1st year and favorable development of $(7) in subsequent periods.

The framework and illustrations become more intriguing when the results are compiled from successive accident years at successive calendar year-end reporting dates, where there are deviations in the emerging experience from initial expectations. For this illustration, we utilize the notion of an underwriting cycle, where the conditions around pricing and loss trends yield a cyclical pattern of ultimate loss ratios. Chart 4 illustrates the cycle used in subsequent examples, in terms of its "peaks and valleys" and the time-period from peak-to-valley and valley-to-peak.

Chart 4
Cycle Assumptions for Accident Year Loss Ratios
and a Constant Expected Loss Ratio over Time

Over the entire period, we assume that the initial expected loss ratio is a constant 65%, with actual loss ratios spanning from 50% to 80% over a 24 year period. That is, a starting loss ratio of 65% increases to 80% over a 6-year period, decreases to 50% over a 12-year period, and then returns to 65% over the next 6 years. With a constant premium volume of $154 each year, the expected losses are $100, with actual losses ranging from $77 (when the loss ratio is 50%) to $123 (when the loss ratio is 80%). Appendix C shows the assumptions for premium volume and loss ratios by accident period, as well as the accident period loss payment and reporting patterns.

The results that are shown in the following tables and charts reflect a model where company management ("Mgmt") consistently forms a best-estimate of ultimate and the corresponding reserves based on the paid BF approach. This reflects a tempered approach in terms of its responsiveness to the emerged claims data from the initial to subsequent valuations. Management's estimates are compared to an actuary's estimate, which is consistently based on the reported BF approach. Therefore, the actuary's estimates reflect a tendency for greater responsiveness to the emerging claims experience than management's.[3]

[3] The reader is reminded of the Disclaimer within the Abstract for this paper. The author's use of the illustrative preferences for method selection by "management' and "an actuary" is intended solely to facilitate the description of the scenario and the potential implications of different method selections on one stakeholder's view of the relative position of another stakeholder's estimate for unpaid claims liabilities, rather than referring to the two stakeholders as "Stakeholder A" and "Stakeholder B."

Table 7 shows the array of estimates for the first three accident periods for the first three calendar periods, in order to provide the reader with a view on the mechanics of the model, before showing the overall results once the illustration reaches steady-state in terms of a rolling set of 10 accident years contributing to a calendar year's result.

Table 7
Projected Ultimate Losses by Method and Selected by Stakeholders
Accident Years 1 to 3 at Calendar Year-ends 1 to 3

AY	Age	EL	CL	P-BF	R-BF	Mgmt	Actuary
1	1	100	104	100	101	100	101
	2	100	104	101	102	101	102
	3	100	104	101	103	101	103
2	1	100	108	101	103	101	103
	2	100	108	102	104	102	104
3	1	100	112	101	104	101	104

For accident year 1, the assumed expected loss is $100 (65% loss ratio) and the true ultimate is assumed to be $104 (68% loss ratio, and indicated by the CL at each age). At the first valuation, management's pick for ultimate losses is based on the paid BF ($100), which is slightly higher (rounding) than the expected losses of $100. The actuary's pick ($101) is a bit more responsive to the emerging experience.

At the second valuation for accident year 1, management's estimate increases to $101, while the actuary's estimate increases to $102. These changes represent prior year development in the calendar year when the change in estimate is made.

Table 8 shows the progression of the respective estimates, for the current accident period and for changes in the estimates for prior periods.

Table 8
Progression of Ultimate Loss Estimates by Accident Year by Calendar Year

		Calendar Year				Calendar Year		
	AY	1	2	3		1	2	3
		Ultimate				Prior Year Development		
Mgmt	1	100	101	101			0	1
	2		101	102				1
	3			101				
					Sum =>		0	1
Actuary	1	101	102	103			1	1
	2		103	104				2
	3			104				
					Sum =>		1	2

Each estimate of ultimate for the current accident period is shown in the boxed-cells in the left-portion of the table. The change in estimates for prior accident periods during a calendar period are shown and compiled (shaded cells) in the right-portion of the table.

Management's current accident year estimates are less than the actuary's estimates, due to the lesser response of the paid BF approach to emerging claims data than that of the reported BF approach. Thus, relative to the recognition of the ultimate losses from the actuary's picks, management's recognition of ultimate losses is delayed. For instance, for accident year 1, ultimate losses of $104 will need to be recognized. By the third valuation, management has recognized $101 while the actuary's estimate is $103; management will have subsequent development of $3, while the actuary's estimate will develop by $1.

Table 9 shows the components of calendar year results over the 1st ten years of the model.

Table 9
Illustration of Current Accident Year and Calendar Year Incurred Losses
Years 1 to 10

	Current AY		Change in Prior		Calendar Year	
	Ult	Ult	Ult	Ult	Ult	Ult
Year	Mgmt	Actuary	Mgmt	Actuary	Mgmt	Actuary
1	100	101	0	0	100	101
2	101	103	0	1	101	103
3	101	104	1	2	103	106
4	102	105	3	4	104	109
5	102	107	5	6	107	113
6	102	108	7	8	109	117
7	102	107	10	11	112	118
8	102	105	12	12	114	117
9	101	104	13	12	115	116
10	101	103	14	10	115	113

Over the 1st ten years, the actuary's loss picks for the current accident year are higher than management's. (Recall that years 1 to 10 reflect ultimate loss ratios greater than initially expected.) Still, both the actuary and management underestimate the true ultimates, as evidenced by the adverse development of prior years' estimates in calendar year results. Table 10 shows the results as the company reaches a 'steady state' in years 10 to 20.

Table 10
Illustration of Current Accident Year and Calendar Year Incurred Losses
Years 10 to 20

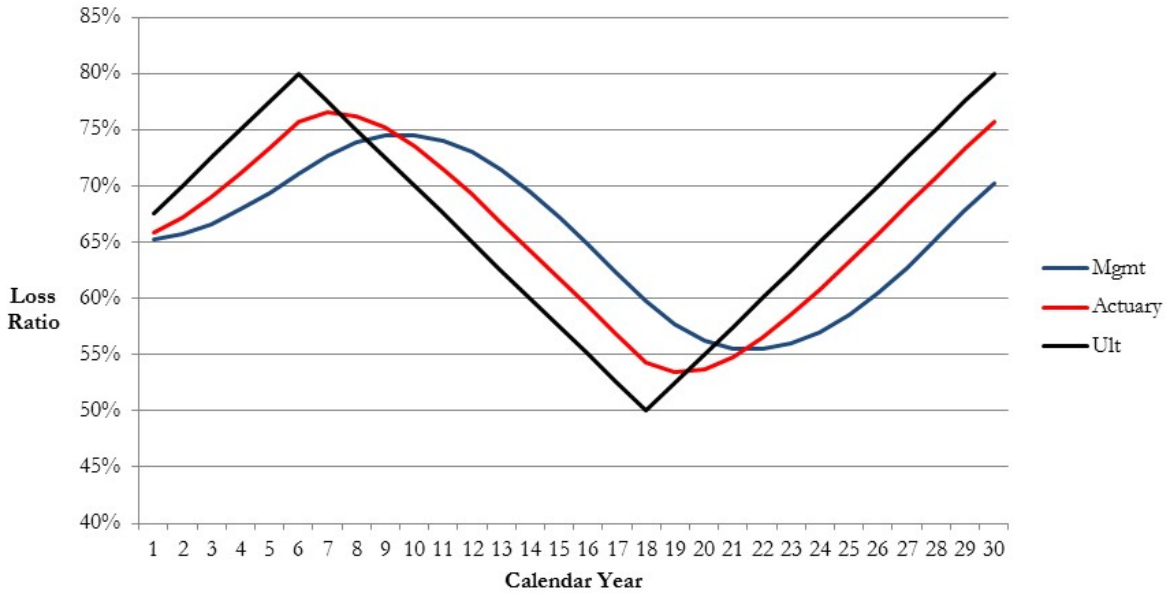
	Current AY		Change in Prior		Calendar Year	
	Ult	Ult	Ult	Ult	Ult	Ult
Year	Mgmt	Actuary	Mgmt	Actuary	Mgmt	Actuary
10	101	103	14	10	115	113
11	100	101	13	9	114	110
12	100	100	12	7	112	107
13	100	99	10	4	110	103
14	99	97	8	2	107	99
15	99	96	5	(1)	103	95
16	98	95	1	(3)	100	91
17	98	93	(2)	(6)	96	87
18	98	92	(6)	(8)	92	83
19	98	93	(9)	(11)	89	82
20	98	95	(12)	(12)	87	83

At year 12, true ultimate losses return to 100 (65% loss ratio), and both management and the actuary recognize this as their views of current accident year losses. However, the financial results for calendar year 12 are still hurt by adverse development from inadequate funding of prior accident years.

The results at year 15 begin to show favorable development of the actuary's prior years' estimates; it takes until year 17 for management's estimates to show favorable development. Although the true loss ratio for accident year 18 reaches its low at 50% ($77 ultimate loss), that calendar year's incurred losses of $92 reflect management's current accident year estimate of $98, and favorable $(6) development from prior years. The actuary's initial view of the current accident year loss ratio at year 18 is $92, giving a bit more recognition to the emerged favorable experience than management's $98, but both still higher of the ultimate emerged loss of $77.

The relative trends in loss ratios are shown in Chart 5 below. The chart, based on the assumptions underlying the outcomes, conveys the common notion that the cycle of calendar year booked loss ratios often reflects a delayed and tempered view of the cycle of ultimate accident year loss ratios.

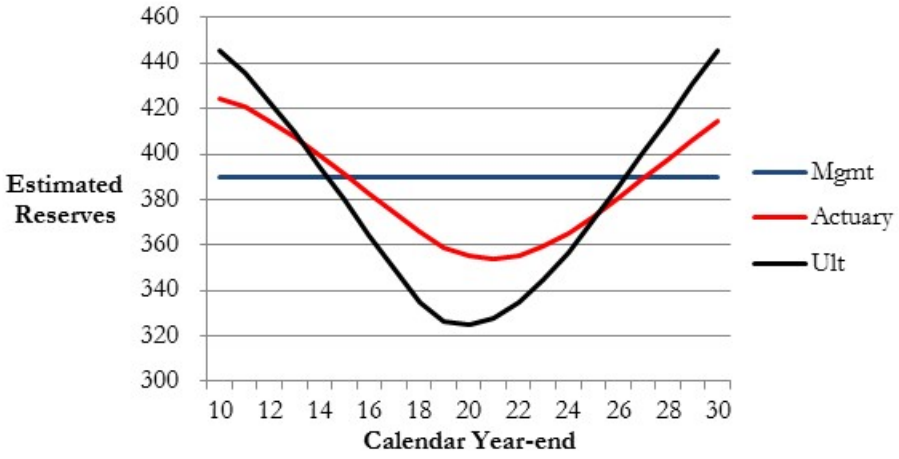
Chart 5
Comparison of Calendar Year Loss Ratios
Based on Alternate Judgments for Ultimate Accident Year Loss Picks



Viewing the model's results in terms of actuarial indicated and management booked loss reserves at successive financial reporting dates provides additional insights as to the potential implications from alternate judgments for the basis of ultimate loss picks. Chart 6 below illustrates the indicated

unpaid claim estimates for years 10 to 30 (the 'steady state' period of the model) as selected by management (based on the paid-BF), by the actuary (based on the reported-BF), and based on the (true or hindsight) ultimate.

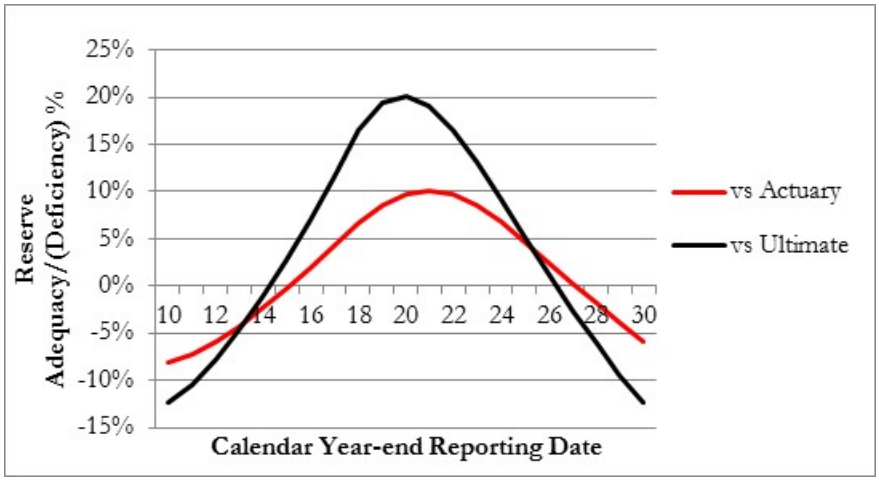
Chart 6
Comparison of Unpaid Claim Estimates
Based on Alternate Judgments for Ultimate Accident Year Loss Picks
At end of Years 10 to 30



Once in a steady state, with constant premium volumes and ELR's, and the paid-BF as the basis for management's picks, the indicated reserves are constant at $390. The actuary's estimates of unpaid claims liability fall and rise over the period shown, with a partial response in selecting ultimates given the deteriorating and improving true claims experience. The hindsight (true) reserves, based on the cyclical accident year ultimates, demonstrate a greater degree of variability, driven by the constant premium volume and rising and falling levels of accident period incurred losses.

The implications of these relative reserve estimates at a point in time, and over time, are further highlighted in Chart 7, which shows the estimated adequacy of management's reserves, in relation to the actuary's indicated reserves at the particular financial reporting date (the red line), and in relation to ultimate (the black line). A 0% reserve adequacy position corresponds to the situation when the booked reserves are equal to another estimate, whether the actuarial indication or the hindsight (true) estimate of unpaid claims. When management's reserve is below the actuary's (or hindsight) estimate, a negative percentage is shown.

Chart 7
Relative Reserve Adequacy Position of Booked Reserves over Time
Based on Alternate Judgments for Ultimate Accident Year Loss Picks
At end of Years 10 to 30



Adequacy Position vs Actuary = (Booked minus Actuary Indication) / (Actuary Indication)

Adequacy Position vs Ultimate = (Booked minus Ultimate Indication) / (Ultimate Indication)

Differences in perspectives for loss picks that may be perceived as 'small' can generate differences in reserve estimates (at a point in time, not only at ultimate) that are relatively large. Management's reserves at the end of calendar year 10 (four years after the peak true loss ratio for accident year 6) are 8% less than the actuary's indication at that time. Eleven years later (at the end of calendar year 21), after loss ratios have improved, management's reserves are 10% greater than the actuary's indication.

A hindsight (ultimate) view of booked reserves is commonly disclosed in a loss reserve runoff schedule in a public insurance company's 10K annual report, or can be derived from manipulations of data presented in Schedule P of insurance companies' statutory-basis annual statement. In Chart 7 above, management's reserves booked at year-end 10 would be ultimately revealed as having been 12% deficient, and the year-end 20 reserves would be revealed to have been 20% redundant.

An integrated view of the model, in terms of its assumptions for cyclical accident year loss ratios, and the hypothetical management's approach to booking accident year losses (based on a paid-BF method), is shown in Chart 8, including the hindsight view of booked reserve adequacy:

Chart 8
Comparison of Loss Ratios and Hindsight Reserve Adequacy

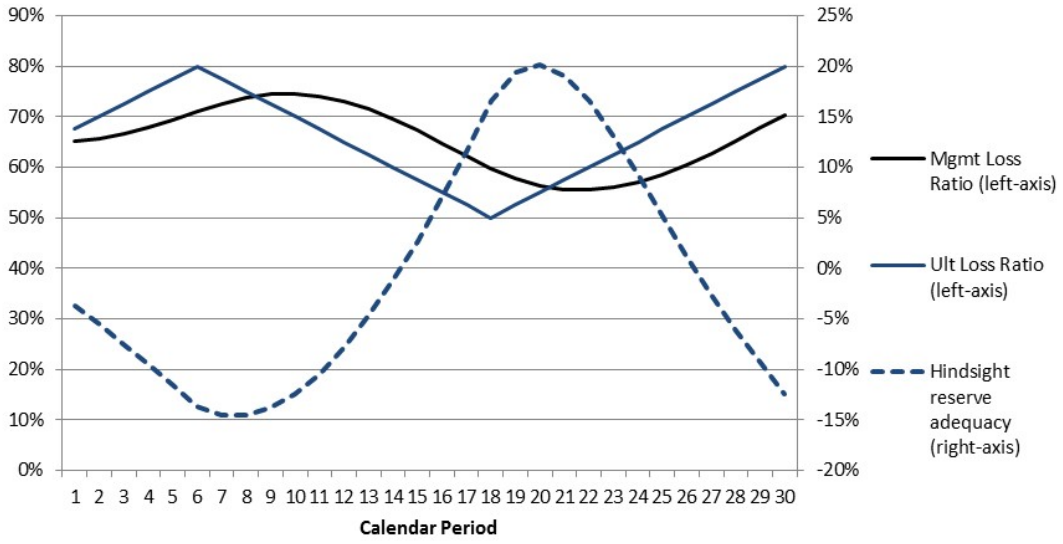


Chart 8 shows the full range of 30 years in the underlying model, including the first 10 years before the steady state is reached in terms of the levels of claim payments and reserves. The chart provides a view on hindsight reserve adequacy over the full range of the assumed cycle.

- A paid BF approach for selecting loss picks creates a delayed recognition of accident year losses, thereby generated a delayed and tempered cycle of calendar year loss ratios, relative to the accident year loss ratio cycle.
- The relative adequacy of loss reserves derived from a paid BF approach, based on the scenario illustrated by the model, ranges from 15% deficient to 20% redundant, in relation to the unpaid claims liabilities from the true ultimate losses. The reserve adequacy cycle is inverted to the true loss ratio cycle, and, in this example, is lagged by 2 periods (driven by the collection of assumptions underlying the model).

5. So, Now What?

I close this paper with a collection of observations, questions, and responsive thoughts (not answers) regarding potential implications of the content in previous sections.

1) The model is simplistic in that it reflects a flat initial expected loss ratio. That is not reality.

Yes, the model is simplistic when viewed from that vantage point. I acknowledge that managements consider the current accident year's experience when setting ultimate losses and the associated reserves at the initial annual reporting. The scenario illustrates a tempered response to the initial expected loss ratio, by way of the paid BF at

all valuations. I believe the model is instructive as is; of course, the model could be enhanced to show some variation in the initial expected losses, although such variation would likely be less than that manifest by true ultimate losses.

2) What happens if differences in stakeholders' styles on the basis of loss picks become "large" in terms of the differences among unpaid claims estimates? Stated another way, at what point are different styles (and differences in reserve estimates) too large to tolerate from an "actuarial reasonability" perspective?

   While a valid and thought-provoking question, it goes beyond the scope of this paper.

3) If a company's management books its best estimate that is different than the company's actuarial indication, what are implications on the level of documentation that is expected?

   Standards of practice for the accounting and actuarial professions require that sufficient documentation of the analysis supporting booked amounts and actuarial indications exists. Where the booked amounts are equivalent to the actuarial indication, then documentation that meets the actuarial standards should be sufficient. Where management considered the actuarial indication and formed its estimate based on alternate judgments regarding assumptions, methods, or basis of picks, management should have sufficient documentation highlighting the areas of difference and the basis (evidence, rationale) for such differences.

4) Is the author suggesting that any rational stakeholder adopt a chain ladder projection at an "early" maturity when the development factor to ultimate is "large?"

   Maybe; that would depend on the relative stability of the historical development experience and consistency in company underwriting and claims operations. The author believes that there is opportunity for actuaries to enhance their measurement and communication of the degree of certainty that can be associated with actuarial indications from different methods and types of data. "Inherent volatility" or "large LDF's" are common reasons for discounting or outright ignoring chain-ladder projections at early maturities. But, has the actuary compiled a history of the various projections over time to assess which tend to perform better than others? Has the actuary tested the performance of methods? This was an area of investigation in Claim Reserving: Performance Testing and the Control Cycle, by Yi Jing, Joseph Lebens, and Stephen Lowe (CAS, 2009). Therein they described a testing approach for evaluating the "skill" of a method, as a "measure of the amount of variation captured by the particular actuarial method." They also wrote that "the control cycle should involve an ongoing assessment of the estimation skill of the actuarial methods currently being employed, and exploration of opportunities to enhance overall estimation skill by implementing better actuarial projection methods."

5) Is the author suggesting that, at some point along the path of an accident year maturing, a particular projection method could be viewed as "wrong" in relation to another method?

   Many individual judgments are made in the course of a reserving analysis and each of these, individually, could be viewed as reasonable, optimistic, conservative, or unreasonable. Generally, the scope of an actuary's professional opinion regarding reserves is on the appropriateness of methods and reasonableness of assumptions and judgments in total (all accident years, all analysis segments), not on individual elements. This is consistent with the actuarial opinion on the loss reserves in aggregate, not for individual claims.

So, my response is "No," in that an individual judgment for a particular method for a particular accident year is likely not the subject of a professional opinion. Still, in this context, consider the following.

Chart 9
Comparison of Ultimate Loss Estimates
Based on Alternate Judgments for Ultimate Accident Year Loss Picks
With a View on a "Reasonability Interval" of the CL-projection

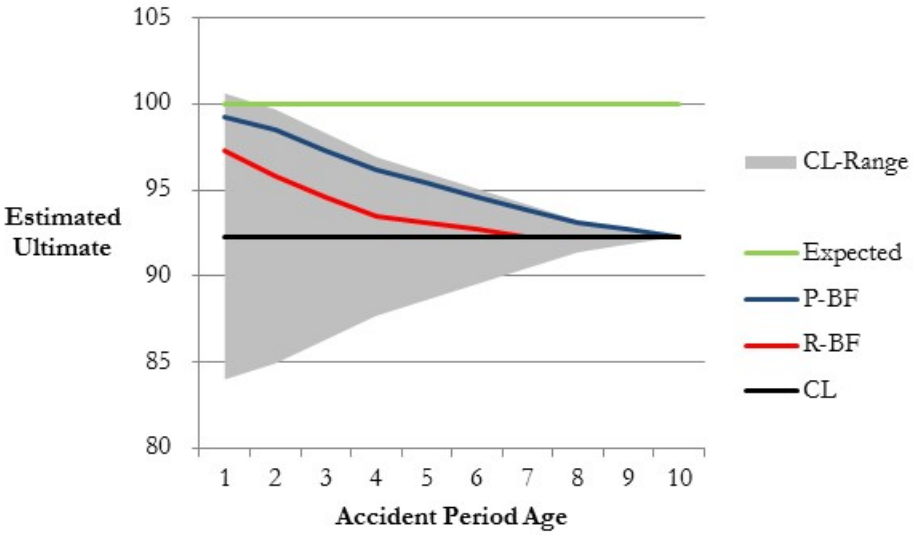


Chart 9 is similar to Chart 3, showing the progression of ultimate loss projections from four basic methods. I have added a shaded area to illustrate a potential range ("reasonability interval") of projected amounts from the Chain Ladder method. The range decreases in breadth over time as the accident year matures. Based on the graphic, the Expected loss pick at the 1st or 2nd valuation would be within, albeit at the high-end of the CL-range. By the 3rd valuation, the Expected amount would not fall within the CL-range. At that point, would a pick based on Expected loss be "wrong?"

6) Is the author suggesting that the stakeholders document their styles for how they generally form their picks?

A documented reserving policy for an insurance company is an element of good governance around reserves, as the reserves are a significant accounting estimate in the financial statements. The company actuary, management, and the Board (audit committee) should ensure a common understanding of their own, and each other's, perspectives, tendencies, principles, and objectives (that is, styles). Likewise, company stakeholders and key personnel from the external audit firm should ensure understanding of each other's perspectives.

A documented reserving policy will help to describe management's view on "why now?" in terms of a response to the emerging claim experience, whether for reporting on results for a quarter for which detailed actuarial re-projections are available or not, and also in response to one or more individual large loss events.

The reserve decision making-process is fluid as internal and external conditions evolve and change over time. Therefore, an overly prescriptive policy is not realistic, desirable, or appropriate.

6. Conclusions

The styles of management and an actuary for selecting loss picks do matter. An articulated policy surrounding how management selects its estimate is good governance to recognize that this selection process does matter and is not subject to whim. Documentation of the selection process of management promotes transparency for stakeholders and is a check that the policy has been followed. It also provides, through transparency, a check on the bounds of how large style differences can become as quantified by the extent of differences from actuarial loss picks. Further, the actuary should ensure that documentation of the actuarial process is in compliance with actuarial standards of practice.

Actuaries and management should communicate, up front, and share their views on how they each think about the degree of responsiveness to the emerging data that their loss picks will likely reflect. When the reserving styles of the various stakeholders are in-sync, the periodic discussions around the period's claims experience and forming views on indications and booked reserves are smoother and less contentious, as compared to when the styles are out-of-sync.

Acknowledgements

The author acknowledges and appreciates the contributions of Kathlyn Herrick and Lisa Slotznick to this paper. Kathlyn assisted with the compilation of industry data and preparing some of the charts displayed herein. Kathlyn and Lisa reviewed an early draft of the paper and enhanced the readability of the paper. Responsibility for any remaining typos, errors, or lack of clarity remains with the author.

Biography of the Author

Mark Littmann is a principal in the Actuarial Services practice of PricewaterhouseCoopers LLP, the US member firm of the PwC global network of firms. He leads actuarial teams supporting external financial audits of insurance organizations and corporations with self-insured exposures and providing consulting services regarding reserving valuations and processes and other analytical applications. He has a degree in Mathematics and Economics from Valparaiso University. He is a Fellow of the CAS and a Member of the American Academy of Actuaries.

Appendix A

US P&C Industry Booked Loss Ratios
Accident Years 1996 to 2009
At 12-month & at 72-month Valuations
(plus 108-month valuation for GL-Occurrence)

		1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	Average
PAL	at 12 months	76%	73%	71%	75%	79%	78%	76%	71%	68%	67%	66%	69%	69%	73%	72%
	at 72 months	71%	69%	70%	75%	80%	78%	75%	67%	64%	63%	63%	67%	67%	70%	70%
	Ratio	0.93	0.94	0.98	1.00	1.01	1.00	0.98	0.95	0.94	0.94	0.96	0.97	0.97	0.97	0.97
CAL	at 12 months	77%	78%	77%	78%	77%	73%	67%	64%	62%	61%	62%	62%	62%	63%	69%
	at 72 months	81%	84%	87%	92%	89%	78%	67%	60%	57%	58%	58%	61%	61%	60%	71%
	Ratio	1.05	1.08	1.13	1.18	1.15	1.07	1.00	0.95	0.93	0.95	0.94	0.97	0.98	0.96	1.02
CMP	at 12 months	74%	68%	74%	74%	71%	75%	60%	56%	59%	61%	53%	55%	69%	60%	65%
	at 72 months	76%	68%	79%	80%	81%	76%	59%	52%	53%	56%	47%	50%	65%	60%	64%
	Ratio	1.02	1.01	1.07	1.09	1.13	1.02	0.97	0.92	0.90	0.92	0.90	0.92	0.94	0.99	0.99
GL-Occ	at 12 months	80%	81%	82%	79%	79%	89%	72%	69%	68%	66%	64%	66%	67%	69%	74%
	at 72 months	77%	84%	91%	95%	98%	101%	79%	63%	55%	55%	54%	60%	61%	61%	74%
	Ratio	0.97	1.04	1.11	1.21	1.24	1.13	1.10	0.91	0.80	0.84	0.85	0.91	0.92	0.89	0.99
	at 108 months	78%	86%	99%	102%	101%	102%	80%	63%	53%	53%	52%				
	Ratio	0.98	1.07	1.20	1.29	1.27	1.14	1.11	0.90	0.78	0.81	0.82				

Loss ratios for GL-Occurrence demonstrated a degree of further development from the 72-month valuation to the 108-month valuation.

For personal auto liability (PAL), the average ratio of the loss ratio at 72-months divided by the loss ratio at 12-months) over the 14 accident years was 0.97 (favorable 3%), with observations that spanned from 0.93 to 1.01.

In contrast, the booked loss ratios for General Liability – Occurrence at 72-months, on average, were within 1% of the loss ratios booked at 12-months. On an accident year by accident year basis, however, individual years' ratios were as high as 1.24 and as low as 0.80. At 108-months, the highest and lowest ratios were 1.29 and 0.78.

Appendix B

Numerical Example to Illustrate Degree of Responsiveness
of Alternate Methods to Actual Loss Experience

	EL	BF	BKT	CL
A Premium	125	125	125	
B Reported Losses	51	51	51	51
C IELR	80%	80%		
D LDF (to Ultimate)		1.82	1.82	1.82
E Expected Reported %		55%	55%	55%
F Expected IBNR %		45%	45%	45%
G Expected Loss	**100**	100		
H Expected Reported Loss		55		
I Expected IBNR Loss		45		
J BF Estimated Ultimate		**96**		
K Initial Expected Loss for BKT			96	
L Expected IBNR Loss for BKT			43	
M BKT Estimated Ultimate			**94**	
N CL Estimated Ultimate				**92**
O Response to Actual	0%	55%	80%	100%

Item(s)	Notes
A, B	Assumed data for the illustration
C, D	Assumptions for the key parameters of the methods.
E, F	Derived from D.
G	A multiplied by C
H, I	Derived from E, F, & G
J	B plus I
K	Equal to J
L	Derived from K & F
M	B plus L
N	B multiplied by D
O	Derived as (Difference of method-estimate to the EL-estimate) divided by the (Difference of the EL and the CL-estimate)

Appendix C

Assumptions for Accident Year Premium, ELR's, and Ultimate Loss Ratios
and Accident Year Loss Payment and Reporting Patterns

Supporting the Tables & Charts in Section 4

AY	Premium	ELR	Ult LR	AY Age	Payment	Reporting
1	154	65%	68%	1	**10%**	**35%**
2	154	65%	70%	2	**20%**	**55%**
3	154	65%	73%	3	**35%**	**70%**
4	154	65%	75%	4	**50%**	**85%**
5	154	65%	78%	5	**60%**	**90%**
6	154	65%	80%	6	**70%**	**95%**
7	154	65%	78%	7	**80%**	**100%**
8	154	65%	75%	8	**90%**	**100%**
9	154	65%	73%	9	**95%**	**100%**
10	154	65%	70%	10	**100%**	**100%**
11	154	65%	68%			
12	154	65%	65%			
13	154	65%	63%			
14	154	65%	60%			
15	154	65%	58%			
16	154	65%	55%			
17	154	65%	53%			
18	154	65%	50%			
19	154	65%	53%			
20	154	65%	55%			
21	154	65%	58%			
22	154	65%	60%			
23	154	65%	63%			
24	154	65%	65%			
25	154	65%	68%			

Movement Analysis

Andy Staudt FIA, FCAS, MAAA

Abstract.
Consistent with the requirements of Actuarial Standards of Practice (ASOPs) 36 and 41 (paragraphs 4.5 and 3.5, respectively), this paper derives simple but mathematically sound formulas for explaining differences in estimates of ultimate from one period to the next. Specifically, the change in ultimate is decomposed into the movement due to loss experience relative to the movement due to changes in assumptions or methods. The approach outlined below is for use in common reserving situations where the Bornhuetter-Ferguson (BF) or Chain-Ladder (CL) methods are used, but can also be easily extended in other circumstances.

Keywords.
Reserving; Communication.

1. INTRODUCTION

In booking the unpaid claims reserve, good governance as well as Actuarial Standards of Practice (ASOPs)[1,2] require that the actuary clearly communicates any material differences in the estimate of ultimate relative to earlier projections. This is so that management has the necessary tools to assess, challenge or validate the actuary's recommendation and make their own determination as to the final carried amount.

In order to do so effectively, the actuary needs to be transparent as to why the estimate of ultimate changed. Did it move as a result of loss experience emerging more or less favorably than expected or did it move because of changes in the underlying methods or assumptions? And in the case of the latter, what impact did these changes have on the final result and why were these changes warranted?

To address these questions, this paper derives simple but mathematically sound formulas for explaining differences in estimates of ultimate loss from one period to the next. Here, the change in ultimate is decomposed into the movement due to loss experience relative to the movement due to changes in assumptions or methods. While most actuaries will already perform this type of analysis in some form (typically via successive substitution of new data and new assumptions into new methods), the "movement analysis" outlined below offers a consistent approach for communicating as well as quantifying change which will work in many practical situations.

[1] Explanation of Material Differences – If a later actuarial communication produced by the same actuary, which opines on the same issue, includes materially different results or expresses a different opinion from the former communication, then the later communication should make it clear that the earlier results or opinion are no longer valid and explain why they have changed. [excerpted from ASOP 41: 3.5]

[2] Changes in Opining Actuary's Assumptions, Procedures, or Methods – If a change occurs in the opining actuary's assumptions, procedures, or methods from those previously employed in providing an opinion on the entity's reserves, and if the actuary believes that the change is likely to have a material effect on the results of the actuary's reserve analysis, then the actuary should disclose the nature of the change. [excepted from ASOP 36: 4.5]

1.1 Outline

The Executive Summary in Section 2 presents the movement analysis in its complete form with all the formulas needed to implement this analysis within a practical setting presented in Table 2.

Section 3 proceeds to develop these formulas iteratively by isolating and quantifying the impact of loss experience as well as certain methodological or assumption changes on estimates of ultimate where the Bornhuetter-Ferguson (Section 3.1) or Chain-Ladder (Section 3.2) methods are used. For readability, the actual derivations of the key formula are contained within a Technical Appendix to this paper. While these sections are not exhaustive as to situations that might arise in practical reserving settings, they can easily enough be extended to other circumstances as will be discussed in Section 3.3. This section should prove useful for understanding the how and why of this analysis intuitively.

Finally, to illustrate this analysis, an example is included in Section 4. Also provided is a workbook which includes the necessary formula to implement this analysis in Excel.

1.2 Notation

The following notation is used within this paper:

- q_k is the percentage of loss developed at time k;
- C_k is the actual loss at time k;
- u is the initial expected loss ratio (IELR);
- P is the premium; and
- U_k is the estimate of ultimate loss at time k.

Using this notation, the Bornhuetter-Ferguson (BF) and Chain-Ladder (CL) estimates of ultimate loss at time k can be written as:

Table 1. BF and CL projections of ultimate loss.

Method	Formulation
BF method	$U_k = C_k + uP(1-q_k)$
CL method	$U_k = C_k / q_k$

Further, "hats" are used to indicate updated assumptions. For instance, where q_k should be taken to be the original assumption of the percentage of loss developed at time k, $\hat{q}_k$ would be the revised assumption as to the percentage of loss developed at time k.

2. EXECUTIVE SUMMARY

In Section 3, the movement analysis is derived by iteratively considering each of the following:

- The movement in ultimate as a result of loss experience emerging differently from expectations;
- The movement in ultimate as a result of premiums emerging differently than originally anticipated;
- The movement in ultimate as a result of changes to key assumptions including (i) development patterns and (ii) IELRs; and
- The movement in ultimate by switching between the CL and BF methods.

That said, the table below provides the complete set of equations for producing the movement analysis. As the exact form depends on what the current and prior methods are, the table is split across this dimension with dots "•" used to indicate where the result is invariant to the method. These equations are also programmed into the attached Excel workbook.

Table 2. Movement Analysis.

Movement in ultimate due to:	Method		Formulation $(U_{k+1} - U_k)$
	Prior	**Current**	
Loss experience	BF	•	$= (C_{k+1} - C_k) - uP(q_{k+1} - q_k)$
	CL	•	$= \left((C_{k+1} - C_k) - \frac{C_k}{q_k}(q_{k+1} - q_k)\right) \times \left(\frac{1}{q_{k+1}}\right)$
Change in method	BF	CL	$+ C_{k+1}\left(\frac{1}{q_{k+1}} - 1\right) - uP(1 - q_{k+1})$
	CL	BF	$+ uP(1 - q_{k+1}) - C_{k+1}\left(\frac{1}{q_{k+1}} - 1\right)$
Change in premium	•	BF	$+ (\hat{P} - P)u(1 - q_{k+1})$
Change in development pattern	•	BF	$+ u\hat{P}[(1 - \hat{q}_{k+1}) - (1 - q_{k+1})]$
	•	CL	$+ C_{k+1}\left[\left(\frac{1}{\hat{q}_{k+1}} - 1\right) - \left(\frac{1}{q_{k+1}} - 1\right)\right]$
Change in IELR	•	BF	$+ (\hat{u} - u)\hat{P}(1 - \hat{q}_{k+1})$

3. MOVEMENT ANALYSIS

3.1 The Bornhuetter-Ferguson (BF) Method

Consider the situation where the estimate of ultimate is set equal to the BF method. Here, the estimate of ultimate can change for any of four reasons:

- Loss experience that emerges more or less favorably than expected;
- Premium amounts that are restated;
- Changes in the development pattern; or
- Changes in the IELR.

The following considers each of these in turn.

3.1.1 Movement in ultimate due to loss experience

Assuming that no assumptions are updated, the change in ultimate $U_{k+1} - U_k$ can be written as:

Table 3. Decomposition of movement in ultimate for the BF method.

Movement in ultimate due to:	**Formulation** $(U_{k+1} - U_k)$
Loss experience	$= (C_{k+1} - C_k) - uP(q_{k+1} - q_k)$

The above should be recognizable as the actual vs. expected (AvE) statistic when using the BF method. $C_{k+1} - C_k$ represents actual emergence and $uP(p_{k+1} - p_k)$ represents expected emergence. Indeed, the change in the BF ultimate without any changes in assumptions reduces to the AvE statistic with claims emergence that is more or less favorable than expected flowing entirely through to the change in ultimate.

3.1.2 Movement in ultimate due to change in development pattern

Suppose as a result of loss experience emerging differently from expectations, the development pattern is revised. Using "hats" to indicate updated assumptions, the change in ultimate is written as:

Table 4. Decomposition of movement in ultimate for the BF method.

Movement in ultimate due to:	**Formulation** $(U_{k+1} - U_k)$
Loss experience	$= (C_{k+1} - C_k) - uP(q_{k+1} - q_k)$
Change in development pattern	$+ uP[(1 - \hat{q}_{k+1}) - (1 - q_{k+1})]$

While this derivation is less straightforward than above, observe that the change in ultimate is decomposed into the AvE statistic from Table 3 and a remainder. In this instance, the remainder is just the difference in the estimated reserve at time $k+1$ implied by the current and prior selected development patterns, or the movement in ultimate due to the change in pattern.

It should be noted that while the subscripts k and $k+1$ might indicate points at which the development pattern is selected, this method will work equally well in instances where the selected development pattern is interpolated. For example, if loss development factors are selected over periods from 3-15 months, 15-27 months and so forth, it is no problem to interpolate the pattern as at 6, 9 and 12 months in order to apply the movement analysis to the most recent year over the subsequent three quarters.

3.1.3 Movement in ultimate due to change in IELR

Going one step further, should an adjustment be made to the IELR as well as the development pattern, the change in ultimate is written as:

Table 5. Decomposition of movement in ultimate for the BF method.

Movement in ultimate due to:	**Formulation** $(U_{k+1} - U_k)$
Loss experience	$= (C_{k+1} - C_k) - uP(q_{k+1} - q_k)$
Change in development pattern	$+ uP[(1-\hat{q}_{k+1}) - (1-q_{k+1})]$
Change in IELR	$+ (\hat{u} - u)P(1-\hat{q}_{k+1})$

Again, the change in ultimate can be decomposed into the movement in ultimate due to loss experience, the movement in ultimate due to change in development pattern and a remainder. Here, the remainder is just the difference in the estimated reserve at time $k+1$ implied by change in IELR (using the current development pattern), or the movement in ultimate due to the change in IELR.

As an aside, note that the order in which the development pattern and IELRs are considered matters. This is obvious from the above as the change in IELR is based on the current development pattern. The above order seems reasonable as it might be practice to select the development pattern prior to the IELR; however, it is easy enough to consider these changes in reverse order as:

Table 6. Decomposition of movement in ultimate for the BF method (alternate formulation).

Movement in ultimate due to:	**Formulation** $(U_{k+1} - U_k)$
Loss experience	$= (C_{k+1} - C_k) - uP(q_{k+1} - q_k)$
Change in IELR	$+ (\hat{u} - u)P(1-q_{k+1})$
Change in development pattern	$+ \hat{u}P[(1-\hat{q}_{k+1}) - (1-q_{k+1})]$

3.1.4 Movement in ultimate due to change in premium

The next natural extension is to consider the impact that changes in premiums will have on the estimate of ultimate. Similar to the prior subsection, a decision needs to be made as to the order in which to consider changes in premium relative to other changes. Although there is an argument to consider it prior to loss experience, the below considers it after making an allowance for deviations in loss experience relative to expectation but prior to changes in assumptions. This is so that the AvE statistic will tie to any prospective estimates of loss emergence computed at prior periods.

Table 7. Decomposition of movement in ultimate for the BF method.

Movement in ultimate due to:	**Formulation** $\left(U_{k+1} - U_k\right)$
Loss experience	$= \left(C_{k+1} - C_k\right) - uP\left(q_{k+1} - q_k\right)$
Change in premium	$+ \left(\hat{P} - P\right)u\left(1 - q_{k+1}\right)$
Change in development pattern	$+ u\hat{P}\left[\left(1 - \hat{q}_{k+1}\right) - \left(1 - q_{k+1}\right)\right]$
Change in IELR	$+ \left(\hat{u} - u\right)\hat{P}\left(1 - \hat{q}_{k+1}\right)$

The above equation provides a near-complete decomposition of the movement in ultimate into each of the key drivers of change when using the BF method, with Tables 3-6 only representing partial solutions. In the next sections, we extend these formulas to consider situations when using the CL method, moving between the CL and BF methods and netting down estimates of gross ultimate loss for the impact of reinsurance.

3.2 The Chain-Ladder (CL) Method

The formulas from the prior section can be extended in situations where the CL, rather than BF, method is used as follows:

Table 8. Decomposition of movement in ultimate for the CL method.

Movement in ultimate due to:	**Formulation** $(U_{k+1} - U_k)$
Loss experience	$= \left((C_{k+1} - C_k) - {C_k}/{q_k} (q_{k+1} - q_k) \right) \times \left({1}/{q_{k+1}} \right)$
Change in development pattern	$+ \left({C_{k+1}}/{\hat{q}_{k+1}} - {C_{k+1}}/{q_{k+1}} \right)$

In some regards, while this analysis is simpler as there is only one assumption to consider (the development pattern), it is important to note that the AvE statistic is expressed slightly differently than in the previous section. In contrast to the BF method, deviations between actual and expected loss experience under the CL method do not correspond one-to-one to movements in ultimate; rather they are leveraged by the expected percentage developed at the future period. This makes sense because CL estimates of future losses depend on historical loss experience, whereas BF estimates of future losses are invariant to historical loss experience. The table below outlines these differences.

Table 9. AvE Statistic vs. Movement in Ultimate due to AvE Statistic.

Method	**AvE statistic**	**Movement in ultimate due to loss experience**
BF method	$(C_{k+1} - C_k) - uP(q_{k+1} - q_k)$	$(C_{k+1} - C_k) - uP(q_{k+1} - q_k)$
CL method	$(C_{k+1} - C_k) - {C_k}/{q_k} (q_{k+1} - q_k)$	$\left[(C_{k+1} - C_k) - {C_k}/{q_k} (q_{k+1} - q_k) \right] \times \left[{1}/{q_{k+1}} \right]$

3.3 Extensions

There are a number of extensions to the above analysis, some of which are considered below.

3.3.1 Movement in ultimate due to change in reinsurance recovery rate

While the above analysis could equally apply to gross or net projections, a common approach to netting down gross projections is to assume a recovery rate on the reserves (i.e., the percentage of gross reserves that might be recovered from reinsurers). Using r and $\hat{r}$ to refer to the current and proposed recovery rate with C_k referring to net of reinsurance losses (but P, u and q all still gross of reinsurance), the movement analysis when using the BF method is as follows:

Table 10. Decomposition of movement in ultimate for the BF method.

Movement in ultimate due to:	**Formulation** $(U_{k+1} - U_k)$
Loss experience	$= (C_{k+1} - C_k) - uP(q_{k+1} - q_k)(1-r)$
Change in premium	$+ (\hat{P} - P)u(1-q_{k+1})(1-r)$
Change in development pattern	$+ u\hat{P}[(1-\hat{q}_{k+1}) - (1-q_{k+1})](1-r)$
Change in IELR	$+ (\hat{u} - u)\hat{P}(1-\hat{q}_{k+1})(1-r)$
Change in recovery rate	$+ \hat{u}\hat{P}(1-\hat{q}_{k+1})[(1-\hat{r}) - (1-r)]$

Note that the first four formulas in the above are very similar to those shown in Table 7, but multiplied by $1-r$ and with gross losses replaced by net losses. The movement in ultimate due to change in recovery rate is then just the gross reserve multiplied by the change in recovery rate.

3.3.2 Movement in ultimate due to change in method

Consider the situation of switching between the BF and CL methods, perhaps because losses are believed to be sufficiently developed so that historical loss experience, rather than initial expectations, is more predictive of future emergence. Again, the question of in which order to consider these changes arises. In this situation, as different projection methods utilize different sets of data and assumptions, it makes sense to consider the change in method after any changes due to loss experience, but before changes in premium or assumptions. And when switching from the BF to CL method, this seems logical as the CL method uses neither premiums nor IELRs and thus these items are irrelevant to the change in ultimate.

With that in mind, the change in ultimate is decomposed as:

Table 11. Decomposition of movement in ultimate including change in method (BF to CL).

Movement in ultimate due to:	**Formulation** $\left(U_{k+1}-U_k\right)$
Loss experience	$=\left(C_{k+1}-C_k\right)-uP\left(q_{k+1}-q_k\right)$
Change in method	$+C_{k+1}\left(1/q_{k+1}-1\right)-uP\left(1-q_{k+1}\right)$
Change in development pattern	$+C_{k+1}\left[\left(1/\hat{q}_{k+1}-1\right)-\left(1/q_{k+1}-1\right)\right]$

If moving from the CL to BF method, the movement in ultimate due to loss experience in the above would be set equal to the leveraged AvE statistic described in the previous section, the order of terms in the "change in method" would be reversed and the movement in ultimate due to change in premium, development pattern or IELR would all revert to those shown in Table 7. This is shown below:

Table 12. Decomposition of movement in ultimate including change in method (CL to BF).

Movement in ultimate due to:	**Formulation** $\left(U_{k+1}-U_k\right)$
Loss experience	$=\left(\left(C_{k+1}-C_k\right)-C_k/q_k\left(q_{k+1}-q_k\right)\right)\times\left(1/q_{k+1}\right)$
Change in method	$+uP\left(1-q_{k+1}\right)-C_{k+1}\left(1/q_{k+1}-1\right)$
Change in premium	$+\left(\hat{P}-P\right)u\left(1-q_{k+1}\right)$
Change in development pattern	$+u\hat{P}\left[\left(1-\hat{q}_{k+1}\right)-\left(1-q_{k+1}\right)\right]$
Change in IELR	$+\left(\hat{u}-u\right)\hat{P}\left(1-\hat{q}_{k+1}\right)$

Tables 11 and 12 above provide complete decompositions of the movement in ultimate into each of the key drivers of change. Note that these tables are combined into a complete analysis as presented in the Executive Summary.

3.3.3 Other

There are a number of other common scenarios for which the above can easily be extended including changes in data (i.e., relying on paid vs. incurred data), adjustments to the data, changes in currency, weighting between projection methods and so forth. That said, in practice the results might never be this clean. There could be other adjustments or idiosyncrasies involved (i.e., actuarial judgment) in the selection of ultimate loss that do not easily fall into one or another bucket and thus would be captured in a remaining catch-all residual which should ideally be minimal and explainable.

4. PRACTICAL EXAMPLE

To illustrate the application of this analysis, consider the following example. Tables A and B show exhibits illustrating the projection of ultimate loss at two subsequent year-ends. Here, Items (2), (3) and (5) are assumptions with Items (1) and (4) assumed to come from the data. The estimate of ultimate is then computed as (4) / (3) for the CL method or (4) + (1) x (2) x [1 – (3)] for the BF method. The ultimate loss ratio (ULR) is also shown.

Table A. Estimate of ultimate as at 31 December 2014.

					Selected		
Year	Premium	IELR	Pattern	Loss	Method	Ultimate	ULR
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2011			100%				
2012	10,000	65%	95%	5,916	CL	6,227	62%
2013	10,000	65%	85%	5,108	BF	6,083	61%
2014	10,000	65%	35%	3,337	BF	7,562	76%
Total	**30,000**					**19,872**	**66%**

Table B. Estimate of ultimate as at 31 December 2015.

					Selected		
Year	Premium	IELR	Pattern	Loss	Method	Ultimate	ULR
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
2012	10,000	60%	99%	6,098	CL	6,160	62%
2013	10,000	65%	98%	6,321	CL	6,450	65%
2014	9,000	70%	80%	4,961	BF	6,221	69%
Total	**29,000**					**18,831**	**65%**

Table C then computes the movement analysis by applying the relevant formulas from Table 2. For example, the movement in ultimate due to loss experience for 2014 is solved as:

$$\begin{aligned} &= (C_{k+1} - C_k) - uP(q_{k+1} - q_k) \\ &= (4{,}961 - 3{,}337) - 65\% \times 10{,}000 \times (85\% - 35\%) \\ &= -1{,}626 \end{aligned}$$

While the remaining implementation can be found in the attached Excel workbook, note that there is no residual as the analysis described above fully decomposes the change in ultimate into each of the key drivers.

Table C. Movement Analysis.

	Change in Ultimate			Movement in ultimate due to change in:					
Year	Prior	Current	Change	Experience	Method	Premium	Pattern	IELR	Residual
2012	6,227	6,160	(68)	(129)	0	0	62	0	0
2013	6,083	6,450	367	563	8	0	(204)	0	0
2014	7,562	6,221	(1,341)	(1,626)	0	(98)	293	90	0
Total	**19,872**	**18,831**	**(1,042)**	**(1,192)**	**8**	**(98)**	**150**	**90**	**0**

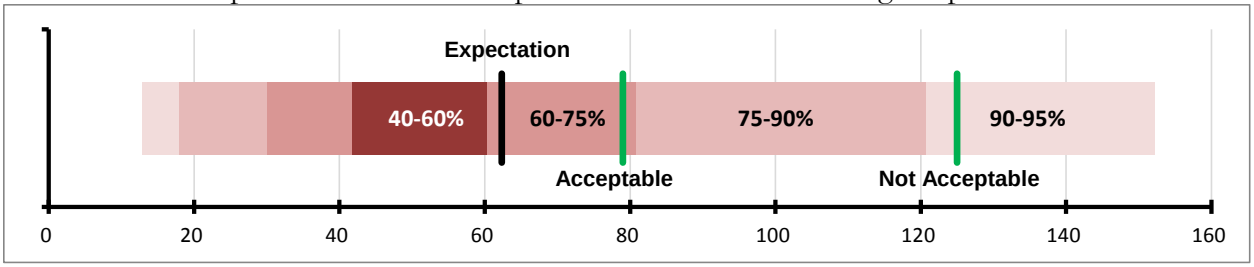
5. CONCLUSION

As actuaries become increasingly influential, there is an additional responsibility to move from opaqueness to transparency. By clearly communicating and quantifying the impacts of certain decisions, we can ensure that management has the appropriate information to assess, challenge or validate our recommendations and make their own determination as to the final carried amount.

The above presented several simple formulas for doing this on a deterministic and retrospective basis in a number of situations that commonly arise in actuarial practice.

That said, there are two useful and practical extensions of the above that are worth highlighting. The first involves moving toward reserve reports that not only isolate the key drivers of change between prior estimates, but also provide prospective estimates as to how losses are expected to emerge in future periods. This should enhance management information as emergence can then be monitored on a regular basis (rather than waiting until the next formal reserve review) and deviations from expectations can be flagged and explored in more detail. In regard to the latter, the other useful extension is to report not just expected emergence, but also to provide a range around that expectation so a determination can be made as to whether or not divergences from expectations are statistically significant.

As an example, consider the below figure which illustrates what this analysis might look like. The black line is the expected loss in the next period with the bars indicating the percentile distribution.



In this instance, emergence of 79 might be acceptable as it falls within the 75th percentile, but emergence around 125 might not be acceptable as it falls above the 90th percentile. In the former instance, the actuary might leave the key assumptions unchanged, but in the latter instance the actuary may wish to modify one or more assumptions as the deviation in claims experience relative to expectation appears to be statistically significant. This is more or less akin to hypothesis testing.

This should be especially doable in Europe as the formal implementation of Solvency II draws near where insurance risk is measured on a one-year basis and thus emergence profiles of loss as well as the distribution around those estimates should be readily available.

Acknowledgment

I would like to thank Xi Wu, Kelly Moore and Ziyi Jiao for their thorough review. This paper is much stronger because of their helpful suggestions and commentary.

Biography of the Author

Andy Staudt is a consultant with Towers Watson in London. He is a Fellow of the Casualty Actuarial Society (FCAS), a member of the American Academy of Actuaries (MAAA) and has a Masters in Statistics from the University of California at Berkeley. He is also a Fellow in the Institute and Faculty of Actuaries (FIA) through mutual recognition and holds a Lloyd's Practising Certificate. He can be reached at: andy.staudt@towerswatson.com or +44 (0) 207 170 3476.

A. TECHNICAL APPENDIX

The following derives some of the key formulas expressed in this paper with the remainder fairly straightforward but tedious to derive and thus omitted here for presentation purposes.

Table (3): $U_{k+1} - U_k = \left[C_{k+1} + uP(1-q_{k+1})\right] - \left[C_k + uP(1-q_k)\right]$

$$
\begin{aligned}
&= \left[C_{k+1} + uP(1-q_{k+1})\right] - \left[C_k + uP(1-q_k)\right] \\
&= C_{k+1} + uP - uPq_{k+1} - C_k - uP + uPq_k \\
&= (C_{k+1} - C_k) - uP(q_{k+1} - q_k)
\end{aligned}
$$

Table (4): $U_{k+1} - U_k = \left[C_{k+1} + uP(1-\hat{q}_{k+1})\right] - \left[C_k + uP(1-q_k)\right]$

$$
\begin{aligned}
&= \left[C_{k+1} + uP(1-\hat{q}_{k+1})\right] - \left[C_k + uP(1-q_k)\right] \\
&= C_{k+1} + uP - uP\hat{q}_{k+1} - C_k - uP + uPq_k \\
&= (C_{k+1} - C_k) - uP\hat{q}_{k+1} + uPq_k + \left[uPq_{k+1} - uPq_{k+1}\right] \\
&= \left[(C_{k+1} - C_k) - uP(q_{k+1} - q_k)\right] + uP(q_{k+1} - \hat{q}_{k+1}) \\
&= \left[(C_{k+1} - C_k) - uP(q_{k+1} - q_k)\right] + uP\left[(1-\hat{q}_{k+1}) - (1-q_{k+1})\right]
\end{aligned}
$$

Table (5): $U_{k+1} - U_k = \left[C_{k+1} + \hat{u}P(1-\hat{q}_{k+1})\right] - \left[C_k + uP(1-q_k)\right]$

$$
\begin{aligned}
&= \left[C_{k+1} + \hat{u}P(1-\hat{q}_{k+1})\right] - \left[C_k + uP(1-q_k)\right] \\
&= C_{k+1} + \hat{u}P - \hat{u}P\hat{q}_{k+1} - C_k - uP + uPq_k \\
&= (C_{k+1} - C_k) + \hat{u}P - \hat{u}P\hat{q}_{k+1} - uP + uPq_k + \left[uPq_{k+1} - uPq_{k+1}\right] + \left[uP\hat{q}_{k+1} - uP\hat{q}_{k+1}\right] \\
&= \left[(C_{k+1} - C_k) - uP(q_{k+1} - q_k)\right] + uP(q_{k+1} - \hat{q}_{k+1}) + (\hat{u}P - uP) - (\hat{u}P\hat{q}_{k+1} - uP\hat{q}_{k+1}) \\
&= \left[(C_{k+1} - C_k) - uP(q_{k+1} - q_k)\right] + uP\left[(1-\hat{q}_{k+1}) - (1-q_{k+1})\right] + (\hat{u} - u)P(1-\hat{q}_{k+1})
\end{aligned}
$$

Table (6): $U_{k+1} - U_k = \left[C_{k+1} + \hat{u}P(1-\hat{q}_{k+1})\right] - \left[C_k + uP(1-q_k)\right]$ (alternative formulation of Table 5)

$$
\begin{aligned}
&= \left[C_{k+1} + \hat{u}P(1-\hat{q}_{k+1})\right] - \left[C_k + uP(1-q_k)\right] \\
&= C_{k+1} + \hat{u}P - \hat{u}P\hat{q}_{k+1} - C_k - uP + uPq_k \\
&= (C_{k+1} - C_k) + \hat{u}P - \hat{u}P\hat{q}_{k+1} - uP + uPq_k + \left[uPq_{k+1} - uPq_{k+1}\right] + \left[\hat{u}Pq_{k+1} - \hat{u}Pq_{k+1}\right] \\
&= \left[(C_{k+1} - C_k) - uP(q_{k+1} - q_k)\right] + (\hat{u}P - uP) - (\hat{u}Pq_{k+1} - uPq_{k+1}) + \hat{u}P(q_{k+1} - \hat{q}_{k+1}) \\
&= \left[(C_{k+1} - C_k) - uP(q_{k+1} - q_k)\right] + (\hat{u} - u)P(1-q_{k+1}) + \hat{u}P\left[(1-\hat{q}_{k+1}) - (1-q_{k+1})\right]
\end{aligned}
$$

Table (7): $U_{k+1} - U_k = \left[C_{k+1} + \hat{u}\hat{P}\left(1-\hat{q}_{k+1}\right)\right] - \left[C_k + uP\left(1-q_k\right)\right]$

$$
\begin{aligned}
&= \left[C_{k+1} + \hat{u}\hat{P}\left(1-\hat{q}_{k+1}\right)\right] - \left[C_k + uP\left(1-q_k\right)\right] \\
&= C_{k+1} + \hat{u}\hat{P} - \hat{u}\hat{P}\hat{q}_{k+1} - C_k - uP + uPq_k \\
&= \left(C_{k+1} - C_k\right) + \hat{u}\hat{P} - \hat{u}\hat{P}\hat{q}_{k+1} - uP + uPq_k + \left[u\hat{P}q_{k+1} - u\hat{P}q_{k+1}\right] + \left[u\hat{P}\hat{q}_{k+1} - u\hat{P}\hat{q}_{k+1}\right] + \left[u\hat{P} - u\hat{P}\right] + \left[uPq_{k+1} - uPq_{k+1}\right] \\
&= \left[\left(C_{k+1} - C_k\right) - uP\left(q_{k+1} - q_k\right)\right] + \left(u\hat{P} - uP\right) - \left(u\hat{P}q_{k+1} - uPq_{k+1}\right) + u\hat{P}\left(q_{k+1} - \hat{q}_{k+1}\right) + \left(\hat{u}\hat{P} - u\hat{P}\right) - \left(\hat{u}\hat{P}\hat{q}_{k+1} - u\hat{P}\hat{q}_{k+1}\right) \\
&= \left[\left(C_{k+1} - C_k\right) - uP\left(q_{k+1} - q_k\right)\right] + \left(\hat{P} - P\right)u\left(1-q_{k+1}\right) + u\hat{P}\left[\left(1-\hat{q}_{k+1}\right) - \left(1-q_{k+1}\right)\right] + \left(\hat{u} - u\right)\hat{P}\left(1-\hat{q}_{k+1}\right)
\end{aligned}
$$

Table (8): $U_{k+1} - U_k = C_{k+1}/\hat{q}_{k+1} - C_{k+1}/q_k$

$$
\begin{aligned}
&= C_{k+1}/\hat{q}_{k+1} - C_k/q_k \\
&= C_{k+1}/\hat{q}_{k+1} - C_k/q_k + \left[C_k/q_{k+1} - C_k/q_{k+1}\right] + \left[C_{k+1}/q_{k+1} - C_{k+1}/q_{k+1}\right] \\
&= \left[C_{k+1}/q_{k+1} - C_k/q_k + C_k/q_{k+1} - C_k/q_{k+1}\right] + \left(C_{k+1}/\hat{q}_{k+1} - C_{k+1}/q_{k+1}\right) \\
&= \left[C_{k+1} - C_k\, q_{k+1}/q_k + C_k - C_k\right] \times \left(1/q_{k+1}\right) + \left(C_{k+1}/\hat{q}_{k+1} - C_{k+1}/q_{k+1}\right) \\
&= \left[\left(C_{k+1} - C_k\right) - C_k/q_k\left(q_{k+1} - q_k\right)\right] \times \left(1/q_{k+1}\right) + \left(C_{k+1}/\hat{q}_{k+1} - C_{k+1}/q_{k+1}\right)
\end{aligned}
$$

Table (10): $U_{k+1} - U_k = \left[C_{k+1} + \hat{u}\hat{P}\left(1-\hat{q}_{k+1}\right)\left(1-\hat{r}\right)\right] - \left[C_k + uP\left(1-q_k\right)\left(1-r\right)\right]$

Omitted.

Table (11): $U_{k+1} - U_k = \left[C_{k+1}/\hat{q}_{k+1}\right] - \left[C_k + uP\left(1-q_k\right)\right]$

$$
\begin{aligned}
&= \left[C_{k+1}/\hat{q}_{k+1}\right] - \left[C_k + uP\left(1-q_k\right)\right] \\
&= C_{k+1}/\hat{q}_{k+1} - C_k - uP + uPq_k + \left[C_{k+1} - C_{k+1}\right] + \left[uPq_{k+1} - uPq_{k+1}\right] + \left[C_{k+1}/q_{k+1} - C_{k+1}/q_{k+1}\right] \\
&= \left[\left(C_{k+1} - C_k\right) - uP\left(q_{k+1} - q_k\right)\right] + \left(C_{k+1}/q_{k+1} - C_{k+1}\right) - uP + uPq_{k+1} + C_{k+1}\left(1/\hat{q}_{k+1} - 1/q_{k+1}\right) \\
&= \left[\left(C_{k+1} - C_k\right) - uP\left(q_{k+1} - q_k\right)\right] + \left[C_{k+1}\left(1/q_{k+1} - 1\right) - uP\left(1-q_{k+1}\right)\right] + C_{k+1}\left[\left(1/\hat{q}_{k+1} - 1\right) - \left(1/q_{k+1} - 1\right)\right]
\end{aligned}
$$

Table (12): $U_{k+1} - U_k = \left[C_{k+1} + \hat{u}\hat{P}\left(1-\hat{q}_{k+1}\right)\right] - C_k/q_k$

Omitted.

The Market Value Margin Within The Distribution-Free Chain Ladder Model - A Way To Account For Calendar Year Effects And Aggregating Lines Of Business

Daniel Burren, PhD, MSc, Actuary SAA

Abstract

Under European and Swiss solvency directives, general insurance companies have to calculate a market value margin (aka risk margin or MVM) for the prediction uncertainty of reserves over each accounting year and until the end of the runoff. The prediction uncertainty is generally split into a process error and an estimation error. In the distribution-free chain ladder framework, [10] derived analytical formulas for the prediction uncertainty over accounting years and showed that they add up to the total runoff uncertainty as given by the Mack error. We suggest a way to modify their methodology in order to account for calendar year uncertainties like a legal reform. Further, we derive the minimum and the maximum market value margin that can result with our modification, which is useful to quantify model uncertainty. Besides, we highlight the simplifications and omissions of the presented ways to infer the MVM. Finally, we discuss aggregating different lines of business. The presented formulas can be calculated in a spreadsheet.

Keywords. market value margin, distribution-free chain ladder model, reserving risk, calendar year effects, SST, Solvency II

1. INTRODUCTION

In Europe and Switzerland, insurance companies are regulated by the Solvency II directive (scheduled to be in full effect on 1 January 2016) and the Swiss Solvency Test (SST, in use since 2006). A comparison of Solvency II with the SST can be found in [4]. These two regulatory frameworks ask insurance companies to back their liabilities based on a one-year distribution of assets and liabilities. In addition, companies have to calculate the market consistent value of technical provisions which is defined as best estimate reserves (defined as the expected present value of future cash flows) plus the market value margin (MVM).

The MVM (aka risk margin) of the general insurance runoff (also called the reserve risk) is the focus of this paper. In this context, the MVM is a margin for the prediction uncertainty of the ultimate claim liabilities. Predictions are usually updated annually when new information is incorporated. These updates have an effect on the result of the insurance company and therefore need to be taken account of in solvency considerations. The prediction uncertainty is generally split into a process error and an estimation error. The process error represents random variations not explained by the model of the reserving actuary. The estimation error represents updates in the estimates of the model's parameters. In [3] and [5] the MVM is defined as the cost of the present value of future solvency capital requirements which will have to be put up during the runoff of the

portfolio of assets and liabilities for the in-force book of business one year in the future.

A mathematically consistent calculation of the MVM is a complicated task which usually requires the application of numerical methods. Analytical approximations have been proposed, many of which rely on Bayesian statistics. [11] and [15], for example, describe how to infer the MVM within a Bayesian log-normal model. [13] derived, within a Gamma-Gamma Bayes chain ladder model, three approximations for the MVM whereof two can be computed analytically. The one-year view in the context of a Bayes chain ladder model was discussed by [1].

Bayesian models have the advantage that they include, in a natural way, the estimation error (also referred to as parameter uncertainty). Further, they can be similar to a classical chain ladder in the sense that the expected ultimate claim is given by a product formula involving factors and the latest cumulative payment (or incurred liability). However, Bayesian methods require selecting and calibrating prior distributions and justification of these selections is sometimes difficult. This might be a reason why the distribution-free chain ladder model, discussed in [6], still is one of the most popular reserving methods. Based on the distribution-free chain ladder model, [10] and [12] independently derived formulas which can be used to calculate the MVM. These formulas generalize the one-year solvency view presented in [9].

We take the methodology of [10] as a starting point and discuss a modification in order to account for calendar year effects like a legal reform or inflation. We propose a straight forward correction for the process error. Regarding the estimation error, the reserving actuary might have an idea when relevant information about parameters will become available (maybe the timing of the legal reform is known) and therefore can judge in which years the estimation error will be high. We show how to incorporate this judgment. Further, we provide a result useful to quantify the error of the actuary's judgment, that is to say we derive the minimum and maximum MVM which can possibly result based on different considerations of the estimation error. Finally, we discuss aggregating different lines of business.

The remainder of the paper is organized as follows. Section 2 reviews the classical chain ladder assumption, introduces the MVM and contains a literature review. Section 3 discusses our approach to accounting year effects, derives the minimum and maximum MVM and provides a numerical example. Section 4 treats the issue of aggregating different lines of business. Section 5 concludes.

2. BACKGROUND AND METHODS

2.1 Definitions and Assumptions

This section introduces the notation, revises the classical chain ladder (CL), aka distribution-free chain ladder model, and introduces our assumptions.

We write $C_{i,j}$ for the cumulative payments (or incurred liabilities) for accident years $i \in \{0, \dots, I\}$ and development years $j \in \{0, \dots, I\}$ and suppose that there is a $J \leq I$ such that $C_{i,J}=C_{i,J+1}=\dots=C_{i,I}$ for all i. If we refer to triangle we mean the following set

$$F_I = \{C_{i,j} : 0 \leq i \leq I, 0 \leq j \leq I, i + j \leq I\}$$

ordered as in Table 1 of Section 3.1. We denote the accounting years by $k \in \{0, \dots, I\}$ meaning that I refers to today and $I+k$ to the year k years in the future. We use 'accounting year' and 'calendar year' as synonyms. We define

$$F_{I+k} = \{C_{i,j} : 0 \leq i \leq I, 0 \leq j \leq I, i + j \leq I + k\}.$$

We assume stochastic independence between cumulative claims $C_{i,j}$ of different accident years i and that there exist constants $f_j>0$ and $\sigma_j>0$ and random variables $\varepsilon_{i,j}$ such that

$$C_{i,j} = f_j \, C_{i,j-1} + \sigma_{j-1}\sqrt{C_{i,j-1}} \, \varepsilon_{i,j} \tag{2.1}$$

where $\varepsilon_{i,j}$ are conditionally, given $S_0 = \{C_{i,0} : 0 \leq i \leq I\}$, independent with expectation $E[\varepsilon_{i,j} | S_0]=0$, $E[\varepsilon_{i,j}^2 | S_0]=1$ and distribution guaranteeing $C_{i,j}>0$ with probability one. These assumptions imply the assumptions of the distribution-free chain ladder model, see e.g. [16].

We write $\hat{f}_j^{I+k}$ and $\hat{\sigma}_j^{I+k}$ for the estimators of f_j and σ_j given all information up to accounting year $I+k$ and define, for $k \leq j+1$,

$$\hat{f}_j^{I+k} = \frac{\sum_{i=0}^{I+k-j-1} C_{i,j+1}}{\sum_{i=0}^{I+k-j-1} C_{i,j}} \tag{2.2}$$

and $\hat{f}_j^{I+k} = \hat{f}_j^{I+j+1}$ for $k>j+1$ (since we focus on the runoff of past accident years only, the estimators remain unchanged for $k>j+1$). We observe that for $k=0$ the classical chain ladder factors are obtained. We obtain the $\hat{\sigma}_j^{I+k}$s as suggested in [6]. The estimated chain ladder ultimate claim is

$$\hat{C}_{i,J}^I = C_{i,I-i} \prod_{j=I-i}^{J} \hat{f}_j^I \tag{2.3}$$

and we abbreviate

$$C_J = \sum_{i=0}^{I} C_{i,J} \text{ and } \hat{C}_J^I = \sum_{i=0}^{I} \hat{C}_{i,J}^I.$$

If $C_{i,j}$ are cumulative payments then the liabilities estimated today to remain outstanding in accounting year $I+k$ are, for $k=0,\ldots,J-1$,

$$\hat{R}_k^I = \sum_{i=I-J+1+k}^{I} \left(\hat{C}_{i,J}^I - \hat{C}_{i,I-i+k}^I\right) = \hat{C}_J^I - \sum_{i=0}^{I} \hat{C}_{i,\min(I-i+k,J)}^I. \tag{2.4}$$

Accordingly, $\hat{R}_0^I$ are the chain ladder reserves estimated in the current accounting year (which equals accident year I).

Finally, the claims development result (*CDR*) of accident year i in accounting year $k \in \{0, \ldots, J\}$ is

$$CDR_{i,k+1} = E\left[C_{i,J}\middle|F_{I+k}\right] - E[C_{i,J}|F_{I+k+1}]. \tag{2.5}$$

with the sigma-algebras defined before. We have $CDR_{i,k+1}=0$ for $i \leq I+k-J$. We write

$$CDR_k = \sum_{i=I-J+k}^{I} CDR_{i,k}. \tag{2.6}$$

The claims development result reflects how the valuation of the ultimate claim changes over a one year period. These changes are due to prediction updates as new information is incorporated. The prediction uncertainty is caused by two risk factors:

1. $\varepsilon_{i,j}$ in (2.1), referred to as the **process error**
2. updates of the chain ladder factors $\hat{f}_j^{I+k}$ in (2.2), referred to as the **estimation error.**

The MVM is the cost of the present value of future capital required to back adverse movements of the CDR caused by these two risk factors - we introduce its formal definition in the next chapter.

2.1.1 A Remark About Implicit Assumptions

We highlight that the CDR as defined in (2.5) does not consider discounting of liabilities. In the standard models of Solvency II and the SST adverse changes in discount factors (adverse meaning that they lead to higher best estimate reserves) are captured by the market risk (which is not the topic of this paper) and, in our understanding, they should also be taken into account in the MVM. Ignoring discounting further implies that the timing of the claims payments has no influence on the best estimate of discounted ultimate liability. However, the CDR as defined above is analogous to how it is defined in existing literature. Indeed, all papers we cite abstract from discounting. We leave

it to future research to introduce the missing risk factors like stochastic discount factors, uncertainty in claim payments and potential mismatches in asset-liability cash-flows.

2.2 The MVM of the Runoff

We next introduce the MVM formally. For this purpose, we first define the following quantities:

Definition 1

- deterministic (F_I-measurable) discount factors $D_{I,k}$, $k=0,\dots,J$ giving the value, in accounting year I, of a unit of money received in year $I+k$
- a risk measure $\rho()$ which quantifies the amount of capital needed to back adverse movements in the CDR (2.5)
- the cost c of capital (6% in Solvency II and the SST)

Assuming, as do [10], that we do not need to put up capital for adverse movements in the MVM itself, the MVM in accounting year I is

$$MVM = c\sum_{k=1}^{J} D_{I,k}\rho(CDR_k). \tag{2.7}$$

We repeat our remark in Section 2.1.1 namely that the MVM as just defined is based on variations of the *nominal* best estimate ultimate liability and therefore neither fluctuations in future discount rates nor the timing of the claims payments play a role. The MVM can be thought of as the present value of dividends required to compensate an investor for providing the risk capital to back the runoff risk.

If we knew the true chain ladder factors f_j then the estimation error would be zero. In this case we would only have to take care of the process error and we could easily calculate a variance (or standard deviation) risk measure for (2.7). We would obtain, for $1\le k\le J$,

$$Var(CDR_k|F_I) = \sum_{i=I-J+k}^{I} Var\big(CDR_{i,k}|F_I\big) = \sum_{i=I-J+k}^{I} (E[C_{i,J}|F_I])^2 \frac{\sigma^2_{I+k-i-1}/f^2_{I+k-i-1}}{E[C_{i,I+k-i-1}|F_I]}$$

which can be estimated by

$$\widehat{Var}(CDR_k|F_I) = \sum_{i=I-J+k}^{I} \big(\hat{C}^I_{i,J}\big)^2 \frac{(\hat{\sigma}^I_{I+k-i-1})^2/\big(\hat{f}^I_{I+k-i-1}\big)^2}{\hat{C}^I_{i,I+k-i-1}}. \tag{2.8}$$

For $k=1$ this corresponds to the estimator of the process error in [8] and [9]. Taking the sum over all k gives the process variance of the total runoff which is one term of the Mack error (the

other being the estimation error of the total runoff). Note that it would not matter if instead of $Var(CDR_k|F_I)$ we used $E[Var(CDR_k|F_{I+k-1})|\ F_I]$ for the risk measure, as we show in the next lemma.

Lemma 1. Suppose the true chain ladder factors are known. The classical CL assumptions imply

$$Var(CDR_k|F_I) = E[Var(CDR_k|F_{I+k-1})|F_I], \qquad k = 1, \dots, J.$$

Proof. By independence of the accident years it is sufficient to prove the equality for $CDR_{i,k}$. The total variance formula gives us

$$\begin{aligned} Var\big(CDR_{i,k}\big|F_I\big) &= E\big[Var\big(CDR_{i,k}\big|F_{I+k-1}\big)|F_I\big] + Var\big(E\big[CDR_{i,k}|F_{I+k-1}\big]|F_I\big) \\ &= E\big[Var\big(CDR_{i,k}\big|F_{I+k-1}\big)|F_I\big] \end{aligned}$$

where the second equality follows because definition (2.5) implies

$$E\big[CDR_{i,k}|F_{I+k-1}\big] = 0, \qquad 1 \le k \le J.$$

Q.E.D.

Unfortunately, the true chain ladder factors are unknown and the estimation error needs to be taken into account. We next review how this has been done in existing literature.

2.3 A Brief Literature Review

Instead of the notation of the original papers we use the notation introduced earlier, in particular c and $D_{I,k}$ as given in Definition 1. Indeed, the discount factors $D_{I,k}$ are omitted in the cited literature and we introduced them to be consistent with (2.7) which defines the MVM as the present value of future dividends.

As an example of a paper using Bayesian methods (which allows a natural treatment of the estimation error) and since it introduces notation, we cite [13]. They employed the standard deviation as a risk measure and discussed the following three ways to estimate the MVM.

A. Regulatory Solvency Proxy

$$MVM = c \sum_{k=1}^{J} D_{I,k} \frac{\hat{R}_k^I}{\hat{R}_0^I}\ \phi\ Stdev(CDR_1|F_I)$$

where $\hat{R}_k^I$ are the reserves (estimated to remain in accounting year k) as obtained with the Bayesian methodology of [13], $\phi > 0$ is a loading and the CDR_1 is as defined in (2.6).

B. Split of Total Uncertainty Approach

$$MVM = c \sum_{k=1}^{J} D_{I,k} \, \phi \, Stdev(CDR_k | F_I).$$

The name comes from the property that the total uncertainty about the ultimate $C_{i,J}$ can be split into single one-year uncertainties for different accounting years as follows

$$Var(C_{i,J}) = \sum_{k=1}^{J+i-I-1} Var\left(CDR_{i,k} | F_I\right).$$

C. Expected Stand Alone Measure

$$MVM = c \sum_{k=1}^{J} D_{I,k} \, \phi \, E[Stdev(CDR_k | F_{I+k-1}) | F_I].$$

[13] derived analytic formulas for A and B and relied on simulations to solve C. They discussed a fourth approach to calculate the MVM for which instead of $\rho(CDR_K)$ they considered $\rho(CDR_K + MVM_k - MVM_{k-1})$ with MVM_k being the MVM calculated in accounting year $I+k$. This means that a markup for the MVM is included. While this approach is certainly more realistic - dividends can be thought of as a liability too - the computation becomes complicated and they had to rely on simulations. Fortunately, a numerical example in their paper supports B to be a good approximation for their fourth approach. As a side note, we remark that Lemma 1 and Jensen's inequality imply that for the distribution-free chain ladder model with known parameters approach B would yield a larger MVM than approach C.

[10] derived the prediction uncertainties for the CDR in the distribution-free chain ladder model. They computed the mean square errors of prediction (MSEP) for the CDRs as defined by

$$MSEP_{CDR_k | F_{I+k-1}}(0) = E[(CDR_k - 0)^2 | F_k], \qquad k = 1, \dots, J$$

with CDR_k as in (2.6), and derived an estimator for the expected value at time I of the MSEP. They then defined the MVM for a variance risk measure given by

$$MVM = c \sum_{k=1}^{J} D_{I,k} \, \phi \, E\left[MSEP_{CDR_k | F_{I+k-1}}(0) | F_I\right] \tag{2.9}$$

and the MVM for a standard deviation motivated risk measure given by

$$MVM = c \sum_{k=1}^{J} D_{I,k} \, \phi \sqrt{E\left[MSEP_{CDR_k | F_{I+k-1}}(0) | F_I\right]} \tag{2.10}$$

where $\phi > 0$ is a loading. This is a generalization of [9] who suggested to use $MSEP_{CDR_1|F_I}(0)$ for a one-year view of solvency considerations. [10] further showed that

$$MSEP_{C_{i,J}|F_I}\left(\hat{C}_{i,J}^I\right) = \sum_{k=1}^{J+i-I} \phi \, E\left[MSEP_{CDR_{i,k}|F_{I+k-1}}(0)|F_I\right] \tag{2.11}$$

with $\hat{C}_{i,J}^I$ defined in (2.3). The left-hand side is the Mack error as introduced in [6]. Hence, the total runoff uncertainty as given by the Mack error splits across accounting years and so their approach is similar to B of [13] stated earlier. Both, the Mack error and $E[MSEP_{CDR_k|F_{I+k-1}}(0)|F_I]$ can be written as a sum of two terms corresponding to the process error variance and the estimation error. Not surprisingly, the process variance in $E[MSEP_{CDR_k|F_{I+k-1}}(0)|F_I]$ equals (2.8).

3. CONSIDERING ACCOUNTING YEAR UNCERTAINTIES

The distribution-free chain ladder is probably the most popular reserving method. There is therefore a good chance that the formulas of [10] for the MVM will become popular, too. Moreover, these formulas can be computed in a spreadsheet, simulations are not required, and they are even implemented in a new package for the statistical software R, see [2]. There are however situations where a modified approach to the MVM is preferable. Suppose, for example, that we are at the dawn of a legal reform which will affect the chain ladder factors. Regarding the process error, a method to filter out accounting year effects (as, for example, described in [14] or chapter 3 of [7]) could be employed and accordingly modified development factors f_j and σ_js (potentially depending on accident years) could be used in (2.8). A correction of this kind would not be enough for the estimation error for the following reason. For any accident year i, the squared estimation error associated with accounting year k is proportional to

$$\frac{1}{\sum_{l=0}^{i-1} C_{l,I-i+k}}$$

(see (1.4) in [10]) which means that claims of all prior accident years reduce the estimation error. We doubt whether this is meaningful when dealing with legal reforms or other uncertain accounting year effects. The following algorithm provides an alternative way.

Algorithm.

1. Compute an error for the entire runoff given current information F_I and taking into account the legal reform (a possible solution could involve simulations assuming appropriate distributions on the parameter space). We denote the resulting quantity by

$$\widehat{MSEP}_{\sum_{i=I-J+1}^{I} C_{i,J}|F_I}\left(\sum_{i=I-J+1}^{I}\hat{C}_{i,J}^{I}\right).$$

2. Compute the total squared estimation error (SEE) according to the difference

$$SEE = \widehat{MSEP}_{\sum_{i=I-J+1}^{I} C_{i,J}|F_I}\left(\sum_{i=I-J+1}^{I}\hat{C}_{i,J}^{I}\right) - \sum_{k=1}^{J}\widehat{Var}(CDR_k|F_I)$$

using (2.8) for the process error with the mentioned modification for accounting year effects.

3. Split the total estimation error across accounting years according to

$$SEE_k = \varpi_k SEE, \qquad 1 \le k \le J \tag{3.1}$$

with weights $\varpi_k \ge 0$ and $\sum_{k=1}^{J}\varpi_k = 1$ calibrated in a way to reflect the timing of the legal reform (actuarial judgment may be required).

4. Approximate future 'accounting year' prediction uncertainties by

$$E\left[MSEP_{CDR_k|F_{I+k-1}}(0)|F_I\right] \approx \widehat{MSEP}_k = SEE_k + \widehat{Var}(CDR_k|F_I), 1 \le k \le J.$$

Use these quantities in (2.9). This is the end of the algorithm.

We remark the following.

a) The total uncertainty still splits over accounting years, i.e.

$$\widehat{MSEP}_1 + \cdots + \widehat{MSEP}_J = \widehat{MSEP}_{\sum_{i=I-J+1}^{I} C_{i,J}|F_I}\left(\sum_{i=I-J+1}^{I}\hat{C}_{i,J}^{I}\right)$$

b) If $C_{i,j}$ are cumulative payments then

$$\varpi_k = \frac{(\hat{R}_{k-1}^{I})^2}{\sum_{l=0}^{J-1}(\hat{R}_l^{I})^2} \tag{3.2}$$

with $\hat{R}_k^I$ as defined in (2.4), yields a regulatory solvency proxy similar to how it is defined in approach A of [13] (see our literature review). In this case, all coefficients of variation given by $\sqrt{SEE_k}/\hat{R}_{k-1}^{I}$ are equal.

c) Instead of doing step 3, the estimation error could be calculated directly for each accounting year. However, this might require nested simulations which we expect to be computationally more involved than calculating the error for the entire runoff as suggested in step 1.

There is no reason why the regulatory solvency proxy should describe the estimation uncertainty

due to reforms. Indeed, suitable weights ϖ might be hard to find and even though the total estimation error is unaffected by these weights, the MVM generally depends on them. The next proposition highlights this dependency for risk measures as defined in (2.9) and (2.10).

Proposition 1. Define

$$MVM_m = c\sum_{k=1}^{J} D_{I,k}\rho_m\big(\widehat{MSEP}(\xi_k)\big), \qquad m \in \{1,2\}$$

with

$$\widehat{MSEP}(\xi_k) = \xi_k + \widehat{Var}(CDR_k|F_I),\ \rho_1(x) = \phi x,\ \rho_2(x) = \phi\sqrt{x}$$

for positive numbers ξ_k, a loading $\phi > 0$ and c and $D_{I,k}$ as given in Definition 1 and $\widehat{Var}(CDR_k|F_I)$ describes the process error given in (2.8) with the mentioned modification for accounting year effects.

The solution to the maximization problem

$$\max_{\xi_k\in\Omega,k=1,\dots,J} MVM_m, \tag{3.3}$$

with Ω being the set of positive numbers ξ_k satisfying $\sum_{k=1}^{J}\xi_k = SEE$ is as follows.

- Let *m=1*. Then $\xi_{k^*} = SEE$ where k^* is the index of the largest $D_{I,k}$ (or one of the largest if there is more than one maximum $D_{I,k}$), and $\xi_k = 0$ for all other k solves (3.3).
- Let *m=2*. Define, for $k \in \{1,\dots,J\}$,

$$\xi_k^* = \frac{D_{I,k}^2}{\sum_{j=1}^{J} D_{I,j}^2}\left(SEE + \sum_{j=1}^{J}\widehat{Var}\big(CDR_j\big|F_I\big)\right) - \widehat{Var}(CDR_k|F_I).$$

If $\xi_k^* \geq 0$ for all $k \in \{1,\dots,J\}$ then these ξ_k^*s solve (3.3). If $\exists k$ with $\xi_k^* < 0$ then

$$\xi_k^c = \frac{D_{I,k}^2}{\sum_{j\in \mathrm{P}} D_{I,j}^2}\left(SEE + \sum_{j\in \mathrm{P}}\widehat{Var}(CDR_j|F_I)\right) - \widehat{Var}(CDR_k|F_I), \qquad if\ k \in \mathrm{P}$$

and $\xi_k^c = 0$ if $k \notin \mathrm{P}$, where P is the set of all indices k for which $\xi_k^c > 0$, solves (3.3).

The solution to the minimization problem

$$\min_{\xi_k\in\Omega,k=1,\dots,J} MVM_m, \tag{3.4}$$

with Ω as in (3.3), is as follows.

- Let *m=1*. Then $\xi_{k^*} = SEE$, where k^* is the index of the smallest $D_{I,k}$ and $\xi_k = 0$ for all other k solves (3.4).

- Let *m=2*. Define k^* to be the index of the smallest

$$\frac{D_{I,k}}{\sqrt{\widehat{Var}(CDR_k|F_I)}}$$

Then $\xi_{k^*} = SEE$ and $\xi_k = 0$ for all other k solves (3.4).

Proof. The proof for *m=1* is obvious. Consider *m=2*. Ignoring the positivity constraints $\xi_k \geq 0$, the Lagrangian of the maximization problem is

$$L = \sum_{k=1}^{J} D_{I,k} \sqrt{\xi_k + \widehat{Var}(CDR_k|F_I)} + \lambda \left(SEE - \sum_{k=1}^{J} \xi_k \right)$$

Thanks to a negative definite Hessian, the first order conditions, given by

$$\frac{D_{I,k}}{2\sqrt{\xi_k + Var(\widehat{CDR}_k|F_I)}} = \lambda \; \forall \; k \in \{1, \dots, J\}, \qquad SEE = \sum_{k=1}^{J} \xi_k,$$

are sufficient for a maximum and therefore the ξ_{k^*}s solve (3.3) if they are all positive. If this is not the case, the Kuhn-Tucker conditions provide the maximum. The solution to (3.4) is obvious.
Q.E.D.

We think that the previous proposition is useful, be it for reporting purposes if the regulator asks about the impact of the selected weights in (3.1) or be it for budget-planning to have an idea how much resources should be spent on calculating the MVM. That is to say the actuary can provide to the company management a range within which the MVM obtained with a more accurate method will fall. The next corollary readily follows from the proposition.

Corollary 1. Consider the maximization problem (3.3) for the standard deviation risk measure (meaning *m=2*). If all discount factors $D_{I,k}$ are equal to 1 then the resulting prediction uncertainties of calendar years with a positive estimation error (where $\xi_k^c > 0$) are identical and smaller than the prediction uncertainty of any other calendar year.

We next provide a numerical example before we proceed with the final topic about aggregation.

3.1 Numerical Example

We borrow an example from [10] and compare their prediction uncertainties to what we obtain based on our Proposition 1 abstracting from discounting, that is to say $D_{I,k}=1$ for all accounting

years k. Our intention is to highlight the impact on the MVM of different weights selected in (3.1) and used to split the estimation error across the runoff. We therefore use (2.8) without any modification for accounting year effects which means that the process errors are identical across the different prediction uncertainties.

Table 1 contains the data, the estimated chain ladder factors and the sigmas obtained with the estimator in [6].

Table 1: Cumulative claims payments $C_{i,j}$ and estimated parameters $\hat{f}_j^I$ and $\hat{\sigma}_j^I$.

i\j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	13'109	20'355	21'337	22'043	22'401	22'658	22'997	23'158	23'492	23'664	23'699	23'904	23'960	23'992	23'994	24'001	24'002
1	14'457	22'038	22'627	23'114	23'238	23'312	23'440	23'490	23'964	23'976	24'048	24'111	24'252	24'538	24'540	24'55	
2	16'075	22'672	23'753	24'052	24'206	24'757	24'786	24'807	24'823	24'888	24'996	25'401	25'681	25'705	25'732		
3	15'682	23'464	24'465	25'052	25'529	25'708	25'752	25'770	25'835	26'075	26'082	26'146	26'150	26'167			
4	16'551	23'706	24'627	25'573	26'046	26'115	26'283	26'481	26'701	26'718	26'724	26'728	26'735				
5	15'499	23'796	24'866	25'317	26'139	26'154	26'175	26'205	26'764	26'818	26'836	26'959					
6	14'629	21'645	22'826	23'599	24'992	25'434	25'476	25'549	25'604	25'709	25'723						
7	17'585	26'288	27'623	27'939	28'335	28'638	28'715	28'759	29'525	30'302							
8	17'419	25'941	27'066	27'761	28'043	28'477	28'721	28'878	28'948								
9	16'665	25'370	26'909	27'611	27'729	27'861	29'830	29'844									
10	15'471	23'745	25'117	26'378	26'971	27'396	27'480										
11	15'103	23'393	26'809	27'691	28'061	29'183											
12	14'540	22'642	23'571	24'127	24'210												
13	14'590	22'336	23'440	24'029													
14	13'967	21'515	22'603														
15	12'930	20'111															
16	12'539																
$\hat{f}_j^I$	1.51105	1.05369	1.02681	1.01709	1.01284	1.01098	1.00292	1.01098	1.00698	1.00141	1.00574	1.00386	1.00359	1.00042	1.00035	1.00004	
$\hat{\sigma}_j^I$	29.4993	16.8327	2.6478	5.4444	3.2477	11.4238	0.1882	2.6132	2.0863	0.0542	0.9079	0.5397	0.7202	0.0079	0.0002	0.0000	

Table 2 presents the following quantities: the estimated prediction uncertainties $\sqrt{E[MSEP_{CDR_k|F_{I+k-1}}(0)|F_I]}$ calculated with the formulas in [10] (column 2), the approximate prediction uncertainties $\sqrt{\widehat{MSEP}_k}$ obtained with the solvency proxy (3.2) (column 3), $\sqrt{\widehat{MSEP}_k}$ resulting from solving the optimization problems (3.3) and (3.4) using the standard deviation i.e. $m=2$ (columns 4 and 5) - given that all discount factors are identical the optimization problems would not have unique solutions for the variance risk measure - and the rooted process error variance (2.8) (column 6); in "Total" we find the rooted sums of all squared elements in the respective columns - by construction it is identical for columns 1 to 4 and corresponds to the Mack error of the entire runoff, for column 6 it corresponds to the rooted process error variance of the entire runoff. The squared estimation error of the total runoff is given by $(3233.7)^2-(2454.7)^2=(2105.0)^2$ and it is this

Table 2: Prediction uncertainties by accounting year over the runoff

k	M.&W. 2015	Solvency Proxy	Maximum	Minimum	$\sqrt{\widehat{Var}(CDR_k\|\mathcal{F}_I)}$
1	1842.9	2077.2	1338.7	2494.6	1338.7
2	1485.1	1419.2	1080.7	1080.7	1080.7
3	1208.3	1118.1	885.2	885.2	885.2
4	1071.1	981.7	834.2	834.2	834.2
5	901.1	831.3	733.2	733.2	733.2
6	785.3	730.7	705.8	669.0	669.0
7	525.2	475.7	705.8	424.1	424.1
8	476.3	438.2	705.8	409.6	409.6
9	366.4	336.6	705.8	320.7	320.7
10	269.3	243.3	705.8	234.0	234.0
11	245.0	229.4	705.8	225.4	225.4
12	180.5	171.3	705.8	170.1	170.1
13	130.1	126.9	705.8	126.6	126.6
14	13.7	13.5	705.8	13.3	13.3
15	2.0	2.0	705.8	1.9	1.9
16	0.3	0.3	705.8	0.3	0.3
Total	3233.7	3233.7	3233.7	3233.7	2454.7

quantity that we split across accounting years according to (3.1) in order to obtain the values in columns 2 to 4. We observe that the prediction uncertainties obtained with [10] are not very different from our solvency proxy. Further, we see that the prediction uncertainties in column "Maximum" are identical for accounting years $k\geq 6$ and equal to the process error $\sqrt{\widehat{Var}(CDR_k|F_I)}$ for $k<6$ which is consistent with Corollary 1. Finally, column "Minimum" shows that the minimum MVM (for $m=2$) is obtained if the entire estimation error is attributed to $k=1$ leaving only the process error for the remaining ks.

In order to quantify the MVM for each approach in Table 2 we assume a cost of capital of $c=6\%$ and a loading of $\phi = 3$ (this calibration corresponds to [13]), that is to say we have

$$MVM = 6\% \sum_{k=1}^{J} 3\sqrt{\widehat{MSEP}_k}$$

The results are in Table 3 where "MVM" shows the monetary values and "Rel. to min." the values relative to the "Minimum."

Table 3: MVM

	M.&W. 2015	Solvency Proxy	Maximum	Minimum
MVM	1'710	1'655	2'274	1'552
Rel to min.	110%	107%	147%	100%

Hence, the error due to wrong weights cannot be larger than 47% of the smallest MVM possible.

4. AGGREGATING LINES OF BUSINESS

Before concluding we discuss aggregation. Suppose that we would like to use correlations to aggregate lines of business in order to obtain the MVM on a company level. Regarding dependencies between lines of business, we need to answer the following questions

- What are the correlations between the process errors?
- What are the correlations between the estimation errors?
- What are the correlations between the estimated claims development results (CDR)?
- What are the correlations between the ultimate liabilities?

These questions cannot be answered independently. For example, if we define correlations for the yearly process errors and for the estimation errors, then the correlations between the CDRs and the correlations between the ultimate liabilities are determined. And it is not difficult to show that the correlations between the ultimate liabilities will be smaller, in absolute value, than the correlations between the CDRs. Or if we define the correlations between the ultimate liabilities, then this will likely imply time-varying correlations for the CDRs. We therefore suggest to reflect well before deciding on a dependency structure and be clear when documenting about it - which can only help in order to fulfill regulatory reporting requirements.

As a side note, we remark that if two individual triangles satisfy the classical chain ladder assumptions then an aggregated triangle obtained by adding up the individual triangles will, in general, no longer satisfy the classical chain ladder assumptions.

5. CONCLUSIONS

We discussed the market value margin (MVM) for a general insurance runoff based on the distribution-free chain ladder model and suggested an easy way to modify the approach of [10] in order to take accounting year effects into consideration. Further, we showed that different splits of the estimation error over the runoff lead to different MVMs even if the estimation error of the total runoff is unchanged. We derived the splits which minimize and maximize the MVM which could be useful to quantify model uncertainty. Finally, we argued that one has to be careful when estimating an aggregated MVM for two lines of business because the correlations between quantities like the process errors, the estimation errors, the claims development results and the ultimate liabilities depend on each other.

We believe that our results are helpful in daily actuarial practice. We leave it to future research to shed light on how the MVM is affected by risk factors like stochastic discount rates and other factors which we mentioned but omitted in our analysis.

Acknowledgment

I thank three referees for very useful comments on an earlier version of this paper. I am further grateful to Andreas Gadmer from SIGNAL IDUNA Reinsurance and to Luca Valli from Endurance for a stimulating exchange on the presented subject and to Caroline Schädler for excellent proof reading.

6. REFERENCES

[1] Bühlmann H, De Felice M, Gisler A, Moriconi F, Wüthrich, M, "Recursive Credibility Formula for Chain Ladder Factors and the Claims Development Result," *Astin Bulletin*, 2009, Vol. 39, No. 1, 275-306.

[2] Carrato A, Gesmann M, Murphy D,Wüthrich M, ZhangW "Claims reserving with R: ChainLadder-0.2.0 Package Vignette, " *CRAN. ran.r-project.org*, 2015.

[3] Federal Office of Private Insurance, "The Swiss Experience with Market Consistent Technical Provisions - the Cost of Capital Approach," www.finma.ch/archiv/bpv/e/themen/00506/00552/00727/index.html, 2006.

[4] Gisler A "The Insurance Risk in the SST and in Solvency II: Modelling and Parameter Estimation," *ASTIN-Colloquium, Helsinki*, 2009.

[5] CEIOPS-SEC-40-10, "QIS5 Calibration Paper," http://ec.europa.eu/internal market/insurance/docs/solvency/qis5/ceiopscalibration-paper en.pdf 2010.

[6] Mack, T "Distribution-free calculation of the standard error of chain ladder reserve estimates," *Astin Bulletin*, 1993, Vol. 23, No. 2, 213-225.

[7] Mack, T "Schadenversicherungsmathematik," *Schriftreihe Angewandte Versicherungsmathematik*, 2002, Vol. 28, No. 2. Auflage, 2002.

[8] Merz M, Wüthrich M "Prediction error of the expected claims development result in the chain ladder method," *Bulletin Swiss Assoc. Act.*, 2007, Vol. 1, 117-137.

[9] Merz M, Wüthrich M "Modeling the claims development result for solvency purposes," *CAS E-Forum Fall*, 2008, 542-568.

[10] Merz M, Wüthrich M "Claims run-off uncertainty: the full picture", *Swiss Finance Institute Research Paper*, 2015, No. 14-69.

[11] Robert, Y C "Market value margin calculations under the cost of capital approach within a Bayesian chain ladder framework," *Insurance Mathematics and Economics*, 2013, Vol. 53, 216-229.

[12] Röhr, A "Chain ladder and error propagation", *CAE Fall Meeting,* 2014.
[13] Salzmann, R, Wüthrich M, "Cost-of-capital margin for a general insurance liability runoff," *Astin Bulletin*, 2010, Vol. 40, No. 2, 415-451.
[14] Taylor, G C "Separation of Inflation and other Effects from the Distribution of Non-Life Insurance Claim Delays" *Astin Bulletin*, 1977, Vol. 9, No. 1-2, 219-230.
[15] Wüthrich M, Embrechts P, Tsanakas A, " Risk margin for a non-life insurance run-off," *Statistics and Risk Modeling*, 2011, Vol. 28, 299-317.
[16] Wüthrich M, Merz M, "Stochastic claims reserving methods in insurance", *John Wiley & Sons Ltd, West Sussex, England*, 2008.

Abbreviations and notations

c, cost of capital
CDR, claims development result
CL, chain ladder
$D_{I,k}$, discount factors
E[ ], expectation operator
MSEP, mean squared error of prediction
MVM, market value margin
Var( ), variance
ϕ, a loading
ρ, a risk measure

Biography of the Author

Daniel Burren is director and actuary at ProMaSta Pte Ltd (www.promasta.com), an actuarial consultancy company in Singapore. Daniel has a MSc in mathematical statistics and a PhD in economics from the University of Bern, Switzerland, is a fully qualified actuary of the Swiss Association of Actuaries and has published in peer reviewed journals including an article in Insurance Mathematics and Economics. Contact details: danielburren@gmail.com, dburren@promasta.com, www.danielburren.ch, +65 8777 1373.

Complex Random Variables

Leigh J. Halliwell, FCAS, MAAA

Abstract: Rarely have casualty actuaries needed, much less wanted, to work with complex numbers. One readily could wisecrack about imaginary dollars and creative accounting. However, complex numbers are well established in mathematics; they even provided the impetus for abstract algebra. Moreover, they are essential in several scientific fields, most notably in electromagnetism and quantum mechanics, the two fields to which most of the sparse material about complex random variables is tied. This paper will introduce complex random variables to an actuarial audience, arguing that complex random variables will eventually prove useful in the field of actuarial science. First, it will describe the two ways in which statistical work with complex numbers differs from that with real numbers, viz., in transjugation versus transposition and in rank versus dimension. Next, it will introduce the mean and the variance of the complex random vector, and derive the distribution function of the standard complex normal random vector. Then it will derive the general distribution of the complex normal multivariate and discuss the behavior and moments of complex lognormal variables, a limiting case of which is the unit-circle random variable $W = e^{i\Theta}$ for real Θ uniformly distributed. Finally, it will suggest several foreseeable actuarial applications of the preceding theory, especially its application to linear statistical modeling. Though the paper will be algebraically intense, it will require little knowledge of complex-function theory. But some of that theory, viz., Cauchy's theorem and analytic continuation, will arise in an appendix on the complex moment generating function of a normal random multivariate.

Keywords: Complex numbers, matrices, and random vectors; augmented variance; lognormal and unit-circle distributions; determinism; Cauchy-Riemann; analytic continuation

1. INTRODUCTION

Even though their education has touched on algebra and calculus with complex numbers, most casualty actuaries would be hard-pressed to cite an actuarial use for numbers of the form $x+iy$. Their use in the discrete Fourier transformation (Klugman [1998], §4.7.1) is notable; however, many would view this as a trick or convenience, rather than as indicating any further usefulness. In this paper we will develop a probability theory for complex random variables and vectors, arguing that such a theory will eventually find actuarial uses. The development, lengthy and sometimes arduous, will take the following steps. Sections 2-4 will base complex matrices in certain real-valued matrices called "double-real." This serves the aim of our presentation, namely, to analogize from real-valued random variables and vectors to complex ones. Transposition and dimension in the real-valued

realm become transjugation and rank in the complex. These differences figure into the standard quadratic form of Section 5, where also the distribution of the standard complex normal random vector is derived. Section 6 will elaborate on the variance of a complex random vector, as well as introduce "augmented variance," i.e., the variance of dyad whose second part is the complex conjugate of the first. Section 7 derives of the formula for the distribution of the general complex normal multivariate. Of special interest to many casualty actuaries should be the treatment of the complex lognormal random vector in Section 8, an intuition into whose behavior Section 9 provides on a univariate or scalar level. Even further simplification in the next two sections leads to the unit-circle random variable, which is the only random variable with widespread deterministic effects. In Section 12 we adapt the linear statistical model to complex multivariates. Finally, Section 13 lists foreseeable applications of complex random variables. However, we believe their greatest benefit resides not in their concrete applications, but rather in their fostering abstractions of thought and imagination. Three appendices delve into mathematical issues too complicated for the body of paper. Those who work on an advanced level with lognormal random variables should read Appendix A ("Real-Valued Lognormal Random Vectors"), regardless of their interest in complex random variables.

2. INVERTING COMPLEX MATRICES

Let $m \times n$ complex matrix Z be composed of real and imaginary parts X and Y, i.e., $\mathrm{Z} = \mathrm{X} + i\mathrm{Y}$. Of course, X and Y also must be $m \times n$. Since only square matrices have inverses, our purpose here requires that $m = n$. Complex matrix $\mathrm{W} = \mathrm{A} + i\mathrm{B}$ is an inverse of Z if and only if $\mathrm{ZW} = \mathrm{WZ} = \mathrm{I}_n$, where I_n is the $n \times n$ identity matrix. Because such an inverse must be unique, we may say that $\mathrm{Z}^{-1} = \mathrm{W}$. Under what conditions does W exist?

First, define the conjugate of Z as $\overline{\mathrm{Z}} = \mathrm{X} - i\mathrm{Y}$. Since the conjugate of a product equals the product of the conjugates,[1] if Z is non-singular, then $\overline{\mathrm{Z}}\overline{\mathrm{Z}^{-1}} = \overline{\mathrm{Z}\mathrm{Z}^{-1}} = \overline{\mathrm{I}_n} = \mathrm{I}_n$. Similarly, $\overline{\mathrm{Z}^{-1}}\overline{\mathrm{Z}} = \mathrm{I}_n$. Therefore, $\overline{\mathrm{Z}}$ too is non-singular, and $\overline{\mathrm{Z}}^{-1} = \overline{\mathrm{Z}^{-1}}$. Moreover, if Z is non-singular, so too are $i^n\mathrm{Z}$ and $i^n\overline{\mathrm{Z}}$. Therefore, the invertibility of $\mathrm{X} + i\mathrm{Y}$, $-\mathrm{Y} + i\mathrm{X}$, $-\mathrm{X} - i\mathrm{Y}$, $\mathrm{Y} - i\mathrm{X}$, $\mathrm{X} - i\mathrm{Y}$, $\mathrm{Y} + i\mathrm{X}$, $-\mathrm{X} + i\mathrm{Y}$, and $-\mathrm{Y} - i\mathrm{X}$ is true for all eight or true for none. Invertibility is no respecter of the real and imaginary parts.

Now if the inverse of Z is $\mathrm{W} = \mathrm{A} + i\mathrm{B}$, then $\mathrm{I}_n = (\mathrm{X} + i\mathrm{Y})(\mathrm{A} + i\mathrm{B}) = (\mathrm{A} + i\mathrm{B})(\mathrm{X} + i\mathrm{Y})$. Expanding the first equality, we have:

$$\begin{aligned}\mathrm{I}_n &= (\mathrm{X} + i\mathrm{Y})(\mathrm{A} + i\mathrm{B}) \\ &= \mathrm{XA} + i\mathrm{YA} + i\mathrm{XB} + i^2\mathrm{YB} \\ &= \mathrm{XA} + i\mathrm{YA} + i\mathrm{XB} - \mathrm{YB} \\ &= (\mathrm{XA} - \mathrm{YB}) + i(\mathrm{YA} + \mathrm{XB})\end{aligned}$$

Therefore, $\mathrm{ZW} = \mathrm{I}_n$ if and only if $\mathrm{XA} - \mathrm{YB} = \mathrm{I}_n$ and $\mathrm{YA} + \mathrm{XB} = 0_{\mathrm{n}\times\mathrm{n}}$. We may combine the last two equations into the partitioned-matrix form:

$$\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}\begin{bmatrix} \mathrm{A} \\ \mathrm{B} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_n \\ 0 \end{bmatrix}$$

Since $\mathrm{YA} + \mathrm{XB} = 0$ if and only if $-\mathrm{YA} - \mathrm{XB} = 0$, another form just as valid is:

$$\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}\begin{bmatrix} -\mathrm{B} \\ \mathrm{A} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathrm{I}_n \end{bmatrix}$$

We may combine these two forms into the balanced form:

$$\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}\begin{bmatrix} \mathrm{A} & -\mathrm{B} \\ \mathrm{B} & \mathrm{A} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_n & 0 \\ 0 & \mathrm{I}_n \end{bmatrix} = \mathrm{I}_{2n}$$

[1] If Z and W are conformable to multiplication:
$\overline{\mathrm{ZW}} = (\mathrm{X} + i\mathrm{Y})(\mathrm{A} + i\mathrm{B}) = \overline{\mathrm{XA} - \mathrm{YB} + i(\mathrm{XB} + \mathrm{YA})} = \mathrm{XA} - \mathrm{YB} - i(\mathrm{XB} + \mathrm{YA}) = (\mathrm{X} - i\mathrm{Y})(\mathrm{A} - i\mathrm{B}) = \overline{\mathrm{Z}}\,\overline{\mathrm{W}}$

Therefore, $\mathrm{ZW} = \mathrm{I}_n$ if and only if $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}\begin{bmatrix} \mathrm{A} & -\mathrm{B} \\ \mathrm{B} & \mathrm{A} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_n & 0 \\ 0 & \mathrm{I}_n \end{bmatrix}$. By a similar expansion of the last equality above, $\mathrm{WZ} = \mathrm{I}_n$ if and only if $\begin{bmatrix} \mathrm{A} & -\mathrm{B} \\ \mathrm{B} & \mathrm{A} \end{bmatrix}\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_n & 0 \\ 0 & \mathrm{I}_n \end{bmatrix}$. Hence, we conclude that the $n \times n$ complex matrix $\mathrm{Z} = \mathrm{X} + i\mathrm{Y}$ is non-singular, or has an inverse, if and only if the $2n \times 2n$ real-valued matrix $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ is non-singular. Moreover, if $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}^{-1}$ exists, it will have the form $\begin{bmatrix} \mathrm{A} & -\mathrm{B} \\ \mathrm{B} & \mathrm{A} \end{bmatrix}$ and Z^{-1} will equal $\mathrm{A} + i\mathrm{B}$.

3. COMPLEX MATRICES AS DOUBLE-REAL MATRICES

That the problem of inverting an $n \times n$ complex matrix resolves into the problem of inverting a $2n \times 2n$ real-valued matrix suggests that with complex numbers one somehow gets "two for the price of one." It even hints of a relation between the general $m \times n$ complex matrix $\mathrm{Z} = \mathrm{X} + i\mathrm{Y}$ and the $2m \times 2n$ complex matrix $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$. If X and Y are $m \times n$ real matrices, we will call $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ a double-real matrix. The matrix is double in two senses; first, in that it involves two (same-sized and real-valued) matrices X and Y, and second, in that its right half is redundant, or reproducible from its left.

Returning to the hint above, we easily see an addition analogy:

$$\mathrm{X}_1 + i\mathrm{Y}_1 + \mathrm{X}_2 + i\mathrm{Y}_2 \Leftrightarrow \begin{bmatrix} \mathrm{X}_1 & -\mathrm{Y}_1 \\ \mathrm{Y}_1 & \mathrm{X}_1 \end{bmatrix} + \begin{bmatrix} \mathrm{X}_2 & -\mathrm{Y}_2 \\ \mathrm{Y}_2 & \mathrm{X}_2 \end{bmatrix}$$

And if Z_1 is $m\times n$ and Z_2 is $n\times p$, so that the matrices are conformable to multiplication, then $Z_1Z_2 = (X_1 + iY_1)(X_2 + iY_2) = (X_1X_2 - Y_1Y_2) + i(X_1Y_2 + Y_1X_2)$. This is analogous with the double-real multiplication:

$$\begin{bmatrix} X_1 & -Y_1 \\ Y_1 & X_1 \end{bmatrix}\begin{bmatrix} X_2 & -Y_2 \\ Y_2 & X_2 \end{bmatrix} = \begin{bmatrix} X_1X_2 - Y_1Y_2 & -X_1Y_2 - Y_1X_2 \\ X_1Y_2 + Y_1X_2 & X_1X_2 - Y_1Y_2 \end{bmatrix}$$

Rather trivial is the analogy between the $m\times n$ complex zero matrix and the $2m\times 2n$ double-real zero matrix $\begin{bmatrix} 0_{m\times n} & -0 \\ 0 & 0 \end{bmatrix}$, as well as that between the $n\times n$ complex identity matrix and the $2n\times 2n$ double-real identity matrix $\begin{bmatrix} I_n & -0 \\ 0 & I_n \end{bmatrix}$.

The general $2m\times 2n$ double-real matrix may itself be decomposed into quasi-real and quasi-imaginary parts: $\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix} = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} + \begin{bmatrix} 0 & -Y \\ Y & 0 \end{bmatrix}$. And in the case of square matrices ($m = n$) this extends to the form $\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix} = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} + \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}\begin{bmatrix} Y & 0 \\ 0 & Y \end{bmatrix}$, wherein the double-real matrix $\begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}$ is analogous with the imaginary unit, inasmuch as:

$$\begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}\begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix} = \begin{bmatrix} -I_n & 0 \\ 0 & -I_n \end{bmatrix} = (-1)\begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix}$$

Finally, one of the most important theorems of linear algebra is that every $m\times n$ complex matrix $Z = X + iY$ may be reduced by invertible transformations to "canonical form" (Healy [1986], 32-34). In symbols, for every Z there exist non-singular matrices U and V such that:

$$\mathrm{U}_{m\times m}\mathrm{Z}_{m\times n}\mathrm{V}_{n\times n}=\begin{bmatrix}\mathrm{I}_r & 0\\ 0 & 0\end{bmatrix}_{m\times n}$$

The $m\times n$ real matrix on the right side of the equation consists entirely of zeroes except for r instances of one along its main diagonal. Since invertible matrix operations can reposition the ones, it is further stipulated that the ones appear as a block in the upper-left corner. Although many reductions of Z to canonical form exist, the canonical forms themselves must all contain the same number of ones, r, which is defined as the rank of Z. Providing the matrices with real and complex parts, we have:

$$\begin{aligned}\mathrm{U}_{m\times m}\mathrm{Z}_{m\times n}\mathrm{V}_{n\times n} &= (\mathrm{P}+i\mathrm{Q})(\mathrm{X}+i\mathrm{Y})(\mathrm{R}+i\mathrm{S})\\ &= (\mathrm{PXR}-\mathrm{QYR}-\mathrm{PYS}-\mathrm{QXS})+i(\mathrm{PXS}-\mathrm{QYS}+\mathrm{PYR}+\mathrm{QXR})\\ &= \mathrm{A}+i\mathrm{B}\\ &= \begin{bmatrix}\mathrm{I}_r & 0\\ 0 & 0\end{bmatrix}+i\,0_{m\times n}\end{aligned}$$

The double-real analogue to this is:

$$\begin{bmatrix}\mathrm{P} & -\mathrm{Q}\\ \mathrm{Q} & \mathrm{P}\end{bmatrix}\begin{bmatrix}\mathrm{X} & -\mathrm{Y}\\ \mathrm{Y} & \mathrm{X}\end{bmatrix}\begin{bmatrix}\mathrm{R} & -\mathrm{S}\\ \mathrm{S} & \mathrm{R}\end{bmatrix}=\begin{bmatrix}\mathrm{A} & -\mathrm{B}\\ \mathrm{B} & \mathrm{A}\end{bmatrix}=\begin{bmatrix}\begin{bmatrix}\mathrm{I}_r & 0\\ 0 & 0\end{bmatrix} & 0_{m\times n}\\ 0_{m\times n} & \begin{bmatrix}\mathrm{I}_r & 0\\ 0 & 0\end{bmatrix}\end{bmatrix}$$

As shown in the previous section, $\begin{bmatrix}\mathrm{P} & -\mathrm{Q}\\ \mathrm{Q} & \mathrm{P}\end{bmatrix}$ is non-singular, or invertible, if and only if $\mathrm{U}=\mathrm{P}+i\mathrm{Q}$ is non-singular; the same is true for $\begin{bmatrix}\mathrm{R} & -\mathrm{S}\\ \mathrm{S} & \mathrm{R}\end{bmatrix}$. Therefore, the rank of the double-real analogue of a complex matrix is twice the rank of the complex matrix. Moreover, the $2r$ instances of one correspond to r quasi-real and r quasi-imaginary instances. It is not possible for the contribution to the rank of a matrix to be real without its being imaginary, and *vice versa*.

To conclude this section, there are extensive analogies between complex and double-real matrices, analogies so extensive that one who lacked either the confidence or the software to work with complex numbers could probably do a work-around with double-real matrices.[2]

4. COMPLEX MATRICES AND VARIANCE

$\Sigma = Var[\mathbf{x}] = E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]$ is a real-valued $n\times n$ matrix, whose jk^{th} element is the covariance of the j^{th} element of x with the k^{th} element. Since the covariance of two real-valued random variables is symmetric, Σ must be a symmetric matrix. But a realistic Σ must have one other property, viz., non-negative definiteness (NND). This means that for every real-valued $n\times 1$ vector ξ, $\xi'\Sigma\xi \geq 0$.[3] This must be true, because $\xi'\Sigma\xi$ is the variance of the real-valued random variable $\xi'\mathbf{x}$:

$$Var[\xi'\mathbf{x}] = E\left[(\xi'\mathbf{x}-\xi'\mu)(\xi'\mathbf{x}-\xi'\mu)'\right] = E\left[\xi'(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\xi\right] = \xi'E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]\xi = \xi'\Sigma\xi$$

Although variances of real-valued random variables may be zero, they must not be negative. Now if $\xi'\Sigma\xi > 0$ for all $\xi \neq 0_{n\times 1}$, the variance Σ is said to be positive-definite (PD). Every invertible NND matrix must be PD. Moreover, every NND matrix may be expressed as the product of some real matrix and its transpose, the most common method for doing this being the Cholesky

---

[2] The representation of the complex scalar $z = x + iy$ as the real 2×2 matrix $\begin{bmatrix} x & -y \\ y & x \end{bmatrix}$ is a common theme in modern algebra (e.g., section 7.2 of the Wikipedia article "Complex number"). We have merely extended the representation to complex matrices. Our representation is even more meaningful when expressed in the Kronecker-product form $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix} = \begin{bmatrix} \mathrm{X} & 0 \\ 0 & \mathrm{X} \end{bmatrix} + \begin{bmatrix} 0 & -\mathrm{Y} \\ \mathrm{Y} & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \mathrm{X} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \otimes \mathrm{Y}.$ Due to certain properties of the Kronecker product (cf. Judge [1988], Appendix A.15), all the analogies of this section would hold even in the commuted form $\mathrm{X} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mathrm{Y} \otimes \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. In practical terms this means that it matters not whether the form is 2×2 of $m\times n$ or $m\times n$ of 2×2.

[3] More accurately, $\xi'\Sigma\xi \geq [0]$, since the quadratic form $\xi'\Sigma\xi$ is a 1×1 matrix. The relevant point is that 1×1 real-valued matrices are as orderable as their real-valued elements are. Appendices A.13 and A.14 of Judge [1988] provide introductions to quadratic forms and definiteness that are sufficient to prove the theorems used herein.

decomposition (Healy [1986], §7.2). Accordingly, if Σ is NND, then $\mathrm{A}'\Sigma\mathrm{A} \geq 0$ for any conformable real-valued matrix A. Finally, if Σ is PD and real-valued $n \times r$ matrix A is of full column rank, i.e., $rank(\mathrm{A}_{n\times r}) = r$, then the $r \times r$ matrix $\mathrm{A}'\Sigma\mathrm{A}$ is PD.

In the remainder of this section we will show how the analogy between $\mathrm{X} + i\mathrm{Y}$ and $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ leads to a proper definition of the variance of a complex random vector. We start by considering $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ as a real-valued variance matrix. In order to be so, first it must be symmetric:

$$\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}' = \begin{bmatrix} \mathrm{X}' & \mathrm{Y}' \\ -\mathrm{Y}' & \mathrm{X}' \end{bmatrix} = \begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$$

Hence, $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ is symmetric if and only if $\mathrm{X}' = \mathrm{X}$ and $\mathrm{Y}' = -\mathrm{Y}$. In words, X is symmetric and Y is skew-symmetric. Clearly, the main diagonal of a skew-symmetric matrix must be zero. But of greater significance, if a and b are real-valued $n \times 1$ vectors:

$$\mathrm{a}'\mathrm{Y}\mathrm{b} = (\mathrm{a}'\mathrm{Y}\mathrm{b})_{1\times 1} = (\mathrm{a}'\mathrm{Y}\mathrm{b})' = \mathrm{b}'\mathrm{Y}'\mathrm{a} = \mathrm{b}'(-\mathrm{Y})\mathrm{a} = -\mathrm{b}'\mathrm{Y}\mathrm{a}$$

Consequently, if $\mathrm{b} = \mathrm{a}$:

$$\mathrm{a}'\mathrm{Y}\mathrm{a} = \frac{\mathrm{a}'\mathrm{Y}\mathrm{a} + \mathrm{a}'\mathrm{Y}\mathrm{a}}{2} = \frac{\mathrm{a}'\mathrm{Y}\mathrm{a} + (-\mathrm{a}'\mathrm{Y}\mathrm{a})}{2} = 0_{1\times 1}$$

Next, considering the specifications on X, Y, a, and b, we evaluate the 2×2 quadratic form:

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}'\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = \begin{bmatrix} a' & b' \\ -b' & a' \end{bmatrix}\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

$$= \begin{bmatrix} a'X + b'Y & b'X - a'Y \\ -b'X + a'Y & a'X + b'Y \end{bmatrix}\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

$$= \begin{bmatrix} a'Xa + b'Ya - a'Yb + b'Xb & -a'Xb - b'Yb + b'Xa - a'Ya \\ -b'Xa + a'Ya + a'Xb + b'Yb & a'Xa + b'Ya - a'Yb + b'Xb \end{bmatrix}$$

$$= \begin{bmatrix} a'Xa + b'Ya - a'Yb + b'Xb & -a'Xb - 0 + b'Xa - 0 \\ -b'Xa + 0 + a'Xb + 0 & a'Xa + b'Ya - a'Yb + b'Xb \end{bmatrix}$$

$$= \begin{bmatrix} a'Xa + b'Ya - a'Yb + b'Xb & -a'Xb + b'Xa \\ -b'Xa + a'Xb & a'Xa + b'Ya - a'Yb + b'Xb \end{bmatrix}$$

$$= \begin{bmatrix} a'Xa + b'Ya - a'Yb + b'Xb & 0 \\ 0 & a'Xa + b'Ya - a'Yb + b'Xb \end{bmatrix}$$

$$= \begin{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}'\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} a \\ b \end{bmatrix}'\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} \end{bmatrix}$$

Therefore, $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}'\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ is PD [or NND] if and only if $\begin{bmatrix} a \\ b \end{bmatrix}'\begin{bmatrix} X & -Y \\ Y & X \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}$ is PD [or NND].

Now the double-real $2n \times 2$ matrix $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ is analogous with the $n \times 1$ complex vector $a + ib$. Its transpose $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}' = \begin{bmatrix} a' & b' \\ -b' & a' \end{bmatrix}$ is analogous with the $1 \times n$ complex vector $a' - ib'$. Moreover, $a' - ib' = (a - ib)' = \overline{a + ib}' = \overline{(a + ib)'} = (a + ib)^*$, where '*' is the combined operation of

transposition and conjugation (order irrelevant).[4] And $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ is analogous with the $n \times n$ complex matrix $\mathrm{X} + i\mathrm{Y}$. Accordingly, the complex analogue of the double-real quadratic form $\begin{bmatrix} \mathrm{a} & -\mathrm{b} \\ \mathrm{b} & \mathrm{a} \end{bmatrix}' \begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix} \begin{bmatrix} \mathrm{a} & -\mathrm{b} \\ \mathrm{b} & \mathrm{a} \end{bmatrix}$ is $(\mathrm{a} + i\mathrm{b})^* (\mathrm{X} + i\mathrm{Y})(\mathrm{a} + i\mathrm{b})$. Moreover, since $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ is symmetric, $(\mathrm{X} + i\mathrm{Y})^* = \mathrm{X}' - i\mathrm{Y}' = \mathrm{X} - i(-\mathrm{Y}) = \mathrm{X} + i\mathrm{Y}$. A matrix equal to its transposed conjugate is said to be Hermetian: matrix Γ is Hermetian if and only if $\Gamma^* = \Gamma$. Therefore, $\Gamma_{n \times n} = \mathrm{X} + i\mathrm{Y}$ is the variance matrix of some complex random variable $\mathbf{z}_{n \times 1} = \mathbf{x} + i\mathbf{y}$ if and only if Γ is Hermetian and $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ is non-negative-definite.[5]

Because $(\mathrm{a} + i\mathrm{b})^* (\mathrm{X} + i\mathrm{Y})(\mathrm{a} + i\mathrm{b}) = \begin{bmatrix} \mathrm{a} \\ \mathrm{b} \end{bmatrix}' \begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix} \begin{bmatrix} \mathrm{a} \\ \mathrm{b} \end{bmatrix} + i \cdot 0 = \begin{bmatrix} \mathrm{a} \\ \mathrm{b} \end{bmatrix}' \begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix} \begin{bmatrix} \mathrm{a} \\ \mathrm{b} \end{bmatrix}$, the definiteness of $\Gamma = \mathrm{X} + i\mathrm{Y}$ is the same as the definiteness of $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$. Therefore, a matrix qualifies as the variance matrix of some complex random vector if and only if it is Hermetian and NND. Just as the variance matrix of a real-valued random vector factors as $\Sigma = \mathrm{A}'\mathrm{A}_{n \times n}$ for some real-valued A, so too the variance matrix of a complex random vector factors as $\Gamma = \mathrm{A}^*\mathrm{A}_{n \times n}$ for some complex A. Likewise, every invertible Hermetian NND matrix must be PD. Due to the skew symmetry of their

[4] The transposed conjugate is sometimes called the "transjugate," which in linear algebra is commonly symbolized with the asterisk. Physicists prefer the "dagger" notation A†, though physicist Hermann Weyl [1950, p. 17] called it the "Hermetian conjugate" and symbolized it as $\tilde{\mathrm{A}}$.

[5] It is superfluous to add 'symmetric' here. For $\Gamma = \mathrm{X} + i\mathrm{Y}$ is Hermetian if and only if $\begin{bmatrix} \mathrm{X} & -\mathrm{Y} \\ \mathrm{Y} & \mathrm{X} \end{bmatrix}$ is symmetric.

complex parts, the main diagonals of Hermetian matrices must be real-valued. If the matrices are NND [or PD], all the elements of their main diagonals must be non-negative [or positive].

Let Γ represent the variance of the complex random vector $\mathbf{z}$. Its jk^{th} element represents the covariance of the j^{th} element of $\mathbf{z}$ with the k^{th} element. Since Γ is Hermetian, $\gamma_{kj} = [\Gamma]_{kj} = [\Gamma^*]_{kj} = [\overline{\Gamma}']_{kj} = [\overline{\Gamma}]_{jk} = [\overline{\Gamma}]_{jk} = \overline{\gamma}_{jk}$. Because of this, it is fitting and natural to define the variance of a complex random vector as:

$$\Gamma = Var[\mathbf{z}] = E\left[(\mathbf{z}-\mu)\overline{(\mathbf{z}-\mu)}'\right] = E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)^*\right]$$

The complex formula is like the real formula except that the second factor in the expectation is transjugated, not simply transposed. This renders Γ Hermetian, since:

$$\Gamma^* = E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)^*\right]^* = E\left[\left\{(\mathbf{z}-\mu)(\mathbf{z}-\mu)^*\right\}^*\right] = E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)^*\right] = \Gamma$$

It also renders Γ NND. For since $(\mathrm{z}-\mu)(\mathrm{z}-\mu)^*$ is NND, its expectation over the probability distribution of $\mathbf{z}$ must also be so. Usually Γ is PD, in which case Γ^{-1} exists.

5. THE EXPECTATION OF THE STANDARD QUADRATIC FORM

The most common quadratic form in $\mathbf{z}_{n\times 1}$ involves the variance of the complex random variable, viz., $(\mathbf{z}-\mu)^*\Gamma^{-1}(\mathbf{z}-\mu)$, where $\Gamma = Var[\mathbf{z}]$. The expectation of this quadratic form equals n, the rank of the variance. The following proof uses the trace function. The trace of a matrix is the sum of its main-diagonal elements, and if A and B are conformable $tr(AB) = tr(BA)$. Moreover, the trace of the expectation equals the expectation of the trace. Consequently:

$$\begin{aligned} E\left[(\mathbf{z}-\mu)^{*}\Gamma^{-1}(\mathbf{z}-\mu)\right] &= E\left[tr\left((\mathbf{z}-\mu)^{*}\Gamma^{-1}(\mathbf{z}-\mu)\right)\right] \\ &= E\left[tr\left(\Gamma^{-1}(\mathbf{z}-\mu)(\mathbf{z}-\mu)^{*}\right)\right] \\ &= tr\left(E\left[\Gamma^{-1}(\mathbf{z}-\mu)(\mathbf{z}-\mu)^{*}\right]\right) \\ &= tr\left(\Gamma^{-1}E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)^{*}\right]\right) \\ &= tr\left(\Gamma^{-1}\Gamma\right) \\ &= tr\left(\mathrm{I}_{n}\right) \\ &= n \end{aligned}$$

The analogies above between complex and double-real matrices might suggest the result to be $2n$. However, for real-valued random variables $E\left[(\mathbf{x}-\mu)^{*}\Sigma^{-1}(\mathbf{x}-\mu)\right]=n$, and the complex case is a superset of the real. So by extension, the complex case must be the same.

But an insight is available into why the value is n, rather than $2n$. Let $\mathbf{x}$ and $\mathbf{y}$ be $n\times 1$ real-valued random vectors. Assume their means to be zero, and their variances to be identity matrices (so zero covariance):

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \left(\mu=\begin{bmatrix}0_{n\times 1}\\ 0_{n\times 1}\end{bmatrix},\quad \Sigma=\begin{bmatrix}\mathrm{I}_{n} & 0\\ 0 & \mathrm{I}_{n}\end{bmatrix}\right)$$

The quadratic form is:

$$\begin{bmatrix}\mathbf{x}' & \mathbf{y}'\end{bmatrix}\Sigma^{-1}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}=\begin{bmatrix}\mathbf{x}' & \mathbf{y}'\end{bmatrix}\begin{bmatrix}\mathrm{I}_{n} & 0\\ 0 & \mathrm{I}_{n}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}=\begin{bmatrix}\mathbf{x}' & \mathbf{y}'\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}=\mathbf{x}'\mathbf{x}+\mathbf{y}'\mathbf{y}=\sum_{j=1}^{n}\left(\frac{\mathbf{x}_j^2}{1}+\frac{\mathbf{y}_j^2}{1}\right)$$

Since the elements have unit variances, the expectation is:

$$E\left[\begin{bmatrix}\mathbf{x}' & \mathbf{y}'\end{bmatrix}\Sigma^{-1}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right]=E\left[\sum_{j=1}^{n}\left(\frac{\mathbf{x}_j^2}{1}+\frac{\mathbf{y}_j^2}{1}\right)\right]=\sum_{j=1}^{n}\left(\frac{E\left[\mathbf{x}_j^2\right]}{1}+\frac{E\left[\mathbf{y}_j^2\right]}{1}\right)=\sum_{j=1}^{n}\left(\frac{1}{1}+\frac{1}{1}\right)=2n$$

Now let $\mathbf{z}$ be the $n\times 1$ complex random vector $\mathbf{x}+i\mathbf{y}$. Since $E\left[\mathbf{x}+i\mathbf{y}\right]=0_{n\times 1}$, the variance of $\mathbf{z}$ is:

$$
\begin{aligned}
\Gamma &= Var[\mathbf{z}] \\
&= Var[\mathbf{x} + i\mathbf{y}] \\
&= Var\left[\begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\right] \\
&= E\left[\begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\left\{\begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\right\}^*\right] \\
&= E\left[\begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^*\begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix}^*\right] \\
&= \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix} E\left[\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^*\right]\begin{bmatrix} \mathrm{I}_n \\ -i\mathrm{I}_n \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix} Var\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\begin{bmatrix} \mathrm{I}_n \\ -i\mathrm{I}_n \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix}\begin{bmatrix} \mathrm{I}_n & 0 \\ 0 & \mathrm{I}_n \end{bmatrix}\begin{bmatrix} \mathrm{I}_n \\ -i\mathrm{I}_n \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \end{bmatrix}\begin{bmatrix} \mathrm{I}_n \\ -i\mathrm{I}_n \end{bmatrix} \\
&= \mathrm{I}_n - i^2\mathrm{I}_n \\
&= 2\mathrm{I}_n
\end{aligned}
$$

The complex quadratic form is:

$$\mathbf{z}^*\Gamma^{-1}\mathbf{z} = \mathbf{z}^*\left(2\mathrm{I}_n\right)^{-1}\mathbf{z} = \frac{\mathbf{z}^*\mathbf{z}}{2} = \sum_{j=1}^{n}\frac{\bar{\boldsymbol{z}}_j\boldsymbol{z}_j}{2} = \sum_{j=1}^{n}\frac{\left(\boldsymbol{x}_j - i\boldsymbol{y}_j\right)\left(\boldsymbol{x}_j + i\boldsymbol{y}_j\right)}{2} = \sum_{j=1}^{n}\frac{\boldsymbol{x}_j^2 + \boldsymbol{y}_j^2}{1+1} = \frac{1}{2}\begin{bmatrix}\mathbf{x}' & \mathbf{y}'\end{bmatrix}\Sigma^{-1}\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix}$$

The complex form is half the real-valued form; hence, its expectation equals *n*. The condensation of the 2*n* real dimensions into *n* complex ones inverts the order of operations:

$$\sum_{j=1}^{n}\left(\frac{\boldsymbol{x}_j^2}{1} + \frac{\boldsymbol{y}_j^2}{1}\right) \Rightarrow \sum_{j=1}^{n}\frac{\boldsymbol{x}_j^2 + \boldsymbol{y}_j^2}{1+1}$$

Within the sigma operator, the sum of two quotients becomes the quotient of two sums. A proof for general variance Γ involves diagonalizing Γ, i.e., that Γ can be eigen-decomposed as $\Gamma = \mathrm{W}\Lambda\mathrm{W}^*$, where Λ is diagonal and $\mathrm{W}\mathrm{W}^* = \mathrm{W}^*\mathrm{W} = \mathrm{I}_n$.[6]

At this point we can derive the standard complex normal distribution. The normal distribution is $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. The standard complex normal random variable is formed from two independent real normal variables whose means equal zero and whose variances equal one half:

$$f_Z(\mathrm{z}) = \frac{1}{\sqrt{2\pi(1/2)}} e^{-\frac{(x-0)^2}{2(1/2)}} \frac{1}{\sqrt{2\pi(1/2)}} e^{-\frac{(y-0)^2}{2(1/2)}} = \frac{1}{\sqrt{\pi}} e^{-x^2} \frac{1}{\sqrt{\pi}} e^{-y^2} = \frac{1}{\pi} e^{-\left(x^2+y^2\right)} = \frac{1}{\pi} e^{-\bar{\mathrm{z}}\mathrm{z}}$$

The distribution of the $n\times 1$ standard complex normal random vector is $f_{\mathbf{z}}(\mathrm{z}) = \frac{1}{\pi^n} e^{\mathrm{z}^*\mathrm{z}}$. A vector so distributed has mean $E[\mathbf{z}] = 0_{n\times 1}$ and variance $Var[\mathbf{z}] = E[\mathbf{z}\bar{\mathbf{z}}'] = \mathrm{I}_n$.

6. COMPLEX VARIANCE, PSEUDOVARIANCE, AND AUGMENTED VARIANCE

Section 4 justified the definition of the variance of a complex random vector as:

$$\Gamma = Var[\mathbf{z}] = E\left[(\mathbf{z}-\mu)\overline{(\mathbf{z}-\mu)}'\right] = E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)^*\right]$$

The naïve formula differs from this by one critical symbol (prime versus asterisk):

$$\mathrm{C} = E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)'\right]$$

[6] Cf. Appendix C for eigen-decomposition and diagonalization. We believe the insight about commuting sums and quotients to be valuable as an abstraction. But of course, a vector of *n* independent complex random variables of unit variance translates into a vector of 2*n* independent real random variables of half-unit variance, and $\sum_{j=1}^{2n} \frac{1}{2} = n$. Because of the half-unit real variance, the formula in the next paragraph for the standard complex normal distribution, lacking any factors of two, is simpler than the formula for the standard real normal distribution.

This naïveté leads many to conclude that $Var[i\mathbf{z}] = i^2 Var[\mathbf{z}] = -Var[\mathbf{z}]$, whereas it is actually:[7]

$$Var[i\mathbf{z}] = E\left[i(\mathbf{z}-\mu)\{i(\mathbf{z}-\mu)\}^*\right] = E\left[i(\mathbf{z}-\mu)\bar{i}(\mathbf{z}-\mu)^*\right] = i\bar{i}E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)^*\right] = i(-i)Var[\mathbf{z}] = Var[\mathbf{z}]$$

Nevertheless, there is a role for the naïve formula, which reduces to:

$$\mathrm{C} = E\left[(\mathbf{z}-\mu)(\mathbf{z}-\mu)'\right] = E\left[(\mathbf{z}-\mu)\overline{(\mathbf{z}-\mu)}'\right] = E\left[(\mathbf{z}-\mu)(\bar{\mathbf{z}}-\bar{\mu})^*\right] = Cov[\mathbf{z},\bar{\mathbf{z}}]$$

Veeravalli [2006], whose notation we follow, calls C the "relation matrix." The Wikipedia article "Complex normal distribution" calls it the "pseudocovariance matrix." Because of the naïveté that leads many to a false conclusion, we prefer the 'pseudo' terminology (better, "pseudovariance") to something as bland as "relation matrix." However, a useful and non-pejorative concept is what we will call the "augmented variance."

The augmented variance is the variance of the complex random vector $\mathbf{z}$ augmented with its conjugate $\bar{\mathbf{z}}$, i.e., the $2n\times 1$ vector $\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}$. Its expectation is $E\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix} = \begin{bmatrix}E[\mathbf{z}]\\ E[\bar{\mathbf{z}}]\end{bmatrix} = \begin{bmatrix}\mu\\ \bar{\mu}\end{bmatrix}$. And its variance is (for brevity we ignore the mean):

$$Var\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix} = E\left[\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}^*\right] = E\left[\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{z}}' & \mathbf{z}'\end{bmatrix}\right] = \begin{bmatrix}Cov[\mathbf{z},\mathbf{z}] & Cov[\mathbf{z},\bar{\mathbf{z}}]\\ Cov[\bar{\mathbf{z}},\mathbf{z}] & Cov[\bar{\mathbf{z}},\bar{\mathbf{z}}]\end{bmatrix}$$

In two ways this matrix is redundant. First, $Cov[\bar{\mathbf{z}},\bar{\mathbf{z}}] = E[\bar{\mathbf{z}}\mathbf{z}'] = \overline{E[\mathbf{z}\bar{\mathbf{z}}']} = \overline{Cov[\mathbf{z},\mathbf{z}]}$; equivalently, $Var[\bar{\mathbf{z}}] = \overline{Var[\mathbf{z}]} = \bar{\Gamma}$. And second, $Cov[\bar{\mathbf{z}},\mathbf{z}] = E[\bar{\mathbf{z}}\mathbf{z}'] = \overline{E[\mathbf{z}\mathbf{z}']} = \overline{Cov[\mathbf{z},\bar{\mathbf{z}}]} = \bar{\mathrm{C}}$. Therefore:

$$Var\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix} = \begin{bmatrix}Cov[\mathbf{z},\mathbf{z}] & Cov[\mathbf{z},\bar{\mathbf{z}}]\\ Cov[\bar{\mathbf{z}},\mathbf{z}] & Cov[\bar{\mathbf{z}},\bar{\mathbf{z}}]\end{bmatrix} = \begin{bmatrix}\Gamma & \mathrm{C}\\ \bar{\mathrm{C}} & \bar{\Gamma}\end{bmatrix}$$

[7] In general, for any complex scalar α, $Var[\alpha\mathbf{z}] = \alpha\bar{\alpha}Var[\mathbf{z}]$.

As with any valid variance matrix, the augmented variance must be Hermetian. Hence, $\begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix} = \begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix}^* = \overline{\begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix}}' = \begin{bmatrix} \overline{\Gamma} & \overline{C} \\ C & \Gamma \end{bmatrix}' = \begin{bmatrix} \Gamma^* & C' \\ C^* & \Gamma' \end{bmatrix}$, from which follow $\Gamma^* = \Gamma$ and $C' = C$.

Moreover, it must be at least NND, if not PD. It is important to note from this that the pseudovariance is an essential part of the augmented **z**; it is possible for two random variables to have the same variance and to covary differently with their conjugates. How a complex random vector covaries with its conjugate is useful information; it is even a parameter of the general complex normal distribution, which we will treat next.

7. THE COMPLEX NORMAL DISTRIBUTION

All the information for deriving the complex normal distribution of $\mathbf{z}_{n\times 1} = \mathbf{x} + i\mathbf{y}$ is contained in the parameters of the real-valued multivariate normal distribution:

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim N\left(\mu_{2n\times 1} = \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix}, \Sigma_{2n\times 2n} = \begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{bmatrix} \right)$$

According to this variance structure, the real and imaginary parts of **z** may covary, as long as the covariance is symmetric: $\Sigma_{\mathbf{yx}} = \Sigma'_{\mathbf{xy}}$. The grand Σ matrix must be symmetric and PD. The probability density function of this multivariate normal is:[8]

$$f_{\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}}(\mathrm{x},\mathrm{y}) = \frac{1}{\sqrt{(2\pi)^{2n}|\Sigma|}} e^{-\frac{1}{2}\begin{bmatrix} \mathrm{x}'-\mu'_{\mathbf{x}} & \mathrm{y}'-\mu'_{\mathbf{y}} \end{bmatrix}\Sigma^{-1}\begin{bmatrix} \mathrm{x}-\mu_{\mathbf{x}} \\ \mathrm{y}-\mu_{\mathbf{y}} \end{bmatrix}} = \frac{1}{\pi^n\sqrt{2^{2n}|\Sigma|}} e^{-\frac{1}{2}\begin{bmatrix} \mathrm{x}'-\mu'_{\mathbf{x}} & \mathrm{y}'-\mu'_{\mathbf{y}} \end{bmatrix}\Sigma^{-1}\begin{bmatrix} \mathrm{x}-\mu_{\mathbf{x}} \\ \mathrm{y}-\mu_{\mathbf{y}} \end{bmatrix}}$$

[8] As derived briefly by Judge [1988, pp 49f]. Chapter 4 of Johnson [1992] is thorough. To be precise, $|\Sigma|$ under the radical should be $\|\Sigma\|$, the absolute value of the determinant of Σ. However, the determinant of a PD matrix must be positive (cf. Judge [1988, A.14(1)]).

Since $\mathbf{z} = \mathbf{x} + i\mathbf{y}$, the augmented vector is $\begin{bmatrix} \mathbf{z} \\ \overline{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} \mathbf{x} + i\mathbf{y} \\ \mathbf{x} - i\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \\ \mathrm{I}_n & -i\mathrm{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \Xi_n \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$. We will call $\Xi_n = \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \\ \mathrm{I}_n & -i\mathrm{I}_n \end{bmatrix}$ the augmentation matrix; this linear function of the real-valued vectors produces the complex vector and its conjugate. An important equation is:

$$\Xi_n \Xi_n^* = \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \\ \mathrm{I}_n & -i\mathrm{I}_n \end{bmatrix} \begin{bmatrix} \mathrm{I}_n & \mathrm{I}_n \\ -i\mathrm{I}_n & i\mathrm{I}_n \end{bmatrix} = \begin{bmatrix} 2\mathrm{I}_n & 0 \\ 0 & 2\mathrm{I}_n \end{bmatrix} = 2\mathrm{I}_{2n}$$

Therefore, Ξ_n has an inverse, viz., one half of its transjugate.

The augmented mean is $E\begin{bmatrix} \mathbf{z} \\ \overline{\mathbf{z}} \end{bmatrix} = \Xi_n \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mu_{\mathbf{x}} + i\mu_{\mathbf{y}} \\ \mu_{\mathbf{x}} - i\mu_{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mu \\ \overline{\mu} \end{bmatrix}$. The augmented variance is:

$$\begin{aligned} Var\begin{bmatrix} \mathbf{z} \\ \overline{\mathbf{z}} \end{bmatrix} &= \Xi_n \Sigma \Xi_n^* \\ &= \begin{bmatrix} \mathrm{I}_n & i\mathrm{I}_n \\ \mathrm{I}_n & -i\mathrm{I}_n \end{bmatrix} \begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{bmatrix} \begin{bmatrix} \mathrm{I}_n & \mathrm{I}_n \\ -i\mathrm{I}_n & i\mathrm{I}_n \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{\mathbf{xx}} + i\Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{xy}} + i\Sigma_{\mathbf{yy}} \\ \Sigma_{\mathrm{xx}} - i\Sigma_{\mathbf{yx}} & \Sigma_{\mathrm{xy}} - i\Sigma_{\mathbf{yy}} \end{bmatrix} \begin{bmatrix} \mathrm{I}_n & \mathrm{I}_n \\ -i\mathrm{I}_n & i\mathrm{I}_n \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{\mathbf{xx}} + \Sigma_{\mathbf{yy}} - i\left(\Sigma_{\mathbf{xy}} - \Sigma_{\mathbf{yx}}\right) & \Sigma_{\mathbf{xx}} - \Sigma_{\mathbf{yy}} + i\left(\Sigma_{\mathbf{xy}} + \Sigma_{\mathbf{yx}}\right) \\ \Sigma_{\mathbf{xx}} - \Sigma_{\mathbf{yy}} - i\left(\Sigma_{\mathbf{xy}} + \Sigma_{\mathbf{yx}}\right) & \Sigma_{\mathbf{xx}} + \Sigma_{\mathbf{yy}} + i\left(\Sigma_{\mathbf{xy}} - \Sigma_{\mathbf{yx}}\right) \end{bmatrix} \\ &= \begin{bmatrix} \Gamma & \mathrm{C} \\ \overline{\mathrm{C}} & \overline{\Gamma} \end{bmatrix} \end{aligned}$$

And so:

$$Var^{-1}\begin{bmatrix} \mathbf{z} \\ \overline{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} \Gamma & \mathrm{C} \\ \overline{\mathrm{C}} & \overline{\Gamma} \end{bmatrix}^{-1} = \left(\Xi_n \Sigma \Xi_n^*\right)^{-1} = \left(\Xi_n^*\right)^{-1} \Sigma^{-1} \left(\Xi_n\right)^{-1}$$

This can be reformulated as $\Xi_n^* Var^{-1}\begin{bmatrix} \mathbf{z} \\ \overline{\mathbf{z}} \end{bmatrix} \Xi_n = \Xi_n^* \begin{bmatrix} \Gamma & \mathrm{C} \\ \overline{\mathrm{C}} & \overline{\Gamma} \end{bmatrix}^{-1} \Xi_n = \Sigma^{-1}$.

We now work these augmented forms into the probability density function:

$$f_{\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}(\mathrm{x},\mathrm{y}) = \frac{1}{\pi^n\sqrt{2^{2n}|\Sigma|}}e^{-\frac{1}{2}\begin{bmatrix}\mathrm{x}'-\mu_{\mathbf{x}}' & \mathrm{y}'-\mu_{\mathbf{y}}'\end{bmatrix}\Sigma^{-1}\begin{bmatrix}\mathrm{x}-\mu_{\mathbf{x}}\\ \mathrm{y}-\mu_{\mathbf{y}}\end{bmatrix}}$$

$$= \frac{1}{\pi^n\sqrt{|2\mathrm{I}_{2n}||\Sigma|}}e^{-\frac{1}{2}\begin{bmatrix}\mathrm{x}'-\mu_{\mathbf{x}}' & \mathrm{y}'-\mu_{\mathbf{y}}'\end{bmatrix}\Xi_n^* Var^{-1}\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}\Xi_n\begin{bmatrix}\mathrm{x}-\mu_{\mathbf{x}}\\ \mathrm{y}-\mu_{\mathbf{y}}\end{bmatrix}}$$

$$= \frac{1}{\pi^n\sqrt{|\Xi_n\Xi_n^*||\Sigma|}}e^{-\frac{1}{2}\left(\Xi_n\begin{bmatrix}\mathrm{x}-\mu_{\mathbf{x}}\\ \mathrm{y}-\mu_{\mathbf{y}}\end{bmatrix}\right)^* Var^{-1}\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}\left(\Xi_n\begin{bmatrix}\mathrm{x}-\mu_{\mathbf{x}}\\ \mathrm{y}-\mu_{\mathbf{y}}\end{bmatrix}\right)}$$

$$= \frac{1}{\pi^n\sqrt{|\Xi_n\Sigma\Xi_n^*|}}e^{-\frac{1}{2}\begin{bmatrix}\bar{\mathrm{z}}'-\bar{\mu}' & \mathrm{z}'-\mu'\end{bmatrix}Var^{-1}\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}\begin{bmatrix}\mathrm{z}-\mu\\ \bar{\mathrm{z}}-\bar{\mu}\end{bmatrix}}$$

$$= \frac{1}{\pi^n\sqrt{\left|Var\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}\right|}}e^{-\frac{1}{2}\begin{bmatrix}\bar{\mathrm{z}}'-\bar{\mu}' & \mathrm{z}'-\mu'\end{bmatrix}Var^{-1}\begin{bmatrix}\mathbf{z}\\ \bar{\mathbf{z}}\end{bmatrix}\begin{bmatrix}\mathrm{z}-\mu\\ \bar{\mathrm{z}}-\bar{\mu}\end{bmatrix}}$$

However, this is not quite the density function of **z**, since the differential volume has not been considered. The correct formula is $f_{\mathbf{z}}(\mathrm{z})dV_{\mathrm{z}} = f_{\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}(\mathrm{x},\mathrm{y})dV_{\mathrm{xy}}$. The differential volume in the xy coordinates is $dV_{\mathrm{xy}} = \left|\prod_{j=1}^{n} dx_j dy_j\right|$. A change of dx_j entails an equal change in the real part of dz_j, even as a change of dy_j entails an equal change in the imaginary part of dz_j. Accordingly, $dV_{\mathrm{z}} = \left|\prod_{j=1}^{n} dx_j\left(i\cdot dy_j\right)\right| = \left|i^n\prod_{j=1}^{n} dx_j dy_j\right| = \left|i^n\right|\left|\prod_{j=1}^{n} dx_j dy_j\right| = 1\cdot\left|\prod_{j=1}^{n} dx_j dy_j\right| = dV_{\mathrm{xy}}$. It so happens that Ξ_n does not distort volume; but this had to be demonstrated.[9]

So finally, the probability density function of the complex random vector **z** is:

[9] This will be abstruse to some actuaries. However, the integration theory is implicit in the change-of-variables technique outlined in Hogg [1984, pp 42-46]. That the $n\times n$ determinant represents "the volume function of an n-dimensional parallelepiped" is beautifully explained in Chapter 4 of Schneider [1973].

$$f_{\mathbf{z}}(\mathrm{z}) = f_{\mathbf{z}}(\mathrm{z})\cdot 1 = f_{\mathbf{z}}(\mathrm{z})\frac{dV_{\mathrm{z}}}{dV_{\mathrm{xy}}} = f_{\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}(\mathrm{x},\mathrm{y})$$

$$= \frac{1}{\pi^n\sqrt{\left|Var\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix}\right|}} e^{-\frac{1}{2}\begin{bmatrix}\overline{\mathrm{z}'-\mu'} & \mathrm{z}'-\mu'\end{bmatrix}Var^{-1}\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix}\begin{bmatrix}\mathrm{z}-\mu\\ \overline{\mathrm{z}-\mu}\end{bmatrix}}$$

$$= \frac{1}{\pi^n\sqrt{\begin{vmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{vmatrix}}} e^{-\frac{1}{2}\begin{bmatrix}\overline{\mathrm{z}'-\mu'} & \mathrm{z}'-\mu'\end{bmatrix}\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{z}-\mu\\ \overline{\mathrm{z}-\mu}\end{bmatrix}}$$

This formula is equivalent to the one found in the Wikipedia article "Complex normal distribution." Although $|\Gamma|\left|\overline{\Gamma}-\overline{\mathrm{C}}\Gamma^{-1}\mathrm{C}\right|$ appears within the radical of that article's formula, it can be shown that $\begin{vmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{vmatrix} = |\Gamma|\left|\overline{\Gamma}-\overline{\mathrm{C}}\Gamma^{-1}\mathrm{C}\right|$. As far as allowable parameters are concerned, μ may be any complex vector. $\begin{vmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{vmatrix}$ is allowed if and only if $\Xi_n^*\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}\Xi_n = 4\Sigma$ is real-valued and PD.

Veeravalli [2006] defines a "proper" complex variable as one whose pseudo[co]variance matrix is $0_{n\times n}$. Inserting zero for C into the formula, we derive the probability density function of a proper complex random variable whose variance is Γ:

$$
\begin{aligned}
f_{\mathbf{z}}(\mathrm{z}) &= \frac{1}{\pi^n\sqrt{\begin{vmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{vmatrix}}} e^{-\frac{1}{2}\begin{bmatrix}\overline{\mathrm{z}}'-\overline{\mu}' & \mathrm{z}'-\mu'\end{bmatrix}\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{z}-\mu\\ \overline{\mathrm{z}}-\overline{\mu}\end{bmatrix}} \\
&= \frac{1}{\pi^n\sqrt{\begin{vmatrix}\Gamma & 0\\ 0 & \overline{\Gamma}\end{vmatrix}}} e^{-\frac{1}{2}\begin{bmatrix}\overline{\mathrm{z}}'-\overline{\mu}' & \mathrm{z}'-\mu'\end{bmatrix}\begin{bmatrix}\Gamma & 0\\ 0 & \overline{\Gamma}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{z}-\mu\\ \overline{\mathrm{z}}-\overline{\mu}\end{bmatrix}} \\
&= \frac{1}{\pi^n\sqrt{|\Gamma||\overline{\Gamma}|}} e^{-\frac{1}{2}\left\{(\overline{\mathrm{z}}'-\overline{\mu}')\Gamma^{-1}(\mathrm{z}-\mu)+(\mathrm{z}'-\mu')\overline{\Gamma}^{-1}(\overline{\mathrm{z}}-\overline{\mu})\right\}} \\
&= \frac{1}{\pi^n\sqrt{|\Gamma||\overline{\Gamma}|}} e^{-\frac{1}{2}\left\{(\overline{\mathrm{z}}'-\overline{\mu}')\Gamma^{-1}(\mathrm{z}-\mu)+\overline{(\mathrm{z}'-\mu')\Gamma^{-1}(\mathrm{z}'-\mu')}\right\}} \\
&= \frac{1}{\pi^n\sqrt{|\Gamma|^2}} e^{-\frac{1}{2}\left\{(\overline{\mathrm{z}}'-\overline{\mu}')\Gamma^{-1}(\mathrm{z}-\mu)+(\overline{\mathrm{z}}'-\overline{\mu}')\Gamma^{-1}(\mathrm{z}-\mu)\right\}} \\
&= \frac{1}{\pi^n|\Gamma|} e^{-(\overline{\mathrm{z}}'-\overline{\mu}')\Gamma^{-1}(\mathrm{z}-\mu)} \\
&= \frac{1}{\pi^n|\Gamma|} e^{-(\mathrm{z}-\mu)^*\Gamma^{-1}(\mathrm{z}-\mu)}
\end{aligned}
$$

The transformations in the last several lines rely on the fact that Γ is Hermetian and PD. Now the standard complex random vector is a proper complex random vector with mean zero and variance I_n. Therefore, in confirmation of Section 5, its density function is $\frac{1}{\pi^n}e^{-\mathrm{z}^*\mathrm{z}}$.

8. THE COMPLEX LOGNORMAL RANDOM VECTOR AND ITS MOMENTS

A complex lognormal random vector is the elementwise exponentiation of a complex normal random vector: $\mathbf{w}_{n\times 1} = e^{\mathbf{z}_{n\times 1}}$. Its conjugate also is lognormal, since $\overline{\mathbf{w}} = \overline{e^{\mathbf{z}}} = e^{\overline{\mathbf{z}}}$. Deriving the probability density function of $\mathbf{w}$ is precluded by the fact that $e: z \to w$ is many-to-one. Specifically, $e^z = w = e^{z+i(2\pi k)}$ for any integer *k*. So unlike the real-valued lognormal random variable, whose density function can be found in Klugman [1998, §A.4.11], an analytic form for the

complex lognormal density is not available. However, even for the real-valued lognormal the density function is of little value; its moments are commonly derived from the moment generating function of the normal variable on which it is based. So too, the moment generating function of the complex normal random vector is available for deriving the lognormal moments.

We hereby define the moment generating function of the complex $n\times1$ random vector $\mathbf{z}$ as $M_{\mathbf{z}}(\mathsf{s}_{n\times1}, \mathsf{t}_{n\times1}) = E\left[e^{\mathsf{s}'\mathbf{z}+\mathsf{t}'\overline{\mathbf{z}}}\right]$. Since this definition may differ from other definitions in the sparse literature, we should justify it. First, because we will take derivatives of this function with respect to s and t, the function must be differentiable. This demands simple transposition in the linear combination, i.e., $\mathsf{s}'\mathbf{z}+\mathsf{t}'\overline{\mathbf{z}}$ rather than the transjugation $\mathsf{s}^*\mathbf{z}+\mathsf{t}^*\overline{\mathbf{z}}$. For transjugation would involve derivatives of the form $\frac{d\overline{s}}{ds}$, which do not exist, as they violate the Cauchy-Riemann condition.[10] Second, even though moments of $\overline{\mathbf{z}}$ are conjugates of moments of $\mathbf{z}$, we will need second-order moments involving both $\mathbf{z}$ and $\overline{\mathbf{z}}$. For this reason both terms must be in the exponent of the moment generating function.

[10] Cf. Appendix D.1.3 of Havil [2003]. Express $f(z = x+iy)$ in terms of real-valued functions, i.e., as $u(x, y)+i\cdot v(x, y)$. The derivative is based on the matrix of real-valued partial derivatives $\begin{bmatrix}\partial u/\partial x & \partial v/\partial x\\ \partial u/\partial y & \partial v/\partial y\end{bmatrix}$. For the derivative to be the same in both directions, the Cauchy-Riemann condition must hold, viz., that $\partial u/\partial x = \partial v/\partial y$ and $\partial v/\partial x = -\partial u/\partial y$. But for $f(z)=\overline{z}=x-iy$, the partial-derivative matrix is $\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}$; hence $\partial u/\partial x \neq \partial v/\partial y$. The Cauchy-Riemann condition becomes intuitive when one regards a valid complex derivative as a double-real 2×2 matrix (Section 3). Compare this with $f(z)=z=x+iy$, whose matrix is $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, which represents the complex number 1.

We start with terminology from Section 7, viz., that $\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix}=\begin{bmatrix}\mathbf{x}+i\mathbf{y}\\ \mathbf{x}-i\mathbf{y}\end{bmatrix}=\begin{bmatrix}\mathrm{I}_n & i\mathrm{I}_n\\ \mathrm{I}_n & -i\mathrm{I}_n\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}=\Xi_n\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ and that $\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\sim N\left(\mu_{2n\times 1}=\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix},\Sigma_{2n\times 2n}=\begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}\right)$. According to Appendix A, the moment generating function of the real-valued normal random vector $\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ is:

$$M_{\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}\left(\begin{bmatrix}\mathrm{a}\\ \mathrm{b}\end{bmatrix}\right)=E\left[e^{\begin{bmatrix}\mathrm{a}' & \mathrm{b}'\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}\right]=e^{\begin{bmatrix}\mathrm{a}' & \mathrm{b}'\end{bmatrix}\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix}+\frac{1}{2}\begin{bmatrix}\mathrm{a}' & \mathrm{b}'\end{bmatrix}\begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}\begin{bmatrix}\mathrm{a}\\ \mathrm{b}\end{bmatrix}}$$

Consequently:

$$\begin{aligned}M_{\mathbf{z}}(\mathrm{s},\mathrm{t})&=E\left[e^{\mathrm{s}'\mathbf{z}+\mathrm{t}'\overline{\mathbf{z}}}\right]\\&=E\left[e^{\begin{bmatrix}\mathrm{s}' & \mathrm{t}'\end{bmatrix}\begin{bmatrix}\mathrm{I}_n & i\mathrm{I}_n\\ \mathrm{I}_n & -i\mathrm{I}_n\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}\right]\\&=E\left[e^{\begin{bmatrix}\mathrm{s}'+\mathrm{t}' & i(\mathrm{s}'-\mathrm{t}')\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}\right]\\&=M_{\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}}\left(\begin{bmatrix}\mathrm{s}+\mathrm{t}\\ i(\mathrm{s}-\mathrm{t})\end{bmatrix}\right)\\&=e^{\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix}+\frac{1}{2}\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}\begin{bmatrix}\mathrm{s}+\mathrm{t}\\ i(\mathrm{s}-\mathrm{t})\end{bmatrix}}\end{aligned}$$

It is so that we could invoke it here that Appendix B went to the trouble of proving that complex values are allowed in this moment generating function.

But in two ways we can simplify this expression. First:

$$\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix}=(\mathrm{s}+\mathrm{t})'\mu_{\mathbf{x}}+i(\mathrm{s}-\mathrm{t})'\mu_{\mathbf{y}}=\mathrm{s}'(\mu_{\mathbf{x}}+i\mu_{\mathbf{y}})+\mathrm{t}'(\mu_{\mathbf{x}}-i\mu_{\mathbf{y}})=\mathrm{s}'\mu_{\mathbf{z}}+\mathrm{t}'\overline{\mu}_{\mathbf{z}}$$

And second, again from Section 7, $Var\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix}=\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}=\Xi_n\Sigma_{2n\times 2n}\Xi_n^*=\Xi_n\begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}\Xi_n^*$, or equivalently, $\frac{\Xi_n^*}{2}\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}\frac{\Xi_n}{2}=\begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}$. Hence:

$$\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}\begin{bmatrix}\mathrm{s}+\mathrm{t}\\ i(\mathrm{s}-\mathrm{t})\end{bmatrix}=\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\frac{\Xi_n^*}{2}\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}\frac{\Xi_n}{2}\begin{bmatrix}\mathrm{s}+\mathrm{t}\\ i(\mathrm{s}-\mathrm{t})\end{bmatrix}$$

On the right side, $\frac{\Xi_n}{2}\begin{bmatrix}\mathrm{s}+\mathrm{t}\\ i(\mathrm{s}-\mathrm{t})\end{bmatrix}=\frac{1}{2}\begin{bmatrix}\mathrm{I}_n & i\mathrm{I}_n\\ \mathrm{I}_n & -i\mathrm{I}_n\end{bmatrix}\begin{bmatrix}\mathrm{s}+\mathrm{t}\\ i(\mathrm{s}-\mathrm{t})\end{bmatrix}=\frac{1}{2}\begin{bmatrix}(\mathrm{s}+\mathrm{t})-(\mathrm{s}-\mathrm{t})\\ (\mathrm{s}+\mathrm{t})+(\mathrm{s}-\mathrm{t})\end{bmatrix}=\begin{bmatrix}\mathrm{t}\\ \mathrm{s}\end{bmatrix}$. And on the left side, $\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\frac{\Xi_n^*}{2}=\frac{1}{2}\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\begin{bmatrix}\mathrm{I}_n & \mathrm{I}_n\\ -i\mathrm{I}_n & i\mathrm{I}_n\end{bmatrix}=\begin{bmatrix}\mathrm{s}' & \mathrm{t}'\end{bmatrix}$. So:

$$\begin{bmatrix}(\mathrm{s}+\mathrm{t})' & i(\mathrm{s}-\mathrm{t})'\end{bmatrix}\begin{bmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{bmatrix}\begin{bmatrix}\mathrm{s}+\mathrm{t}\\ i(\mathrm{s}-\mathrm{t})\end{bmatrix}=\begin{bmatrix}\mathrm{s}' & \mathrm{t}'\end{bmatrix}\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}\begin{bmatrix}\mathrm{t}\\ \mathrm{s}\end{bmatrix}$$

And the simplified expression, based on the mean and the augmented variance of $\mathbf{z}$, is:

$$\begin{aligned}M_{\mathbf{z}}(\mathrm{s},\mathrm{t})&=E\left[e^{\mathrm{s}'\mathbf{z}+\mathrm{t}'\overline{\mathbf{z}}}\right]\\&=e^{\mathrm{s}'\mu_{\mathbf{z}}+\mathrm{t}'\overline{\mu}_{\mathbf{z}}+\frac{1}{2}\begin{bmatrix}\mathrm{s}' & \mathrm{t}'\end{bmatrix}\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}\begin{bmatrix}\mathrm{t}\\ \mathrm{s}\end{bmatrix}}\\&=e^{\mathrm{s}'\mu+\mathrm{t}'\overline{\mu}+\frac{1}{2}(\mathrm{s}'\Gamma\mathrm{t}+\mathrm{s}'\mathrm{C}\mathrm{s}+\mathrm{t}'\overline{\mathrm{C}}\mathrm{t}+\mathrm{t}'\overline{\Gamma}\mathrm{s})}\end{aligned}$$

As in Appendix A, let e_j denote the j^{th} unit vector of $\Re^n$, or even better, of C^n. Then:

$$\begin{aligned}E\left[e^{\mathbf{z}_j}\right]&=M_{\mathbf{z}}(\mathrm{e}_j,0)=e^{\mu_j+\frac{1}{2}\mathrm{C}_{jj}}\\E\left[\overline{e^{\mathbf{z}_j}}\right]&=E\left[e^{\overline{\mathbf{z}}_j}\right]=M_{\mathbf{z}}(0,\mathrm{e}_j)=e^{\overline{\mu}_j+\frac{1}{2}\overline{\mathrm{C}}_{jj}}=\overline{E\left[e^{\mathbf{z}_j}\right]}\end{aligned}$$

Moreover:

$$E\left[e^{\mathbf{z}_j}e^{\mathbf{z}_k}\right]=E\left[e^{\mathbf{z}_j+\mathbf{z}_k}\right]=M_{\mathbf{z}}(\mathrm{e}_j+\mathrm{e}_k,0)=e^{\mu_j+\mu_k+\frac{1}{2}(\mathrm{C}_{jj}+\mathrm{C}_{jk}+\mathrm{C}_{kj}+\mathrm{C}_{kk})}$$

According to Section 6, C is symmetric ($\mathrm{C}' = \mathrm{C}$). This and further simplification leads to:

$$E\left[e^{\mathbf{z}_j} e^{\mathbf{z}_k}\right] = e^{\mu_j + \mu_k + \frac{1}{2}\left(\mathrm{C}_{jj} + \mathrm{C}_{jk} + \mathrm{C}_{kj} + \mathrm{C}_{kk}\right)} = e^{\left(\mu_j + \frac{1}{2}\mathrm{C}_{jj}\right) + \left(\mu_k + \frac{1}{2}\mathrm{C}_{kk}\right) + \frac{1}{2}\left(\mathrm{C}_{jk} + \mathrm{C}_{jk}\right)} = E\left[e^{\mathbf{z}_j}\right] E\left[e^{\mathbf{z}_k}\right] \cdot e^{\mathrm{C}_{jk}}$$

Hence, mindful of the transjugation (*) in the definition of complex covariance, we have:

$$Cov\left[e^{\mathbf{z}_j}, \overline{e^{\mathbf{z}_k}}\right] = E\left[e^{\mathbf{z}_j}\,\overline{\overline{e^{\mathbf{z}_k}}}\right] - E\left[e^{\mathbf{z}_j}\right] E\left[\overline{\overline{e^{\mathbf{z}_k}}}\right] = E\left[e^{\mathbf{z}_j} e^{\mathbf{z}_k}\right] - E\left[e^{\mathbf{z}_j}\right] E\left[e^{\mathbf{z}_k}\right] = E\left[e^{\mathbf{z}_j}\right] E\left[e^{\mathbf{z}_k}\right] \cdot \left(e^{\mathrm{C}_{jk}} - 1\right)$$

In terms of $\mathbf{w} = e^{\mathbf{z}}$ this translates as $Cov[\mathbf{w}, \overline{\mathbf{w}}] = \left(E[\mathbf{w}] E[\mathbf{w}]'\right) \circ \left(e^{\mathrm{C}} - 1_{n \times n}\right)$, in which the '$\circ$' operator represents elementwise multiplication.[11] So too:

$$Cov\left[\overline{e^{\mathbf{z}_j}}, e^{\mathbf{z}_k}\right] = \overline{Cov\left[e^{\mathbf{z}_j}, \overline{e^{\mathbf{z}_k}}\right]} = E\left[\overline{e^{\mathbf{z}_j}}\right] E\left[\overline{e^{\mathbf{z}_k}}\right] \cdot \left(e^{\overline{\mathrm{C}}_{jk}} - 1\right)$$

This translates as $Cov[\overline{\mathbf{w}}, \mathbf{w}] = \left(E[\overline{\mathbf{w}}] E[\overline{\mathbf{w}}]'\right) \circ \left(e^{\overline{\mathrm{C}}} - 1_{n \times n}\right)$.

The remaining combination is the mixed form $E\left[e^{\mathbf{z}_j}\,\overline{e^{\mathbf{z}_k}}\right]$:

$$E\left[e^{\mathbf{z}_j}\,\overline{e^{\mathbf{z}_k}}\right] = E\left[e^{\mathbf{z}_j} e^{\overline{\mathbf{z}}_k}\right] = E\left[e^{\mathbf{z}_j + \overline{\mathbf{z}}_k}\right] = M_{\mathbf{z}}\left(\mathrm{e}_j, \mathrm{e}_k\right) = e^{\mu_j + \overline{\mu}_k + \frac{1}{2}\left(\Gamma_{jk} + \mathrm{C}_{jj} + \overline{\mathrm{C}}_{kk} + \overline{\Gamma}_{kj}\right)}$$

Since Γ is Hermetian, $\overline{\Gamma}_{kj} = \overline{\Gamma}'_{jk} = \Gamma^*_{jk} = \Gamma_{jk}$. Hence:

$$E\left[e^{\mathbf{z}_j}\,\overline{e^{\mathbf{z}_k}}\right] = e^{\mu_j + \overline{\mu}_k + \frac{1}{2}\left(\Gamma_{jk} + \mathrm{C}_{jj} + \overline{\mathrm{C}}_{kk} + \overline{\Gamma}_{kj}\right)} = e^{\left(\mu_j + \frac{1}{2}\mathrm{C}_{jj}\right) + \left(\overline{\mu}_k + \frac{1}{2}\overline{\mathrm{C}}_{kk}\right) + \frac{1}{2}\left(\Gamma_{jk} + \Gamma_{jk}\right)} = E\left[e^{\mathbf{z}_j}\right] E\left[\overline{e^{\mathbf{z}_k}}\right] \cdot e^{\Gamma_{jk}}$$

Therefore, $Cov\left[e^{\mathbf{z}_j}, e^{\mathbf{z}_k}\right] = E\left[e^{\mathbf{z}_j}\,\overline{e^{\mathbf{z}_k}}\right] - E\left[e^{\mathbf{z}_j}\right] E\left[\overline{e^{\mathbf{z}_k}}\right] = E\left[e^{\mathbf{z}_j}\right] E\left[\overline{e^{\mathbf{z}_k}}\right] \cdot \left(e^{\Gamma_{jk}} - 1\right)$, which translates as $Cov[\mathbf{w}, \mathbf{w}] = \left(E[\mathbf{w}] E[\overline{\mathbf{w}}]'\right) \circ \left(e^{\Gamma} - 1_{n \times n}\right)$. By conjugation, $Cov\left[\overline{e^{\mathbf{z}_j}}, \overline{e^{\mathbf{z}_k}}\right] = E\left[\overline{e^{\mathbf{z}_j}}\right] E\left[e^{\mathbf{z}_k}\right] \cdot \left(e^{\overline{\Gamma}_{jk}} - 1\right)$, which translates as $Cov[\overline{\mathbf{w}}, \overline{\mathbf{w}}] = \left(E[\overline{\mathbf{w}}] E[\mathbf{w}]'\right) \circ \left(e^{\overline{\Gamma}} - 1_{n \times n}\right)$.

[11] Elementwise multiplication is formally known as the Hadamard, or Hadamard-Schur, product, of which we will make use in Appendices A and C. Cf. Million [2007].

We conclude this section by expressing it all in terms of $\mathbf{w}_{n\times 1} = e^{\mathbf{z}_{n\times 1}}$. Let z be complex normal with mean μ and augmented variance $Var\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix} = \begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}$. And let D be the $n\times 1$ vector consisting of the main diagonal of C. Then $\overline{\mathrm{D}}$ is the vectorization of the diagonal of $\overline{\mathrm{C}}$. So the augmented mean of $\mathbf{w}$ is $E\begin{bmatrix}\mathbf{w}\\ \overline{\mathbf{w}}\end{bmatrix} = \begin{bmatrix}e^{\mu+\mathrm{D}/2}\\ e^{\overline{\mu}+\overline{\mathrm{D}}/2}\end{bmatrix}$. And the augmented variance of $\mathbf{w}$ is:

$$
\begin{aligned}
Var\begin{bmatrix}\mathbf{w}\\ \overline{\mathbf{w}}\end{bmatrix} &= \begin{bmatrix}Cov[\mathbf{w},\mathbf{w}] & Cov[\mathbf{w},\overline{\mathbf{w}}]\\ Cov[\overline{\mathbf{w}},\mathbf{w}] & Cov[\overline{\mathbf{w}},\overline{\mathbf{w}}]\end{bmatrix}\\
&= \begin{bmatrix}\left(E[\mathbf{w}]E[\overline{\mathbf{w}}]'\right)\circ\left(e^{\Gamma}-1_{n\times n}\right) & \left(E[\mathbf{w}]E[\mathbf{w}]'\right)\circ\left(e^{\mathrm{C}}-1_{n\times n}\right)\\ \left(E[\overline{\mathbf{w}}]E[\overline{\mathbf{w}}]'\right)\circ\left(e^{\overline{\mathrm{C}}}-1_{n\times n}\right) & \left(E[\overline{\mathbf{w}}]E[\mathbf{w}]'\right)\circ\left(e^{\overline{\Gamma}}-1_{n\times n}\right)\end{bmatrix}\\
&= \begin{pmatrix}E[\mathbf{w}]E[\overline{\mathbf{w}}]' & E[\mathbf{w}]E[\mathbf{w}]'\\ E[\overline{\mathbf{w}}]E[\overline{\mathbf{w}}]' & E[\overline{\mathbf{w}}]E[\mathbf{w}]'\end{pmatrix}\circ\left(e^{\begin{bmatrix}\Gamma & \mathrm{C}\\ \overline{\mathrm{C}} & \overline{\Gamma}\end{bmatrix}}-1_{2n\times 2n}\right)\\
&= \left(E\begin{bmatrix}\mathbf{w}\\ \overline{\mathbf{w}}\end{bmatrix}E\begin{bmatrix}\overline{\mathbf{w}} & \mathbf{w}\end{bmatrix}'\right)\circ\left(e^{Var\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix}}-1_{2n\times 2n}\right)\\
&= \left(E\begin{bmatrix}\mathbf{w}\\ \overline{\mathbf{w}}\end{bmatrix}E\begin{bmatrix}\mathbf{w}\\ \overline{\mathbf{w}}\end{bmatrix}^{*}\right)\circ\left(e^{Var\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix}}-1_{2n\times 2n}\right)
\end{aligned}
$$

Scaling all the lognormal means to unity (or setting $\mu = -\mathrm{D}/2$), we can say that the coefficient-of-lognormal-augmented-variation matrix equals $e^{Var\begin{bmatrix}\mathbf{z}\\ \overline{\mathbf{z}}\end{bmatrix}}-1_{2n\times 2n}$, which is analogous with the well-known coefficient of lognormal variation $e^{\sigma^2}-1$.

9. THE COMPLEX LOGNORMAL RANDOM VARIABLE

The previous section derived the augmented mean and variance of the lognormal random vector; this section provides some intuition into it. The complex lognormal random variable, or scalar,

derives from the real-valued normal bivariate $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{bmatrix}\right)$. Zero is not much of a restriction; since $e^{CN(\mu, \nu)} = e^{\mu + CN(0, \nu)} = e^{\mu} \circ e^{CN(0, \nu)}$, the normal mean affects only the scale of the lognormal. The variance is written in correlation form, where $-1 \le \rho \le 1$. As usual, $0 < \sigma, \tau < \infty$.

Define $\begin{bmatrix} Z \\ \overline{Z} \end{bmatrix} = \Xi_1 \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X + iY \\ X - iY \end{bmatrix}$. Its mean is zero, and according to Section 7 its variance (the augmented variance) is:

$$\begin{aligned} Var\begin{bmatrix} Z \\ \overline{Z} \end{bmatrix} &= \begin{bmatrix} 1 & i \\ 1 & -i \end{bmatrix} \begin{bmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 + \tau^2 - i(\rho\sigma\tau - \rho\sigma\tau) & \sigma^2 - \tau^2 + i(\rho\sigma\tau + \rho\sigma\tau) \\ \sigma^2 - \tau^2 - i(\rho\sigma\tau + \rho\sigma\tau) & \sigma^2 + \tau^2 + i(\rho\sigma\tau - \rho\sigma\tau) \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 + \tau^2 & \sigma^2 - \tau^2 + 2i\rho\sigma\tau \\ \sigma^2 - \tau^2 - 2i\rho\sigma\tau & \sigma^2 + \tau^2 \end{bmatrix} \end{aligned}$$

We will say little about non-zero correlation ($\rho \neq 0$); but at this point a digression on complex correlation is apt. The coefficient of correlation between *Z* and its conjugate is:

$$\rho_{Z\overline{Z}} = \frac{\sigma^2 - \tau^2 + 2i\rho\sigma\tau}{\sigma^2 + \tau^2} = \frac{\overline{\sigma^2 - \tau^2 - 2i\rho\sigma\tau}}{\sigma^2 + \tau^2} = \overline{\rho_{\overline{Z}Z}}$$

As a form of covariance, correlation is Hermetian. Moreover:

$$0 \le \rho_{Z\overline{Z}}\overline{\rho_{Z\overline{Z}}} = \rho_{Z\overline{Z}}\rho_{\overline{Z}Z} = \frac{\left(\sigma^2 - \tau^2\right)^2 + 4\rho^2\sigma^2\tau^2}{\left(\sigma^2 + \tau^2\right)^2} \le \frac{\left(\sigma^2 - \tau^2\right)^2 + 4(1)\sigma^2\tau^2}{\left(\sigma^2 + \tau^2\right)^2} = \frac{\left(\sigma^2 + \tau^2\right)^2}{\left(\sigma^2 + \tau^2\right)^2} = 1$$

So, the magnitude of complex correlation is not greater than unity. The imaginary part of the correlation is zero unless some correlation exists between the real and imaginary parts of the underlying bivariate. More interesting are the two limits: $\lim_{\tau^2 \to 0^+} \rho_{Z\overline{Z}} = 1$ and $\lim_{\sigma^2 \to 0^+} \rho_{Z\overline{Z}} = -1$. In the

first case, $\overline{Z} \to Z$ in a statistical sense, and the correlation approaches one. In the second case, $\overline{Z} \to -Z$, and the correlation approaches negative one.

Now if $W = e^Z$, by the formulas of Section 8, $E[W] = e^{0+(\sigma^2-\tau^2+2i\rho\sigma\tau)/2} = e^{(\sigma^2-\tau^2)/2} \cdot e^{i\rho\sigma\tau}$ and $E[\overline{W}] = e^{(\sigma^2-\tau^2)/2} \cdot e^{-i\rho\sigma\tau}$. And the augmented variance is:

$$
\begin{aligned}
Var\left[\frac{W}{\overline{W}}\right] &= \left(E\left[\frac{W}{\overline{W}}\right]E\left[\frac{W}{\overline{W}}\right]^*\right) \circ \left(e^{Var\left[\frac{Z}{\overline{Z}}\right]} - 1_{2\times 2}\right) \\
&= \left(\begin{bmatrix} e^{(\sigma^2-\tau^2)/2} \cdot e^{i\rho\sigma\tau} \\ e^{(\sigma^2-\tau^2)/2} \cdot e^{-i\rho\sigma\tau} \end{bmatrix}\begin{bmatrix} e^{(\sigma^2-\tau^2)/2} \cdot e^{-i\rho\sigma\tau} & e^{(\sigma^2-\tau^2)/2} \cdot e^{i\rho\sigma\tau} \end{bmatrix}\right) \circ \left(e^{\begin{bmatrix} \sigma^2+\tau^2 & \sigma^2-\tau^2+i\cdot 2\rho\sigma\tau \\ \sigma^2-\tau^2-i\cdot 2\rho\sigma\tau & \sigma^2+\tau^2 \end{bmatrix}} - 1_{2\times 2}\right) \\
&= e^{\sigma^2-\tau^2}\begin{bmatrix} 1 & e^{2i\rho\sigma\tau} \\ e^{-2i\rho\sigma\tau} & 1 \end{bmatrix} \circ \begin{bmatrix} e^{\sigma^2+\tau^2}-1 & e^{\sigma^2-\tau^2+2i\rho\sigma\tau}-1 \\ e^{\sigma^2-\tau^2-2i\rho\sigma\tau}-1 & e^{\sigma^2+\tau^2}-1 \end{bmatrix} \\
&= e^{\sigma^2-\tau^2}\begin{bmatrix} e^{\sigma^2+\tau^2}-1 & e^{\sigma^2-\tau^2+4i\rho\sigma\tau}-e^{2i\rho\sigma\tau} \\ e^{\sigma^2-\tau^2-4i\rho\sigma\tau}-e^{-2i\rho\sigma\tau} & e^{\sigma^2+\tau^2}-1 \end{bmatrix} \\
&= \begin{bmatrix} e^{2\sigma^2}-e^{\sigma^2-\tau^2} & e^{2\sigma^2-2\tau^2+4i\rho\sigma\tau}-e^{\sigma^2-\tau^2}\cdot e^{2i\rho\sigma\tau} \\ e^{2\sigma^2-2\tau^2-4i\rho\sigma\tau}-e^{\sigma^2-\tau^2}\cdot e^{-2i\rho\sigma\tau} & e^{2\sigma^2}-e^{\sigma^2-\tau^2} \end{bmatrix}
\end{aligned}
$$

In the first case above, as $\tau^2 \to 0^+$, $E\left[\frac{W}{\overline{W}}\right] \to e^{\sigma^2/2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $Var\left[\frac{W}{\overline{W}}\right] \to e^{\sigma^2}\left(e^{\sigma^2}-1\right)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Since the complex part Y becomes probability-limited to its mean of zero, the complex lognormal degenerates to the real-valued $W = e^X$. The limiting result is oblivious to the underlying correlation ϱ, since $\overline{W} \to W$.

In the second case, as $\sigma^2 \to 0^+$, $E\begin{bmatrix} W \\ \overline{W} \end{bmatrix} \to e^{-\tau^2/2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $Var\begin{bmatrix} W \\ \overline{W} \end{bmatrix} \to e^{-\tau^2}\begin{bmatrix} e^{\tau^2}-1 & e^{-\tau^2}-1 \\ e^{-\tau^2}-1 & e^{\tau^2}-1 \end{bmatrix}$. As in the first case, both $E[\overline{W}] = E[W]$ and the underlying correlation ϱ has disappeared. Nevertheless, the variance shows *W* and its conjugate to differ; in fact, their correlation is the real-valued $\rho_{W\overline{W}} = \left(e^{-\tau^2}-1\right)\Big/\left(e^{\tau^2}-1\right) = -e^{-\tau^2} = -E[W]$. Since $\tau^2 > 0$, $-1 < \rho_{W\overline{W}} < 0$ and $0 < E[\overline{W}] = E[W] < 1$.

Both these cases are understandable from the "geometry" of $W = e^Z = e^{X+iY} = e^X e^{iY}$. The complex exponential function is the basis of polar coordinates; e^X is the magnitude of *W*, and *Y* is the angle of *W* in radians counterclockwise from the real axis of the complex plane. Imagine a canon whose angle and range can be set. In the first case, the angle is fixed at zero, but the range is variable. This makes for a lognormal distribution along the positive real axis. In the second case, the canon's angle varies, but its range is fixed at $e^0 = 1$. This makes all the shots to land on the complex unit circle; hence, their mean lies within the circle, i.e., $|E[W]| < 1$. Moreover, the symmetry of *Y* as $N(0, \tau^2)$-distributed guarantees $E[W]$ to fall on the real axis, or $-1 < E[W] < 1$. Furthermore, since the normal density function strictly decreases in both directions from the mean, more shots land to the right of the imaginary axis than to the left, so $0 < E[W] = e^{-\tau^2} < 1$. A "right-handed" canon, or a canon whose angle is measured clockwise from the real axis, fires $\overline{W} = e^X e^{-iY}$ shots.

A shot from an unrestricted canon will "almost surely" not land on the real axis.[12] If we desire negative values from the complex lognormal random variable, as a practical matter we must extract them from its real or complex parts, e.g., $U = \mathrm{Re}(W)$. One can see in the second case, that as τ^2 grows larger, so too grows larger the probability that $U < 0$. As $\tau^2 \to \infty$, the probability approaches one half. In the limit, the shots are uniformly distributed around the complex unit circle. In this specialized case ($\sigma^2 \to 0^+$ and $\tau^2 \to \infty$), the distribution of $U = \mathrm{Re}(W)$ is

$f_U(u) = \dfrac{1}{\pi\sqrt{1-u^2}}$, for $-1 \le u \le 1$.[13]

This suggests a third case, in which $\tau^2 \to \infty$ while σ^2 remains at some positive amount. An intriguing feature of complex variables is that infinite variance in *Y* leads to a uniform distribution of e^{iY}.[14] So if $W = e^Z = e^X e^{iY}$, $U = \mathrm{Re}(W) = e^X \cos Y$ will be something of a reflected lognormal; both its tails will be as heavy as the lognormal's.[15] In this case:

$$E\left[\frac{W}{\overline{W}}\right] = \lim_{\tau^2\to\infty}\begin{bmatrix} e^{\left(\sigma^2-\tau^2+2i\rho\sigma\tau\right)/2} \\ e^{\left(\sigma^2-\tau^2+2i\rho\sigma\tau\right)/2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$Var\left[\frac{W}{\overline{W}}\right] = \lim_{\tau^2\to\infty}\begin{bmatrix} e^{2\sigma^2} - e^{\sigma^2-\tau^2} & e^{2\sigma^2-2\tau^2+4i\rho\sigma\tau} - e^{\sigma^2-\tau^2}\cdot e^{2i\rho\sigma\tau} \\ e^{2\sigma^2-2\tau^2-4i\rho\sigma\tau} - e^{\sigma^2-\tau^2}\cdot e^{-2i\rho\sigma\tau} & e^{2\sigma^2} - e^{\sigma^2-\tau^2} \end{bmatrix} = \begin{bmatrix} e^{2\sigma^2} & 0 \\ 0 & e^{2\sigma^2} \end{bmatrix}$$

Again, ϱ has disappeared from the limiting distribution; but in this case $\rho_{W\overline{W}} = 0$.

---

[12] For an event almost surely to happen means that its probability is unity; for an event almost surely not to happen means that its probability is zero. The latter case means not that the event will not happen, but rather that the event has zero probability mass. For example, if $X \sim$ Uniform[0, 1], Prob[X=½] = 0. So X almost surely does not equal ½, even though ½ is as possible as any other number in the interval.

[13] For more on this bimodal Arcsine(-1, 1) distribution see Wikipedia, "Arcsine distribution."

[14] The next section expands on this important subject. "Infinite variance in *Y*" means "as the variance of Y approaches infinity." It does not mean that e^{iY} is uniform for a variable *Y* whose variance is infinite, e.g., for a Pareto random variable whose shape parameter is less than or equal to two.

[15] Cf. Halliwell [2013] for a discussion on the right tails of the lognormal and other loss distributions.

In practical work with $U = \mathrm{Re}(W)$,[16] the angular part e^{iY} will be more important than the lognormal range e^{X}. For example, one who wanted the tendency for the larger magnitudes of $U = \mathrm{Re}(W)$ to be positive might set the mean of *Y* at $-\pi/2$ and the correlation ϱ to some positive value. Thus, greater than average values of *Y*, angling off into quadrants 4 and 1 of the complex plane, would correlate with larger than average values of *X* and hence of e^{X}. Of course, $Var[Y] = \tau^2$ would have to be small enough that deviations of $\pm\pi$ from $E[Y] = -\pi/2$ would be tolerably rare. Equivalently, one could set the mean of *Y* at $\pi/2$ and the correlation ϱ to some negative value. As a second example, one who wanted negative values of *U* to be less frequent than positive, might set both the mean of *Y* and ϱ to zero, and set the variance of *Y* so that $\mathrm{Prob}[|Y| > \pi/2]$ is desirably small. Some distributions of *U* for $\tau^2 >> \sigma^2$ are bimodal, as in the specialized case $\sigma^2 \to 0^+$ and $\tau^2 \to \infty$. But less extreme parameters would result in unimodal distributions for *U* over the entire real number line.

10. THE COMPLEX UNIT-CIRCLE RANDOM VARIABLE

In the previous section we claimed that as the variance τ^2 of the normal random variable *Y* approaches infinity, e^{iY} approaches a uniform distribution over the complex unit circle. The explanation and justification of this claim in this section prepare for an important implication in the next.

Let real-valued random variable *Y* be distributed as $\mathrm{N}[\mu, \sigma^2]$, and let $W = e^{iY}$. According to the moment-generating-formula of Section 8, $M_Y(it) = E[e^{itY}] = e^{it\mu + (it)^2\sigma^2/2} = e^{it\mu - t^2\sigma^2/2}$. Although the

[16] In the absence of an analytic distribution, practical work with the complex lognormal would seem to require simulating its values from the underlying normal distribution.

formula applies to complex values of *t*, here we'll restrict it to real values. With $t \in \Re$ $M_Y(it)$ is known as the characteristic function of real variable *Y*. And so:

$$\lim_{\sigma^2 \to \infty} M_Y(it) = \lim_{\sigma^2 \to \infty} e^{it\mu - t^2\sigma^2/2} = e^{it\mu} \lim_{\sigma^2 \to \infty} e^{-t^2\sigma^2/2} = \delta_{t0} = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}$$

It is noteworthy, and indicative of a uniformity of some sort, that μ drops out of the result.

Next, let real-valued random variable Θ be uniformly distributed over $[a, a + 2\pi n]$, where *n* is a positive integer; in symbols, $\Theta \sim \mathrm{U}[a, a + 2\pi n]$. Then:

$$\begin{aligned} M_\Theta(it) &= E\left[e^{it\Theta}\right] \\ &= \int_{\theta=a}^{a+2\pi n} e^{it\theta} \frac{1}{2\pi n} d\theta \\ &= \left. \frac{e^{it\theta}}{2\pi i t n} \right|_a^{a+2\pi n} \\ &= e^{ita} \frac{e^{2\pi i t n} - 1}{2\pi i t n} \\ &= \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } tn = \pm 1, \pm 2, \ldots \\ \neq 0 & \text{if } tn \text{ not integral} \end{cases} \end{aligned}$$

Letting *n* approach infinity, we have:

$$\lim_{n \to \infty} M_\Theta(it) = e^{ita} \lim_{n \to \infty} \frac{e^{2\pi i t n} - 1}{2\pi i t n} = \delta_{t0} = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t = 0 \end{cases}$$

Hence, $\lim_{n \to \infty} M_\Theta(it) = \lim_{\sigma^2 \to \infty} M_Y(it) = \delta_{t0}$. The equality of the limits of the characteristic functions of the random variables implies the identity of the limits of their distributions; hence, the diffuse uniform $\mathrm{U}[a, a + \infty]$ is "the same" as the diffuse normal $\mathrm{N}[\mu, \infty]$.[17]

[17] Quotes are around 'the same' because the limiting distributions are not proper distributions. The notion of diffuse distributions comes from Venter [1996, pp. 406-410], who shows there how different diffuse distributions result in

Indeed, for the limit to be δ_{t0} it is not required that *n* be an integer. But for $\Theta \sim \mathrm{U}[a, a+2\pi n]$, the integral moments of $W = e^{i\Theta}$ are:

$$E[W^j] = E[e^{ij\Theta}] = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } jn = \pm 1, \pm 2, \ldots \\ \neq 0 & \text{if } jn \text{ not integral} \end{cases}$$

So if *n* is an integer, *jn* will be an integer, and all the integral moments of *W* will be zero, except for the zeroth. Therefore, the integral moments of $W = e^{i\Theta}$ are invariant to *n*, as long as the *n* in $2\pi n$, the width of the interval of Θ, is a whole number. Hence, although we hereby define the unit-circle random variable as $e^{i\Theta}$ for $\Theta \sim U[0, 2\pi]$, the choice of $a = 0$ and $n = 1$ is out of convenience, rather than out of necessity. The probability for $e^{i\Theta}$ to be in an arc of this circle of length *l* equals $l/2\pi$.

The integral moments of the conjugate of $\overline{W} = e^{-i\Theta}$ are the same, for $E[\overline{W}^j] = E[\overline{W^j}] = \overline{E[W^j]} = \overline{\delta_{j0}} = \delta_{j0} = E[W^j]$. Alternatively, $E[\overline{W}^j] = E[e^{-ij\Theta}] = \delta_{(-j)0} = \delta_{j0}$. And the *jk*th mixed moment is $E[W^j \overline{W}^k] = E[e^{ij\Theta} e^{-ik\Theta}] = E[e^{i(j-k)\Theta}] = \delta_{(j-k)0} = \delta_{jk}$. Since $E[W] = E[\overline{W}] = 0$, the augmented variance of the unit-circle random variable is:

$$Var\begin{bmatrix} W \\ \overline{W} \end{bmatrix} = E\left[\begin{bmatrix} W \\ \overline{W} \end{bmatrix}\begin{bmatrix} \overline{W} & W \end{bmatrix}\right] = E\begin{bmatrix} W\overline{W} & WW \\ \overline{W}\overline{W} & \overline{W}W \end{bmatrix} = \begin{bmatrix} \delta_{11} & \delta_{20} \\ \delta_{02} & \delta_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathrm{I}_2$$

Hence, $W = e^{i\Theta}$ for $\Theta \sim U[0, 2\pi]$ is not just a unit-circle random variable; having zero mean and unit variance, it is the *standard* unit-circle random variable.

---

different Bayesian estimates. But here every continuous random variable *Y* diffuses through the periodicity of e^{iY} into the same limiting distribution, viz., the Kronecker δ_{t0} (note 31).

Multiplying W by a complex constant $\alpha \neq 0$ affects the radius of the random variable, whose jk^{th} mixed moment is:

$$E\left[(\alpha W)^j \left(\overline{\alpha W}\right)^k\right] = \alpha^j \overline{\alpha}^k E\left[W^j \overline{W}^k\right] = \alpha^j \overline{\alpha}^k \delta_{jk} = \begin{cases} (\alpha\overline{\alpha})^j & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

The augmented variance is $Var\left[\begin{matrix} \alpha W \\ \overline{\alpha W} \end{matrix}\right] = \alpha\overline{\alpha} Var[W] = \alpha\overline{\alpha}\, \mathrm{I}_2$. One may consider α as an instance of a complex random variable A. Due to the independence of A from W, the jk^{th} mixed moment of AW is $E\left[(AW)^j \left(\overline{AW}\right)^k\right] = E\left[A^j \overline{A}^k\right] E\left[W^j \overline{W}^k\right] = E\left[A^j \overline{A}^k\right] \delta_{jk} = E\left[\left(A\overline{A}\right)^j\right] \delta_{jk}$. Its augmented variance is $Var\left[\begin{matrix} AW \\ \overline{AW} \end{matrix}\right] = E\left[A\overline{A}\right] Var[W] = \left\{Var[A] + E[A]E\left[\overline{A}\right]\right\} Var[W]$. Unlike the one-dimensional W, AW can cover the whole complex plane. However, like W, it too possesses the desirable property that $E\left[(AW)^j\right] = \delta_{j0}$.[18]

11. UNIT-CIRCULARITY AND DETERMINISM

The single most important quality of a random variable is its mean. In fact, just having reliable estimates of mean values would satisfy many users of actuarial analyses. Stochastic advances in actuarial science over the last few decades notwithstanding, much actuarial work remains deterministic. Determinism is not the reduction of a stochastic answer $Y = f(X)$ to its mean $E[Y] = E[f(X)]$. Rather, the deterministic assumption is that the expectation of a function of a random variable equals the function of the expectation of the random variable; in symbols,

[18] The existence of the moments $E\left[A^j \overline{A}^k\right]$ needs to be ascertained. In particular, moments for j and k as negative integers will not exist unless $\text{Prob}[A = 0] = 1 - \text{Prob}[A \neq 0] = 1 - \text{Prob}\left[A\overline{A} > 0\right] = 0$.

$E[Y] = E[f(X)] = f(E[X])$. Because this assumption is true for linear *f*, it was felt to be a reasonable or necessary approximation for non-linear *f*.

Advances in computing hardware and software, as well as increased technical sophistication, have made determinism more avoidable and less acceptable. However, the complex unit-circular random variable provides a habitat for the survival of determinism. To see this, let *f* be analytic over the domain of complex random variable *Z*. From Cauchy's Integral Formula (Havil [2003, Appendix D.8 and D.9]) it follows that within the domain of *Z*, *f* can be expressed as a convergent series $f(z) = a_0 + a_1 z + \ldots = a_0 + \sum_{j=1}^{\infty} a_j z^j$. Taking the expectation, we have:

$$E[f(Z)] = a_0 + \sum_{j=1}^{\infty} a_j E[Z^j]$$

But if for every positive integer *j* $E[Z^j] = E[Z]^j$, then:

$$E[f(Z)] = a_0 + \sum_{j=1}^{\infty} a_j E[Z^j] = a_0 + \sum_{j=1}^{\infty} a_j E[Z]^j = f(E[Z])$$

Therefore, determinism conveniently works for analytic functions of random variables whose moments are powers of their means.

Now a real-valued random variable whose moments are powers of its mean would have the characteristic function:

$$M_X(it) = E[e^{itX}] = 1 + \sum_{j=1}^{\infty} \frac{(it)^j}{j!} E[X^j] = 1 + \sum_{j=1}^{\infty} \frac{(it)^j}{j!} E[X]^j = e^{itE[X]} = M_{E[X]}(it)$$

This is the characteristic function of the "deterministic" random variable, i.e., the random variable whose probability is massed at one point, its mean. So determinism with real-valued random

variables requires "deterministic" random variables. But some complex random variables, such as the unit-circle, have the property $E[Z^j] = E[Z]^j$ without being deterministic.

In fact, when $E[Z^j] = E[Z]^j$, not only is $E[f(Z)] = f[E[Z]]$. For positive integer *k*, $f^k(z)$ is as analytic as *f* itself; hence, $E[f^k(Z)] = f^k(E[Z])$. So the determinism with these complex random variables is valid for all moments; nothing is lost.

In Section 10 we saw that for the unit-circle random variable $W = e^{i\Theta}$ and for $j = \pm 1, \pm 2, \ldots$, $E[W^{-j}] = E[W^j] = 0 = E[W]^{|j|}$. Can determinism extend to non-analytic functions which involve the negative moments? For example, let $g(z) = 1/(\eta - z)$, for some complex $\eta \neq 0$. The function is singular at $z = \eta$; but within the disc $\{z : |z/\eta| < 1\} = \{z : |z| < |\eta|\}$ the function equals the convergent series:

$$g(z) = 1/(\eta - z) = \frac{1}{\eta} \cdot \frac{1}{\left(1 - \frac{z}{\eta}\right)} = \frac{1}{\eta}\left\{1 + \frac{z}{\eta} + \left(\frac{z}{\eta}\right)^2 + \ldots\right\} = \frac{1}{\eta} + \frac{z}{\eta^2} + \frac{z^2}{\eta^3} + \ldots$$

Outside the disc, or for $\{z : |z/\eta| > 1\} = \{z : |z| > |\eta|\}$, another convergent series represents the function:

$$g(z) = 1/(\eta - z) = -\frac{1}{z} \cdot \frac{1}{\left(1 - \frac{\eta}{z}\right)} = -\frac{1}{z}\left\{1 + \frac{\eta}{z} + \left(\frac{\eta}{z}\right)^2 + \ldots\right\} = -\frac{1}{z} - \frac{\eta}{z^2} - \frac{\eta^2}{z^3} - \ldots$$

So, if $|\eta| > 1$, then $|W| = 1 < |\eta|$. In this case:

$$\begin{aligned} E[g(W)] &= E\left[\frac{1}{\eta}+\frac{W}{\eta^2}+\frac{W^2}{\eta^3}+\ldots\right] \\ &= \frac{1}{\eta}+\frac{E[W]}{\eta^2}+\frac{E[W^2]}{\eta^3}+\ldots \\ &= \frac{1}{\eta}+\frac{0}{\eta^2}+\frac{0}{\eta^3}+\ldots \\ &= \frac{1}{\eta} \end{aligned}$$

However, if $|\eta|<1$, then $|W|=1>|\eta|$. So in this case:

$$\begin{aligned} E[g(W)] &= E\left[-\frac{1}{W}-\frac{\eta}{W^2}-\frac{\eta^2}{W^3}-\ldots\right] \\ &= -E[W^{-1}]-\eta E[W^{-2}]-\eta^2 E[W^{-3}]-\ldots \\ &= -0-\eta\cdot 0-\eta^2\cdot 0-\ldots \\ &= 0 \end{aligned}$$

Both answers are correct; however, only the first satisfies the deterministic equation $E[g(W)]=g(E[W])=g(0)=1/\eta$.

To understand why the answer depends on whether η is inside or outside the complex unit circle, let us evaluate $E[g(W)]$ directly:

$$E[g(W)]=E[1/(\eta-W)]=E[1/(\eta-e^{i\Theta})]=\int_{\theta=0}^{2\pi}\frac{1}{\eta-e^{i\theta}}\frac{d\theta}{2\pi}$$

The next step is to transform from θ into $z=e^{i\theta}$. So $dz=ie^{i\theta}d\theta=izd\theta$, and the line integral transforms into a contour integral over the unit circle C:

$$
\begin{aligned}
E[g(W)] &= \int_{\theta=0}^{2\pi} \frac{1}{\eta - e^{i\theta}} \frac{d\theta}{2\pi} \\
&= \oint_C \frac{1}{\eta - z} \frac{iz}{iz} \frac{d\theta}{2\pi} \\
&= \oint_C \frac{1}{\eta - z} \frac{1}{z} \frac{dz}{2\pi i} \\
&= \frac{1}{2\pi i} \oint_C \frac{dz}{z(\eta - z)} \\
&= \frac{1}{2\pi i} \oint_C \left(\frac{1}{z} + \frac{1}{\eta - z} \right) \frac{1}{\eta} dz \\
&= \frac{1}{\eta} \left(\frac{1}{2\pi i} \oint_C \frac{dz}{z} + \frac{1}{2\pi i} \oint_C \frac{dz}{\eta - z} \right) \\
&= \frac{1}{\eta} \left(\frac{1}{2\pi i} \oint_C \frac{dz}{z - 0} - \frac{1}{2\pi i} \oint_C \frac{dz}{z - \eta} \right)
\end{aligned}
$$

Now the value of each of these integrals is one if its singularity is within the unit circle *C*, and zero if it is not.[19] Of course, the singularity of the first integral at $z = 0$ lies within *C*; hence, its value is one. The second integral's singularity at $z = \eta$ lies within *C* if and only if $|\eta| < 1$. Therefore:

$$
E[g(W)] = \frac{1}{\eta} \left(\frac{1}{2\pi i} \oint_C \frac{dz}{z - 0} - \frac{1}{2\pi i} \oint_C \frac{dz}{z - \eta} \right) = g(E[W]) \cdot \begin{cases} 1 & \text{if } |\eta| > 1 \\ 0 & \text{if } |\eta| < 1 \end{cases}
$$

So the deterministic equation will hold for one of the Laurent series according to which the domain of the non-analytic function is divided into regions of convergence. Fascinating enough is how the function $\varphi(\eta) = \frac{1}{2\pi i} \oint_C \frac{dz}{z - \eta} = \begin{cases} 1 & \text{if } |\eta| > 1 \\ 0 & \text{if } |\eta| < 1 \end{cases}$ serves as the indicator of a state, viz., the state of being inside or outside the complex unit circle.

[19] Technically, the integral has no value if the singularity lies on *C*; but there are some practical advantages for "splitting the difference" in that case.

12. THE LINEAR STATISTICAL MODEL

Better known as "regression" models, linear statistical models extend readily into the realm of complex numbers. A general real-valued form of such models is presented and derived in Halliwell [1997, Appendix C]:

$$\begin{bmatrix}\mathbf{y}_1\\ \hdashline \mathbf{y}_2\end{bmatrix}=\begin{bmatrix}\mathrm{X}_1\\ \hdashline \mathrm{X}_2\end{bmatrix}\beta+\begin{bmatrix}\mathbf{e}_1\\ \hdashline \mathbf{e}_2\end{bmatrix},\quad Var\begin{bmatrix}\mathbf{e}_1\\ \hdashline \mathbf{e}_2\end{bmatrix}=\left[\begin{array}{c|c}\Sigma_{11} & \Sigma_{12}\\ \hline \Sigma_{21} & \Sigma_{22}\end{array}\right]$$

The subscript '1' denotes observations, '2' denotes predictions. Vector $\mathbf{y}_1$ is observed; the whole design matrix X is hypothesized, as well as the fourfold 'Σ' variance structure. Although the variance structure may be non-negative-definite (NND), the variance of the observations Σ_{11} must be positive-definite (PD). Also, the observation design X_1 must be of full column rank. The last two requirements ensure the existence of the inverses Σ_{11}^{-1} and $\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}$. The best linear unbiased predictor of $\mathbf{y}_2$ is $\hat{\mathbf{y}}_2=\mathrm{X}_2\hat{\boldsymbol{\beta}}+\Sigma_{21}\Sigma_{11}^{-1}\left(\mathbf{y}_1-\mathrm{X}_1\hat{\boldsymbol{\beta}}\right)$. The variance of prediction error $\mathbf{y}_2-\hat{\mathbf{y}}_2$ is $Var\left[\mathbf{y}_2-\hat{\mathbf{y}}_2\right]=\left(\mathrm{X}_2-\Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)\mathrm{Var}\left[\hat{\boldsymbol{\beta}}\right]\left(\mathrm{X}_2-\Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)'+\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Embedded in these formulas are the estimator of β and its variance: $\hat{\boldsymbol{\beta}}=\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1=Var\left[\hat{\boldsymbol{\beta}}\right]\cdot\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1$.

For the purpose of introducing complex numbers into the linear statistical model we will concern ourselves here only the estimation of the parameter β. So we drop the subscripts '1' and '2' and simplify the observation as $\mathbf{y}=\mathrm{X}\beta+\mathbf{e}$, where $Var[\mathbf{e}]=\Gamma$. Again, X must be of full column rank and Γ must be Hermetian PD. According to Section 4, transjugation is to complex matrices what transposition is to real-valued matrices. Therefore, the short answer for a complex model is: $\hat{\boldsymbol{\beta}}=\left(\mathrm{X}^*\Gamma^{-1}\mathrm{X}\right)^{-1}\mathrm{X}^*\Gamma^{-1}\mathbf{y}=Var\left[\hat{\boldsymbol{\beta}}\right]\cdot\mathrm{X}^*\Gamma^{-1}\mathbf{y}$.

However, deriving the solution from the double-real representation in Section 3 will deepen the understanding. The double-real form of the observation is:

$$\begin{bmatrix} \mathbf{y}_r & -\mathbf{y}_i \\ \mathbf{y}_i & \mathbf{y}_r \end{bmatrix} = \begin{bmatrix} \mathrm{X}_r & -\mathrm{X}_i \\ \mathrm{X}_i & \mathrm{X}_r \end{bmatrix} \begin{bmatrix} \beta_r & -\beta_i \\ \beta_i & \beta_r \end{bmatrix} + \begin{bmatrix} \mathbf{e}_r & -\mathbf{e}_i \\ \mathbf{e}_i & \mathbf{e}_r \end{bmatrix}$$

All the vectors and matrices in this form are real-valued. The subscripts '*r*' and '*i*' denote the real and imaginary parts of **y**, X, β, and **e**. Due to the redundancy of double-real representation, we may retain just the left column:

$$\begin{bmatrix} \mathbf{y}_r \\ \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathrm{X}_r & -\mathrm{X}_i \\ \mathrm{X}_i & \mathrm{X}_r \end{bmatrix} \begin{bmatrix} \beta_r \\ \beta_i \end{bmatrix} + \begin{bmatrix} \mathbf{e}_r \\ \mathbf{e}_i \end{bmatrix}$$

Note that if X is real-valued, then $\mathrm{X}_i = 0$, and $\mathbf{y}_r$ and $\mathbf{y}_i$ become two "data panels," each with its own parameter β_r and β_i.[20]

Now let $\Xi_t = \begin{bmatrix} \mathrm{I}_t & i\mathrm{I}_t \\ \mathrm{I}_t & -i\mathrm{I}_t \end{bmatrix}$, the augmentation matrix of Section 7, where *t* is the number of observations. Since the augmentation matrix is non-singular, premultiplying the left-column form by it yields equivalent but insightful forms:

[20] This assumes zero covariance between the error vectors.

$$\begin{bmatrix} \mathrm{I}_t & i\mathrm{I}_t \\ \mathrm{I}_t & -i\mathrm{I}_t \end{bmatrix}\begin{bmatrix} \mathbf{y}_r \\ \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathrm{I}_t & i\mathrm{I}_t \\ \mathrm{I}_t & -i\mathrm{I}_t \end{bmatrix}\begin{bmatrix} \mathrm{X}_r & -\mathrm{X}_i \\ \mathrm{X}_i & \mathrm{X}_r \end{bmatrix}\begin{bmatrix} \beta_r \\ \beta_i \end{bmatrix} + \begin{bmatrix} \mathrm{I}_t & i\mathrm{I}_t \\ \mathrm{I}_t & -i\mathrm{I}_t \end{bmatrix}\begin{bmatrix} \mathbf{e}_r \\ \mathbf{e}_i \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y}_r + i\mathbf{y}_i \\ \mathbf{y}_r - i\mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathrm{X}_r + i\mathrm{X}_i & -\mathrm{X}_i + i\mathrm{X}_r \\ \mathrm{X}_r - i\mathrm{X}_i & -\mathrm{X}_i - i\mathrm{X}_r \end{bmatrix}\begin{bmatrix} \beta_r \\ \beta_i \end{bmatrix} + \begin{bmatrix} \mathbf{e}_r + i\mathbf{e}_i \\ \mathbf{e}_r - i\mathbf{e}_i \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathrm{X} & i\mathrm{X} \\ \overline{\mathrm{X}} & -i\overline{\mathrm{X}} \end{bmatrix}\begin{bmatrix} \beta_r \\ \beta_i \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathrm{X}\beta_r + i\mathrm{X}\beta_i \\ \overline{\mathrm{X}}\beta_r - i\overline{\mathrm{X}}\beta_i \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathrm{X}\beta \\ \overline{\mathrm{X}\beta} \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathrm{X} & 0 \\ 0 & \overline{\mathrm{X}} \end{bmatrix}\begin{bmatrix} \beta \\ \overline{\beta} \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$$

The first insight is that $\overline{\mathbf{y}} = \overline{\mathrm{X}\beta + \mathbf{e}} = \overline{\mathrm{X}}\overline{\beta} + \overline{\mathbf{e}}$ is as much observed as $\mathbf{y} = \mathrm{X}\beta + \mathbf{e}$. The second insight is that $Var\begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$ is an augmented variance, whose general form according to Section 6 is $Var\begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix} = \begin{bmatrix} \Gamma & \mathrm{C} \\ \overline{\mathrm{C}} & \overline{\Gamma} \end{bmatrix}$. Therefore, the general form of the observation of a complex linear model is $\begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathrm{X} & 0 \\ 0 & \overline{\mathrm{X}} \end{bmatrix}\begin{bmatrix} \beta \\ \overline{\beta} \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$, where $Var\begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix} = \begin{bmatrix} \Gamma & \mathrm{C} \\ \overline{\mathrm{C}} & \overline{\Gamma} \end{bmatrix}$. Not only is $\overline{\mathbf{y}}$ as observable as $\mathbf{y}$, but also $\overline{\beta}$ is as estimable as β. Furthermore, although the augmented variance may default to $\mathrm{C} = 0$, the complex linear statistical model does not require $\begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$ to be "proper complex," as defined in Section 7.

Since X is of full column rank, so too must be $\begin{bmatrix} \mathrm{X} & 0 \\ 0 & \overline{\mathrm{X}} \end{bmatrix}$. And since Γ is Hermetian PD, both it and its conjugate $\overline{\Gamma}$ are invertible. But the general form of the observation requires $\begin{bmatrix} \Gamma & \mathrm{C} \\ \overline{\mathrm{C}} & \overline{\Gamma} \end{bmatrix}$ to be

Hermetian PD, hence invertible. A consequence is that both the "determinant" forms $\Gamma - C\overline{\Gamma}^{-1}\overline{C}$ and $\overline{\Gamma} - \overline{C}\Gamma^{-1}C$ are Hermetian PD and invertible. With this background it can be shown, and the reader should verify, that $\begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix}^{-1} = \begin{bmatrix} H & K \\ \overline{K} & \overline{H} \end{bmatrix}$, where $H = \left(\Gamma - C\overline{\Gamma}^{-1}\overline{C}\right)^{-1}$ and $K = -\Gamma^{-1}C\overline{H}$. The important point is that inversion preserves the augmented-variance form.

The solution of the complex linear model $\begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} X & 0 \\ 0 & \overline{X} \end{bmatrix}\begin{bmatrix} \beta \\ \overline{\beta} \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix}$, where $Var\begin{bmatrix} \mathbf{e} \\ \overline{\mathbf{e}} \end{bmatrix} = \begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix}$, is:

$$\begin{aligned}
\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\overline{\boldsymbol{\beta}}} \end{bmatrix} &= \left(\begin{bmatrix} X & 0 \\ 0 & \overline{X} \end{bmatrix}^{*} \begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix}^{-1} \begin{bmatrix} X & 0 \\ 0 & \overline{X} \end{bmatrix} \right)^{-1} \begin{bmatrix} X & 0 \\ 0 & \overline{X} \end{bmatrix}^{*} \begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} \\
&= \left(\begin{bmatrix} X^{*} & 0 \\ 0 & X' \end{bmatrix} \begin{bmatrix} H & K \\ \overline{K} & \overline{H} \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & \overline{X} \end{bmatrix} \right)^{-1} \begin{bmatrix} X^{*} & 0 \\ 0 & X' \end{bmatrix} \begin{bmatrix} H & K \\ \overline{K} & \overline{H} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \overline{\mathbf{y}} \end{bmatrix} \\
&= \begin{bmatrix} X^{*}HX & X^{*}K\overline{X} \\ X'\overline{K}X & X'\overline{H}\overline{X} \end{bmatrix}^{-1} \begin{bmatrix} X^{*}H\mathbf{y} + X^{*}K\overline{\mathbf{y}} \\ X'\overline{K}\mathbf{y} + X'\overline{H}\overline{\mathbf{y}} \end{bmatrix}
\end{aligned}$$

And $Var\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\overline{\boldsymbol{\beta}}} \end{bmatrix} = \begin{bmatrix} X^{*}HX & X^{*}K\overline{X} \\ X'\overline{K}X & X'\overline{H}\overline{X} \end{bmatrix}^{-1}$, which must exist since it is a quadratic form based on the Hermetian PD $\begin{bmatrix} \Gamma & C \\ \overline{C} & \overline{\Gamma} \end{bmatrix}^{-1}$ and the full column rank $\begin{bmatrix} X & 0 \\ 0 & \overline{X} \end{bmatrix}$. Reformulate this as:

$$\begin{bmatrix} X^{*}HX & X^{*}K\overline{X} \\ X'\overline{K}X & X'\overline{H}\overline{X} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\overline{\boldsymbol{\beta}}} \end{bmatrix} = \begin{bmatrix} X^{*}H\mathbf{y} + X^{*}K\overline{\mathbf{y}} \\ X'\overline{K}\mathbf{y} + X'\overline{H}\overline{\mathbf{y}} \end{bmatrix}$$

The conjugates of the two equations in $\hat{\boldsymbol{\beta}}$ and $\hat{\overline{\boldsymbol{\beta}}}$ are the same equations in $\hat{\overline{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{\beta}}$:

$$\begin{bmatrix} X^{*}HX & X^{*}K\overline{X} \\ X'\overline{K}X & X'\overline{H}\overline{X} \end{bmatrix} \begin{bmatrix} \hat{\overline{\boldsymbol{\beta}}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} X^{*}H\mathbf{y} + X^{*}K\overline{\mathbf{y}} \\ X'\overline{K}\mathbf{y} + X'\overline{H}\overline{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} X^{*}HX & X^{*}K\overline{X} \\ X'\overline{K}X & X'\overline{H}\overline{X} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\overline{\boldsymbol{\beta}}} \end{bmatrix}$$

Therefore, $\begin{bmatrix} \overline{\hat{\boldsymbol{\beta}}} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \overline{\hat{\boldsymbol{\beta}}} \end{bmatrix}$. It is well known that the estimator of a linear function of a random variable is the linear function of the estimator of the random variable. But conjugation is not a linear function. Nevertheless, we have just proven that the estimator of the conjugate is the conjugate of the estimator.

13. ACTUARIAL APPLICATIONS OF COMPLEX RANDOM VARIABLES

How might casualty actuaries put complex random variables to work? Since the support of most complex random variables is a plane, rather than a line, their obvious application is bivariate. An example is a random variable whose real part is loss and whose imaginary part is LAE. Another application might pertain to copulas. According to Venter [2002, p. 69], "copulas are joint distributions of unit random variables." One could translate these joint distributions into distributions of complex variables whose support is the complex unit square, i.e., the square whose vertices are the points $z = 0, 1, 1+i, i$. However, for now it seems that real-valued bivariates provide the necessary theory and technique for these purposes.

Actuaries who have applied log-linear models to triangles with paid increments have been frustrated applying them to incurred triangles. The problem is that incurred increments are often negative, and the logarithm of a negative number is not real-valued. This has led Glenn Meyers [2013] to seek modified lognormal distributions whose support includes the negative real numbers. The persistent intractability of the log-linear problem was a major reason for our attention to the lognormal random vector $\mathbf{w}_{n\times 1} = e^{\mathbf{z}_{n\times 1}}$ in Section 8. But to model an incurred loss as the exponential function of a complex number suffers from two drawbacks. First, to model a real-valued loss as $e^{\mathbf{z}} = e^{x} \cdot e^{iy}$ requires *y* to be an integral multiple of π. The mixed random variable e^{X} with probability *p* and

$-e^X$ with probability $1-p$ is not lognormal. No more suitable are such "denatured" random variables as $\mathrm{Re}(e^Z)$. Second, one still cannot model the eminently practical value of zero, because for all z, $e^{\mathbf{z}} \neq 0$.[21] At present it does not appear that complex random variables will give birth to useful distributions of real-valued random variables. Even the unit-circle and indicator random variables of Sections 10 and 11, as interesting as they are in the theory of analytic functions, most likely will engender no distributions valuable to actuarial work.

The complex version of the linear model in Section 12 showed us that conjugates of observations are themselves observations and that conjugates of estimators are estimators of conjugates. Moreover, there we found a use for augmented variance. Nonetheless we are still fairly bound to our conclusion to Section 3, that one who lacked either the confidence or the software to work with complex numbers could probably do a work-around with double-real matrices.

So how can actuarial science benefit from complex random variables? The great benefit will come from new ways of thinking. The first step will be to overcome the habit of picturing a complex number as half real and half imaginary. Historically, it was only after numbers had expanded from rational to irrational that the whole set was called "real." Numbers ultimately are sets; zero is just the empty set. How real are sets? Regardless of their mathematical reality, they are not physically real. Complex numbers were deemed "real" because mathematicians needed them for the solution of polynomial equations. In the nineteenth century this spurred the development of abstract algebra. At first new ways of thinking amount to differences in degree; at some point many develop

---

[21] If $e^a = 0$ for some *a*, then for all z $e^z = e^{z-a+a} = e^{z-a}e^a = e^{z-a} \cdot 0 = 0$. One who sees that $\lim_{x\to-\infty} e^x \cdot e^{iy} = 0 \cdot e^{iy} = 0$ might propose to add the ordinate $\mathrm{Re}(z) = x = -\infty$ to the complex plane. But not only is this proposal artificial; it also militates against the standard theory of complex variables, according to which all points infinitely far from zero constitute one and the same point at infinity.

into differences in kind. One might argue, "Why study Euclidean geometry? It all derives from a few axioms." True, but great theorems (e.g., that the sum of the angles of a triangle is the sum of two right angles) can be a long way from their axioms. A theorem means more than the course of its proof; often there are many proofs of a theorem. Furthermore, mathematicians often work backwards from accepted or desired truths to efficient and elegant sets of axioms. Perhaps the most wonderful thing about mathematics is its "unreasonable effectiveness in the natural sciences," to quote physicist Eugene Wigner. The causality between pure and applied mathematics works in both directions. Therefore, it is likely that complex random variables and vectors will find their way into actuarial science. But it will take years, even decades, and technology and education will have to prepare for it.

14. CONCLUSION

Just as physics divides into different areas, e.g., theoretical, experimental, and applied, so too actuarial science, though perhaps more concentrated on business application, justifiably has and needs a theoretical component. Theory and application cross-fertilize each other. In this paper we have proposed to add complex numbers to the probability and statistics of actuarial theory. With patience, the technically inclined actuary should be able to understand the theory of complex random variables delineated herein. In fact, our multivariate approach may even more difficult to understand than the complex-function theory; but both belong together. Although all complex matrices and operations were formed from double-real counterparts, we believe that the "sum is greater than the parts," i.e., that the assimilation of this theory will lead to higher-order thinking and creativity. In the sixteenth century the "fiction" of $i = \sqrt{-1}$ allowed mathematicians to solve more equations. Although at first complex solutions were deemed "extraneous roots," eventually their practicality became recognized; so that today complex numbers are essential for science and

engineering. Applying complex numbers to probability has lagged; but even now it is part of signal processing in electrical engineering. Knowing how rapidly science has developed with nuclear physics, molecular biology, space exploration, and computers, who would dare to bet against the usefulness of complex random variables to actuarial science by the mid-2030s, when many scientists and futurists expect nuclear fusion to be harnessed and available for commercial purposes?

REFERENCES

[1.] Halliwell, Leigh J., Conjoint Prediction of Paid and Incurred Losses, CAS *Forum*, Summer 1997, 241-380, www.casact.org/pubs/forum/97sforum/97sf1241.pdf.

[2.] Halliwell, Leigh J., "Classifying the Tails of Loss Distributions," CAS *E-Forum*, Spring 2013, Volume 2, www.casact.org/pubs/forum/13spforumv2/Haliwell.pdf.

[3.] Havil, Julian, *Gamma: Exploring Euler's Constant*, Princeton University Press, 2003.

[4.] Healy, M. J. R., *Matrices for Statistics*, Oxford, Clarendon Press, 1986.

[5.] Hogg, Robert V., and Stuart A. Klugman, *Loss Distributions*, New York, Wiley, 1984.

[6.] Johnson, Richard A., and Dean Wichern, *Applied Multivariate Statistical Analysis* (Third Edition), Englewood Cliffs, NJ, Prentice Hall, 1992.

[7.] Judge, George G., Hill, R. C., et al., *Introduction to the Theory and Practice of Econometrics* (Second Edition), New York, Wiley, 1988.

[8.] Klugman, Stuart A., et al., *Loss Models: From Data to Decisions*, New York, Wiley, 1998.

[9.] Meyers, Glenn, "The Skew Normal Distribution and Beyond," *Actuarial Review*, May 2013, p. 15, www.casact.org/newsletter/pdfUpload/ar/AR_May2013_1.pdf.

[10.] Million, Elizabeth, The Hadamard Product, 2007, http://buzzard.ups.edu/courses/2007spring/projects/million-paper.pdf.

[11.] Schneider, Hans, and George Barker, *Matrices and Linear Algebra*, New York, Dover, 1973.

[12.] Veeravalli, V. V., *Proper Complex Random Variables and Vectors*, 2006, http://courses.engr.illinois.edu/ece461/handouts/notes6.pdf.

[13.] Venter, G.G., "Credibility," *Foundations of Casualty Actuarial Science* (Third Edition), Casualty Actuarial Society, 1996.

[14.] Venter, G.G., "Tails of Copulas," *PCAS* LXXXIX, 2002, 68-113, www.casact.org/pubs/proceed/proceed02/02068.pdf.

[15.] Weyl, Hermann, *The Theory of Groups and Quantum Mechanics*, New York, Dover, 1950 (ET by H. P. Robertson from second German Edition 1930).

[16.] Wikipedia contributors, "Arcsine distribution," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Arcsine_distribution (accessed October 2014).

[17.] Wikipedia contributors, "Complex normal distribution," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Complex_normal_distribution (accessed October 2014).

[18.] Wikipedia contributors, "Complex number," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Complex_number (accessed October 2014).

APPENDIX A

REAL-VALUED LOGNORMAL RANDOM VECTORS

Feeling that the treatment of lognormal random vectors in Section 8 would be too long, we have decided to prepare for it in Appendices A and B. According to Section 7, the probability density function of real-valued $n\times1$ normal random vector $\mathbf{x}$ with mean μ and variance Σ is:

$$f_{\mathbf{x}}(\mathrm{x})=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}e^{-\frac{1}{2}(\mathrm{x}-\mu)'\Sigma^{-1}(\mathrm{x}-\mu)}$$

Therefore, $\int_{x\in\Re^n} f_{\mathbf{x}}(\mathrm{x})dV=1$. The single integral over $\Re^n$ represents an n-multiple integral over each x_j from $-\infty$ to $+\infty$; $dV=dx_1\ldots dx_n$.

The moment generating function of $\mathbf{x}$ is $M_{\mathbf{x}}(\mathrm{t})=E\left[e^{t'\mathbf{x}}\right]=E\left[e^{\sum_{j=1}^{n}t_jx_j}\right]$, where t is a suitable $n\times1$ vector.[22] Partial derivatives of the moment generating function evaluated at $\mathrm{t}=0_{n\times1}$ equal moments of $\mathbf{x}$, since:

$$\left.\frac{\partial^{k_1+\ldots k_n}M_{\mathbf{x}}(t)}{\partial^{k_1}x_1\ldots\partial^{k_n}x_n}\right|_{t=0}=E\left[x_1^{k_1}\ldots x_n^{k_n}e^{t'\mathbf{x}}\right]\Big\|_{t=0}=E\left[x_1^{k_1}\ldots x_n^{k_n}\right]$$

But lognormal moments are values of the function itself. For example, if $\mathrm{t}=\mathrm{e}_j$, the j^{th} unit vector, then $M_{\mathbf{x}}(\mathrm{e}_j)=E\left[e^{\mathrm{e}_j'\mathbf{x}}\right]=E\left[e^{X_j}\right]$. Likewise, $M_{\mathbf{x}}(\mathrm{e}_j+\mathrm{e}_k)=E\left[e^{X_j}e^{X_k}\right]$. The moment generating function of $\mathbf{x}$, if it exists, is the key to the moments of $e^{\mathbf{x}}$.

[22] All real-valued t vectors are suitable; Appendix B will extend the suitability to complex t.

The moment generating function of the real-valued multivariate normal **x** is:

$$\begin{aligned} M_{\mathbf{x}}(\mathrm{t}) &= E\left[e^{\mathrm{t}'\mathbf{x}}\right] \\ &= \int_{x\in\Re^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathrm{x}-\mu)'\Sigma^{-1}(\mathrm{x}-\mu)} e^{\mathrm{t}'\mathrm{x}} dV \\ &= \int_{x\in\Re^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}\left\{(\mathrm{x}-\mu)'\Sigma^{-1}(\mathrm{x}-\mu)-2\mathrm{t}'\mathrm{x}\right\}} dV \end{aligned}$$

A multivariate "completion of the square" results in the identity:

$$(\mathrm{x}-\mu)'\Sigma^{-1}(\mathrm{x}-\mu)-2\mathrm{t}'\mathrm{x} = (\mathrm{x}-[\mu+\Sigma\mathrm{t}])'\Sigma^{-1}(\mathrm{x}-[\mu+\Sigma\mathrm{t}])-2\mathrm{t}'\mu-\mathrm{t}'\Sigma\mathrm{t}$$

We leave it for the reader to verify. By substitution, we have:

$$\begin{aligned} M_{\mathbf{x}}(\mathrm{t}) &= \int_{\mathrm{x}\in\Re^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}\left\{(\mathrm{x}-\mu)'\Sigma^{-1}(\mathrm{x}-\mu)-2\mathrm{t}'\mathrm{x}\right\}} dV \\ &= \int_{\mathrm{x}\in\Re^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}\left\{(\mathrm{x}-[\mu+\Sigma\mathrm{t}])'\Sigma^{-1}(\mathrm{x}-[\mu+\Sigma\mathrm{t}])-2\mathrm{t}'\mu-\mathrm{t}'\Sigma\mathrm{t}\right\}} dV \\ &= \int_{\mathrm{x}\in\Re^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathrm{x}-[\mu+\Sigma\mathrm{t}])'\Sigma^{-1}(\mathrm{x}-[\mu+\Sigma\mathrm{t}])} dV \cdot e^{\mathrm{t}'\mu+\mathrm{t}'\Sigma\mathrm{t}/2} \\ &= 1\cdot e^{\mathrm{t}'\mu+\mathrm{t}'\Sigma\mathrm{t}/2} \\ &= e^{\mathrm{t}'\mu+\mathrm{t}'\Sigma\mathrm{t}/2} \end{aligned}$$

The reduction of the integral to unity in the second last line is due to the fact that $\frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(\mathrm{x}-[\mu+\Sigma\mathrm{t}])'\Sigma^{-1}(\mathrm{x}-[\mu+\Sigma\mathrm{t}])}$ is the probability density function of the real-valued $n\times 1$ normal random vector with mean $\mu+\Sigma\mathrm{t}$ and variance Σ. This new mean is valid if it is real-valued, which will be so if t is real-valued. In fact, $\mu+\Sigma\mathrm{t}$ is real-valued if and only if t is real-valued.

So the moment generating function of the real-valued normal multivariate $\mathbf{x} \sim N(\mu, \Sigma)$ is $M_{\mathbf{x}}(t) = e^{t'\mu + t'\Sigma t/2}$, which is valid at least for $t \in \Re^n$. As a check:[23]

$$\frac{\partial M_{\mathbf{x}}(t)}{\partial t} = (\mu + \Sigma t) e^{t'\mu + t'\Sigma t/2} \Rightarrow E[\mathbf{x}] = \left.\frac{\partial M_{\mathbf{x}}(t)}{\partial t}\right|_{t=0} = \mu$$

And for the second derivative:

$$\begin{aligned}
\frac{\partial^2 M_{\mathbf{x}}(t)}{\partial t \partial t'} &= \frac{\partial (\mu + \Sigma t) e^{t'\mu + t'\Sigma t/2}}{\partial t'} \\
&= \left(\Sigma + (\mu + \Sigma t)(\mu + \Sigma t)' \right) e^{t'\mu + t'\Sigma t/2} \\
&\Rightarrow E[\mathbf{x}\mathbf{x}'] = \left.\frac{\partial^2 M_{\mathbf{x}}(t)}{\partial t \partial t'}\right|_{\mathbf{t}=0} = \Sigma + \mu\mu' \\
&\Rightarrow Var[\mathbf{x}] = E[\mathbf{x}\mathbf{x}'] - \mu\mu' = \Sigma
\end{aligned}$$

The lognormal moments follow from the moment generating function:

$$E\left[e^{X_j}\right] = E\left[e^{e_j' x}\right] = M_{\mathbf{x}}(e_j) = e^{e_j'\mu + e_j'\Sigma e_j/2} = e^{\mu_j + \Sigma_{jj}/2}$$

The second moments are conveniently expressed in terms of first moments:

$$\begin{aligned}
E\left[e^{X_j} e^{X_k}\right] &= E\left[e^{(e_j + e_k)' x} \right] \\
&= e^{(e_j + e_k)'\mu + (e_j + e_k)'\Sigma(e_j + e_k)/2} \\
&= e^{\mu_j + \mu_k + (\Sigma_{jj} + \Sigma_{jk} + \Sigma_{kj} + \Sigma_{kk})/2} \\
&= e^{\mu_j + \Sigma_{jj}/2 + \mu_k + \Sigma_{kk}/2 + (\Sigma_{jk} + \Sigma_{kj})/2} \\
&= e^{\mu_j + \Sigma_{jj}/2} \cdot e^{\mu_k + \Sigma_{kk}/2} \cdot e^{(\Sigma_{jk} + \Sigma_{kj})/2} \\
&= e^{\mu_j + \Sigma_{jj}/2} \cdot e^{\mu_k + \Sigma_{kk}/2} \cdot e^{(\Sigma_{jk} + \Sigma_{jk})/2} \\
&= E\left[e^{X_j}\right] E\left[e^{X_k}\right] \cdot e^{\Sigma_{jk}}
\end{aligned}$$

[23] The vector formulation of partial differentiation is explained in Appendix A.17 of Judge [1988].

So, $Cov\left[e^{X_j}, e^{X_k}\right] = E\left[e^{X_j}e^{X_k}\right] - E\left[e^{X_j}\right]E\left[e^{X_k}\right] = E\left[e^{X_j}\right]E\left[e^{X_k}\right]\left(e^{\Sigma_{jk}} - 1\right)$, which is the multivariate equivalent of the well-known scalar formula $CV^2\left[e^X\right] = Var\left[e^X\right] / E\left[e^X\right]^2 = e^{\sigma^2} - 1$. Letting $E\left[e^{\mathbf{x}}\right]$ denote the $n\times 1$ vector whose j^{th} element is $E\left[e^{X_j}\right]$,[24] and $diag\left(E\left[e^{\mathbf{x}}\right]\right)$ as its $n\times n$ diagonalization, we have $Var[\mathbf{x}] = diag\left(E\left[e^{\mathbf{x}}\right]\right)\left\{e^{\Sigma} - 1_{n\times n}\right\}diag\left(E\left[e^{\mathbf{x}}\right]\right)$. Because $diag\left(E\left[e^{\mathbf{x}}\right]\right)$ is diagonal in positive elements (hence, symmetric and PD), $Var[\mathbf{x}]$ is NND [or PD] if and only if $e^{\Sigma} - 1_{n\times n}$ is NND [or PD].[25] Symmetry is no issue here, because for real-valued matrices, Σ is symmetric if and only if $e^{\Sigma} - 1_{n\times n}$ is symmetric.

The relation between Σ and $\mathrm{T} = e^{\Sigma} - 1_{n\times n}$ is merits a discussion whose result will be clear from a consideration of 2×2 matrices. Since $\Sigma_{2\times 2}$ is symmetric, it is defined in terms of three real numbers: $\Sigma = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$. Now Σ is NND if and only if 1) $a \geq 0$, 2) $d \geq 0$, and 3) $ad - b^2 \geq 0$. Σ is PD if and only if these three conditions are strictly greater than zero. If *a* or *d* is zero, by the third condition *b* also must be zero.[26] Since we are not interested in degenerate random variables, which are effectively constants, we will require *a* and *d* to be positive. With this requirement, Σ is NND if and

---

[24] This would follow naturally from the "elementwise" interpretation of e^{A}, i.e., that the exponential function of matrix A is the matrix of the exponential functions of the elements of A. But if A is a square matrix, e^{A} may have the "matrix" interpretation $\mathrm{I}_n + \sum_{j=1}^{\infty} \mathrm{A}^j / j!$.

[25] PD [positive-definite] and NND [non-negative-definite] are defined in Section 4.

[26] Since NND matrices represent variances, $\Sigma = Var\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} Cov[X_1, X_1] & Cov[X_1, X_2] \\ Cov[X_2, X_1] & Cov[X_2, X_2] \end{bmatrix} = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$. The fact that *a* or *d* equals 0 implies that *b* equals 0 means that a random variable can't covary with another random variable unless it covaries with itself.

only if $b^2 \le ad$, and PD if and only if $b^2 < ad$. Since *a* and *d* are positive, so too is *ad*, as well as the geometric mean $\gamma = \sqrt{ad}$. So Σ is NND if and only if $-\gamma \le b \le \gamma$, and PD if and only if $-\gamma < b < \gamma$. It is well-known that $\min(a, d) \le \gamma \le \frac{a+d}{2} \le \max(a, d)$ with equality if and only if $a = d$.

Now the same three conditions determine the definiteness of $\mathrm{T} = e^{\Sigma} - 1_{2\times 2} = \begin{bmatrix} e^a - 1 & e^b - 1 \\ e^b - 1 & e^d - 1 \end{bmatrix}$.

Since we required *a* and *d* to be positive, both $e^a - 1$ and $e^d - 1$ are positive. This leaves the definiteness of T dependent on the relation between $(e^b - 1)(e^b - 1)$ and $(e^a - 1)(e^d - 1)$. We will next examine this relation according to the three cases $b = 0$, $b > 0$, and $b < 0$, all of which must be subject to $-\gamma \le b \le \gamma$.

First, if $b = 0$, then $-\gamma < b < \gamma$ and Σ is PD. Furthermore, $(e^b - 1)(e^b - 1) = 0 < (e^a - 1)(e^d - 1)$. Therefore, in this case, the lognormal transformation $\Sigma \to \mathrm{T} = e^{\Sigma} - 1_{2\times 2}$ is from PD to PD. And zero covariance in the normal pair produces zero covariance in the lognormal pair. In fact, since zero covariance between normal bivariates implies independence (cf. §2.5.7 of Judge [1988]), the lognormal bivariates also are independent.

In the second case, $b > 0$, or more fully, $0 < b \le \gamma$. Σ is PD if and only if $b < \gamma$. Define the function $\varphi(x) = \ln\left((e^x - 1)/x\right)$ for positive real *x* (or $x \in \Re^+$). A graph will show that the function strictly increases, i.e., $\varphi(x_1) < \varphi(x_2)$ if and only if $x_1 < x_2$. Moreover, the function is concave upward. This means that the line segment between points $(x_1, \varphi(x_1))$ and $(x_2, \varphi(x_2))$ lies above

the curve $\varphi(x)$ for intermediate values of *x*. In particular, $\frac{\varphi(x_1)+\varphi(x_2)}{2} \geq \varphi\left(\frac{x_1+x_1}{2}\right)$.

Equivalently, for all $x_1, x_2 \in \Re^+$, $2\varphi\left(\frac{x_1+x_1}{2}\right) \leq \varphi(x_1)+\varphi(x_2)$ with equality if and only if $x_1 = x_2$.

Therefore, since *a* and *d* are positive, $2\varphi\left(\frac{a+d}{2}\right) \leq \varphi(a)+\varphi(d)$. And since $0 < \gamma = \sqrt{ad} \leq \frac{a+d}{2}$,

$2\varphi(\gamma) \leq 2\varphi\left(\frac{a+d}{2}\right) \leq \varphi(a)+\varphi(d)$. So $2\varphi(\gamma) \leq \varphi(a)+\varphi(d)$ with equality if and only if $a = d$.

Furthermore, since in this case $0 < b \leq \gamma$, $2\varphi(b) \leq 2\varphi(\gamma) \leq \varphi(a)+\varphi(d)$. Hence, $2\varphi(b) \leq \varphi(a)+\varphi(d)$. The equality prevails if and only if $b = \gamma = a = d$, or if and only if $a = b = d$.

If $a = b = d$ then Σ is not PD; otherwise Σ is PD. Hence:

$$2\ln\left(\frac{e^b-1}{b}\right) = 2\varphi(b) \leq \varphi(a)+\varphi(d) = \ln\left(\frac{e^a-1}{a}\right)+\ln\left(\frac{e^d-1}{d}\right)$$

The inequality is preserved by exponentiation:

$$\left(\frac{e^b-1}{b}\right)^2 \leq \left(\frac{e^a-1}{a}\right)\left(\frac{e^d-1}{d}\right)$$

This leads at last to the inequality:

$$\left(e^b-1\right)^2 = b^2\left(\frac{e^b-1}{b}\right)^2 \leq \frac{b^2}{ad}\left(e^a-1\right)\left(e^d-1\right) = \left(\frac{b}{\gamma}\right)^2\left(e^a-1\right)\left(e^d-1\right) \leq 1\cdot\left(e^a-1\right)\left(e^d-1\right)$$

Therefore, in this case $\left(e^b-1\right)^2 \leq \left(e^a-1\right)\left(e^d-1\right)$ with equality if and only if $a = b = d$. This means that if $b > 0$, the lognormal transformation $\Sigma \to \mathrm{T} = e^{\Sigma} - 1_{2\times 2}$ is from NND to NND. But T is NND only if $\left(e^b-1\right)^2 = \left(e^a-1\right)\left(e^d-1\right)$, or only if $a = b = d$. Otherwise, T is PD. So, when $b > 0$, $\mathrm{T} = e^{\Sigma} - 1_{2\times 2}$ is PD except when all four elements of Σ have the same positive value. Even

the NND matrix $\Sigma = \begin{bmatrix} a & \sqrt{ad} \\ \sqrt{ad} & d \end{bmatrix}$ log-transforms into a PD matrix, unless $a = d$. So all PD and most NND normal variances transform into PD lognormal variances. A NND lognormal variance indicates that at least one element of the normal random vector is duplicated.

In the third and final case, $b < 0$, or more fully, $-\gamma \le b < 0$. This is equivalent to $0 < -b \le \gamma$, or to the second case with $-b$. In that case, $\left(e^{-b} - 1\right)^2 \le \left(e^a - 1\right)\left(e^d - 1\right)$ with equality if and only if $a = -b = d$. But from this, as well as from the fact that $0 < e^{2b} < e^{2 \cdot 0} < 1$, it follows:

$$\left(e^b - 1\right)^2 = e^{2b}\left(1 - e^{-b}\right)^2 = e^{2b}\left(e^{-b} - 1\right)^2 \le e^{2b}\left(e^a - 1\right)\left(e^d - 1\right) < 1 \cdot \left(e^a - 1\right)\left(e^d - 1\right)$$

So in this case the inequality is strict: $\left(e^b - 1\right)^2 < \left(e^a - 1\right)\left(e^d - 1\right)$, and T is PD. Therefore, if $b < 0$, the lognormal transform $\Sigma \to \mathrm{T} = e^{\Sigma} - 1_{2\times 2}$ is PD, even if Σ is NND.

To summarize, the lognormal transformation $\Sigma \to \mathrm{T} = e^{\Sigma} - 1_{2\times 2}$ is PD if Σ is PD. Even when Σ is not PD, but merely NND, T is almost always PD. Only when Σ is so NND as to conceal a random-variable duplication is its lognormal transformation NND.

The Hadamard (elementwise) product and Schur's product theorem allow for an understanding of the general lognormal transformation $\Sigma \to \mathrm{T} = e^{\Sigma} - 1_{n\times n}$. Denoting the elementwise n^{th} power of Σ as $\Sigma^{\circ n} = \overbrace{\Sigma \circ \ldots \circ \Sigma}^{n \text{ factors}}$, we can express elementwise exponentiation as $e^{\Sigma} = \sum_{j=0}^{\infty} \Sigma^{\circ j} / j!$. So $\mathrm{T} = e^{\Sigma} - 1_{n\times n} = \sum_{j=1}^{\infty} \Sigma^{\circ j} / j!$. According to Schur's theorem (§3 of Million [2007]), the Hadamard

product of two NND matrices is NND.[27] Since Σ is NND, its powers $\Sigma^{\circ j}$ are NND, as well as the terms $\Sigma^{\circ j}/j!$. Being the sum of a countable number of NND matrices, T also must be NND.[28] But if just one of the terms of the sum is PD, the sum itself must be PD. Therefore, if Σ is PD, then T also is PD.

Now the kernel of $m \times n$ matrix A is the set of all $\mathrm{x} \in \Re^n$ such that $\mathrm{Ax} = 0$, or $ker(\mathrm{A}) = \{\mathrm{x} : \mathrm{Ax} = 0\}$. The kernel is a linear subspace of $\Re^n$ and its dimensionality is $n - rank(\mathrm{A})$. By the Cholesky decomposition the NND matrix U can be factored as $\mathrm{U}_{n\times n} = \mathrm{W}'\mathrm{W}_{n\times n}$. The quadratic form in U is $\mathrm{x}'\mathrm{Ux} = \mathrm{x}'\mathrm{W}'\mathrm{Wx} = (\mathrm{Wx})'(\mathrm{Wx})$. If $\mathrm{x}'\mathrm{Ux} = 0$, then $\mathrm{Wx} = 0_{n\times 1}$, and $\mathrm{Ux} = \mathrm{W}'\mathrm{Wx} = \mathrm{W}'0_{n\times 1} = 0_{n\times 1}$. Conversely, if $\mathrm{Ux} = 0_{n\times 1}$, then $\mathrm{x}'\mathrm{Ux} = 0$. So the kernel of NND matrix U is precisely the solution set of $\mathrm{x}'\mathrm{Ux} = 0$, i.e., $\mathrm{x}'\mathrm{Ux} = 0$ if and only if $x \in ker(\mathrm{U})$. Therefore, the kernel of $\mathrm{T} = \sum_{j=1}^{\infty} \Sigma^{\circ j}/j!$ is the intersection of the kernels of $\Sigma^{\circ j}$, or $ker(\mathrm{T}) = \bigcap_{j=1}^{\infty} ker(\Sigma^{\circ j})$. It is possible for this intersection to be of dimension zero, i.e., for it to equal $\{0_{n\times 1}\}$, even though the kernel of no $\Sigma^{\circ j}$ is. Because of the accumulation of intersections in $\sum_{j=1}^{k\to\infty} \Sigma^{\circ j}/j!$, the lognormal transformation of NND matrix Σ tends to be "more PD" than Σ itself.

[27] Appendix C provides a proof of this theorem.

[28] For the quadratic form of a sum equals the sum of the quadratic forms: $\mathrm{x}'\left(\sum_k \mathrm{U}_k\right)\mathrm{x} = \sum_k \mathrm{x}'\mathrm{U}_k\mathrm{x}$. In fact, if the coefficients a_j are non-negative, then $\sum_{j=1}^{\infty} a_j \Sigma^{\circ j}$ is NND, provided that the sum converges, as e^{Σ} does.

We showed above that if Σ is PD, then $T = e^{\Sigma} - 1_{n\times n}$ is PD. But even if Σ is NND, T will be PD, unless Σ contains a 2×2 subvariance $\begin{bmatrix} \Sigma_{jj} & \Sigma_{jk} \\ \Sigma_{kj} & \Sigma_{kk} \end{bmatrix}$ whose four elements are all equal.

APPENDIX B

THE NORMAL MOMENT GENERATING FUNCTION AND COMPLEX ARGUMENTS

In Appendix A we saw that the moment generating function of the real-valued normal multivariate $\mathbf{x} \sim N(\mu, \Sigma)$, viz., $M_{\mathbf{x}}(t) = e^{t'\mu + t'\Sigma t/2}$, is valid at least for $t \in \Re^n$. The validity rests on the identity $\int_{x \in \Re^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-[\mu+\Sigma t])' \Sigma^{-1} (x-[\mu+\Sigma t])} dV = 1$ for real-valued $\xi = \mu + \Sigma t$. But in Section 8 we must know the value of $\varphi(\xi) = \int_{x \in \Re^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\xi)' \Sigma^{-1} (x-\xi)} dV$ when ξ is a complex $n \times 1$ vector. So in this appendix, we will prove that for all $\xi \in C^n$, $\varphi(\xi) = 1$.

The proof begins with diagonalization. Since Σ is symmetric and PD, Σ^{-1} exists and is symmetric and PD. According to the Cholesky decomposition (Healy [1986, §7.2]), there exists a real-valued $n \times n$ matrix W such that $W'W = \Sigma^{-1}$. Due to theorems on matrix rank, W must be non-singular, or invertible. So the transformation $y = Wx$ is one-to-one. And letting $\zeta = W\xi$, we have $y - \zeta = W(x - \xi)$. Moreover, the volume element in the y coordinates is:

$$dV_y = \|W\| dV_x = \sqrt{|W|^2} dV_x = \sqrt{|W'||W|} dV_x = \sqrt{|W'W|} dV_x = \sqrt{|\Sigma^{-1}|} dV_x = \frac{dV_x}{\sqrt{|\Sigma|}}.$$

Hence:

$$
\begin{aligned}
\varphi(\xi) &= \int_{x\in\Re^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(x-\xi)'\Sigma^{-1}(x-\xi)} dV \\
&= \int_{x\in\Re^n} \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(x-\xi)'W'W(x-\xi)} dV_x \\
&= \int_{x\in\Re^n} \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}[W(x-\xi)]'[W(x-\xi)]} \frac{dV_x}{\sqrt{|\Sigma|}} \\
&= \int_{y\in\Re^n} \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}(y-\zeta)'(y-\zeta)} dV_y \\
&= \int_{y\in\Re^n} \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\sum_{j=1}^{n}(y_j-\zeta_j)^2} dV \\
&= \int_{y\in\Re^n} \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{j=1}^{n}(y_j-\zeta_j)^2} dV \\
&= \prod_{j=1}^{n} \int_{y_j=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{j=1}^{n}(y_j-\zeta_j)^2} dy_j \\
&= \prod_{j=1}^{n} \psi(\zeta_j)
\end{aligned}
$$

In the last line $\psi(\zeta) = \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{x=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\zeta)^2} dx$. Obviously, if ζ is real-valued, $\psi(\zeta)=1$. So the issue of the value of a moment generating function of a complex variable resolves into the issue of the "total probability" of a unit-variance normal random variable with a complex mean.[29]

To evaluate $\psi(\zeta)$ requires some complex analysis with contour integrals. First, consider the standard-normal density function with a complex argument: $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$. By function-composition rules, since both z^2 and the exponential function are "entire" functions (i.e., analytic

[29] We deliberately put 'total probability' in quotes because the probability density function with complex ζ is not proper; it may produce negative and even complex values for probability densities.

over the whole complex plane), so too is $f(z)$. Therefore, $\oint_C f(z)dz = 0$ for any closed contour *C* (cf. Appendix D.7 of Havil [2003]). Let *C* be the parallelogram traced from vertex $z = b$ to vertex $z = a$ to vertex $z = a - \zeta$ to vertex $z = b - \zeta$ and finally back to vertex $z = b$. Therefore:

$$\begin{aligned} 0 &= \oint_C f(z)dz \\ &= \int_b^a f(z)dz + \int_a^{a-\zeta} f(z)dz + \int_{a-\zeta}^{b-\zeta} f(z)dz + \int_{b-\zeta}^{b} f(z)dz \end{aligned}$$

The line segments along which the second and fourth integrals proceed are finite; their common length is $L = |(a-\zeta)-a| = |-\zeta| = |\zeta| = |b-(b-\zeta)|$, where $|\zeta| = \sqrt{\zeta\bar{\zeta}} \geq 0$. By the triangle inequality $\left|\int_a^{a-\zeta} f(z)dz\right| \leq \int_a^{a-\zeta} |f(z)dz| = \int_a^{a-\zeta} |f(z)||dz|$. But $|f(z)|$ is a continuous real-valued function, so over a closed interval it must be upper-bounded by some positive real number *M*. Hence, $\left|\int_a^{a-\zeta} f(z)dz\right| \leq \int_a^{a-\zeta} |f(z)||dz| \leq \int_a^{a-\zeta} Sup(|f(z)|)|dz| = Sup(|f(z \in [a, a-\zeta])|)\int_a^{a-\zeta} |dz| = M(a)\cdot L$. Likewise,

$$\left|\int_{b-\zeta}^{b} f(z)dz\right| \leq Sup(|f(z \in [b, b-\zeta])|)\int_{b-\zeta}^{b} |dz| = M(b)\cdot L.$$

Now, in general:

$$\begin{aligned}
\left|f(z)\right| &= \sqrt{f(z)\overline{f(z)}} \\
&= \sqrt{\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \cdot \overline{\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}}} \\
&= \frac{\sqrt{e^{-\frac{z^2}{2}} \cdot \overline{e^{-\frac{z^2}{2}}}}}{\sqrt{2\pi}} \\
&\propto \sqrt{e^{-\frac{z^2}{2}} \cdot \overline{e^{-\frac{z^2}{2}}}} \\
&\propto \sqrt{e^{-\frac{z^2+\overline{z^2}}{2}}} \\
&\propto \sqrt{e^{-\mathrm{Re}\left(z^2\right)}}
\end{aligned}$$

Therefore, since $a \in \Re$:

$$\begin{aligned}
\lim_{a\to-\infty} M(a)\cdot L &\propto \lim_{a\to-\infty} Sup\left(\sqrt{e^{-\mathrm{Re}\left(z^2\right)}};\, z \in [a,\, a-\zeta]\right)\cdot L \\
&\propto \lim_{a^2\to+\infty} Sup\left(\sqrt{e^{-a^2\,\mathrm{Re}\left(\left(\frac{z}{a}\right)^2\right)}};\, \frac{z}{a} \in \left[1,\, 1-\frac{\zeta}{a}\right]\right)\cdot L \\
&\propto \sqrt{e^{-\infty\,\mathrm{Re}(1)}}\cdot L \\
&\propto 0
\end{aligned}$$

Similarly, $\lim_{b\to+\infty} M(b)\cdot L \propto 0$. So in the limit as $a \to -\infty$ and $b \to +\infty$ on the real axis, the second and fourth integrals approach zero. Accordingly:

$$\begin{aligned}
0 &= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \{0\} \\
&= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \left\{\int_b^a f(z)dz + \int_a^{a-\zeta} f(z)dz + \int_{a-\zeta}^{b-\zeta} f(z)dz + \int_{b-\zeta}^{b} f(z)dz\right\} \\
&= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \left\{\int_b^a f(z)dz + \int_{a-\zeta}^{b-\zeta} f(z)dz\right\}
\end{aligned}$$

And so, $\lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{a-\zeta}^{b-\zeta} f(z)dz = -\lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_b^a f(z)dz = \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_a^b f(z)dz = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 1$

So, at length:

$$\begin{aligned}
\psi(\zeta) &= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{x=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\zeta)^2} dx \\
&= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{x=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}((z+\zeta)-\zeta)^2} d(z+\zeta) \\
&= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{z=a-\zeta}^{b-\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} d(z+\zeta) \\
&= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{z=a-\zeta}^{b-\zeta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\
&= \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{z=a-\zeta}^{b-\zeta} f(z) dz \\
&= 1
\end{aligned}$$

So, working backwards, what we proved for one dimension, viz., $\psi(\zeta) = \int_{x=-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\zeta)^2} dx = 1$,

applies *n*-dimensionally: for all $\xi \in C^n$, $\varphi(\xi) = \int_{\mathrm{x}\in\Re^n} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(\mathrm{x}-\xi)'\Sigma^{-1}(\mathrm{x}-\xi)} dV = 1$. Therefore, even

for complex t, $M_{\mathbf{x}}(\mathrm{t}) = e^{\mathrm{t}'\mu+\mathrm{t}'\Sigma\mathrm{t}/2}$. Complex values are allowable as arguments in the moment generating function of a real-valued normal vector. This result is critical to Section 8.

Though we believe the contour-integral proof above to be worthwhile for its instructional value, a simple proof comes from the powerful theorem of analytic continuation (cf. Appendix D.12 of Havel [2003]). This theorem concerns two functions that are analytic within a common domain. If the functions are equal over any smooth curve within the domain, no matter how short,[30] then they

[30] The length of the curve must be positive; equality at single points, or punctuated equality, does not qualify.

are equal over all the domain. Now $\psi(\zeta) = \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{x=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\zeta)^2} dx$ is analytic over all the complex plane. And for all real-valued ζ, $\psi(\zeta) = 1$. So $\psi(\zeta)$ and $f(\zeta) = 1$ are two functions analytic over the complex plane and identical on the real axis. Therefore, by analytic continuation $\psi(\zeta)$ must equal one for all complex ζ. Analytic continuation is analogous with the theorem in real analysis that two smooth functions equal over any interval are equal everywhere. Analytic continuation derives from the fact that a complex derivative is the same in all directions. It is mistaken to regard the real and imaginary parts of the derivative as partial derivatives, as if they applied respectively to the real and imaginary axes of the independent variable. Rather, the whole derivative applies in every direction.

APPENDIX C

EIGEN-DECOMPOSITION AND SCHUR'S PRODUCT THEOREM

Appendix A quoted Schur's Product Theorem, viz., that the Hadamard product of non-negative-definite (NND) matrices is NND. Million [2007] proves it as Theorem 3.4; however, we believe our proof in this appendix to be simpler; moreover, it affords a review of eigen-decomposition. Those familiar with eigen-decomposition may skip to the last paragraph.

Let Γ be an $n\times n$ Hermetian NND matrix. As explained in Section 4, 'Hermetian' means that $\Gamma = \Gamma^*$; 'NND' means that for every complex $n\times n$ vector z (or $\mathbf{z} \in C^n$), $\mathbf{z}^*\Gamma\mathbf{z} \geq 0$. Positive definiteness [PD] is a stricter condition, in which $\mathbf{z}^*\Gamma\mathbf{z} = 0$ if and only if $\mathbf{z} = 0_{n\times 1}$.

Complex scalar λ and non-zero vector ν form an "eigenvalue-eigenvector" pair of Γ, if $\Gamma\nu = \lambda\nu$. Since $\nu = 0_{n\times 1}$ is excluded as a trivial solution, vector ν can be scaled to unity, or $\nu^*\nu = 1$. But $\Gamma\nu = \lambda\nu$ if and only if $(\Gamma - \lambda \mathrm{I}_n)\nu = 0_{n\times 1}$. If $\Gamma - \lambda \mathrm{I}_n$ is non-singular, or invertible, then:

$$\nu = \mathrm{I}_n\nu = (\Gamma - \lambda \mathrm{I}_n)^{-1}(\Gamma - \lambda \mathrm{I}_n)\nu = (\Gamma - \lambda \mathrm{I}_n)^{-1} 0_{n\times 1} = 0_{n\times 1}$$

Hence, allowable eigenvectors require for $\Gamma - \lambda \mathrm{I}_n$ to be singular, or for its determinant $|\Gamma - \lambda \mathrm{I}_n|$ to be zero. Since the determinant is an n^{th}-degree equation (with complex coefficients based on the elements of Γ) in λ, it has n root values of λ, not necessarily distinct. So the determinant can be factored as $f(\lambda) = |\Gamma - \lambda \mathrm{I}_n| = (\lambda - \lambda_1)\ldots(\lambda - \lambda_n)$. Since $f(\lambda_j) = |\Gamma - \lambda_j \mathrm{I}_n| = 0$, there exist non-zero solutions to $(\Gamma - \lambda \mathrm{I}_n)\nu = 0_{n\times 1}$. So for every eigenvalue there is a non-zero eigenvector, even a non-zero eigenvector of unit magnitude.

The first important result is that the eigenvalues of Γ are real-valued and non-negative. Consider the j^{th} eigenvalue-eigenvector pair, which satisfies the equation $\Gamma\nu_j = \lambda_j \nu_j$. Therefore, $\nu_j^* \Gamma \nu_j = \lambda_j \nu_j^* \nu_j$. Since Γ is NND, $\nu_j^* \Gamma \nu_j$ is real-valued and non-negative. Also, $\nu_j^* \nu_j$ is real-valued and positive. Therefore, their quotient λ_j is a real-valued and non-negative scalar. Furthermore, if Γ is PD, $\nu_j^* \Gamma \nu_j$ is positive, as well as λ_j.

The second important result is that eigenvectors paired with unequal eigenvalues are orthogonal. Let the two unequal eigenvalues be $\lambda_j \neq \lambda_k$. Because the eigenvalues are real-valued, $\overline{\lambda}_j = \lambda_j$. The eigenvector equations are $\Gamma\nu_j = \lambda_j \nu_j$ and $\Gamma\nu_k = \lambda_k \nu_k$. The following string of equations relies on Γ's being Hermetian (so $\Gamma = \Gamma^*$):

$$\begin{aligned}
\left(\lambda_j - \lambda_k\right)\nu_j^* \nu_k &= \lambda_j \nu_j^* \nu_k - \lambda_k \nu_j^* \nu_k \\
&= \overline{\lambda}_j \nu_j^* \nu_k - \lambda_k \nu_j^* \nu_k \\
&= \left(\lambda_j \nu_k^* \nu_j\right)^* - \lambda_k \nu_j^* \nu_k \\
&= \left(\nu_k^* \Gamma \nu_j\right)^* - \nu_j^* \Gamma \nu_k \\
&= \nu_j^* \Gamma^* \nu_k - \nu_j^* \Gamma \nu_k \\
&= \nu_j^* \Gamma \nu_k - \nu_j^* \Gamma \nu_k \\
&= 0
\end{aligned}$$

Because $\lambda_j - \lambda_k \neq 0$, the eigenvectors must be orthogonal, or $\nu_j^* \nu_k = 0$. If all the eigenvectors are distinct, the eigenvectors form an orthogonal basis of C^n. But even if not, the kernel of each eigenvalue, or $ker\left(\Gamma - \lambda_j \mathrm{I}_n\right) = \left\{\mathrm{z} \in C^n : \Gamma\mathrm{z} = \lambda_j \mathrm{z}\right\}$ is a linear subspace of C^n whose rank or dimensionality equals the multiplicity of the root λ_j in the characteristic equation $f(\lambda) = \left|\Gamma - \lambda \mathrm{I}_n\right|$. This means that the number of mutually orthogonal eigenvectors paired with an eigenvalue equals

how many times that eigenvalue is a root of its characteristic equation. Consequently, there exist n eigenvalue-eigenvector pairs (λ_j, ν_j) such that $\Gamma\nu_j = \lambda_j\nu_j$ and $\nu_j^*\nu_k = \delta_{ij}$.[31]

Now, define W as the partitioned matrix $\mathrm{W}_{n\times n} = [\nu_1 \quad \cdots \quad \nu_n]$. The jk^{th} element of $\mathrm{W}^*\mathrm{W}$ equals $\nu_j^*\nu_k = \delta_{ij}$; hence, $\mathrm{W}^*\mathrm{W} = \mathrm{I}_n$. A matrix whose transjugate is its inverse is called "unitary," as is W.[32] Furthermore, define Λ as the $n\times n$ diagonal matrix whose jj^{th} element is λ_j. Then:

$$\Gamma\mathrm{W} = \Gamma[\nu_1 \quad \cdots \quad \nu_n] = [\Gamma\nu_1 \quad \cdots \quad \Gamma\nu_n] = [\lambda_1\nu_1 \quad \cdots \quad \lambda_n\nu_n] = [\nu_1 \quad \cdots \quad \nu_n]\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix} = \mathrm{W}\Lambda$$

And so, $\Gamma = \Gamma\mathrm{I}_n = \Gamma\mathrm{W}\mathrm{W}^* = \mathrm{W}\Lambda\mathrm{W}^*$, and Γ is said to be "diagonalized." Thus have we shown, assuming the theory of equations,[33] the third important result, viz., that every NND Hermetian matrix can be diagonalized. Other matrices can be diagonalized; the NND [or PD] consists in the fact that all the eigenvalues of this diagonalization are non-negative [or positive].

The fourth and final "eigen" result relies on the identity $\mathrm{W}^*\nu_j = \mathrm{e}_j$, which just extracts the j^{th} columns of each side of $\mathrm{W}^*\mathrm{W} = \mathrm{I}_n$. As in Appendix A, e_j is the j^{th} unit vector. Therefore:

---

[31] The Kronecker delta, δ_{ij}, is the function $\mathrm{IF}(i = j, 1, 0)$.

[32] To be precise, at this point W* is only the left-inverse of W. But by matrix-rank theorems, the rank of W equals n, so W has a unique full inverse W^{-1}. Then $\mathrm{W}^* = \mathrm{W}^*\mathrm{I}_n = \mathrm{W}^*(\mathrm{W}\mathrm{W}^{-1}) = (\mathrm{W}^*\mathrm{W})\mathrm{W}^{-1} = \mathrm{I}_n\mathrm{W}^{-1} = \mathrm{W}^{-1}$.

[33] The theory of equations guarantees the existence of the roots of the n^{th}-degree equation $f(\lambda) = |\Gamma - \lambda\mathrm{I}_n|$. Appendix D.9 of Havel [2003] contains a quick and lucid proof of this, the fundamental theorem of algebra.

$$
\begin{aligned}
\Gamma &= W\Lambda W^* \\
&= W\left(\sum_{j=1}^{n}\lambda_j e_j e_j^*\right)W^* \\
&= W\left(\sum_{j=1}^{n}\lambda_j \left(W^*\nu_j\right)\left(W^*\nu_j\right)^*\right)W^* \\
&= W\left(\sum_{j=1}^{n}\lambda_j W^*\nu_j\nu_j^* W\right)W^* \\
&= WW^*\left(\sum_{j=1}^{n}\lambda_j \nu_j\nu_j^*\right)WW^* \\
&= I_n\left(\sum_{j=1}^{n}\lambda_j \nu_j\nu_j^*\right)I_n \\
&= \sum_{j=1}^{n}\lambda_j \nu_j\nu_j^*
\end{aligned}
$$

The form $\sum_{j=1}^{n}\lambda_j \nu_j\nu_j^*$ is called the "spectral decomposition" of Γ (§7.4 of Healy [1986]), which plays the leading role in the following succinct proof of Schur's product theorem.

If Σ and T are two $n\times n$ Hermetian NND definite matrices, we may spectrally decompose them as $\Sigma = \sum_{j=1}^{n}\lambda_j \nu_j\nu_j^*$ and $T = \sum_{j=1}^{n}\kappa_j \eta_j\eta_j^*$, where all the λ and $\varkappa$ scalars are non-negative. Accordingly:

$$
\begin{aligned}
(\Sigma \circ \mathrm{T})_{jk} &= (\Sigma)_{jk}(\mathrm{T})_{jk} \\
&= \left(\sum_{r=1}^{n} \lambda_r \nu_r \nu_r^*\right)_{jk} \left(\sum_{s=1}^{n} \kappa_s \eta_s \eta_s^*\right)_{jk} \\
&= \left(\sum_{r=1}^{n} \lambda_r (\nu_r)_j (\overline{\nu}_r)_k\right)\left(\sum_{s=1}^{n} \kappa_s (\eta_s)_j (\overline{\eta}_s)_k\right) \\
&= \sum_{r=1}^{n}\sum_{s=1}^{n} \lambda_r \kappa_s (\nu_r)_j (\eta_s)_j (\overline{\nu}_r)_k (\overline{\eta}_s)_k \\
&= \sum_{r=1}^{n}\sum_{s=1}^{n} \lambda_r \kappa_s (\nu_r \circ \eta_s)_j (\overline{\nu}_r \circ \overline{\eta}_s)_k \\
&= \sum_{r=1}^{n}\sum_{s=1}^{n} \lambda_r \kappa_s (\nu_r \circ \eta_s)_j \left(\overline{\nu_r \circ \eta_s}\right)_k \\
&= \sum_{r=1}^{n}\sum_{s=1}^{n} \lambda_r \kappa_s \left((\nu_r \circ \eta_s)(\nu_r \circ \eta_s)^*\right)_{jk} \\
&= \left(\sum_{r=1}^{n}\sum_{s=1}^{n} \lambda_r \kappa_s (\nu_r \circ \eta_s)(\nu_r \circ \eta_s)^*\right)_{jk}
\end{aligned}
$$

Hence, $\Sigma \circ \mathrm{T} = \sum_{r=1}^{n}\sum_{s=1}^{n} \lambda_r \kappa_s (\nu_r \circ \eta_s)(\nu_r \circ \eta_s)^*$. Since each matrix $(\nu_r \circ \eta_s)(\nu_r \circ \eta_s)^*$ is Hermetian NND, and each scalar $\lambda_r \kappa_s$ is non-negative, $\Sigma \circ \mathrm{T}$ must be Hermetian NND. Therefore, the Hadamard product of NND matrices is NND.

The Gauss-Markov Theorem: Beyond the BLUE

Leigh J. Halliwell, FCAS, MAAA

Abstract: Until now the Gauss-Markov theorem has been the handmaid of least squares; it has served as a proof that the least-squares method produces the Best Linear Unbiased Estimator (BLUE). This theoretical paper shows that it can be, and should be, reformulated as the solution to the problem of the minimization of a quadratic form subject to a linear constraint. The whole theory of linear statistical modeling, from basic to complicated, receives a clean and efficient development on the basis of this reformulation; estimates and predictions based thereon are BLUE from the start, rather than BLUE by subsequent proof. With an intermediate-level background in matrix algebra the reader will understand the frequent interpretations of this development in terms of an n-dimensional projective geometry. Because this paper elevates BLUE to its true role, "Beyond the BLUE" really means "To the True BLUE."

Keywords: Gauss-Markov, BLUE, linear model, projection, distance metric

1. INTRODUCTION

The many treatments of the Gauss-Markov theorem (e.g., Judge [1988, 202-206], Halliwell [2007, Appendix B], and Wikipedia) lead one to believe that the theorem is no more than a proof that $\hat{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$ is best linear unbiased estimator (BLUE) of β in the model $\mathbf{y} = \mathrm{X}\beta + \mathbf{e}$, where $Var[\mathbf{e}] = \sigma^2\mathrm{I}$. In this capacity the theorem is impressive enough; however, with a little abstraction it becomes much more, as we shall see in the following eleven sections.

2. STATEMENT OF THE THEOREM AND ITS PROOF

The Gauss-Markov theorem is essentially the solution to a constrained-optimization problem, more exactly, to the problem of minimizing a quadratic form subject to a linear constraint. Here is our formulation of the theorem:

> The Gauss-Markov Theorem: If symmetric $\Sigma_{n\times n}$ is positive-definite and $\mathrm{A}_{m\times n}$ is of full-row rank, then $\Phi(\mathrm{W}) = \mathrm{W}'\Sigma^{-1}\mathrm{W}$ can be minimized subject to the linear constraint $\mathrm{A}_{m\times n}\mathrm{W}_{n\times p} = \mathrm{B}_{m\times p}$. The value $\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$ uniquely minimizes Φ at $\Phi(\mathrm{W}^*) = \mathrm{B}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$.

To prove the theorem, we take for granted two theorems about positive-definite matrices.[1] First, positive-definite matrices have inverses; the inverses also are positive-definite. Therefore, symmetric Σ^{-1} exists, and is positive-definite. Second, if $\mathrm{A}_{m\times n}$ is of full-row rank and $\mathrm{T}_{n\times n}$ is positive-definite, then ATA' is positive-definite. From these it follows that $\mathrm{A}\Sigma\mathrm{A}'$ is positive-definite and invertible; hence, $\mathrm{W}^* = \Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}$ exists. Moreover, W^* satisfies the constraint, since $\mathrm{AW}^* = \mathrm{A}\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B} = \mathrm{I}_m\mathrm{B} = \mathrm{B}$.

Now if W_1 satisfies the constraint, then:

$$\begin{aligned}\mathrm{W}_1'\Sigma^{-1}\mathrm{W}^* &= \mathrm{W}_1'\Sigma^{-1}\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}\\ &= \mathrm{W}_1'\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}\\ &= \left(\mathrm{AW}_1\right)'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}\\ &= \mathrm{B}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}\end{aligned}$$

And since W^* is an allowable instance of W_1, we have the following chain of equalities:

$$\mathrm{W}_1'\Sigma^{-1}\mathrm{W}^* = \mathrm{B}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B} = \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^* = \left(\mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\right)' = \left(\mathrm{W}_1'\Sigma^{-1}\mathrm{W}^*\right)' = \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}_1$$

As the heart of the Gauss-Markov proof:

$$\begin{aligned}\Phi\left(\mathrm{W}_1\right) - \Phi\left(\mathrm{W}^*\right) &= \mathrm{W}_1'\Sigma^{-1}\mathrm{W}_1 - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\\ &= \mathrm{W}_1'\Sigma^{-1}\mathrm{W}_1 - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^* - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^* + \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\\ &= \mathrm{W}_1'\Sigma^{-1}\mathrm{W}_1 - \mathrm{W}_1'\Sigma^{-1}\mathrm{W}^* - \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}_1 + \mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*\\ &= \left(\mathrm{W}_1 - \mathrm{W}^*\right)'\Sigma^{-1}\left(\mathrm{W}_1 - \mathrm{W}^*\right)\\ &\geq 0_{p\times p}\end{aligned}$$

[1] For a review of positive-definite and non-negative-definite (or positive-semi-definite) matrices see Judge [1988, Appendix A.14] and Halliwell [1997, Appendix A].

The last line is to be taken in a matrix-definite sense, viz., that the difference $\Phi(\mathrm{W}_1)-\Phi(\mathrm{W}^*)$ is the non-negative-definite matrix $(\mathrm{W}_1-\mathrm{W}^*)'\Sigma^{-1}(\mathrm{W}_1-\mathrm{W}^*)$. And because Σ^{-1} is positive-definite, the difference equals the zero matrix ($0_{p\times p}$) if and only if $\mathrm{W}_1=\mathrm{W}^*$. Therefore, $\mathrm{W}^*=\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$ uniquely minimizes $\Phi(\mathrm{W})=\mathrm{W}'\Sigma^{-1}\mathrm{W}$ subject to $\mathrm{AW}=\mathrm{B}$. Furthermore, the minimum is $\Phi(\mathrm{W}^*)=\mathrm{B}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$.

3. GEOMETRICAL INTERPRETATION WITH A DISTANCE METRIC

A geometrical interpretation of the theorem will prove helpful. Again, let W_1 satisfy the constraint, and let $\mathrm{W}^*=\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$. From the chain of equalities, we derive:

$$(\mathrm{W}_1-\mathrm{W}^*)'\Sigma^{-1}\mathrm{W}^*=\mathrm{W}_1'\Sigma^{-1}\mathrm{W}^*-\mathrm{W}^{*\prime}\Sigma^{-1}\mathrm{W}^*=0_{p\times p}=(0_{p\times p})'=\mathrm{W}^{*\prime}\Sigma^{-1}(\mathrm{W}_1-\mathrm{W}^*)$$

These are unusual quadratic forms. The usual quadratic form is $\mathrm{y}'\Sigma\mathrm{x}$, where the factors before and after Σ^{-1} are $1\times n$ and $n\times 1$ vectors. Here the form is $\mathrm{Y}'\Sigma^{-1}\mathrm{X}$, where the factors before and after Σ^{-1} are $p\times n$ and $n\times p$ matrices, and the integer p may exceed one.

But for now, consider the usual quadratic form in the special case that $\Sigma=\mathrm{I}_n$. Actuaries know that $\mathrm{x}'\mathrm{I}_n\mathrm{x}=\mathrm{x}'\mathrm{x}=\sum_{i=1}^{n}x_i^2$ is the square of the distance from the origin of $\Re^n$ to x (or the area of a square the length of whose sides is that distance). Less well known is that $\mathrm{y}'\mathrm{x}=\mathrm{x}'\mathrm{y}=\sum_{i=1}^{n}x_iy_i$ represents the area of a rectangle, the length of one of whose sides is

the length of the projection of one vector onto the other. Most will recognize, however, an equivalent interpretation, viz., that $\mathrm{y}'\mathrm{x} = \mathrm{x}'\mathrm{y} = 0$ if and only if $\mathrm{x} \perp \mathrm{y}$. The standard (Euclidean) definition of the distance from x to y is $d(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{y}-\mathrm{x})'(\mathrm{y}-\mathrm{x})}$. It has the three properties of a metric on $\Re^n$:

1. $d(\mathrm{x},\mathrm{y}) \geq 0;\ d(\mathrm{x},\mathrm{y}) = 0 \Leftrightarrow \mathrm{x} = \mathrm{y}$ non – negativity; trivially zero
2. $d(\mathrm{y},\mathrm{x}) = d(\mathrm{x},\mathrm{y})$ symmetry
3. $d(\mathrm{x},\mathrm{y}) + d(\mathrm{y},\mathrm{z}) \geq d(\mathrm{x},\mathrm{z})$ triangle inequality

But for any positive-definite matrix $\Sigma_{n\times n}$, one can define a valid "Σ metric" on $\Re^n$ as $d_\Sigma(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{y}-\mathrm{x})'\Sigma^{-1}(\mathrm{y}-\mathrm{x})}$, which is valid in that it possesses these three properties.[2] The matrix Σ represents a combination of scaling and rotating the axes of $\Re^n$.

So what is special in the Gauss-Markov theorem about $\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$? Adapting the concept of perpendicularity to a metric, we have:

$$(\mathrm{W}_1 - \mathrm{W}^*)'\Sigma^{-1}\mathrm{W}^* = 0_{p\times p}$$

[2] Some confusion results from using the inverse of Σ in the quadratic form; one must think twice to determine whether something is a Σ metric or a Σ^{-1} metric. However, consider the usual formula for the ellipse whose major semi-axis is two units and minor semi-axis is one: $(x_1/2)^2 + (x_2/1)^2 = 1^2$. As a quadratic form this would be:

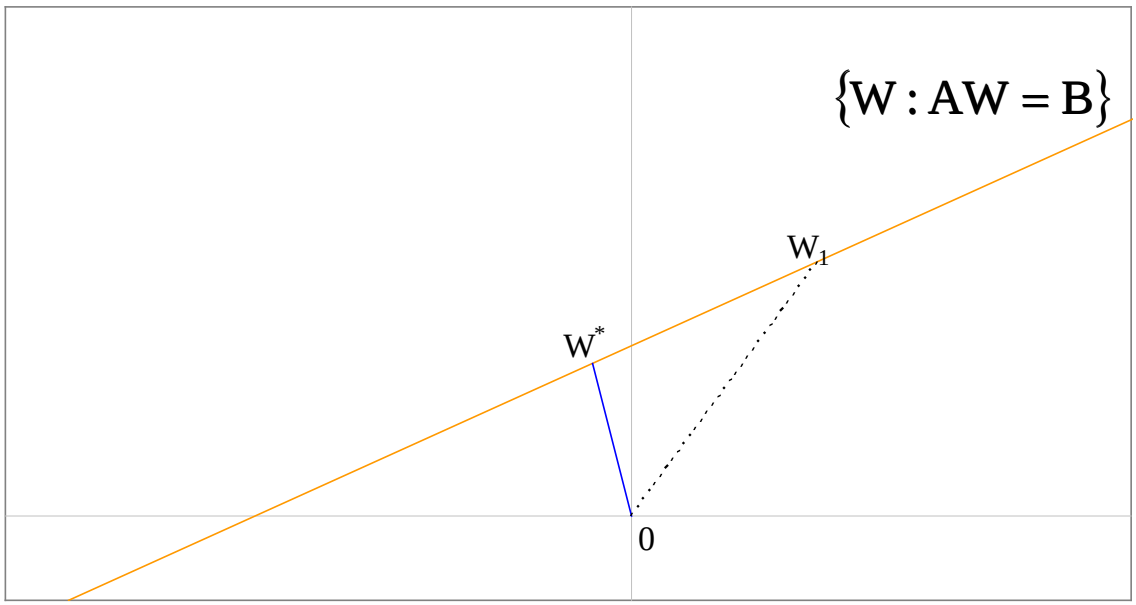
$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1^2$$

It seemed more natural to call this a [2 0; 0 1] metric (as if to say, "Two units on the first axis count as one unit on the second."), rather than to call it a [½ 0; 0 1] metric. This ellipse is the set of points in $\Re^2$ whose distance from the origin is one unit according to the [2 0; 0 1] metric. It may help some readers to know that $d_\Sigma(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{y}-\mathrm{x})'\Sigma^{-1}(\mathrm{y}-\mathrm{x})}$ is called the "Mahalanobis distance" (cf. Wikipedia), in whose definition the Σ matrix is inverted. Appendix A provides a proof of the triangle inequality, as well as a justification of the geometric interpretation of $\mathrm{x}'\mathrm{y}$ as the product of the length of x and the length of the projection of y onto x.

This means that according to the Σ metric W^* is perpendicular to $W_1 - W^*$ (and vice versa). In mathematical notation, $W^* \underset{\Sigma}{\perp} (W_1 - W^*)$. The heart of the Gauss-Markov theorem, expressed above as $\Phi(W_1) - \Phi(W^*) = (W_1 - W^*)' \Sigma^{-1} (W_1 - W^*)$, is really just the Pythagorean theorem adapted to the Σ metric:

$$W^{*\prime} \Sigma^{-1} W^* + (W_1 - W^*)' \Sigma^{-1} (W_1 - W^*) = W_1' \Sigma^{-1} W_1$$

W^* is the element of the constraint set closest to the origin according to the Σ metric. The following diagram clarifies this:



The orange line represents the constraint set.[3] The origin, W^*, and W_1 form the Σ-right triangle, of which W_1 is the hypotenuse, and W^* and $W_1 - W^*$ are the legs. The salient point is that the Σ area of the square with side W^* is less than or equal to that of the square with side W_1, or $W^{*\prime} \Sigma W^* \le W_1' \Sigma W_1$, and equal if and only if $W_1 = W^*$. This is valid even

[3] Since the constraint on W is linear, the constraint set is a hyperplane (technically, an affine apace). The Gauss-Markov theorem requires a *linear* constraint; constraints involving curvature are inadmissible.

when the area concept is abstracted from a non-negative scalar to a non-negative-definite matrix.

This ends the geometric interpretation. Gauss-Markov reasoning happens whenever a quadratic form is to be minimized subject to a linear constraint. Gauss-Markov/BLUE proofs are abstractions of what we all learned in plane Geometry, viz., that the shortest distance from a point to a straight line is along a line segment perpendicular to the line. Lines are abstracted into linear constraints and distance is abstracted into a Σ metric.

It is hardly necessary to memorize the formula for W^*. With the following heuristic reasoning one can derive it on the fly. Since $\mathrm{A}_{m\times n}$ is of full-row rank (or of rank *m*), the $m\times m$ matrix AA' is invertible. In fact, as stated above, for any positive-definite $\mathrm{T}_{n\times n}$, ATA' is invertible. Thus, there is a family of "right inverses" of A that have the form $\mathrm{TA}'(\mathrm{ATA}')^{-1}$. W^* will be the matrix product of one of these right inverses and B, i.e., $\mathrm{W}^* = \mathrm{TA}'(\mathrm{ATA}')^{-1}\mathrm{B}$. Since we seek to minimize $\mathrm{W}'\Sigma^{-1}\mathrm{W}$, distance is measured according to a $\mathrm{T} = \Sigma$ metric. According to this metric $\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$ is the element of the constraint set closest to the origin.

4. PROJECTION INTO THE CONSTRAINT SPACE

In the interest of economy and precision, let us introduce some more formalism. Our 'W' variables denote elements of $\Re^{n\times p}$, the real space of $n\times p$ dimensions. Let us use 'Ω' to denote the constraint set: $\Omega = \left\{\mathrm{W}\in\Re^{n\times p} : \mathrm{AW} = \mathrm{B}\right\}$. Obviously, $\Omega\subseteq\Re^{n\times p}$; but it is not empty under the assumption that $\mathrm{A}_{m\times n}$ is of full-row rank. In fact, we have just seen that

$\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$ is the element of Ω closest to the origin of $\Re^{n\times p}$ according to the Σ metric. We may say that $\mathrm{W}^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B}$ is the Σ projection of the origin into Ω. In general, what is the Σ projection of *any* element of $\Re^{n\times p}$ into Ω?

Using '*P*' for projection, we define $P(\mathrm{W}_0; \Omega, \Sigma)$ as the function which projects $\mathrm{W}_0 \in \Re^{n\times p}$ into Ω according to the Σ metric. As before, Ω is the non-empty solution set of the linear constraint $\mathrm{AW} = \mathrm{B}$, and Σ is positive-definite. When these parameters are understood, we will use the abbreviation $P(\mathrm{W}_0)$. So $P(\mathrm{W}_0; \Omega, \Sigma)$ is *an* element of Ω that minimizes the Σ-metric distance from W_0 to Ω. Equivalently, it minimizes the quadratic form $\Phi(\mathrm{W}) = (\mathrm{W} - \mathrm{W}_0)'\Sigma^{-1}(\mathrm{W} - \mathrm{W}_0)$ subject to $\mathrm{AW} = \mathrm{B}$.

We could argue from scratch as in Section 2, but the following analysis is more insightful. The constraint $\mathrm{AW} = \mathrm{B}$ is equivalent to $\mathrm{A}(\mathrm{W} - \mathrm{W}_0) = \mathrm{B} - \mathrm{AW}_0$. So the projection problem is to minimize $(\mathrm{W} - \mathrm{W}_0)'\Sigma^{-1}(\mathrm{W} - \mathrm{W}_0)$ subject to $\mathrm{A}(\mathrm{W} - \mathrm{W}_0) = \mathrm{B} - \mathrm{AW}_0$. This is the Gauss-Markov problem with two changes in variables:

$$\begin{aligned} \mathrm{W} &\rightarrow \mathrm{W} - \mathrm{W}_0 \\ \mathrm{B} &\rightarrow \mathrm{B} - \mathrm{AW}_0 \end{aligned}$$

Hence, the Gauss-Markov theorem states that $(\mathrm{W} - \mathrm{W}_0)^* = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B} - \mathrm{AW}_0)$ uniquely minimizes $\Phi(\mathrm{W} - \mathrm{W}_0) = (\mathrm{W} - \mathrm{W}_0)'\Sigma^{-1}(\mathrm{W} - \mathrm{W}_0)$. But since W_0 is a constant, $(\mathrm{W} - \mathrm{W}_0)^* = \mathrm{W}^* - \mathrm{W}_0$, or $\mathrm{W}^* = \mathrm{W}_0 + (\mathrm{W} - \mathrm{W}_0)^* = \mathrm{W}_0 + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B} - \mathrm{AW}_0)$. So there is not just *an* element of projection, but a unique element:

$$\begin{aligned} P(\mathrm{W};\Omega,\Sigma) &= \mathrm{W} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{B}-\mathrm{AW}) \\ &= \{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B} \end{aligned}$$

As a check:

$$\begin{aligned} \mathrm{A}P(\mathrm{W};\Omega,\Sigma) &= \mathrm{A}\{\mathrm{I}_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W} + \mathrm{A}\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B} \\ &= \{\mathrm{A} - \mathrm{A}\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W} + \mathrm{A}\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B} \\ &= \{\mathrm{A} - \mathrm{I}_m\mathrm{A}\}\mathrm{W} + \mathrm{I}_m\mathrm{B} \\ &= \mathrm{B} \end{aligned}$$

Hence, for all $\mathrm{W} \in \Re^{n\times p}$, $\mathrm{A}P(\mathrm{W}) \in \Omega$. So P is a mapping from $\Re^{n\times p}$ into Ω, i.e., $P : \Re^{n\times p} \to \Omega$. In particular, the mapping of the origin is:

$$P(0_{n\times p}) = \{\mathrm{I}_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}0 + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B} = \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{B},$$

which is the 'W^*' of the theorem itself. Accordingly, we may employ the formulation $P(\mathrm{W}) = \{\mathrm{I}_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathrm{W} + P(0)$.

P maps element $\mathrm{W} \in \Re^{n\times p}$ to the closest element of constraint set Ω according to the Σ metric. Geometrically, P sends a Σ perpendicular from W into Ω; in symbols, $P(\mathrm{W}) - \mathrm{W} \underset{\Sigma}{\perp} \mathrm{W}_1 - P(\mathrm{W})$, for every $\mathrm{W}_1 \in \Omega$, as the following algebra shows:

$$
\begin{aligned}
&\left(P(\mathrm{W})-\mathrm{W}\right)'\Sigma^{-1}\left(\mathrm{W}_1-P(\mathrm{W})\right)\\
&\quad=\left(\left\{\mathrm{I}_n-\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{A}\right\}\mathrm{W}+P(0)-\mathrm{W}\right)'\Sigma^{-1}\left(\mathrm{W}_1-P(\mathrm{W})\right)\\
&\quad=\left(P(0)-\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{AW}\right)'\Sigma^{-1}\left(\mathrm{W}_1-P(\mathrm{W})\right)\\
&\quad=\left(\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}-\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{AW}\right)'\Sigma^{-1}\left(\mathrm{W}_1-P(\mathrm{W})\right)\\
&\quad=\left(\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\left(\mathrm{B}-\mathrm{AW}\right)\right)'\Sigma^{-1}\left(\mathrm{W}_1-P(\mathrm{W})\right)\\
&\quad=\left(\mathrm{B}-\mathrm{AW}\right)'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{A}\Sigma\,\Sigma^{-1}\left(\mathrm{W}_1-P(\mathrm{W})\right)\\
&\quad=\left(\mathrm{B}-\mathrm{AW}\right)'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{A}\left(\mathrm{W}_1-P(\mathrm{W})\right)\\
&\quad=\left(\mathrm{B}-\mathrm{AW}\right)'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\left(\mathrm{AW}_1-\mathrm{A}P(\mathrm{W})\right)\\
&\quad=\left(\mathrm{B}-\mathrm{AW}\right)'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\left(\mathrm{B}-\mathrm{B}\right)\\
&\quad=0
\end{aligned}
$$

Because of the first property of a metric (zero-triviality), within the restricted domain Ω, P is the identity mapping. Hence, not only is P a mapping *into* the constraint set Ω; it is also a mapping *onto* Ω. Nonetheless, we will prove it algebraically. If $\mathrm{W}\in\Omega$:

$$
\begin{aligned}
P(\mathrm{W})&=\left\{\mathrm{I}_n-\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{A}\right\}\mathrm{W}+\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}\\
&=\mathrm{W}-\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\left(\mathrm{AW}\right)+\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}\\
&=\mathrm{W}-\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\left(\mathrm{B}\right)+\Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{B}\\
&=\mathrm{W}
\end{aligned}
$$

Conversely, if $P(\mathrm{W})=\mathrm{W}$, then $\mathrm{AW}=\mathrm{A}P(\mathrm{W})=\mathrm{B}$ and $\mathrm{W}\in\Omega$. Therefore, P is a many-to-one mapping from $\Re^{n\times p}$ onto constraint set Ω, and an element of $\Re^{n\times p}$ belongs to Ω if and only if P acts upon it as an identity mapping.

Just as the Σ-metric right inverse $\Sigma A'(A\Sigma A')^{-1}$ is conspicuous in the formula $P(0_{n\times p}) = \Sigma A'(A\Sigma A')^{-1}B$, so too is it conspicuous in the formula for what we will call the "Σ-projection matrix" $I_n - \Sigma A'(A\Sigma A')^{-1}A$.[4] Since Σ is positive-definite, it can be Cholesky-decomposed as $\Sigma = QQ'$ for some non-singular $Q_{n\times n}$. Then the matrix can be factored as $I_n - \Sigma A'(A\Sigma A')^{-1}A = Q\{I_n - Q'A'(AQQ'A')^{-1}AQ\}Q^{-1} = QMQ^{-1}$. So the rank of the matrix is the rank of $M = I_n - Q'A'(AQQ'A')^{-1}AQ$. But M is a (symmetric) idempotent matrix (i.e., $M = MM'$), and the rank of an idempotent matrix equals its trace (Judge [1988, Appendix A.4 and A.12] and Halliwell [1997, Appendix B, 317; also Note 3]). Employing basic theorems about the trace operator, we derive:

$$\begin{aligned}
rank\left(I_n - \Sigma A'(A\Sigma A')^{-1}A\right) &= rank\left(Q\{I_n - Q'A'(AQQ'A)^{-1}AQ\}Q^{-1}\right) \\
&= rank\left(I_n - Q'A'(AQQ'A)^{-1}AQ\right) \\
&= Tr\left(I_n - Q'A'(AQQ'A)^{-1}AQ\right) \\
&= Tr(I_n) - Tr\left(Q'A'(AQQ'A)^{-1}AQ\right) \\
&= Tr(I_n) - Tr\left((AQQ'A)^{-1}AQQ'A'\right) \\
&= Tr(I_n) - Tr(I_m) \\
&= n - m
\end{aligned}$$

So every column of the elements of constraint set Ω has m fewer degrees of freedom than the columns of $\Re^{n\times p}$; in a sense, the dimensionality of Ω is $\Re^{(n-m)\times p}$. This can be surmised from the full-row rank of $A_{m\times n}$, which imposes m independent restrictions on the elements of $\Re^{n\times p}$ that belong to Ξ. This will prove useful in Section 9, in which we will treat linear statistical models with parameter constraints.

[4] The projection matrix shows to its greatest effect in the homogeneous form, i.e., in the differential form, $P(W_2) - P(W_1) = \{I_n - \Sigma A'(A\Sigma A')^{-1}A\}(W_2 - W_1)$.

5. VARIANCE AS A METRIC

In this section we will show how the variance of a random vector serves as its natural metric. Let $\mathbf{x}$ be an $n\times 1$ random vector with mean μ and non-degenerate variance Σ. Since the variance is non-degenerate, the variance of every non-zero linear combination of its elements is positive, i.e., Σ is positive-definite and Σ^{-1} exists. The variance of a random vector is a measure of its ability to differ from its mean. So the distances of random vectors from their means should somehow be invariant, when their variances serve as their distance metrics.

The square of the Σ-metric distance of $\mathbf{x}$ from its mean is $d_\Sigma^2(\mu,\mathbf{x})=(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)$. And by definition, the variance of $\mathbf{x}$ is $Var[\mathbf{x}]=\Sigma=E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]$. Using again the trace-operator theorems of the previous section, we find:

$$\begin{aligned}
E\left[d_\Sigma^2(\mu,\mathbf{x})\right] &= E\left[(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right] \\
&= Tr\left(E\left[(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right]\right) \\
&= E\left[Tr\left((\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right)\right] \\
&= E\left[Tr\left(\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right)\right] \\
&= Tr\left(E\left[\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]\right) \\
&= Tr\left(\Sigma^{-1}E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\right]\right) \\
&= Tr\left(\Sigma^{-1}\Sigma\right) \\
&= Tr\left(I_n\right)=n
\end{aligned}$$

Hence, a random vector's variance is its natural metric, according to which its expected squared distance from its mean equals its dimensionality *n*, the degrees of its freedom.[5]

6. PROJECTIONS OF RANDOM VECTORS

As in the previous section, let **x** be an *n*×1 random vector with mean μ and variance Σ, i.e., $\mathbf{x} \sim (\mu, \Sigma)$. Since the Gauss-Markov theorem has to do with abstract projective geometry, we may inquire about the moments of the Σ projection of **x** into the constraint space $\Omega = \{\mathrm{x} \in \Re^n : \mathrm{A}_{m\times n}\mathrm{x}_{n\times 1} = \mathrm{b}_{m\times 1}\}$.

The Σ projection is $P(\mathbf{x}; \Omega, \Sigma) = \{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathbf{x} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b}$. Therefore:

$$\begin{aligned} E[P(\mathbf{x}; \Omega, \Sigma)] &= E\left[\{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mathbf{x} + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b}\right] \\ &= \{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}E[\mathbf{x}] + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b} \\ &= \{I_n - \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\}\mu + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{b} \\ &= \mu + \Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathrm{b} - \mu) \end{aligned}$$

The variance follows from the standard formula $Var[\mathrm{Q}\mathbf{x}] = \mathrm{Q}Var[\mathbf{x}]\mathrm{Q}'$:

[5] Moreover, if **x** is multivariate normal, or if **x** ~ N(μ, Σ), then $(\mathbf{x} - \mu)' \Sigma^{-1}(\mathbf{x} - \mu) \sim \chi^2_n$ (Judge [1988], §2.5.9). The multivariate normal distribution is unique in that its probability distribution is a function of its variance metric: $f_{\mathbf{x}}(\mathrm{x}) \propto e^{-\frac{1}{2}(\mathrm{x}-\mu)'\Sigma^{-1}(\mathrm{x}-\mu)}$. Most likely, this is the ultimate reason why normality is preserved under any linear transformation.

$$
\begin{aligned}
Var\left[P(\mathbf{x};\Omega,\Sigma)\right] &= Var\left[\left\{I_n - \Sigma A'(A\Sigma A')^{-1}A\right\}\mathbf{x} + \Sigma A'(A\Sigma A')^{-1}b\right] \\
&= Var\left[\left\{I_n - \Sigma A'(A\Sigma A')^{-1}A\right\}\mathbf{x}\right] \\
&= \left\{I_n - \Sigma A'(A\Sigma A')^{-1}A\right\}Var[\mathbf{x}]\left\{I_n - \Sigma A'(A\Sigma A')^{-1}A\right\}' \\
&= \left\{I_n - \Sigma A'(A\Sigma A')^{-1}A\right\}\Sigma\left\{I_n - A'(A\Sigma A')^{-1}A\Sigma\right\} \\
&= \Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma + \Sigma A'(A\Sigma A')^{-1}A\Sigma A'(A\Sigma A')^{-1}A\Sigma \\
&= \Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma + \Sigma A'(A\Sigma A')^{-1}A\Sigma \\
&= \Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma
\end{aligned}
$$

Therefore, if $\mathbf{x} \sim (\mu, \Sigma)$, then $P(\mathbf{x};\Omega,\Sigma) \sim \left(P(\mu;\Omega,\Sigma), \Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma\right)$. As a check, $E[A\mathbf{x}] = AP(\mu) = b$ and $Var[A\mathbf{x}] = A\left(\Sigma - \Sigma A'(A\Sigma A')^{-1}A\Sigma\right)A' = 0_{m\times m}$. This will prove useful in Section 11.

7. PARAMETER ESTIMATION IN THE LINEAR STATISTICAL MODEL

Now let us apply our Gauss-Markov theorem to the linear statistical model. First, and as an easy start, we will apply it to derive the best linear unbiased estimator (BLUE) of the parameter β in the model $\mathbf{y} = X_{t\times k}\beta_{k\times 1} + \mathbf{e}$, where $Var[\mathbf{e}] = \Sigma_{t\times t}$. X is of full-column rank, and Σ is positive-definite. The estimator is linear in $\mathbf{y}$, or $\hat{\boldsymbol{\beta}} = W'\mathbf{y}$.[6] Because it is unbiased for all β, $E[\hat{\boldsymbol{\beta}}] = E[W'\mathbf{y}] = W'X\beta = \beta$. So matrix W' is constrained according to the equation $W'X = I_k$, which transposes as $X'W = I_k$. The best of the unbiased estimators minimizes $Var[\hat{\boldsymbol{\beta}}] = Var[W'\mathbf{y}] = W'\Sigma W$. So the problem is to minimize

[6] For a reason immediately to become apparent, we use here the transpose of W.

$\Phi(\mathrm{W}) = \mathrm{W}'\Sigma\mathrm{W} = \mathrm{W}'\left(\Sigma^{-1}\right)^{-1}\mathrm{W}$ subject to $\mathrm{X}'\mathrm{W} = \mathrm{I}_k$. The correspondences between the theorem and this model are:

Theorem	←	Model
Σ	←	Σ^{-1}
W	←	W
A	←	X'
B	←	I_k

X' is of full-row rank, because X is of full-column rank. Hence, according to the theorem, $\mathrm{W}^* = \Sigma^{-1}\mathrm{X}''\left(\mathrm{X}'\Sigma^{-1}\mathrm{X}''\right)^{-1}\mathrm{I}_k = \Sigma^{-1}\mathrm{X}\left(\mathrm{X}'\Sigma^{-1}\mathrm{X}\right)^{-1}$. So $\hat{\boldsymbol{\beta}} = \mathrm{W}^{*'}\mathbf{y} = \left(\mathrm{X}'\Sigma^{-1}\mathrm{X}\right)^{-1}\mathrm{X}'\Sigma^{-1}\mathbf{y}$ and $Var\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{W}^{*'}\Sigma\mathrm{W}^* = \mathrm{I}_k'\left(\mathrm{X}'\Sigma^{-1}\mathrm{X}''\right)^{-1}\mathrm{I}_k = \left(\mathrm{X}'\Sigma^{-1}\mathrm{X}\right)^{-1}$. Accordingly, $\hat{\boldsymbol{\beta}} = Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{X}'\Sigma^{-1}\mathbf{y}$.

8. PREDICTION IN THE LINEAR STATISTICAL MODEL

The goal of most linear modeling is not to estimate the parameter β, but rather to estimate[7] quantities which eventually will be observed. Although the model makes such quantities dependent on the parameter, the parameter itself is usually hypothetical and never to be observed. With partitioning between the observed $\mathbf{y}_1$ and the to-be-predicted $\mathbf{y}_2$ (hence, containing missing values) the general form of the linear statistical model is:

$$\begin{bmatrix}\mathbf{y}_1 \\ \hline \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathrm{X}_1 \ _{(t_1\times k)} \\ \hline \mathrm{X}_2 \ _{(t_2\times k)}\end{bmatrix}\beta_{k\times 1} + \begin{bmatrix}\mathbf{e}_1 \\ \hline \mathbf{e}_2\end{bmatrix}, \text{where } Var\begin{bmatrix}\mathbf{e}_1 \\ \hline \mathbf{e}_2\end{bmatrix} = \left[\begin{array}{c|c}\Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22}\end{array}\right] = \Sigma$$

Not only must Σ be symmetric and non-negative-definite, Σ_{11} must be positive-definite, and X_1 must be of full-column rank. We seek the best linear-in-$\mathbf{y}_1$, unbiased estimator

[7] More accurately, the goal is to predict – we seek the best linear unbiased prediction. But we will continue to call this BLUE, because BLUP already has a different technical meaning in statistics (*Wikipedia*, "Best linear unbiased prediction").

(BLUE) of $\mathbf{y}_2$, i.e., $\hat{\mathbf{y}}_2 = \mathrm{W}'\mathbf{y}_1$ for some matrix W', which depends only on the partitions of the design X and variance Σ matrices. Because the estimator is unbiased for all β, $0 = E[\mathbf{y}_2 - \hat{\mathbf{y}}_2] = E[\mathbf{y}_2 - \mathrm{W}'\mathbf{y}_1] = (\mathrm{X}_2 - \mathrm{W}'\mathrm{X}_1)\beta$. Thus the estimator is unbiased if and only if $\mathrm{W}'\mathrm{X}_1 = \mathrm{X}_2$. By transposition, $\mathrm{X}_1'\mathrm{W} = \mathrm{X}_2'$, where W is $t_1 \times t_2$.

But now there is a complication in being "best." Predicting $\mathbf{y}_2$ as $\hat{\mathbf{y}}_2$, we will err by the amount $\mathbf{y}_2 - \hat{\mathbf{y}}_2$. So it is the prediction-error variance that we must minimize:

$$
\begin{aligned}
Var[\mathbf{y}_2 - \hat{\mathbf{y}}_2] &= Var[\mathbf{y}_2 - \mathrm{W}'\mathbf{y}_1] \\
&= Var\left[\begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix}\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}\right] \\
&= \begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix} Var\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}\begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix}' \\
&= \begin{bmatrix} -\mathrm{W}' & I_{t_2} \end{bmatrix}\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\begin{bmatrix} -\mathrm{W} \\ I_{t_2} \end{bmatrix} \\
&= \mathrm{W}'\Sigma_{11}\mathrm{W} - \mathrm{W}'\Sigma_{12} - \Sigma_{21}\mathrm{W} + \Sigma_{22}
\end{aligned}
$$

Although we can ignore the constant Σ_{22} in the minimization, we cannot ignore the second and third terms, which are linear in W.

The key here is to apply the one-to-one transform $\mathrm{V} \leftrightarrow \mathrm{W} - \Sigma_{11}^{-1}\Sigma_{12}$. The transformation of the constraint set is $\{\mathrm{W} : \mathrm{X}_1'\mathrm{W} = \mathrm{X}_2'\} = \{\mathrm{V} : \mathrm{X}_1'\mathrm{V} = \mathrm{X}_1'\mathrm{W} - \mathrm{X}_1'\Sigma_{11}^{-1}\Sigma_{12} = \mathrm{X}_2' - \mathrm{X}_1'\Sigma_{11}^{-1}\Sigma_{12}\}$. So expressed in terms of V:

$$
\begin{aligned}
Var\left[\mathbf{y}_2 - \hat{\mathbf{y}}_2\right] &= \mathrm{W}'\Sigma_{11}\mathrm{W} - \mathrm{W}'\Sigma_{12} - \Sigma_{21}\mathrm{W} + \Sigma_{22} \\
&= \left(\mathrm{V} + \Sigma_{11}^{-1}\Sigma_{12}\right)'\Sigma_{11}\left(\mathrm{V} + \Sigma_{11}^{-1}\Sigma_{12}\right) - \left(\mathrm{V} + \Sigma_{11}^{-1}\Sigma_{12}\right)'\Sigma_{12} - \Sigma_{21}\left(\mathrm{V} + \Sigma_{11}^{-1}\Sigma_{12}\right) + \Sigma_{22} \\
&= \left(\mathrm{V}' + \Sigma_{21}\Sigma_{11}^{-1}\right)\Sigma_{11}\left(\mathrm{V} + \Sigma_{11}^{-1}\Sigma_{12}\right) - \left(\mathrm{V}' + \Sigma_{21}\Sigma_{11}^{-1}\right)\Sigma_{12} - \Sigma_{21}\left(\mathrm{V} + \Sigma_{11}^{-1}\Sigma_{12}\right) + \Sigma_{22} \\
&= \mathrm{V}'\Sigma_{11}\mathrm{V} + \mathrm{V}'\Sigma_{12} + \Sigma_{21}\mathrm{V} + \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&\quad - \mathrm{V}'\Sigma_{12} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \Sigma_{21}\mathrm{V} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{22} \\
&= \mathrm{V}'\Sigma_{11}\mathrm{V} + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)
\end{aligned}
$$

This transformation is a matrix version of completing the square. We can now apply the Gauss-Markov theorem to the problem of minimizing $\mathrm{V}'\Sigma_{11}\mathrm{V} = \mathrm{V}'\left(\Sigma_{11}^{-1}\right)^{-1}\mathrm{V}$ subject to $\mathrm{X}_1'\mathrm{V} = \mathrm{X}_2' - \mathrm{X}_1'\Sigma_{11}^{-1}\Sigma_{12}$. The correspondences are:

Theorem	←	Model
Σ	←	Σ_{11}^{-1}
W	←	V
A	←	X_1'
B	←	$\mathrm{X}_2' - \mathrm{X}_1'\Sigma_{11}^{-1}\Sigma_{12}$

As before, the conditions are met; X_1' is of full-row rank since X_1 is of full-column rank. Hence, $\mathrm{V}^* = \Sigma_{11}^{-1}\mathrm{X}_1\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\left(\mathrm{X}_2' - \mathrm{X}_1'\Sigma_{11}^{-1}\Sigma_{12}\right)$, and $\hat{\mathbf{y}}_2 = \mathrm{W}^{*'}\mathbf{y}_1$, where:

$$
\begin{aligned}
\mathrm{W}^{*'} &= \left(\mathrm{V}^* + \Sigma_{11}^{-1}\Sigma_{12}\right)' \\
&= \mathrm{V}^{*'} + \Sigma_{21}\Sigma_{11}^{-1} \\
&= \left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1} + \Sigma_{21}\Sigma_{11}^{-1}
\end{aligned}
$$

The minimized prediction-error variance is:

$$
\begin{aligned}
Var\left[\mathbf{y}_2 - \hat{\mathbf{y}}_2\right] &= \mathrm{V}^{*'}\Sigma_{11}\mathrm{V}^* + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right) \\
&= \left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)' + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)
\end{aligned}
$$

Introducing the estimator $\hat{\boldsymbol{\beta}} = \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1$ and its variance $Var\left[\hat{\boldsymbol{\beta}}\right] = \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}$ allows us to simplify:

$$\begin{aligned}
\hat{\mathbf{y}}_2 &= \mathrm{W}^{*'}\mathbf{y}_1 \\
&= \left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1 \\
&= \left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)\hat{\boldsymbol{\beta}} + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1 \\
&= \mathrm{X}_2\hat{\boldsymbol{\beta}} + \Sigma_{21}\Sigma_{11}^{-1}\left(\mathbf{y}_1 - \mathrm{X}_1\hat{\boldsymbol{\beta}}\right) \\
Var\left[\mathbf{y}_2 - \hat{\mathbf{y}}_2\right] &= \left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)Var\left[\hat{\boldsymbol{\beta}}\right]\left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)' + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)
\end{aligned}$$

Combining this completing-the-square technique with the Gauss-Markov theorem makes for a cleaner and more elegant proof than that in Halliwell [1997, Appendix C, 328-330].

9. LINEAR STATISTICAL MODELS WITH PARAMETER CONSTRAINTS

Here we will impose upon the model of Section 8 a constraint on β, viz., $\mathrm{R}_{j\times k}\beta_{k\times 1} = \mathrm{r}_{j\times 1}$. The rows of R must be linearly independent, i.e., R must be of full-row rank. The constraint set $\left\{\beta \in \Re^k : \mathrm{R}\beta = \mathrm{r}\right\}$ is non-empty because right inverses of R exist, most obviously $\mathrm{R}'\left(\mathrm{RR}'\right)^{-1}$. Hence, $\beta = \mathrm{R}'\left(\mathrm{RR}'\right)^{-1}\mathrm{r}$ exists and satisfies the constraint.

Two procedures are commonly employed to solve β-constrained linear models. The first is to reduce the parameter dimension according to the equation $\beta = \beta_0 + \mathrm{S}_{k\times(k-j)}\gamma_{(k-j)\times 1}$, for some matrix S (of full-column rank) such that $\mathrm{RS} = 0_{j\times(k-j)}$, as done in Halliwell [1997, Appendix B, 321-324]. This is the purist approach to the problem, but it requires an understanding of eigen-decomposition, cannot be performed in Excel without add-ins, and may suffer from the numerical-analysis problem of deciding when small eigenvalues should

be zeroed. The second procedure is to employ the Lagrange multiplier (Judge [1988, §6.2, 235-237]) to minimize $\Lambda(\beta,\lambda_{j\times 1}) = (\mathbf{y}-\mathrm{X}\beta)'\Sigma^{-1}(\mathbf{y}-\mathrm{X}\beta)+2\lambda'(\mathrm{R}\beta-\mathrm{r})$. But a third procedure (Halliwell [1998, Appendix C]) to us is the most convincing.

This procedure is to treat the β constraint as the limit of $\mathbf{r}=\mathrm{R}\beta+\boldsymbol{\eta}$ as $Var[\boldsymbol{\eta}]\to 0_{j\times j}$. We could have specified the variance as $\sigma^2\mathrm{I}_j$, and the limit as $\sigma^2\to 0$; but for the sake of generality we will let $Var[\boldsymbol{\eta}]=\mathrm{H}$ be any positive-definite matrix. So we can form the following augmented linear model, which satisfies the conditions of Section 8:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{r}\\ \hline \mathbf{y}_2\end{bmatrix}=\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\\ \hline \mathrm{X}_2\end{bmatrix}\beta+\begin{bmatrix}\mathbf{e}_1\\ \boldsymbol{\eta}\\ \hline \mathbf{e}_2\end{bmatrix},\ \text{where}\ Var\begin{bmatrix}\mathbf{e}_1\\ \boldsymbol{\eta}\\ \hline \mathbf{e}_2\end{bmatrix}=\left[\begin{array}{cc:c}\Sigma_{11} & 0 & \Sigma_{12}\\ 0 & \mathrm{H} & 0\\ \hdashline \Sigma_{21} & 0 & \Sigma_{22}\end{array}\right]$$

The parameter estimator, which depends on H, is:

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\mathrm{H}) &= \left(\begin{bmatrix}\mathrm{X}_1' & \mathrm{R}'\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathrm{X}_1' & \mathrm{R}'\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{r}\end{bmatrix}\\ &= \left(\begin{bmatrix}\mathrm{X}_1' & \mathrm{R}'\end{bmatrix}\begin{bmatrix}\Sigma_{11}^{-1} & 0\\ 0 & \mathrm{H}^{-1}\end{bmatrix}\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathrm{X}_1' & \mathrm{R}'\end{bmatrix}\begin{bmatrix}\Sigma_{11}^{-1} & 0\\ 0 & \mathrm{H}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{r}\end{bmatrix}\\ &= \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1+\mathrm{R}'\mathrm{H}^{-1}\mathrm{R}\right)^{-1}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1+\mathrm{R}'\mathrm{H}^{-1}\mathbf{r}\right)\\ &= Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1+\mathrm{R}'\mathrm{H}^{-1}\mathbf{r}\right)\end{aligned}$$

Therefore, according to the formulas of the previous section, the predictor is:

$$\begin{aligned}\hat{\mathbf{y}}_2(\mathrm{H}) &= \mathrm{X}_2\hat{\boldsymbol{\beta}}(\mathrm{H})+\begin{bmatrix}\Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\left(\begin{bmatrix}\mathbf{y}_1\\ \mathbf{r}\end{bmatrix}-\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\end{bmatrix}\hat{\boldsymbol{\beta}}(\mathrm{H})\right)\\ &= \mathrm{X}_2\hat{\boldsymbol{\beta}}(\mathrm{H})+\Sigma_{21}\Sigma_{11}^{-1}\left(\mathbf{y}_1-\mathrm{X}_1\hat{\boldsymbol{\beta}}(\mathrm{H})\right)\end{aligned}$$

And the variance of the prediction error is:

$$Var\left[\mathbf{y}_2 - \hat{\mathbf{y}}_2(\mathrm{H})\right] = \left(\mathrm{X}_2 - \begin{bmatrix}\Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\end{bmatrix}\right) Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\left(\mathrm{X}_2 - \begin{bmatrix}\Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{X}_1\\ \mathrm{R}\end{bmatrix}\right)'$$
$$+\left(\Sigma_{22} - \begin{bmatrix}\Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \mathrm{H}\end{bmatrix}^{-1}\begin{bmatrix}\Sigma_{12}\\ 0\end{bmatrix}\right)$$
$$= \left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right) Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\left(\mathrm{X}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathrm{X}_1\right)' + \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

These two formulas depend on H only insofar as $\hat{\boldsymbol{\beta}}$ depends on H. Therefore, it remains for us to determine $\boldsymbol{\beta}^* = \lim_{\mathrm{H}\to 0}\hat{\boldsymbol{\beta}}(\mathrm{H})$.

We start with $Var\left[\boldsymbol{\beta}^*\right] = \lim_{\mathrm{H}\to 0}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1 + \mathrm{R}'\mathrm{H}^{-1}\mathrm{R}\right)^{-1}$. The following proof makes use of the theorem $\left(\mathrm{A}+\mathrm{BDC}\right)^{-1} = \mathrm{A}^{-1} - \mathrm{A}^{-1}\mathrm{B}\left(\mathrm{D}^{-1} + \mathrm{CA}^{-1}\mathrm{B}\right)^{-1}\mathrm{CA}^{-1}$ (cf. Judge [1988, A.7, 938]; the inverses must exist, as they do here): [8]

$$\begin{aligned}
Var\left[\boldsymbol{\beta}^*\right] &= \lim_{\mathrm{H}\to 0}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1 + \mathrm{R}'\mathrm{H}^{-1}\mathrm{R}\right)^{-1}\\
&= \lim_{\mathrm{H}\to 0}\left\{\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1} - \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\left[\mathrm{H} + \mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\right\}\\
&= \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1} - \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\left[0 + \mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\\
&= \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1} - \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\left[\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\\
&= Var\left[\hat{\boldsymbol{\beta}}\right] - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]
\end{aligned}$$

The variance of the constrained estimator is neatly expressed in terms of the variance of the unconstrained estimator $Var\left[\hat{\boldsymbol{\beta}}\right] = \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}$. As a check:

$$Var\left[\mathrm{R}\boldsymbol{\beta}^*\right] = \mathrm{R}Var\left[\boldsymbol{\beta}^*\right]\mathrm{R}' = \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}' - \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}' = 0$$

[8] In the following formulas the existence of the inverse of $\mathrm{R}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{R}' = \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'$ is guaranteed, since the variance matrix is positive-definite and R is of full-row rank.

In order to take the limit of $\hat{\boldsymbol{\beta}}(\mathrm{H})$ we need the following intermediate result:

$$
\begin{aligned}
Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{R}'\mathrm{H}^{-1} &= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\left\{\mathrm{I}_k - \mathrm{R}'\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\right\}\mathrm{R}'\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\left\{\mathrm{R}' - \mathrm{R}'\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\mathrm{I}_j - \left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\mathrm{I}_j - \left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'-\mathrm{H}\right]\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\mathrm{I}_j - \mathrm{I}_j + \left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{H}\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left\{\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1}\mathrm{H}\right\}\mathrm{H}^{-1} \\
&= \left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\left[\mathrm{H}+\mathrm{R}\left(\mathrm{X}'\Sigma_{11}^{-1}\mathrm{X}\right)^{-1}\mathrm{R}'\right]^{-1} \\
&= Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{H}+\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\boldsymbol{\beta}^* &= \lim_{\mathrm{H}\to 0}\hat{\boldsymbol{\beta}}(\mathrm{H}) \\
&= \lim_{\mathrm{H}\to 0} Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \mathrm{R}'\mathrm{H}^{-1}\mathbf{r}\right) \\
&= \lim_{\mathrm{H}\to 0} Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \lim_{\mathrm{H}\to 0} Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{R}'\mathrm{H}^{-1}\mathbf{r} \\
&= Var\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \left(\lim_{\mathrm{H}\to 0} Var\left[\hat{\boldsymbol{\beta}}(\mathrm{H})\right]\mathrm{R}'\mathrm{H}^{-1}\right)\left(\operatorname*{plim}_{\mathrm{H}\to 0}\mathbf{r}\right) \\
&= Var\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \left(Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\right)\left(\operatorname*{plim}_{Var[\boldsymbol{\eta}]\to 0}\left[\mathrm{R}\beta+\boldsymbol{\eta}\right]\right) \\
&= Var\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + \left(Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\right)\mathrm{R}\beta \\
&= Var\left[\boldsymbol{\beta}^*\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{Var\left[\hat{\boldsymbol{\beta}}\right] - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\right\}\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{X}_1'\Sigma_{11}^{-1}\mathbf{y}_1 + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r}
\end{aligned}
$$

As a check:

$$\begin{aligned}
\mathrm{R}\boldsymbol{\beta}^* &= \mathrm{R}\left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{R} - \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + \mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= \left\{\mathrm{R} - \mathrm{I}_j\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + \mathrm{I}_j\mathrm{r} \\
&= \mathrm{r}
\end{aligned}$$

And so, to summarize, the solution of a β-constrained linear model is the solution of the unconstrained model with the substitution of $\boldsymbol{\beta}^*$ for $\hat{\boldsymbol{\beta}}$, where:

$$\boldsymbol{\beta}^* = \left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r}$$

$$Var\left[\boldsymbol{\beta}^*\right] = Var\left[\hat{\boldsymbol{\beta}}\right] - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]$$

10. PARAMETER CONSTRAINTS AS PROJECTIONS

The formulas above for $\boldsymbol{\beta}^*$ and $Var\left[\boldsymbol{\beta}^*\right]$ may seem cumbersome, perhaps even repugnant. However, they become perspicuous when interpreted as a projection. From Section 6 we take the projection formula $P(\mathbf{x}; \{\mathrm{A}\mathrm{x} = \mathrm{b}\}, \Sigma) = \left\{I_n - \Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{A}\right\}\mathbf{x} + \Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{b}$, where $\Sigma = Var[\mathbf{x}]$. But now let the constraint space Ω be $\left\{\beta \in \Re^{\mathrm{k}} : \mathrm{R}\beta = \mathrm{r}\right\}$. In this case:

$$\begin{aligned}
\boldsymbol{\beta}^* &= \left\{\mathrm{I}_k - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}\right\}\hat{\boldsymbol{\beta}} + Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{r} \\
&= P\left(\hat{\boldsymbol{\beta}}; \{\mathrm{R}\beta = \mathrm{r}\}, Var\left[\hat{\boldsymbol{\beta}}\right]\right)
\end{aligned}$$

Hence, the constrained parameter estimator is the projection of the unconstrained estimator according to the metric of the variance of the unconstrained estimator. Just to corroborate, we see that the variance of the constrained estimator,

$Var\left[\boldsymbol{\beta}^*\right] = Var\left[\hat{\boldsymbol{\beta}}\right] - Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\left[\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]\mathrm{R}'\right]^{-1}\mathrm{R}Var\left[\hat{\boldsymbol{\beta}}\right]$ accords with the projection variance $Var\left[P(\mathbf{x};\Omega,\Sigma)\right] = \Sigma - \Sigma\mathrm{A}'\left(\mathrm{A}\Sigma\mathrm{A}'\right)^{-1}\mathrm{A}\Sigma$.

Similarly to how we argued in Section 4, $Var\left[\hat{\boldsymbol{\beta}}\right]$ can be Cholesky-decomposed as $Var\left[\hat{\boldsymbol{\beta}}\right] = \mathrm{QQ}'$ for some non-singular $\mathrm{Q}_{k\times k}$. So the constrained variance can be factored as $Var\left[\boldsymbol{\beta}^*\right] = \mathrm{Q}\left\{\mathrm{I}_k - \mathrm{Q}'\mathrm{R}'\left(\mathrm{RQQ}'\mathrm{R}'\right)^{-1}\mathrm{RQ}\right\}\mathrm{Q}' = \mathrm{QMQ}'$, and its rank is that of the idempotent matrix $\mathbf{M}$, whose rank equals its trace. Again, to continue as in Section 4:

$$Tr(\mathrm{M}) = Tr(\mathrm{I}_k) - Tr\left(\mathrm{Q}'\mathrm{R}'\left(\mathrm{RQQ}'\mathrm{R}'\right)^{-1}\mathrm{RQ}\right) = k - Tr\left(\left(\mathrm{RQQ}'\mathrm{R}'\right)^{-1}\mathrm{RQQ}'\mathrm{R}'\right) = k - j$$

Therefore, $rank\left(Var\left[\boldsymbol{\beta}^*\right]\right) = k - j = rank\left(Var\left[\hat{\boldsymbol{\beta}}\right]\right) - j$. The parameter constraint reduces the parameter variance by j degrees of freedom. In words, $\mathrm{R}\boldsymbol{\beta}^*$ is a degenerate random variable, or a constant. Certainly it is, since by the constraint $\mathrm{R}\boldsymbol{\beta}^* = \mathrm{r}$.

All this shows that the solution of a parameter-constrained model is equivalent to the projection of the solution of an unconstrained model. There seems to be a certain commutativity between constraining/projecting and solving.

11. INFORMATION AS PROJECTION

We start with the equation of Section 7: $\mathbf{y} = \mathrm{X}_{t\times k}\beta_{k\times 1} + \mathbf{e}$, where $Var\left[\mathbf{e}\right] = \Sigma_{t\times t}$. However, let us suppose that β is known and needs no estimation. Our best prediction of $\mathbf{y}$ is $\mathrm{X}\beta$, whose prediction-error variance is $Var\left[\mathbf{y} - \hat{\mathbf{y}}\right] = Var\left[\mathbf{y} - \mathrm{X}\beta\right] = Var\left[\mathbf{e}\right] = \Sigma$. At this stage we

are saying nothing more than $\mathbf{y} \sim (\mathrm{X}\beta, \Sigma)$. But furthermore suppose that we have observed $\mathrm{A}_{m\times t}\mathbf{y}$, where A is of full-row rank. Name the observed value $\mathbf{b}_{m\times 1}$. The problem is to predict **y** after the observation.

Since $\mathbf{y} = \mathrm{X}\beta + \mathbf{e}$, $\mathbf{b} = \mathrm{A}\mathbf{y} = \mathrm{AX}\beta + \mathrm{A}\mathbf{e}$, where $Var[\mathrm{A}\mathbf{e}] = \mathrm{A}\Sigma\mathrm{A}'$. Since **y** is the same in the observation as in the prediction, the observation covaries with the prediction; in fact, $Cov[\mathrm{A}\mathbf{y}, \mathbf{y}] = Cov[\mathrm{A}\mathbf{e}, \mathbf{e}] = \mathrm{A}\Sigma$. We can predict **y** according to the parameter-constrained model of Section 9:

$$\begin{bmatrix} \mathbf{b} \\ \beta_0 \\ \hline \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathrm{A}\,\mathrm{X} \\ \mathrm{I}_k \\ \hline \mathrm{X} \end{bmatrix} \beta + \begin{bmatrix} \mathrm{A}\mathbf{e} \\ 0 \\ \hline \mathbf{e} \end{bmatrix}, \text{ where } Var \begin{bmatrix} \mathrm{A}\mathbf{e} \\ 0 \\ \hline \mathbf{e} \end{bmatrix} = \left[\begin{array}{ccc} \mathrm{A}\Sigma\mathrm{A}' & 0 & \mathrm{A}\Sigma \\ 0 & 0 & 0 \\ \hline \Sigma\mathrm{A}' & 0 & \Sigma \end{array}\right]$$

Although this will work, a simpler and more appealing model can be constructed if one allows for zero-dimensional matrices.[9] Because all $m\times 0$ and $0\times n$ matrices are of rank zero, $\mathrm{A}_{m\times 0}\mathrm{B}_{0\times n} = 0_{m\times n}$. This is nothing more than the nullity of the empty summation operator, i.e., $(\mathrm{AB})_{ij} = \sum_{k=1}^{0} (\mathrm{A})_{ik}(\mathrm{B})_{kj} = 0$. The simpler model is:

$$\begin{bmatrix} \mathbf{b} - \mathrm{AX}\beta \\ \hline \mathbf{y} - \mathrm{X}\beta \end{bmatrix} = \begin{bmatrix} \mathrm{X}_{1\ (m\times 0)} \\ \hline \mathrm{X}_{2\ (t\times 0)} \end{bmatrix} \gamma_{0\times 1} + \begin{bmatrix} \mathrm{A}\mathbf{e} \\ \hline \mathbf{e} \end{bmatrix}, \text{ where } Var \begin{bmatrix} \mathrm{A}\mathbf{e} \\ \hline \mathbf{e} \end{bmatrix} = \left[\begin{array}{cc} \mathrm{A}\Sigma\mathrm{A}' & \mathrm{A}\Sigma \\ \hline \Sigma\mathrm{A}' & \Sigma \end{array}\right]$$

Its solution begins with:

[9] It is a windfall for a matrix language to allow for zeros in the dimensions of its arrays, as do APL, J, and R. SAS/IML does not; at least it did not in the late 1990s (version 7), when the author last used it.

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{0\times 1} &= Var[\hat{\boldsymbol{\gamma}}]\mathrm{X}_1'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathbf{b}-\mathrm{AX}\beta)\\ &= \left(\mathrm{X}_1'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathbf{b}-\mathrm{AX}\beta)\\ &= (0\times 0)^{-1}(0\times m)\cdot(m\times m)^{-1}(m\times 1)\\ &= (0\times 0)^{-1}(0\times 1)\end{aligned}$$

The only thing to give pause here is the inverse of the 0×0 matrix. But the space of real 0-vectors, $\Re^0$, contains just one element, viz., the origin. It is closed under addition and multiplication (0+0 = 0×0 = 0), and 0 serves as its identity element. So in $\Re^0$, $0^{-1}=0$. Hence, $(0\times 0)^{-1}=(0\times 0)$. Therefore, $Var[\hat{\boldsymbol{\gamma}}]=(0\times 0)$ and $\hat{\boldsymbol{\gamma}}_{0\times 1}=(0\times 1)$.[10] Finally:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathrm{X}\beta+\widehat{\mathbf{y}-\mathrm{X}\beta}\\ &= \mathrm{X}\beta+\mathrm{X}_2\hat{\boldsymbol{\gamma}}+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\{\mathbf{b}-\mathrm{AX}\beta\}-\mathrm{X}_1\hat{\boldsymbol{\gamma}})\\ &= \mathrm{X}\beta+0_{t\times 1}+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\{\mathbf{b}-\mathrm{AX}\beta\}-0_{m\times 1})\\ &= \mathrm{X}\beta+\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}(\mathbf{b}-\mathrm{AX}\beta)\end{aligned}$$

The variance of its prediction error is:

$$\begin{aligned}Var[\mathbf{y}-\hat{\mathbf{y}}] &= Var[(\mathbf{y}-\mathrm{X}\beta)-(\hat{\mathbf{y}}-\mathrm{X}\beta)]\\ &= \left(\mathrm{X}_2-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{X}_1\right)Var[\hat{\boldsymbol{\gamma}}]\left(\mathrm{X}_2-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{X}_1\right)'+\Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\\ &= (t\times 0)(0\times 0)(0\times t)+\Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\\ &= 0_{t\times t}+\Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\\ &= \Sigma-\Sigma\mathrm{A}'(\mathrm{A}\Sigma\mathrm{A}')^{-1}\mathrm{A}\Sigma\end{aligned}$$

Except for slight notational differences, this solution is the same as that of Section 6. As long as no parameter needs to be estimated (or the parameter dimension is 0×1), the linear statistical model treats "*m* dimensions" of prior information as a projection into a subspace of $t-m$ dimensions.

[10] To elaborate on the previous footnote, we have verified that APL, J, and R yield these results. Therefore, they correctly treat $(0\times 0)^{-1}$ as (0×0).

12. COMBINING ESTIMATES

It is not uncommon for an actuary linearly to combine two or more unbiased estimators of the same quantity. Of course, it is desirable for the combination to be best. In the simplest situation of independent scalar estimators, the best combination uses weights inversely proportional to the variances of the estimators. But with Gauss-Markov theorem one can determine the best linear combination of vector estimators, even if they are not independent.

To frame the problem, suppose that we have n unbiased estimators $\hat{\mathbf{y}}_i$ of the same $t\times 1$ vector $\mathbf{y}$, as well as their $t\times t$ prediction-error variances $\Sigma_{ii} = Var\left[\mathbf{y} - \hat{\mathbf{y}}_i\right]$. Suppose also that we have the $t\times t$ prediction-error covariances $\Sigma_{ij} = Cov\left[\mathbf{y} - \hat{\mathbf{y}}_i, \mathbf{y} - \hat{\mathbf{y}}_j\right]$. Frequently the covariances are $0_{t\times t}$, but there are realistic exceptions. Stack the estimators and block their (co)variances:

$$\hat{\mathbf{Y}}_{nt\times 1} = \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \vdots \\ \hat{\mathbf{y}}_n \end{bmatrix}, \quad Var\left[\mathbf{Y} - \hat{\mathbf{Y}}\right]_{nt\times nt} = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix}$$

The variance matrix must be non-negative-definite; but we will assume it to be positive definite, hence invertible. If the weight given to $\hat{\mathbf{y}}_i$ is the $t\times t$ matrix W_i', the combined estimator will be $\hat{\mathbf{y}} = \sum_{i=1}^{n} \mathrm{W}_i'\hat{\mathbf{y}}_i = \begin{bmatrix} \mathrm{W}_1' & \cdots & \mathrm{W}_n' \end{bmatrix}_{t\times nt} \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \vdots \\ \hat{\mathbf{y}}_n \end{bmatrix} = \mathrm{W}'\hat{\mathbf{Y}}$. In order for this estimator to be unbiased, $\begin{bmatrix} \mathrm{W}_1' & \cdots & \mathrm{W}_n' \end{bmatrix} \begin{bmatrix} \mathrm{I}_t \\ \vdots \\ \mathrm{I}_t \end{bmatrix} = \mathrm{W}' \begin{bmatrix} \mathrm{I}_t \\ \vdots \\ \mathrm{I}_t \end{bmatrix} = \mathrm{I}_t$. The transpose of this constraint is

$\left[\mathrm{I}_t \quad \ldots \quad \mathrm{I}_t\right]_{t\times nt}\mathrm{W}_{nt\times t} = \mathrm{I}_t$. The best combination will minimize the combined prediction-error variance $\mathrm{W}'Var\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\mathrm{W}$. Posing the problem in the proper form, we seek to minimize $\mathrm{W}'Var\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\mathrm{W} = \mathrm{W}'\left(Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\right)^{-1}\mathrm{W}$ subject to $\left[\mathrm{I}_t \quad \ldots \quad \mathrm{I}_t\right]_{t\times nt}\mathrm{W}_{nt\times t} = \mathrm{I}_t$. According to the Gauss-Markov theorem:

$$\mathrm{W}^* = Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots \\ \mathrm{I}_t\end{bmatrix}\left(\left[\mathrm{I}_t \quad \ldots \quad \mathrm{I}_t\right]Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots \\ \mathrm{I}_t\end{bmatrix}\right)^{-1}\mathrm{I}_t$$

$$= Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots \\ \mathrm{I}_t\end{bmatrix}\left(\left[\mathrm{I}_t \quad \ldots \quad \mathrm{I}_t\right]Var^{-1}\left[\mathbf{Y}-\hat{\mathbf{Y}}\right]\begin{bmatrix}\mathrm{I}_t\\ \vdots \\ \mathrm{I}_t\end{bmatrix}\right)^{-1}$$

If the covariances $\Sigma_{i\neq j}$ are zero, this simplifies to:

$$\mathrm{W}^* = \begin{bmatrix}\Sigma_{11}^{-1}\\ \vdots \\ \Sigma_{nn}^{-1}\end{bmatrix}\left(\sum_{i=1}^{n}\Sigma_{ii}^{-1}\right)^{-1}$$

It is recognizable as the matrix version of the well known rule of weighting independent scalar estimates inversely proportionally to their variances.[11] Appendix B will provide a simple example of covarying estimates, and will outline its importance to conjoint modeling, or to modeling in which ultimate paid and incurred losses must be equal.

[11] Unlike scalar weighting, a matrix-weighted average can fall outside its extremes, e.g.:
$\begin{bmatrix}0.50 & 0.25\\ 0.25 & 0.40\end{bmatrix}\begin{bmatrix}400\\ 420\end{bmatrix}+\begin{bmatrix}0.50 & -0.25\\ -0.25 & 0.60\end{bmatrix}\begin{bmatrix}440\\ 400\end{bmatrix}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}425\\ 398\end{bmatrix}$, yet $398 \notin [400,420]$. This is due to non-zero off-diagonal amounts ($\pm$ 0.25) in the weighting matrices. In practice, such amounts are relatively small, and the matrix-weighted averages lie within their extremes. Cf. Judge [1988, 287].

13. CONCLUSION

The Gauss-Markov theorem is truly profound. It provides a lucid basis for solving a wide range of modeling and estimation problems, even within the rudimentary matrix functionality of Excel. As the many sections of this paper have demonstrated, it deserves to be liberated from being an appendage to the least-squares approach to linear statistical modeling.[12]

[12] For a brief history of the least-squares method and the true relation of the Gauss-Markov theorem to it see Appendix C.

REFERENCES

[1.] Halliwell, Leigh, "Conjoint Prediction of Paid and Incurred Losses," *1997 Loss Reserving Discussion Papers*, Casualty Actuarial Society, 1997, 241-379, www.casact.org/pubs/forum/97sforum/97sf1241.pdf.

[2.] Halliwell, Leigh, "Statistical Models and Credibility," *CAS Forum* (Winter 1998), www.casact.org/pubs/forum/98wforum/98wf061.pdf.

[3.] Halliwell, Leigh, "Chain-Ladder Bias: Its Reason and Meaning," *Variance*, 1:2, 2007, 214-247, www.variancejournal.org/issues/01-02/214.pdf.

[4.] Judge, George G., Hill, R. C., *et al.*, *Introduction to the Theory and Practice of Econometrics* (Second Edition), New York, John Wiley & Sons, 1988.

[5.] Wikipedia contributors, "Best linear unbiased prediction," *Wikipedia*, http://en.wikipedia.org/wiki/Best_linear_unbiased_prediction (accessed September 2015).

[6.] Wikipedia contributors, "Gauss–Markov theorem," *Wikipedia*, http://en.wikipedia.org/wiki/Gauss-Markov_theorem (accessed September 2015).

[7.] Wikipedia contributors, "Least squares," *Wikipedia,* http://en.wikipedia.org/wiki/Least_squares (accessed September 2015)

[8.] Wikipedia contributors, "Mahalanobis distance," *Wikipedia*, http://en.wikipedia.org/wiki/Mahalanobis_distance (accessed September 2015).

APPENDIX A

GEOMETRIC MATTERS CONCERNING VECTORS IN $\Re^n$

In this appendix we will interpret the vector dot product and prove the triangle inequality. For $\mathrm{x},\mathrm{y} \in \Re^n$ the dot product $\mathrm{x} \bullet \mathrm{y} = x'y = \sum_{i=1}^{n} x_i y_i$. However, this is easily generalized with a Σ metric as $\mathrm{x}'\Sigma^{-1}\mathrm{y}$. The Σ-metric triangle inequality is:

$$\sqrt{(\mathrm{x}+\mathrm{y})'\Sigma^{-1}(\mathrm{x}+\mathrm{y})} \le \sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}} + \sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}$$

As for the dot product, let $\hat{\mathrm{y}}$ be the Σ projection of y onto x. If $\mathrm{x}_{n\times 1} \neq 0$, then x is of full-column rank, $\mathrm{x}'\Sigma^{-1}\mathrm{x}$ is 1×1 positive-definite, and $(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}$ exists. So vector y will Σ-project as some multiple of x, or $\hat{\mathrm{y}} = \mathrm{x}\beta$. From Section 7, $\beta = (\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}\mathrm{x}'\Sigma^{-1}\mathrm{y}$. Hence, $\hat{\mathrm{y}} = \mathrm{x}(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}\mathrm{x}'\Sigma^{-1}\mathrm{y}$. Accordingly, $\mathrm{x}'\Sigma^{-1}\mathrm{y} = \mathrm{x}'\Sigma^{-1}\mathrm{x}(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}\mathrm{x}'\Sigma^{-1}\mathrm{y} = \mathrm{x}'\Sigma^{-1}\hat{\mathrm{y}}$. So the Σ-metric dot product of two vectors is equal to Σ-metric dot product of one vector and the Σ projection of the other onto it. Although $(\mathrm{x}'\Sigma^{-1}\mathrm{x})^{-1}$ does not exist if $\mathrm{x} = 0$, we know that the projection of any vector onto 0 is 0. Hence, our geometric interpretation of the dot product is valid for all x and y. For the Euclidean metric $\Sigma = \mathrm{I}_n$, the projection is the perpendicular, and $|\hat{\mathrm{y}}| = |\mathrm{y}|\cos\theta$, where θ is the angle between the two vectors with vertex at 0. From this follows the well-known formula $\mathrm{x} \bullet \mathrm{y} = x'y = |\mathrm{x}||\mathrm{y}|\cos\theta$.

In preparation for the triangle inequality, since Σ is positive-definite, the following 2×2 symmetric matrix is non-negative-definite:

$$\left[\mathrm{x}_{n\times 1} \quad \mathrm{y}_{n\times 1}\right]'\Sigma^{-1}\left[\mathrm{x} \quad \mathrm{y}\right]=\begin{bmatrix}\mathrm{x}'\\ \mathrm{y}'\end{bmatrix}\Sigma^{-1}\left[\mathrm{x} \quad \mathrm{y}\right]=\begin{bmatrix}\mathrm{x}'\Sigma^{-1}\mathrm{x} & \mathrm{x}'\Sigma^{-1}\mathrm{y}\\ \mathrm{y}'\Sigma^{-1}\mathrm{x} & \mathrm{y}'\Sigma^{-1}\mathrm{y}\end{bmatrix}$$

It is a theorem that the determinant of a non-negative-definite matrix is non-negative; but we can readily prove it here for the 2×2 case. Such a matrix can be Cholesky factored as $\begin{bmatrix}a & 0\\ b & d\end{bmatrix}\begin{bmatrix}a & b\\ 0 & d\end{bmatrix}$, for real numbers *a*, *b*, and *c*. This equals $\begin{bmatrix}a^2 & ab\\ ba & b^2+d^2\end{bmatrix}$, whose determinant is $a^2\left(b^2+d^2\right)-a^2b^2=a^2d^2$, which must be greater than or equal to zero.

Now let '~' stand for the relationship in the triangle inequality:

$$\sqrt{\left(\mathrm{x}+\mathrm{y}\right)'\Sigma^{-1}\left(\mathrm{x}+\mathrm{y}\right)}\sim\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}+\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}$$

Because the quantities under all the radical signs are non-negative, the following transformations will not affect the relationship:

$$\begin{aligned}
\left(\mathrm{x}+\mathrm{y}\right)'\Sigma^{-1}\left(\mathrm{x}+\mathrm{y}\right) &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x}+2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}+\mathrm{y}'\Sigma^{-1}\mathrm{y}\\
\mathrm{x}'\Sigma^{-1}\mathrm{x}+\mathrm{x}'\Sigma^{-1}\mathrm{y}+\mathrm{y}'\Sigma^{-1}\mathrm{x}+\mathrm{y}'\Sigma^{-1}\mathrm{y} &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x}+2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}+\mathrm{y}'\Sigma^{-1}\mathrm{y}\\
\mathrm{x}'\Sigma^{-1}\mathrm{y}+\mathrm{y}'\Sigma^{-1}\mathrm{x} &\sim 2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}\\
2\,\mathrm{x}'\Sigma^{-1}\mathrm{y} &\sim 2\sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}\\
\mathrm{x}'\Sigma^{-1}\mathrm{y} &\sim \sqrt{\mathrm{x}'\Sigma^{-1}\mathrm{x}}\sqrt{\mathrm{y}'\Sigma^{-1}\mathrm{y}}\\
\left(\mathrm{x}'\Sigma^{-1}\mathrm{y}\right)^2 &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x}\cdot\mathrm{y}'\Sigma^{-1}\mathrm{y}\\
\mathrm{x}'\Sigma^{-1}\mathrm{y}\cdot\mathrm{y}'\Sigma^{-1}\mathrm{x} &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x}\cdot\mathrm{y}'\Sigma^{-1}\mathrm{y}\\
0 &\sim \mathrm{x}'\Sigma^{-1}\mathrm{x}\cdot\mathrm{y}'\Sigma^{-1}\mathrm{y}-\mathrm{x}'\Sigma^{-1}\mathrm{y}\cdot\mathrm{y}'\Sigma^{-1}\mathrm{x}
\end{aligned}$$

But the expression on the right of the last line is the determinant of our 2×2 non-negative-definite matrix. Therefore '~' is '≤'. Thus have we proven the triangle inequality in $\mathfrak{R}^n$ for every valid Σ metric.

APPENDIX B

COVARYING ESTIMATORS AND CONJOINT PREDICTION

This appendix furnishes a simple, but not too uncontrived, example of combining estimators that are not independent. Let $X_i \sim [\mu, \sigma^2]$ be independent random variables. Our task will be to estimate the mean μ. However, we must estimate it from two known statistics, $Y_1 = (X_1 + X_2 + X_3)/3$ and $Y_2 = (X_3 + X_4)/2$. Four X variables have been melded into two Y variables: $Y_1 \sim [\mu, \sigma^2/3]$ and $Y_2 \sim [\mu, \sigma^2/2]$. But since X_3 is common to both, they are not independent; rather, $Cov[Y_1, Y_2] = Cov[X_3/3, X_3/2] = \sigma^2/6$. So the first two moments of the **y** vector are:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \sigma^2 \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix} \right)$$

The problem is to minimize $\mathrm{W}'\sigma^2 \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix} \mathrm{W}$ subject to $\begin{bmatrix} 1 & 1 \end{bmatrix} \mathrm{W}_{2\times 1} = \mathrm{I}_1$. By the Gauss-Markov theorem (the σ^2 cancels, so it's omitted):

$$\mathrm{W}^* = \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-1} \mathrm{I}_1$$

$$= \frac{1}{5/36} \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \frac{1}{5/36} \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-1}$$

$$= \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1/2 & -1/6 \\ -1/6 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 1/3 \\ 1/6 \end{bmatrix} (1/2)^{-1} = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$$

So the minimal variance results from combining in a 2:1 = 10:5 ratio. One who ignored the covariance would weight them in a 3:2 = 9:6 ratio, underweighting the first and overweighting the second. The minimal variance itself is:

$$\begin{bmatrix}2/3 & 1/3\end{bmatrix}\sigma^2\begin{bmatrix}1/3 & 1/6\\ 1/6 & 1/2\end{bmatrix}\begin{bmatrix}2/3\\ 1/3\end{bmatrix} = \sigma^2\frac{5}{18} = \sigma^2\cdot 0.2\overline{7}$$

Since $\sigma^2/4 < \sigma^2\cdot 0.2\overline{7} =< \sigma^2/3$, the informational value of the two Y statistics lies in between the informational values of three and four independent X statistics.

As for conjoint prediction, the following model combines submodels *a* and *b*:

$$\begin{bmatrix}\mathbf{y}_{a1}\\ \mathbf{y}_{b1}\\ \hline \mathbf{y}_{a2}\\ \mathbf{y}_{b2}\end{bmatrix} = \left[\begin{array}{cc} \mathrm{X}_{a1} & 0\\ 0 & \mathrm{X}_{b1}\\ \hline \mathrm{X}_{a2} & 0\\ 0 & \mathrm{X}_{b2}\end{array}\right]\begin{bmatrix}\beta_a\\ \beta_b\end{bmatrix} + \begin{bmatrix}\mathbf{e}_{a1}\\ \mathbf{e}_{b1}\\ \hline \mathbf{e}_{a2}\\ \mathbf{e}_{b2}\end{bmatrix}, \text{where } Var\begin{bmatrix}\mathbf{e}_{a1}\\ \mathbf{e}_{b1}\\ \hline \mathbf{e}_{a2}\\ \mathbf{e}_{b2}\end{bmatrix} = \left[\begin{array}{cc:cc} \Sigma_{a11} & 0 & \Sigma_{a12} & 0\\ 0 & \Sigma_{b11} & 0 & \Sigma_{b12}\\ \hdashline \Sigma_{a21} & 0 & \Sigma_{a22} & 0\\ 0 & \Sigma_{b21} & 0 & \Sigma_{b22}\end{array}\right]$$

One who works through the formulas of Section 8 will find that the solution of the combination is identical to the combination of the separate solutions (Halliwell [1998, Appendix A]). Were it not for this good fortune, one would have to model everything in order to model anything. So this combination is trivial; although the submodels are written down together, they interact neither in the design matrix nor in the variance structure. But conjoint prediction (Halliwell [1997]) makes use of the fact that paid losses (model *a*) and incurred losses (model *b*) must ultimately be equal by exposure period. This constrains the variance matrix; the sums of the paid and the incurred errors of each exposure period must be equal. But additionally, it imposes a restriction on the parameters. The *a priori*, or prior-to-any-observation, expected values are $E[\mathbf{y}_a] = \mathrm{X}_a\beta_a$ and $E[\mathbf{y}_b] = \mathrm{X}_b\beta_b$. The exposure-period sums of these paid and incurred vectors must also be equal. A "semi-conjoint"

model adds the appropriate β constraint to the trivial combination: $\mathrm{R}\beta = \left(\mathrm{Q} \begin{bmatrix} \mathrm{X}_a & 0 \\ 0 & \mathrm{X}_b \end{bmatrix} \right) \cdot \begin{bmatrix} \beta_a \\ \beta_b \end{bmatrix} = 0 = \mathrm{r}$; consequently, $\hat{\boldsymbol{\beta}}_a$ and $\hat{\boldsymbol{\beta}}_b$ will covary. This parameter covariance will introduce covariance, or off-diagonal blocks, into $Var\begin{bmatrix} \mathbf{y}_{a2} - \hat{\mathbf{y}}_{a2} \\ \mathbf{y}_{b2} - \hat{\mathbf{y}}_{b2} \end{bmatrix}$.

According to Section 12, one may best combine the semi-constrained solutions $\begin{bmatrix} \hat{\mathbf{y}}_{a2} \\ \hat{\mathbf{y}}_{b2} \end{bmatrix}$ and $Var\begin{bmatrix} \mathbf{y}_{a2} - \hat{\mathbf{y}}_{a2} \\ \mathbf{y}_{b2} - \hat{\mathbf{y}}_{b2} \end{bmatrix}$ according to the linear constraint that exposure-period sums of paid and incurred losses are equal. Equivalently, in terms of Section 11, one can project the semi-constrained solution into the subspace of the constraint. Although a proof of this has so far eluded us, it works with examples. So the Gauss-Markov theorem seems to allow modeling temporarily to ignore variance restrictions in order to arrive at a tentative solution that can rather easily be collapsed by the hitherto ignored restrictions into the desired solution. This is the meaning of the sentence at the end of Section 10: "There seems to be a certain commutativity between constraining/projecting and solving." Conjoint prediction by collapsing a semi-conjoint model is much easier than fully conjoint prediction; it requires no eigen-decomposition, and is amenable to a spreadsheet solution.

APPENDIX C

LEAST-SQUARES VERSUS GAUSS-MARKOV

Many, probably most, actuaries think in terms of linear regression, rather than in terms of linear modeling. The standard linear-regression problem begins with t observed quantities y_j. Each observation is associated with a k-tuple of known variables $(x_{j1},\ldots,x_{jk})$, on which the observation is believed linearly to depend, i.e., $y_j = x_{j1}\beta_1 + \ldots + x_{jk}\beta_k$. Of course, if $t = k$ and the k-tuples are linearly independent, one is merely solving simultaneous equations for the β_j. The regression problem arises when $t > k$, and the equations are approximate: $y_j \approx x_{j1}\beta_1 + \ldots + x_{jk}\beta_k$. One then needs to find the values of β_j that make $x_{j1}\beta_1 + \ldots + x_{jk}\beta_k$ most closely approximate the y_j. A reasonable method, called "least squares," is to find the β_j that minimize the sum of the squared errors, i.e., to minimize $f(\beta_1,\ldots,\beta_k) = \sum_{j=1}^{t}\left(y_j - x_{j1}\beta_1 + \ldots + x_{jk}\beta_k\right)^2$. This is a problem well within the capability of a first-year calculus student.

The least-squares criterion for fitting, or "regressing," the best line to data first appeared in print in 1805, when Legendre published his Nouvelles méthodes pour la détermination des orbites des comètes. Earlier, in 1801, Gauss had applied the method to predict the reappearance of Ceres, which had just been discovered and then lost. However, he did not publish the method until 1806 in his Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium. Apparently, he did not refer to Legendre; and in the ensuing controversy over priority Gauss insisted that he had worked out the method at least as early

as 1795, when at the age of eighteen he entered the University of Göttingen.[13] The relevant point of this interesting story is that in this early astronomical setting, the least-squares method was not statistical modeling. It was applied to deterministically moving objects (comets and the newly discovered asteroids). All uncertainty stemmed from the imprecision of the astronomers. But for the first time it was realized that many economical but "fuzzy" observations could be more useful than one costly but "sharp" observation.

Gradually the approximate equations were turned into exact ones with random error terms: $\boldsymbol{y}_j = x_{j1}\beta_1 + \ldots + x_{jk}\beta_k + \boldsymbol{e}_j$. Gauss himself in 1822 stated the optimality of the least-squares method, an early form of BLUE. So today we talk of the "Gauss-Markov" theorem because Gauss started it. But the linear algebra and statistical theory that developed after his death in 1855 culminated in the work of Andrey Markov (1856-1922). Even today it is common for students to be introduced into linear modeling by way of least squares; many texts still refer to the matrix formula $\boldsymbol{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}$ as the OLS ("Ordinary Least Squares") estimator.

How does the least-squares method differ from our version of the Gauss-Markov theorem? To put it in modern terms, both deal with estimating the β parameter in the model of Section 7: $\mathbf{y} = \mathrm{X}\beta + \mathbf{e}$, where $Var[\mathbf{e}] = \Sigma$. But instead of finding the $t \times k$ matrix W that will make estimator $\hat{\boldsymbol{\beta}} = \mathrm{W}'\mathbf{y}$ unbiased and of minimal variance, as per the Gauss-Markov theorem, the least-squares method seeks the value of β for which Xβ most closely

[13] Wikipedia "Least squares" gives an excellent account of this history, which is also recounted in many histories of mathematics. However, there is slight disagreement about some of the dates. Most historians cede the priority to Gauss.

approximates $\mathbf{y}$. "Closeness" here requires distance as measured by the Σ metric. So the least-squares problem is to minimize $f(\beta) = (\mathbf{y} - \mathrm{X}\beta)' \Sigma^{-1} (\mathbf{y} - \mathrm{X}\beta)$. As with the Gauss-Markov theorem, X must be of full-column rank and Σ must be positive-definite. But the two approaches are not logically equivalent. Although they yield the same answer, BLUE is *a posteriori* to the least-squares answer, whereas it is *a priori* to the Gauss-Markov.

The usual approach to the minimization is by means of multivariate calculus:

$$\frac{\partial f}{\partial \beta} = -2\mathrm{X}'\Sigma^{-1}(\mathbf{y} - \mathrm{X}\beta)$$

$$\frac{\partial^2 f}{\partial \beta \partial \beta'} = 2\mathrm{X}'\Sigma^{-1}\mathrm{X}$$

Setting the first derivative to $0_{k\times 1}$, we derive $\tilde{\boldsymbol{\beta}} = (\mathrm{X}'\Sigma^{-1}\mathrm{X})^{-1}\mathrm{X}'\Sigma^{-1}\mathbf{y}$. Since the second derivative is positive-definite, the critical value $\tilde{\boldsymbol{\beta}}$ is a minimum, as desired. However, since vector differentiation is unfamiliar to many (cf. Judge [1988, Appendix A.16]), we will solve the problem algebraically:

$$\begin{aligned}
f(\beta) &= (\mathbf{y} - \mathrm{X}\beta)' \Sigma^{-1} (\mathbf{y} - \mathrm{X}\beta) \\
&= \left(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}} + \mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta]\right)' \Sigma^{-1} \left(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}} + \mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta]\right) \\
&= \left(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}\right)' \Sigma^{-1} \left(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}\right) + \left(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}\right)' \Sigma^{-1} \mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta] + [\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1} \left(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}\right) + [\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1}\mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta] \\
&= f(\tilde{\boldsymbol{\beta}}) + 2[\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1} \left(\mathbf{y} - \mathrm{X}\tilde{\boldsymbol{\beta}}\right) + [\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1}\mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta] \\
&= f(\tilde{\boldsymbol{\beta}}) + 2[\tilde{\boldsymbol{\beta}} - \beta]' \left\{\mathrm{X}'\Sigma^{-1}\mathbf{y} - \mathrm{X}'\Sigma^{-1}\mathrm{X}\tilde{\boldsymbol{\beta}}\right\} + [\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1}\mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta] \\
&= f(\tilde{\boldsymbol{\beta}}) + 2[\tilde{\boldsymbol{\beta}} - \beta]' \left\{\mathrm{X}'\Sigma^{-1}\mathbf{y} - \mathrm{X}'\Sigma^{-1}\mathrm{X}(\mathrm{X}'\Sigma^{-1}\mathrm{X})^{-1}\mathrm{X}'\Sigma^{-1}\mathbf{y}\right\} + [\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1}\mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta] \\
&= f(\tilde{\boldsymbol{\beta}}) + 2[\tilde{\boldsymbol{\beta}} - \beta]' \left\{\mathrm{X}'\Sigma^{-1}\mathbf{y} - \mathrm{X}'\Sigma^{-1}\mathbf{y}\right\} + [\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1}\mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta] \\
&= f(\tilde{\boldsymbol{\beta}}) + [\tilde{\boldsymbol{\beta}} - \beta]' \mathrm{X}'\Sigma^{-1}\mathrm{X}[\tilde{\boldsymbol{\beta}} - \beta] \\
&\geq f(\tilde{\boldsymbol{\beta}})
\end{aligned}$$

As in Section 2, the last line is to be taken in a matrix-definite sense. Moreover, since $\mathbf{X}'\Sigma^{-1}\mathbf{X}$ is positive-definite, the inequality is strict except for $\beta = \tilde{\boldsymbol{\beta}}$. Therefore, $\tilde{\boldsymbol{\beta}}$ uniquely minimizes $f(\beta) = (\mathbf{y} - \mathbf{X}\beta)' \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta)$. Geometrically, the least-squares method drops a Σ perpendicular from $\mathbf{y}$ to the linear subspace swept by $\mathbf{X}\beta$.

But now that we have a "least-squares" estimator, we must check its "BLUE-ness." This is the meaning of the sentence above, that BLUE is *a posteriori* to the least-squares answer. Of course, from our *a priori* Gauss-Markov approach, we already know it to be BLUE, since it is identical to the Section 7 formula $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}$. If $\tilde{\boldsymbol{\beta}}$ were not identical to $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ would be either biased or not as good as $\hat{\boldsymbol{\beta}}$; it would lack either the 'B' or the 'U' of BLUE.

Finally, despite the historical development from least squares to Gauss-Markov, this is neither a "distinction without a difference" nor a matter of taste. Developing the theory of linear statistical modeling from our Gauss-Markov theorem allows us cleanly to solve problems that the least-squares approach can solve only with difficulty, if at all – such problems as predicting (Section 8), constraining (Section 9), projecting (Section 10), incorporating prior information (Section 11), and combining estimates (Section 12).

Credibility for Pricing Loss Ratios and Loss Costs

Uri Korn, FCAS, MAAA

Abstract
This paper discusses how credibility can be applied to pricing loss ratios and loss costs. A method is also presented that can perform a credibility weighted allocation of losses without changing the overall average, which often occurs when applying credibility. Finally, it is shown how Generalized Linear Mixed Models can be used to credibility weight loss ratios while taking multiple dimensions into account. Workarounds are shown for some common pitfalls, and it is explained how to implement these models in spreadsheets.

Keywords. Bühlmann-Straub Credibility, Bayesian Credibility, Loss Ratios, Loss Costs, Generalized Linear Mixed Models

1. INTRODUCTION

When doing any type of actuarial analysis, credibility is an issue that must be frequently dealt with. However, many seemingly simple credibility applications are difficult to apply in practice. For loss ratios and loss costs, seemingly simple concepts such as how to calculate the Bühlmann-Straub parameters or how to perform a credibility weighted allocation are difficult to apply in practice.

In this paper, we discuss these and other practical issues that arise when using credibility with loss ratios and loss costs. For our discussion, we will use the term loss ratio for brevity, but everything mentioned is applicable to loss costs as well. For loss costs, "exposures" should be substituted for "premium" for everything below.

1.1 Outline

We will start our discussion with Bühlmann-Straub credibility and how to apply it to loss ratios and loss costs. The following section discusses the recommended method for calculating loss ratios for pricing studies when credibility is being performed. Section 4 introduces a credibility model that ensures that the credibility weighted results always tie to the original loss ratio. This method is especially useful for performing a credibility weighted allocation of a selected loss ratio. And finally, section 5 discusses the use of mixed models to perform the credibility weighting. It also discusses dealing with some common pitfalls and shows how to implement these models in a spreadsheet or other environment.

2. BÜHLMANN-STRAUB CREDIBILITY

The first topic that will be discussed is how to calculate the Bühlmann-Straub parameters. This includes calculation of the within variance and the between variance. The formulas for each are shown below (Dean

2005).

$$\hat{EPV} = \frac{\sum_{g=1}^{G}\sum_{n=1}^{N_g} W_{gn}(X_{gn} - \bar{X}_g)^2}{\sum_{g=1}^{G}(N_g - 1)} \quad \textbf{(2.1)}$$

$$\hat{VHM} = \frac{\sum_{g=1}^{G} W_g(\bar{X}_g - \bar{X})^2 - (G-1)\hat{EPV}}{W - \frac{\sum_{g=1}^{G} W_g^2}{W}} \quad \textbf{(2.2)}$$

Where EPV is the expected value of the process variance, or the "within variance", and VHM is the variance of the hypothetical means, or the "between variance". W is the weight, G is the number of groups, N is the number of periods, X_{gn} is the indication for group g in period n, $\bar{X}_g$ is the average for group g across all periods, and $\bar{X}$ is the average across all groups and periods.

2.1 The Within Variance

For loss ratios, we will assume that the variance of total losses is proportional to the premium, which implies that the variance of a loss ratio is proportional to the inverse of premium (since calculating the variance of the latter involves dividing the former by the square of the premium). A closer look shows that this must be the case, since the variance of total losses for two (uncorrelated) accounts is equal to the sum of the individual variances. Assuming any other relationship between premium and variance will not agree with this result and will lead to inconsistencies. Similarly, for loss costs, we will assume that the variance is inversely proportional to the exposures.

The next question is what data should be used for calculating this parameter. The answer is that it should be based off of the observed experience, although this is not as straightforward as it sounds. The variance should <u>not</u> be based off of the final selected estimates for each year by using a Bornhuetter-Ferguson method; doing so artificially reduces the variance since each year is moved closer to the a priori estimate and so does not represent the true volatility in the data. Instead, we recommend using an approach similar to the Cape Cod method that compares actual paid or reported losses to used premiums, which are premiums divided by the loss development factor. If the data is capped and excess ratios are used to produce final uncapped loss ratios, then the excess ratios should be applied to the premiums as well to produce used, capped premium, since this reflects the premium relevant to the capped losses. If we want, we can also reflect the fact that some of volatility observed in the loss ratios is due to yearly changes that are not captured in trend or rate changes. We can take this into account and give older, less predictive years less weight by

applying an exponential decay factor to the weights as well. This will be discussed further later on. Doing this will reflect the level of credibility inherent in each year and group, and this is the weight that should be used in the formulas above. Dividing capped paid or reported losses by used, capped premium is mathematically equivalent to using the chain ladder estimates for the ultimate loss ratios multiplied or divided by one minus the excess ratios, depending on how the excess ratios are expressed. So, we are essentially using chain ladder ultimate loss ratios with weights for each year as described. Using this method, we can analyze the actual experience that has emerged and the volatility estimates will be appropriate.

Note that the within variance formula above multiplies the differences squared by the weights, but does not divide by the total of the weights afterwards. This is because the within variance used in the Bühlmann-Straub formula is really more accurately described as a within variance factor, and not the actual within variance for anything in particular. This can be seen from the Bühlmann-Straub credibility formula as well; rearranging the formula below shows that this parameter is divided by the dollar amount to come up with the final within variance.

$$Z = \frac{N}{N + W/A} = \frac{1}{1 + \frac{W/N}{A}} = \frac{1}{1 + V/A}$$

Where N is the weight, W is the within variance from the Bühlmann-Straub formula, or the within variance factor as we will call it, V is the actual within variance, and A is the between variance.

The within variance formula (2.1) assumes that the product of the weights and the square differences from the mean all have the same expected value. The square differences from the mean represent the variance component. So, by taking the average of these values as the within variance factor, this formula essentially assumes that the variances of each year multiplied by the weight are consistent, which is the same as assuming that the weights are proportional to the variance.

Lastly, we will note that formula (2.1) takes an average of the within variance factors by segment, only weighting by the number of years, but not the premium volume. If one wishes, one can modify the formula and use a weighted average by premium volume instead.

2.2 The Between Variance

The second parameter, the between variance, is even more volatile and difficult to calculate than the first. When constructing a hierarchical model with multiple levels, for smaller, lower down levels, if the estimates of this parameter appear unreasonable, assumptions can be made for how each level's between variance relates to the levels above it, and it can then be judgmentally selected accordingly. This parameter is easier to calculate with more groups, and so it can also be calculated between finer segmentations than being used, and then judgmentally adjusted as well. The formula shown above (2.2) can sometimes return negative values,

which means that the indicated between variance is zero.

This formula also assumes that the within variance factor is the same for all groups. Using the logic we discussed, the following formula can be used when the within variance factor is assumed to differ among segments. Caution should be used when doing this however; the within variance is difficult to calculate due to data volatility. It is normally best to use an average across segments for everything. This should only be done in some special cases where the within variance is expected to be significantly different between groups, such as when working with primary and excess data together.

$$\hat{VHM} = \frac{\sum_{g=1}^{G} W_g[(\bar{X}_g - \bar{X})^2 - \frac{(G-1)}{G}\frac{\hat{EPV}_g}{W_g}]}{W - \frac{\sum_{g=1}^{G} W_g^2}{W}} \qquad \textbf{(2.3)}$$

3. CALCULATION OF LOSS RATIOS

For the loss ratios used in any credibility method, we recommend using similar guidelines as mentioned, at least as a starting point. This is not essential however, except for when working with mixed models, which will be discussed later. To recap, the loss ratios for each year are equal to the capped paid or reported losses divided by the used, capped premium, which is the premium divided by the LDF and then multiplied or divided by one minus the excess ratio (ignoring trend and on-leveling). As we mentioned, this is equivalent to multiplying the capped loss ratios by the LDF and then multiplying or dividing by one minus the excess ratio.

If the losses are capped, the loss ratios produced from the credibility procedure should be adjusted to reflect the fact that we are only analyzing a portion of the losses. The final loss ratio should be taken as a weighted average of the credibility weighted result and the overall average loss ratio with weights of one minus the excess ratio and the excess ratio, respectively (assuming that the excess ratio is expressed as a percentage of total losses). This approach assumes that the excess portion that we are not analyzing is running the same as the average for all segments. (It is also acceptable, however, to assume that the excess portion for each segment is running the same as the capped portion and to skip this adjustment, if one desires.)

An additional factor should also be applied to the weight for each year so that more recent years which have more predictive power for the going forward loss ratio receive more weight. This factor is needed since a Bornhuetter-Ferguson method is normally done using the a priori loss ratio obtained from the Cape Cod method. This step uses the a priori loss ratio, but effectively gives even more weight to the recent years,

which have the most predictive power for the going forward loss ratio. Since we are skipping this step, we need another way to give some more weight to the latest years. An exponential decay factor of 0.95 should give similar indications as the full Cape Cod/Bornhuetter-Ferguson method, depending on the LDFs, and a higher or lower factor can be used based on the perceived rate that the business is changing. So, to recap, the weights should be the used premium multiplied or divided by one minus the excess ratio, multiplied by the yearly decay factor. To keep the weights given to each segment appropriate, the total premium should be used as the weight when combining across segments, and the weights mentioned should only be used when aggregating results across years. If the same premium is used for multiple LDF segmentations, the used premium can be calculated using the implied LDF, that is, the total calculated chain ladder ultimate divided by the paid or reported losses.

There are many advantages to this approach. The first is that loss ratios produced in this fashion are a good representation of the actual experience for each year, and the weights correspond to the amount of credibility inherent in each year's estimate; this makes the data well suited for a credibility routine. Second, it is easier to streamline and automate than a Bornhuetter-Ferguson or other similar method, especially when there are many segmentations in the data. Third, it makes it easier to apply assumptions at finer levels of detail than the Bornhuetter-Ferguson method. Lastly, the final weights given to each year are more explicit instead of being implied from the loss development pattern.

There is sometimes some confusion that a Bornhuetter-Ferguson method already performs credibility weighting. This is only true from a reserving perspective, but not from a going forward profitability point of view. A Bornhuetter-Ferguson method gives more weight to the a priori loss ratio for more recent, greener years for which the IBNR for those years are more uncertain. But from a going forward perspective, even if all losses came in instantaneously and there was no need for any loss development, there would still be a need to credibility weight results because of the volatility inherent in the experience. For complete years, the amount of credibility for each year depends on the premium volume. For incomplete years, it is the premium multiplied by the percentage of the year that we have already observed. (The variance is really slightly higher because of the uncertainty in the estimation of the LDFs, but accounting for this would just give more weight to older years, which is counter-intuitive.) So, for a going forward, pricing perspective, if credibility is being applied, we recommend not using the Bornhuetter-Ferguson method and sticking with the original chain ladder method with weights as described. Doing this is non-essential, however, as we mentioned, except for mixed models. But regardless of which methods are used to calculate the actual loss ratios, the Bornhuetter-Ferguson results are <u>not</u> appropriate for the calculation of the within variance.

We will illustrate one way of performing this method with an example: we are developing a book of business that contains segments and sub-segments that we wish to perform credibility on in a hierarchical fashion. We group the data at the sub-segment level. We then calculate three values for each sub-segment for

each year: the on-level premium; the trended, uncapped loss ratio; and the weight. The calculation of the on-level premium is straightforward. The calculation of the latter two is shown in these two formulas (assuming that our analysis is performed on capped, reported losses and that excess ratios are expressed as a percentage of total losses):

$$Loss\ Ratio = \frac{Capped\ Reported\ Losses}{On\text{-}Level\ Premium} \times LDF \times Trend\ Factor / (1 - Excess\ Ratio) \quad \textbf{(3.1)}$$

$$Initial\ Weight = On\text{-}Level\ Premium / LDF \times (1 - Excess\ Ratio) \times Yearly\ Decay\ Factor \quad \textbf{(3.2)}$$

Using these initial weights in the credibility calculation would cause improper weights being given to each segment and sub-segment that are not based on the total premiums of each. To use the total premium as the weights, but still perform the Cape Cod approach as we described above, we apply an off-balance factor for each sub-segment and calculate the final weights used as follows: (Subscripts are used in the below for added clarity; they were ignored in 3.1 and 3.2 for brevity.)

$$Off\text{-}Balance\ Factor_{\text{Sub-Segment}} = \frac{\sum^{All\ Years} On\text{-}Level\ Premium_{\text{Sub-Segment}}}{\sum_{All\ Years} Initial\ Weights_{\text{Sub-Segment}}} \quad \textbf{(3.3)}$$

$$Final\ Weight_{\text{Sub-Segment},Year} = Intial\ Weight_{\text{Sub-Segment},Year} \times Off\text{-}Balance\ Factor_{\text{Sub-Segment}} \quad \textbf{(3.4)}$$

The final loss ratio to use as the input for each sub-segment is calculated by taking the weighted average of the yearly loss ratios using this as the weight. With this approach, summing up the results by segment and year and then calculating the segment loss ratios will tie to the sum of the sub-segment loss ratios, which is clearly a desired condition. These final weights can also be used as the base for final weights in a Generalized Linear Mixed Model (GLMM) or a Bayesian credibility model and both the regression weights and the relative credibility by year will be appropriate. (A further step is really needed for GLMMs, which will be discussed later.) Once the credibility procedure is run, the final selected loss ratios are equal to:

$$LR_{\text{Sub-Segment}} = Credibility\ Loss\ Ratio_{\text{Sub-Segment}} \times (1 - XSR_{\text{Sub-Segment}}) + Average\ LR \times XSR_{\text{Sub-Segment}} \quad \textbf{(3.5)}$$

Where *LR* is the loss ratio and *XSR* is the excess ratio. As mentioned, it is also acceptable to skip this last step. As a compromise, instead of using the overall average loss ratio, the loss ratio from the corresponding segment can be used as well.

As a slight alternative, it is also possible to develop losses and calculate the initial weights at the policy level. Results can then be rolled up into sub-segments by adding the ultimate losses and the initial weights from each policy. The final weights can then be calculated at this level, although it is possible to calculate them at the policy level as well. Doing this yields the same results, but allows for more flexibility in the segmentation structure used for credibility and also makes it easier to use assumptions, such as LDFs and excess ratios, at the policy level.

As mentioned, this approach produces data that fits very nicely into a credibility procedure. Another benefit is that the segmentation structure has less of an impact on the final results than a similar Bornhuetter-Ferguson method.

4. THE TUG-OF-WAR CREDIBILITY METHOD

We will introduce a credibility method that ensures that the average of the resulting credibility weighted results matches the original. This method is well suited for performing a credibility weighted allocation but has other uses that will be discussed. We will focus on loss ratios, although this method can be applied to other items as well.

A frequent problem with applying credibility to loss ratios, is that the average of the credibility weighted results often does not match the original. This causes practical issues since now we must either change our originally selected overall estimate or else the sum of the segments will not tie to the combined. A common solution is to apply an off-balance factor that forces the average of the credibility weighted loss ratios to equal the original overall average, but doing so often produces questionable results, especially when the segments are small and when this off-balance factor is large.

These problems will be demonstrated with the following example: We are analyzing a book of business with a total premium volume of $200 million, which consists of one very large segment with $100 million of premium and a bunch of smaller segments that in total make up the other $100 million. The total loss ratio is judged to be 70%, and we wish to produce credibility weighted loss ratios for each of the segments. The loss ratio of the large segment is 90% and is almost fully credible. The smaller segments have an average loss ratio of 50%, and because of their size, have almost no credibility. If we calculated credibility weighted loss ratios for each segment, the large segment would end up with a loss ratio close to 90%, and the smaller segments would be assigned loss ratios close to the overall mean, which is 70%. Each of these results seem to make sense at the individual level, but summing up all the parts, our average loss ratio for the book is now

around 80%, much higher than the originally estimated 70%. If we applied an off-balance factor to each of the loss ratios, the factor would be equal to 0.7 / 0.8 = 0.875. The large segment would now have a loss ratio of 0.9 x 0.875 = 78.75%, and each of the smaller segments would have loss ratios of 0.7 x 0.875 = 61.25%. The combined average loss ratio is now 70%, as expected, but the results by segment are no longer reasonable. The large, almost fully credible segment is not given enough credibility, only around 50%, and the smaller segments are given way too much.

However, if we took a closer look at the above, the results before the off-balance factor may be problematic as well. If the total loss ratio is 70% and there is one large, nearly fully credible segment with a loss ratio of 90%, then this should imply that the total loss ratio of the smaller segments is 50%. In fact, if we conducted our analysis removing the large segment, this is what we would expect to see. The average loss ratio of the smaller segments can be deduced from what we know about the larger segment. Neither method above takes this into account since they both look at each segment individually, ignoring the results of the other segments.

4.1 Using Bayesian Credibility

To implement this method, we will be using a simple Bayesian credibility model that does not require any special software to run. The results of this model are also consistent with Bühlmann-Straub credibility as will be shown. The reason for using the Bayesian version is because the Bühlmann-Straub method only produces a point estimate, whereas we need to know the entire distribution so that we can find the most optimal solution subject to the constraint that the results must tie to the original overall number. This can only be done using the Bayesian version.

We will be using a normal distribution to model loss ratios, although with variances that differ for each observation. Note that this assumption is not the same as assuming that these items are normally distributed; we are only assuming that each individual loss ratio has a normal distribution on what its possible outcomes might have been. In this way, it is more similar to kernel smoothing than to assuming a distribution. Assuming normality with variances inversely proportional to the dollar amount also produces the same results as taking a weighted average by the dollar amounts, and so is consistent with traditional actuarial analysis.

We will also be assuming that the prior distribution (that is, the credibility complement, in Bayesian terms) is normal as well, which is the common assumption. This is a conjugate prior and the resulting posterior distribution (that is, the credibility weighted result) will also be normal. Only when we assume normality for both the observations and the prior, Bayesian credibility produces the same results as Bühlmann-Straub credibility. The mean of this posterior normal distribution is equal to the weighted average of the actual and prior means, with weights equal to the inverse of the variances of each. As for the variance, the inverse of the variance is equal to the sum of the inverses of the within and between variances (Bolstad 2007). The

variance of the item being credibility weighted is comparable to the within variance, and the variance of the prior is comparable to the between variance. This means that the resulting credibility assigned is equal to the inverse of the within variance divided by the sum of the inverses of both the within variance and the between variance. Using some algebra:

$$Z = \frac{1/V}{1/V + 1/A} \times \frac{V}{V} = \frac{1}{1 + V/A}$$

Where V is the within variance and A is the between variance (or equivalently the variance of the prior distribution). Examining the Bühlmann-Straub credibility formula again, where W is the within variance factor:

$$Z = \frac{N}{N + W/A} = \frac{1}{1 + \frac{W/N}{A}} = \frac{1}{1 + V/A}$$

So, it can be seen that when using normal distributions, Bayesian credibility is equivalent to Bühlmann-Straub credibility. The likelihood formula for this Bayesian model is:

$$\begin{aligned} &N(\text{Credibility Result, Actual Result, Within Variance}) \\ &+ N(\text{Credibility Result, Credibility Complement, Between Variance}) \end{aligned} \qquad \textbf{(4.1)}$$

Where $N(A, B, C)$ is the logarithm of the probability density function (PDF) of a normal distribution at A with a mean of B and variance of C. Maximizing the likelihood of this formula will produce the mentioned result. As an alternative, it is also possible to use the formulas for the mean and variance of the posterior normal distribution that we mentioned. (As a practical issue when programming, it may be necessary to set a minimum on the PDF values so that they are not too close to zero, which can cause problems with logarithms.)

This simple Bayesian model can be solved using only Maximum Likelihood Estimation (MLE). Since the resulting posterior distribution is normally distributed, the mode of this distribution is equal to the mean, as is known. This means that the MLE, which returns the mode, will also be returning the mean in this case.

4.2 Implementing the Method

To implement the Tug-of-War method, we maximize the likelihood of the credibility weighted loss ratios, while constraining the parameters so that the resulting average will match the original.

To do this, we start with initial parameters that represent the relative amount of the total losses allocated to each segment. We then use these initial parameters to calculate percentages that will always add up to one

by taking the initial parameter for each segment and dividing by the sum of all of the initial parameters. We then convert this into a loss ratio by multiplying each percentage by the total amount of losses across all segments and then divide by the premium for each segment. We then calculate the likelihood for each loss ratio using the Bayesian credibility formula shown above. Since each loss ratio affects all of the others, we need to weight the likelihood of each segment to account for this. The weights used for each segment should be the premium. (To use weights in MLE, each log-likelihood should be multiplied by the weight.) The initial parameters are set using an optimization routine that maximizes the total likelihood.

In practice, it helps if the initial parameters are on a logarithmic scale so that negative numbers do not cause problems with negative loss ratios. The parameter of one of the segments can be fixed at a number such as zero or another value, since the real number of parameters is one less than the number of segments since the sum of the percentages must equal one. Also, to help ensure that the maximization routine converges to the correct solution, good starting values should be chosen; these can be obtained from the regular Bühlmann-Straub indications.

There are multiple ways to implement the above scheme. Another way is to set the percentages of one of the segments to one minus the sum of the rest, although this can sometimes result in negative percentages. Another version that is sometimes helpful is to use relativities instead of percentages. In this version, the initial parameters are the initial relativities (on a logarithmic scale). The average relativity is then calculated by taking a weighted average of these initial relativities using the premium as the weight. The final relativities are then set to the initial relativities divided by the average. This will ensure that the resulting average matches the original. Note that in both versions, credibility is calculated on the loss ratios themselves and not on the percentages or the relativities.

To review, the steps are as follows:

1) $Initial\ Parameters\ (Set\ by\ Maximization\ Routine)$

2) $Relative\ Percent\ of\ Losses_i = \exp(Initial\ Parameter_i)$

3) $Percent\ of\ Losses_i = \dfrac{Relative\ Percent\ of\ Losses_i}{\sum Relative\ Percent\ of\ Losses}$

4) $Loss\ Ratio_i = Percent\ of\ Losses_i \times Total\ Losses / Premium_i$

5) $\text{Log-}\ Likelihood_i = \text{Log-}\ Likelihood\ (Loss\ Ratio_i) \times Premium_i$

6) $Total\ \text{Log-}\ Likelihood = \sum \text{Log-}\ Likelihood_i$

If implementing the relativities version, the steps are slightly different:

1) $Initial\ Parameters\ (Set\ by\ Maximization\ Routine)$

2) $Initial\ Relativity_i = \exp(Initial\ Parameter_i)$

3) $Average\ Relativity = \dfrac{\sum Initial\ Relativities_i \times Premium_i}{\sum Premium_i}$

4) $Relativity_i = Initial\ Relativity_i / Average\ Relativity$

5) $Loss\ Ratio_i = Relativity_i \times Overall\ Loss\ Ratio$

6) $\text{Log-}Likelihood_i = \text{Log-}Likelihood(Loss\ Ratio_i) \times Premium_i$

7) $Total\ \text{Log-}Likelihood = \sum \text{Log-}Likelihood_i$

We named this method the Tug-of-War method because each loss ratio tries to maximize its own likelihood, and because of the constraint that the resulting average must equal the original, each loss ratio "tugs" on all of the others as they fight for the highest likelihood that they can achieve. This method produces better results than the application of a simple off-balancing factor, since the likelihood is maximized over all possible combinations that tie to the original average, and so the best tying result is selected. The complement for each segment is essentially revised based on available information from the other segments. It should be noted though that if the off-balance is small, there may not be much benefit to using this more complicated method, and the use of a simpler off-balancing factor may be preferable.

For both the overall loss ratio used as the complement of credibility as well as the individual segment loss ratios used in this model, they can be either actual loss ratios dictated solely from the experience, or they can be selected with some degree of judgment. If the overall loss ratio used is a selected loss ratio, and the segment loss ratios are from the experience, this method is essentially performing a credibility-weighted allocation of the selected loss ratio. A hierarchical model can also be built where the overall loss ratio used for each level is the credibility weighted result from the previous level. Alternatively, if the segment loss ratios are judgmentally selected, and the overall is set to the average of these loss ratios, then this method performs a credibility weighting on the selected loss ratios.

As another similar option, it is possible to use the actual, experience dictated loss ratios for both the overall and the segments and have this method take care of all the selections via credibility weighting, since with a good credibility method there is less need to manually select loss ratios. Adjustments can be made afterwards though, if needed. A hierarchical model can be built similar to the above, as well. It is suggested to use loss ratios and weights as was explained above, but any reasonable method can be used as long as the within and between variances are calculated correctly.

A couple of examples of applying this method are shown below for illustration. The first is very similar to the one given above but shows the actual estimate produced from applying this method in practice.

	Total	Segment 1	Segment 2	Segment 3	Segments 4 - 21
Total Premium	20M	10M	500K	500K	500K
Loss Ratio	70%	100%	40%	40%	40%
Within Standard Deviation	3.9%	31.6%	31.6%	31.6%	31.6%
Between Standard Deviation	10%				
Tug-of-War LR	70%	93.2%	46.8%	46.8%	46.8%
Implied Credibility		77.2%	77.2%	77.2%	77.2%
Bühlmann-Straub LR	81.7%	96.1%	67.3%	67.3%	67.3%
Bühlmann-Straub Credibility		87.0%	9.1%	9.1%	9.1%

Note that with this method, the large segment receives slightly less credibility than it does using the Bühlmann-Straub method. This is because the result of this large segment affects not only its own loss ratio, but all of the other segments as well.

The next example is nearly identical except that one of the smaller segments, segment 3, has a higher loss ratio of 80%. The details are shown below. The point of this example is to show that negative credibilities are possible since the large segment with ten million in premium and a very high loss ratio essentially lowers the complement of credibility for the remaining segments, since, as we have mentioned, we would expect to see an overall lower loss ratio if we performed the analysis without this large segment. Note though that the resulting Tug-of-War loss ratio for this segment still comes out higher than the other small segments, as expected.

	Total	Segment 1	Segment 2	Segment 3	Segments 4 - 21
Total Premium	20M	10M	500K	500K	500K
Loss Ratio	71%	100%	40%	**80%**	40%
Within Standard Deviation	3.9%	31.6%	31.6%	31.6%	31.6%
Between Standard Deviation	10%				
Tug-of-War LR	71%	93.4%	48.4%	**52.1%**	48.4%
Implied Credibility		77.2%	72.8%	**-210.3%**	72.8%
Bühlmann-Straub LR	82.3%	96.2%	68.2%	71.8%	68.2%
Credibility		87.0%	9.1%	9.1%	9.1%

These results and implied credibilities will be explained more in the next section as well.

4.3 Understanding the Results

The loss ratios resulting from this method can sometimes be difficult to interpret at first glance. Even though the correlation between the resulting loss ratios from this method and the Bühlmann-Straub method are usually very high, the relationship between the credibility numbers is less apparent at first. In the simple examples shown in the previous section, it was relatively easy to understand the results, but more realistic scenarios can be more difficult to interpret.

As we explained above, the complement of credibility is effectively changed with this method as it takes all of the information about the expected average loss ratio and the other segment's loss ratios into account. A segment's loss ratio is impacted by the other segments' loss ratios since they provide information and can be used to imply something about our current loss ratio. The amount of impact other loss ratios affect each other is related to how credible each loss ratio is. Using this logic, we can produce a formula to derive what the effective complement for each segment's loss ratio is. We do this by starting with the total losses for the entire book and subtracting out the amount of losses from all of the other segments using the Bühlmann-Straub derived loss ratios. But subtracting out all of these losses would be giving the effect that segments have on each other too much weight. To account for the partial credibility of these loss ratios, we subtract out only a portion of the losses; for this fraction, we use the calculated Bühlmann-Straub credibilities as an approximation. We then divide by the appropriate premium volume to convert these losses into loss ratios.

With this formula, each group receives a different effective complement based on the loss ratios and relative weights of all of the other segments. The formula is as follows:

$$Complement = \frac{Total\ Premium \times Average\ LR - \sum_{i=\ All\ Other\ Segments} Premium_i \times Cred\ LR_i \times Z_i}{Total\ Premium - \sum_{i=\ All\ Other\ Segments} Premium_i \times Z_i} \quad \textbf{(4.2)}$$

Where *Cred LR* is the Bühlmann-Straub loss ratio and *Z* is the credibility. The implied credibility from this new effective complement can be calculated by inverting the credibility formula and solving for *Z*, which results in the following:

$$Z = \frac{LR_{Cred} - LR_{Complement}}{LR_{Segment} - LR_{Complement}} \quad \textbf{(4.3)}$$

These resulting credibilities will not match the Bühlmann-Straub credibilities exactly, but the correlation is usually very high, and these can be used to help explain the results.

As mentioned, some of the resulting loss ratios may not fall in between the (original) complement and the initially indicated loss ratio. Even though we can understand and explain the results, this may still be undesirable. A simple solution is to just select different loss ratios for these segments. This occurs most often with smaller segments and so the impact to the overall average will be small. Another solution is to apply a penalty to the likelihood to help keep the results within range. One way to do this is to subtract from the likelihood the product of the amount that the loss ratio is out of the range by some small penalty constant. (This should be done within the parenthesis before the likelihood is multiplied by the premium so that the penalty is multiplied by the premium volume as well; this seemed to work best. Also, the penalty should usually be less than one or two.) This approach will not guarantee that the loss ratios remain within the range, but it will help push them closer and make being outside of the range less likely. Note, however, that using a penalty puts more constraints on the loss ratios and often lowers the correlation between the implied credibilities and the original and so may make the other loss ratios more difficult to explain.

4.4 Using Classical Credibility

Even though this method requires the within and between variance parameters, it can also be implemented in a classical credibility-like (or limited fluctuation) fashion, if desired. Even though classical credibility has some guidelines for selecting different credibility thresholds, such as having the estimate not

deviate by more than 5% from the true value 90% of the time etc., any such selections for these parameters are mostly arbitrary. That is not to say that there are any problems with using classical credibility; it is just important to realize the need for judgmental estimates and not assume that the results are more objective than they really are. Classical credibility can provide reasonable credibility weighted results in a small amount of time, which in itself is a lot to say in its support.

The Bühlmann-Straub credibility formula is $N / (N + K)$. This formula will assign 50% credibility when $N = K$, and so K can be thought of as the criteria for half credibility. The premium threshold for this can be judgmentally selected. Alternatively, if one is more comfortable with choosing a full credibility threshold, a threshold can be selected for approximate full credibility, and we can then rearrange the classical credibility formula of $Z = \sqrt{X / K}$ to $K_{Half\ \ Credibility\ \ Criteria} = 0.25\, K_{Full\ \ Credibility\ \ Criteria}$ to convert this into a rough half credibility threshold, although of course, this will not be exact.

Using this, the within variance for a segment can be set as the inverse of the dollar amount being used as the weight multiplied by a factor. The between variance can be set to the inverse of the dollar amount that should receive half credibility multiplied by the same factor. The actual factor used has no impact; it is just needed to put the variances on an appropriate scale so that the method can converge.

5. USING GENERALIZED LINEAR MIXED MODELS

5.1 Credibility Weighting Loss Ratios and Loss Costs

As an alternative to the methods presented above, it is also possible to use a Generalized Linear Mixed Model (GLMM) to credibility weight loss ratios and loss costs. See Klinker (2011) for an introduction to these models. Using a GLMM with an identity-link and a normal distribution will produce the same results as applying Bühlmann-Straub credibility. Besides for the benefits it offers of easily allowing hierarchical and multidimensional models, using a GLMM automates the calculations of the within and between variances.

A problem, however, arises when using premiums as the base for the weights, since a GLMM assumes that a weight represents a number of observations. Because of this, using premium will almost always result in assigning full credibility to everything since each premium dollar will be counted as an observation and so the number of observations will be very high[1]. Using an alternative weight, such as claim counts, does not fulfill the desired objective of weighting by premiums, since GLMMs use the same weights for credibility as they do for the regression. Weighting by counts may also cause a bias if there are some segments with high frequency, low severity claims that have a high loss ratio and vice versa, for example. One solution is to multiply the weights by an additional constant equal to the total number of reported claims across all

1 The referenced paper actually shows an example using premium as weights but this appears to be an error.

segments divided by the sum of the original weights so that the new sum of the weights across all segments will be equal to the total number of reported claims. This will allow us to weight by premium volume but still keep the total weight consistent with the number of observations overall. This approach produces reasonable credibility estimates when applied in practice. To summarize, using this method in addition to what we discussed above, the weights should be equal to the following:

$$Premium \;/\; LDF \times (1 - Excess\ Ratio) \times Yearly\ Weight\ Factor \times K \tag{5.1}$$

Where K is the factor that we mentioned[2]. Note that there is only one K factor for all of the data and it has the same value for every segment, regardless of the actual number of claim counts for each. Using this approach, it is possible to build hierarchical and multidimensional credibility models using GLMMs.

Using a GLMM also allows us to use a log-link when credibility weighting, which sometimes produces better results than an identity-link when there are extreme values, as there often are with volatile data, but not always; both ways can be tested to see which produces better results. To avoid errors caused from taking the logarithm of zero, observations with loss ratios of zero should be modified to a very small number slightly above zero, such as 0.00001. Also, even without a log-link, loss ratios with zero weights should be removed so as to not cause errors, which will occur with some GLMM implementations if left in.

5.2 Multidimensional Credibility Models

With GLMMs, it is also possible to build a multidimensional credibility model in which each dimension is assigned a relativity, and each relativity is credibility weighted back towards zero. For multidimensional models, multiplicative relativities usually behave much better and are recommended.

Assuming we have two dimensions and we wish to perform credibility weighting on the relativities of each, there are two main types of models we can build, and another that is a compromise of these two approaches, as will be explained. For this section, we will assume that the two dimensions we are dealing with are industry and territory.

The first type of model is a true two dimensional model where the resulting loss ratios are the product of the two relativities. This assumes that territory relativities are the same for each industry (and vice versa). So if a particular territory is higher than average overall, it will be higher for every single industry by the exact same amount. A positive of this model is that it leverages the credibility of each territory across all industries. But this is a down side as well since it assumes the relativities are always the same, which they will not always be.

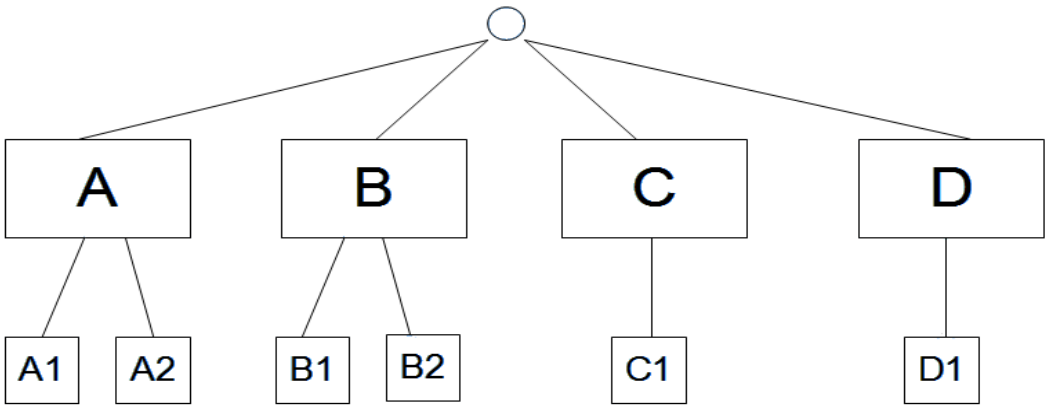
2 Note that this additional factor is not needed for Bayesian models.

The second type of model we can build is really a hierarchical model. For example, we can put territory under industry and first perform credibility at the industry level. We then perform a separate territory credibility calculation for each industry. This will allow territory relativities to differ by industry, but does not leverage the credibility of a territory across industries. Which of these two models to choose depends on our perception of how different the territory relativities are across industries and how volatile our data is. Both of these models can be calculated using GLMMs.

A compromise model can also be built that leverages the credibility of each territory across the industries, but also allows each industry's territory relativities to differ based on the amount of credibility within each cell. In this model, the territory relativities for each industry are effectively credibility weighted back towards the overall territory relativity, which itself is credibility weighted back towards zero. This is the ideal model that combines the best points of each of the above models. Such a model can be built using a GLMM with both industry and territory included as random effects (that is, included as components of the model that take credibility into account), and the interaction of these two added as a third random effect[3]. This type of model is very powerful at producing results at fine levels of detail even when the data is very thin and volatile.

5.3 Uneven Hierarchical Models

When building a model to perform credibility weighting, sometimes we can encounter a data structure where each group has a different number of levels. For example, suppose we are building a hierarchical model on groups and subgroups that looks like the following:



Groups A and B each have two children, while groups C and D only have one, and so really do not have

3 In R, an interaction is added by using a colon between variables. Using the lme4 package, a random effect has the syntax, *(1 | group)* where *group* is the variable we create a random effect on. To do as described, the syntax would be: *(1 | industry) + (1 | territory) + (1 | industry: territory)*

any sub-groupings. This can create problems when building a GLMM since if random effects are assigned to the subgroups C1 and D1, the groups C and D will each have two credibility coefficients that do the same thing, effectively giving double credibility to these groups. Really, we would want the coefficients for C1 and D1 to be given values of zero. This is in fact what happens when building a GLM on this type of data, but not with a GLMM.

This type of model can be built using a GLMM if we add the subgroup random effects as slope coefficients instead of regular intercept coefficients. To explain, most random effects modify the intercept and add or subtract an amount from the intercept, which is same as adding or subtracting this term from the entire equation. But it is also possible to have a random effect behave like a slope parameter instead[4]. Doing this, the coefficients of the random effects are multiplied by a data variable in the equation. Using this, we can create a new variable that is one if its subgroup has any siblings, meaning that it is not the only child of its parent, and zero if it has no siblings. If we create the subgroup random effect as a slope on this variable, it will not allow the nodes C1 and D1 to have non-zero values, and the model will behave as expected.

Similarly, if building the "compromise" model described in the previous section where we gave the example of constructing a model by industry and territory, this unevenness of levels may need to be accounted for as well. A regular model will give double credibility if there is a territory with only one industry, or an industry only under a certain territory. Instead of siblings, we refer to these relationships as cousins. To account for this, similar binary variables can be setup in the data that indicate whether any cousins exist, and the random effects can be added as slope parameters to these variables as described.

5.4 Implementing Mixed Models in Spreadsheets

GLMM credibility models that are either additive or multiplicative can also be implemented in spreadsheets fairly easily using maximum likelihood estimation. To do this, we first determine the formula of the loss ratios, such as *log(Fitted LR) = intercept + territory + industry*, which would create a multiplicative model with territory and industry relativities. To build a regular GLM without credibility weighting, the log-likelihood should be calculated as follows:

$$\sum N\left(Fitted\ LR_i, Actual\ LR_i, \frac{Within\ Variance\ Factor}{Premium\ Base_i}\right) \qquad \textbf{(5.2)}$$

Where *N(A, B, C)* is the logarithm of the Normal PDF at *A* with a mean of *B* and a variance of *C*. The

4 The syntax shown in the previous footnote will create a random effect on the intercept. To create a random effect on the slope, the syntax is: *(0 + variable | group)* where *variable* is the variable we are creating the slope on. (The "*0 +*" is needed here to let it know not to create the random effect on the intercept as well. If we left this part out, random effects would be added both to the intercept and as a slope.)

intercept and territory and industry coefficients should all be determined using a routine that maximizes the total log-likelihood.

A GLMM can be implemented using the simple Bayesian model we described above since MLE parameters are assumed to be approximately normally distributed, and so the posterior distribution should be approximately normal as well. To calculate the log-likelihood for the GLMM, we add the following to formula (5.2):

$$\sum_{t=\text{all territories}} N(Coefficient_t, 0, Between\,Variance\,Territories) + \sum_{i=\text{all industries}} N(Coefficient_i, 0, Between\,Variance\,Industries) \quad \textbf{(5.3)}$$

Using zero as the mean for the prior distributions effectively weights everything back towards the intercept, which is what performs the credibility weighting. The between variances are difficult to calculate, limiting the advantage of this approach however. They can be estimated by looking at the variances of each parameter while controlling for all of the other parameters, possibly by using a GLM.

A plus side is that the Tug-of-War method can be implemented. We suggest using the relativities version shown above and implementing as follows, assuming a multiplicative model: The log-likelihood for the relativity coefficients should be calculated first using formulas similar to (5.3). The exponent of the log-relativities should be taken to calculate the actual relativities for each combination of dimensions and the weighted average overall relativity should be calculated. Revised relativities should then be computed by dividing each relativity by the average and the final loss ratios can be calculated by multiplying these relativities by the average loss ratio. The log-likelihood for each loss ratio can then be taken and added to the overall total. This method will ensure that the overall average of the credibility weighted results ties to the original. It is also possible to ensure that the average of each loss ratio across a particular dimension, industry for example, ties the original average industry loss ratios as well. This can be done by calculating the average relativities across each industry and dividing each relativity by the average relativity for each industry. Ensuring that the averages of more than one dimension tie to the originals puts too many constraints on the solution and is not recommended[5].

5 We ignored the log-likelihood weights in this discussion. One option is to leave out the weights even though this may cause the Tug-of-War method to not work as well. Another option is to apply weights to each relativity log-likelihood equal to the total premium for each item across all of the other dimensions, and weights to each loss ratio log-likelihood equal to the premium of each. This will help the Tug-of-War part of the method work better but slightly violates Bayes' formula which is the formula we are using for the credibility weighting.

6. ACCOUNTING FOR MIX CHANGES AND NON-RENEWALS

The last topic we will discuss is non-renewals and business mix changes. Very often, to improve a book of business, some accounts or segments perceived to be under-performing will be non-renewed, and actuaries are often asked to quantify the impact of these actions. One method (which is often favored by the underwriters) is to completely eliminate all non-renewed business from the experience and calculate predictions on this cleaned up data. But doing so does not account for the credibility of the non-renewed business. To give an extreme example, assume all policies have a loss on average of once every five years and are completely identical in terms of expected losses. If after a couple of years, all accounts with a loss are non-renewed, the historical loss ratio on the remaining business will clearly look much better, but the book really has not changed at all. The expected going forward loss ratio is exactly the same. The same example can be applied to business mix changes as well.

Instead, when calculating the benefit, we suggest incorporating credibility in most cases. (In some cases, however, a major change has truly been made and a unique segment has been non-renewed for which the overall loss ratio of the book does not serve as a good credibility complement; in these situations, it may not make sense to incorporate credibility.) If a particular segment is non-renewed, credibility weighted loss ratios can be produced by segment using one of the methods described above, and the difference to the total loss ratio can be calculated both with and without this particular segment to determine the effect. If accounts with the highest frequency or loss ratios are non-renewed, credibility weighted loss ratios can be calculated by frequency or loss ratio band and the effect can be determined. If just a bunch of poor accounts are non-renewed, a hierarchical model that properly reflects the segmentations in the book of business can be built that goes all the way down to the policy level, and the result of excluding these policies can be determined as well. Although this last case may be the most difficult to model. The same applies to mix changes. Credibility weighted loss ratios can be produced per segment and the total weighted average loss ratio can be calculated before and after the change to help judge the effect on the overall book.

7. CONCLUSION

As pricing actuaries, we are relied upon to help make many important strategic and quantitative decisions. Without a good credibility mechanism, a choice often needs to be made between not giving enough detail and giving enough detail but not accurately. Applying credibility allows us to balance these two demands and provide enough detail and do so accurately.

REFERENCES

Bolstad, W., "Introduction to Bayesian Statistics (Second Edition)," 2007, p.106-108, 207-209

Dean, C., "Topics in Credibility Theory," Education and Examination Committee of the Society of Actuaries, 2005, https://www.soa.org/files/pdf/c-24-05.pdf

Klinker, F., "Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting," Casualty Actuarial Society Forum, 2011, p. 1-25, http://www.casact.org/pubs/forum/11wforumpt2/klinker.pdf

Biography of the Author

Uri Korn is an AVP & Actuary at Axis Insurance serving as the Research and Development support for all commercial lines of insurance. Prior to that, he was a Supervising Actuary at AIG in the Casualty pricing department. His work and research experience includes practical applications of credibility, trend estimation, increased limit factors, non-aggregated loss development methods, and Bayesian models. Uri Korn is a Fellow of the Casualty Actuarial Society and a member of the American Academy of Actuaries.

Incorporating Model Error into the Actuary's Estimate of Uncertainty

Jamie Mackay, and Dave Otto, FCAS MAAA

Abstract

Current approaches to measuring uncertainty in an unpaid claim estimate often focus on parameter risk and process risk but do not account for model risk. This paper introduces simulation-based approaches to incorporating model error into an actuary's estimate of uncertainty. The first approach, called Weighted Sampling, aims to incorporate model error into the uncertainty of a single prediction. The next two approaches, called Rank Tying and Model Tying, aim to incorporate model error in the uncertainty associated with aggregating across multiple predictions. Examples are shown throughout the paper and issues to consider when applying these approaches are also discussed.

Keywords

Model uncertainty, model risk, model error, parameter risk, process risk, model variance, parameter variance, process variance, mean squared error, unpaid claim estimate, uncertainty, reserve variability, bias, simulation, scaling, weighted sampling, rank tying, model tying.

Table of Contents

1 Introduction

One of the core practices performed by property and casualty actuaries is the estimation of unpaid claims, which according to Actuarial Standard of Practice Number 43 (ASOP 43), *Property/Casualty Unpaid Claim Estimates*, is defined as:

> Unpaid Claim Estimate – The actuary's estimate of the obligation for future payment resulting from claims due to past events.

Estimates by their nature are subject to uncertainty and our profession has strived to communicate the uncertainty inherent in unpaid claim estimates to the users of our services. In the past, communications were mostly verbal in the sense that they warned the user of the risk that the actual outcome may vary, perhaps materially, from any estimate, but were rarely accompanied by a quantification of the magnitude of this uncertainty. More recently, actuaries have developed approaches to measure uncertainty and have included this information in their communications.

ASOP 43 suggests that there are three sources of uncertainty in an unpaid claim estimate.

> Section 3.6.8 Uncertainty – "When the actuary is measuring uncertainty, the actuary should consider the types and sources of uncertainty being measured and choose the methods, models and assumptions that are appropriate for the measurement of such uncertainty…Such types and sources of uncertainty surrounding unpaid claim estimates may include uncertainty due to model risk, parameter risk, and process risk." (emphasis added)

ASOP 43 defines each risk as follows:

> 2.7 Model Risk – "The risk that the methods are not appropriate to the circumstances or the models are not representative of the specified phenomenon."
>
> 2.8 Parameter Risk – "The risk that the parameters used in the methods or models are not representative of future outcomes."
>
> 2.10 Process Risk – "The risk associated with the projection of future contingencies that are inherently variable, even when the parameters are known with certainty."

Common approaches to measuring uncertainty, such as the Bootstrapping approach described by England and Verrall (1999, 2002 and 2006) and England (2001) and the distribution-free methodology described by Thomas Mack (1993), are based on the premise that a single model in isolation is representative of the unpaid claims process, and as a result, uncertainty is measured only for parameter and process risk. We believe that circumstances exist in current practice where model risk is evident in the uncertainty surrounding an unpaid claim estimate, and as a result, this paper introduces methodologies to incorporate its impact. These methodologies leverage existing approaches that measure parameter and process risk by supplementing their results with the inclusion of model risk. Examples are shown throughout the paper that, to the extent practical, are based on a single case study which is discussed in more detail in Appendix A.

1.1 Background

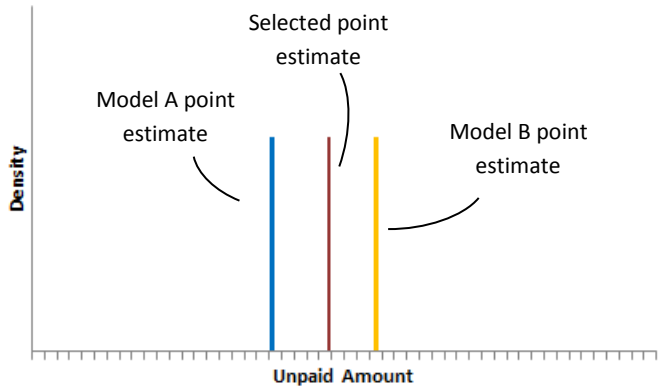
The genesis of this paper and the methodologies presented herein are the result of a dilemma that the authors observed when estimating uncertainty associated with an unpaid claim estimate. This dilemma is perhaps best explained through a hypothetical example.

Consider a hypothetical situation where an actuary uses two actuarial projection methodologies (i.e. models) to estimate unpaid claims for a book of business: Model A and Model B, which both produce a point estimate. Based on the actuary's expertise and professional judgment, the actuary selects the central estimate (colloquially referred to as a "best estimate") to be the straight average of the two point estimates. In other words:

$$Central\ Estimate = \frac{(Model\ A\ Point\ Estimate + Model\ B\ Point\ Estimate)}{2}$$

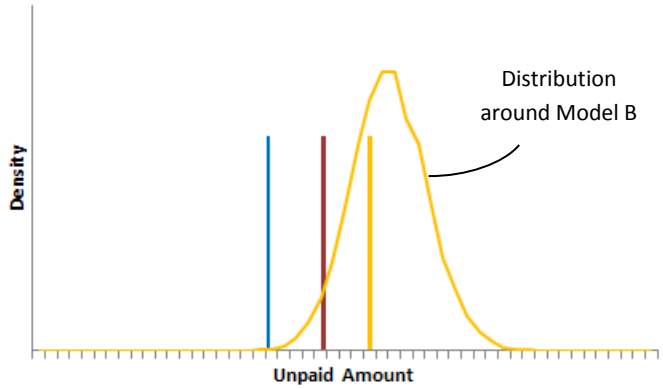
Graphically, these point estimates are shown in Figure 1.

Figure 1. Actuarial central estimate



In order to convey uncertainty in this example, the actuary uses Model B as the basis for estimating uncertainty and observes the following distribution in Figure 2.

Figure 2. Distribution around Model B



If it is assumed that the distribution in Figure 2 is intended to represent the range of uncertainty in the actuary's estimate, then a couple of observations raise concern:

- The actuarial central estimate is not centrally located within the distribution; and
- The distribution implies that the point estimate from Model A is an unlikely outcome, which conflicts with the actuary's professional judgment to equally weight the point estimates from Model A with Model B in selecting a central estimate.

This example is not unique in that it is common for an actuary to estimate unpaid claims with more than one model and it is rare for different models to produce point estimates that are equivalent. Furthermore, current approaches to estimating uncertainty tend to model uncertainty within the context of a single model, which often is not equivalent to the actuary's selected central estimate.

2 Scaling

One approach to dealing with this dilemma is to shift the distribution about Model B so that the mean of the distribution is set equal to the actuary's selected central estimate. This approach, referred herein as scaling, can be done additively, which maintains the same variance, or multiplicatively, which maintains the same coefficient of variation, where:

> For each point, x_i, within a distribution with mean equal to $\bar{x}$, the corresponding scaled points, x'_i, in the distribution are equal to:

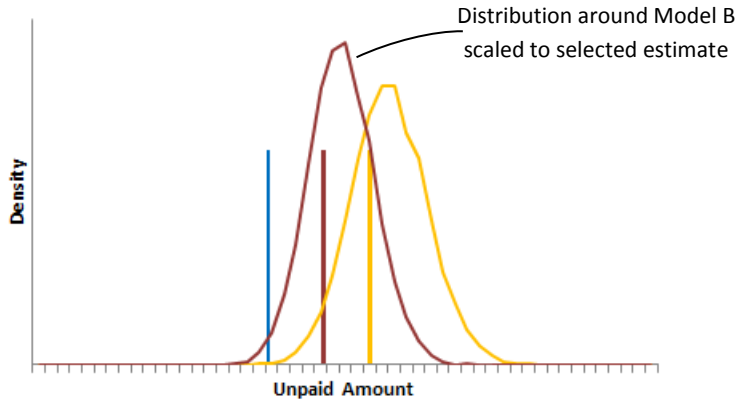
$$Additive\ Scaling: x'_i = x_i + [central\ estimate - \bar{x}]$$

$$Multiplicative\ Scaling: x'_i = x_i \frac{[central\ estimate]}{\bar{x}}$$

Scaling a distribution can be a suitable approach when the magnitude of scaling is immaterial, however, this approach tends to produce unsatisfactory results as the magnitude of the difference between the

point estimates increase. For example, consider the hypothetical results before and after scaling multiplicatively to the actuarial central estimate in Figure 3.

Figure 3. Scaling



In this situation, the mean of the implied distribution after scaling reconciles with the actuarial central estimate, however, the point estimate from Model A continues to appear as an outlier. While this example may be an exaggeration, it highlights a dilemma that an actuary faces when the indications from various models diverge.

3 Mean Squared Error

In order to address this dilemma it may be helpful to explore uncertainty in an estimate from a mathematical perspective. [Authors note: The mathematical terms and formulas in this section are used only for the purpose of establishing a theoretical foundation for uncertainty and its relationship with model error. The approaches introduced afterward for incorporating model error do not rely on these formulas and this section of the paper, however, these formulas are believed to be useful for understanding the basic concepts of uncertainty.]

Uncertainty, as used in the context of this paper, implies that the actual outcome may turn out to be different from our estimate (i.e. prediction). In statistics, the Mean Squared Error (MSE) measures this difference. Consider an outcome as a random variable, y and a prediction, $\hat{y}$. The mean squared error is:

$$E[(y - \hat{y})^2]$$

Expanding this term through additive properties yields:

$$E[(y - \hat{y})^2] = E[(y - \hat{y} + (E[y] - E[y]) + (E[\hat{y}] - E[\hat{y}]))^2]$$

Reordering yields

$$E[(y - \hat{y})^2] = E\left[\left((y - E[y]) - (\hat{y} - E[\hat{y}]) + E[y] - E[\hat{y}]\right)^2\right]$$

If we assume y and $\hat{y}$ are independent, then the formula reduces to

$$E[(y - \hat{y})^2] = E[(y - E[y])^2] + E[(\hat{y} - E[\hat{y}])^2] + (E[y] - E[\hat{y}])^2$$

Appendix B derives this formula in more detail. This equation as it is currently structured highlights a key relationship: the mean squared error equals the sum of process variance, parameter variance and squared bias, where:

$$\boldsymbol{Process\ Variance} = \boldsymbol{Var(y)} = \boldsymbol{E}\left[(\boldsymbol{y} - \boldsymbol{E}[\boldsymbol{y}])^2\right];$$

$$\boldsymbol{Parameter\ Variance} = \boldsymbol{Var}\ (\hat{\boldsymbol{y}}) = \boldsymbol{E}\left[(\hat{\boldsymbol{y}} - \boldsymbol{E}[\hat{\boldsymbol{y}}])^2\right]; \boldsymbol{and}$$

$$\boldsymbol{Squared\ Bias} = \ (\boldsymbol{Bias}(\boldsymbol{y}, \hat{\boldsymbol{y}}))^2 = (\boldsymbol{E}[\boldsymbol{y}] - \boldsymbol{E}[\hat{\boldsymbol{y}}])^2.$$

These terms are discussed further below.

3.1 Process Variance

$\boldsymbol{Var(y)} = \boldsymbol{E}\left[(\boldsymbol{y} - \boldsymbol{E}[\boldsymbol{y}])^2\right]$

The formula for process variance uses the terms y and $E[y]$. The variable y is the actual outcome we are trying to predict, which is presumed to be a random variable that is generated from a distribution with mean equal to $E[y]$. In other words, process variance measures the variance of actual outcomes.

Insurance is believed to be a stochastic process (or nearly stochastic in the sense that the sheer number of conditions which contribute to an actual outcome makes it appear random simply because we are unable to account for all of that information) and the variability inherent in a single outcome occurring is measured by process variance. Consider the flipping of a coin where the probability of a "head" occurring is equal to the probability of a "tail." Despite this knowledge of the underlying probabilities, we are still unable to accurately predict the outcome from a single flip of the coin because there is an element of randomness to any single observation. The estimation of unpaid claims in insurance is similar in that the actual outcome to which we are predicting is a single observation that is one of many probable outcomes which could occur.

3.2 Parameter Variance

$\boldsymbol{Var}\ (\hat{\boldsymbol{y}}) = \boldsymbol{E}[(\hat{\boldsymbol{y}} - \boldsymbol{E}[\hat{\boldsymbol{y}}])^2]$

The formula for parameter variance uses the terms $\hat{y}$ and $E[\hat{y}]$ where the variable $\hat{y}$ is the prediction. Actuaries make predictions of unpaid claims through the application of projection methodologies that attempt to model the overall insurance process using parameters that are estimated from a data sample. Generally speaking, not every point within the distribution of probable predictions from a model is a suitable candidate for an actuarial prediction. Our goal as actuaries is to parameterize the model such that the resulting prediction, $\hat{y}$, is central to the distribution, however, this prediction may not be equal to the true underlying mean of the model, $E[\hat{y}]$, because of our uncertainty in estimating

the model's parameters from the data sample. Parameter variance is also called estimation variance because this term of the MSE measures the uncertainty in the estimation of the model parameters.

3.3 Squared Bias

$$(\boldsymbol{Bias}(\boldsymbol{y},\hat{\boldsymbol{y}}))^{\boldsymbol{2}} = (\boldsymbol{E}[\boldsymbol{y}] - \boldsymbol{E}[\hat{\boldsymbol{y}}])^{\boldsymbol{2}}$$

In statistics, a prediction, $\hat{y}$, is considered unbiased if the expected value of the prediction is equal to the expected value of the outcome, y, to which we are trying to predict. Otherwise, statistical bias exists and is measured through this term of the mean squared error. Squared bias is relevant when attempting to estimate the parameters of the MSE, which is beyond the scope of this paper. Some methods of estimation, such as maximum likelihood techniques, may produce biased estimates and will require squared bias to be incorporated into the MSE but for simplicity of discussion we will assume squared bias is equal to zero and we will not address it further in this paper when discussing the MSE.

3.4 Estimating the MSE – Single Model

Although the formula for the mean squared error provides theoretical insights into the components of uncertainty in a prediction, it remains a quandary to apply in an actuarial context since it requires us to be able to measure statistical properties (namely mean and variance) of outcomes that could occur, which are unknown. In many industries, the statistical properties of actual outcomes can be derived by observing a sufficiently large number of trials, but unfortunately, the unpaid claim process is not a repeatable exercise.

One way actuaries have dealt with this predicament is by estimating uncertainty on the condition that a particular actuarial projection methodology (i.e. model) in isolation is representative of the random variable, y. In other words, if the unknown distribution of probable outcomes, $f(y)$, is defined by the distribution of probable predictions from Model A, represented as $f_A(y)$, such that:

$$f(y) = f_A(y)$$

then

$$[MSE|f(y) = f_A(y)] = E[(y_A - \hat{y}_A)^2]$$

where,

y_A is the actual outcome, y, generated from Model A, and

$\hat{y}_A$ is the prediction, $\hat{y}$, from Model A.

Under this conditional assumption, process variance can be defined as the distribution of probable outcomes generated from Model A and parameter variance can be defined as the variance in actuarial estimates generated from Model A.

An interesting observation is that the distribution of uncertainty corresponding to the MSE represents a range that is at least as wide and most likely wider than the range of probable outcomes (i.e. process variance) since it must also incorporate the uncertainty associated with the actuary's estimate of the

model's parameters (i.e. parameter variance). In other words the distribution of uncertainty, such as the one shown for Model B in Figure 2, represents the actuary's estimate of potential outcomes conditional on the particular model (i.e. process variance) and the data sample used to estimate the model's parameters (i.e. parameter variance).

3.5 Estimating the MSE – Multiple Models

In isolation, a distribution derived from a single model has intuitive appeal since it represents the only information available. In practice, however, it is uncommon for an actuary's analysis of unpaid claims to be comprised of evaluating only a single model in isolation. ASPOP 43 states:

> Section 3.6.1 Methods and Models – "The actuary should consider the use of multiple methods or models appropriate to the purpose, nature and scope of the assignment and the characteristics of the claims, unless in the actuary's professional judgment, reliance upon a single method or model is reasonable given the circumstances. If for any material component of the unpaid claim estimate the actuary does not use multiple methods or models, the actuary should disclose and discuss the rationale for this decision in the actuarial communication."

Therefore, if multiple models are utilized by the actuary to estimate unpaid claims it seems prudent that the measure of uncertainty recognize the additional knowledge gained from the application of more than one model. As previously hypothesized in Section 1.1, if an actuary uses two models to estimate unpaid claims for a book of business, Model A and Model B with corresponding distributions of probable predictions that could be used to define the distribution of outcomes, $f_A(y)$ and $f_B(y)$ respectively, then two alternatives for estimating the MSE are:

$$[MSE|f(y) = f_A(y)] = E[(y_A - \hat{y}_A)^2]$$

$$[MSE|f(y) = f_B(y)] = E[(y_B - \hat{y}_B)^2]$$

However, it is very likely that

$$f_A(y) \neq f_B(y)$$

and hence the actuary is left with two conflicting solutions for the MSE in this example. If both models are believed to be reasonable representations of $f(y)$, then it may not be appropriate to assume that only one is representative of $f(y)$ because of the ramification it implies with the other model.

$$If\ f(y) = f_A(y), then\ f(y) \neq f_B(y)$$

And likewise

$$If\ f(y) = f_B(y), then\ f(y) \neq f_A(y)$$

Perhaps both models are reasonable representations of $f(y)$ but each model suffers from some unknown function of inaccuracy that we will characterize as model error, such that

$$Model\ Error\ of\ Model\ A = \xi_A = f(y) - f_A(y)$$

$$Model\ Error\ of\ Model\ B = \xi_B = f(y) - f_B(y)$$

Then the introduction of model error can be used to explain the inconsistency between models:

$$f(y) = f_A(y) + \xi_A = f_B(y) + \xi_B$$

Unfortunately, we revert to the predicament of defining uncertainty with unknown terms since model error is unknown. If we use Model A and its corresponding model error to define the distribution, $f(y)$, then:

$$[MSE|f(y) = f_A(y) + \xi_A] = E[(y_A - \hat{y}_A)^2] + \Xi_A$$

is equal to

$$[MSE|f(y) = f_B(y) + \xi_B] = E[(y_B - \hat{y}_B)^2] + \Xi_B$$

where Ξ_A represents the unknown inaccuracy in the MSE as a result of model error in Model A (i.e. ξ_A) and Ξ_B represents the unknown inaccuracy in the MSE as a result of model error in Model B (i.e ξ_B).

If the distribution of uncertainty reflects the uncertainty in outcomes defined by a particular model (i.e. process variance) and the uncertainty associated with estimating that model's parameters (parameter variance) it seems reasonable to incorporate the additional uncertainty associated with the potential error in the underlying model (i.e. model error). Otherwise, the actuary's estimate of uncertainty may be incomplete.

Model error and its corresponding impact on the MSE are both unknown, however, as a general rule the actuary strives to minimize model error. Nevertheless, some model error may remain because it is not possible or practical to identify and correct for it. In the context of selecting a central point estimate, the actuary must choose a single number and oftentimes that number will be based on a weighted average of the reasonable indications from multiple models rather than being set equal to the estimate from any single model. The philosophy underlying this approach, which is akin to hedging one's bet, is that a weighted average of models results in a corresponding unknown model error that is **preferred** to relying on the unknown model error of any single model.

This same philosophy is proposed as our approach to incorporating model error into the actuary's distribution of uncertainty. Revisiting our previous hypothetical that an actuary uses two models to estimate unpaid claims for a book of business, Model A and Model B, and after minimizing model error in Model A and Model B to the extent appropriate the actuary uses expertise and professional judgment to assign weight to the point estimates from these models in accordance with their perceived value as a reasonable predictor such that:

$$Central\ Estimate = w\hat{y}_A + (1-w)\hat{y}_B$$

where

$$0 \leq w \leq 1;$$

$$\hat{y}_A = the\ prediction\ from\ Model\ A; and$$

$$\hat{y}_B = the\ prediction\ from\ Model\ B$$

Then, the MSE and corresponding distribution of uncertainty expressed as a weighted average of predictions from Model A and Model B where each model is separately considered in isolation as representative of the random variable, y,

$$[MSE|f(y) = wf_A(y) + (1-w)f_B(y)]$$

is preferred to the MSE and corresponding distribution conditional only on Model A

$$[MSE|f(y) = f_A(y)]$$

or the MSE and corresponding distribution conditional only on Model B

$$[MSE|f(y) = f_B(y)]$$

if the unknown model error inherent in this weighted averaging of models, $w(\xi_A) + (1-w)(\xi_B)$, is preferred to relying solely on the unknown model error inherent in Model A,ξ_A, or the unknown model error inherent in Model B,ξ_B.

It should be noted that the word "preferred" is used rather than a mathematical relationship such as "less than" in the context of this discussion because this is a philosophical approach. Ideally, we wish to develop a solution that eliminates model error but in the absence of being able to do so, a reasonable alternative is to attempt to recognize our uncertainty in whatever model error remains.

4 Model Error

Before progressing further, it may be helpful to differentiate model error from other types of error. Previously, model risk was defined as "the risk that the methods are not appropriate to the circumstances or the models are not representative of the specified phenomenon."

Many actuarial projection methodologies (i.e. models) can be shown to have no model error when applied in a controlled environment under specific limitations; however, these conditions rarely exist, if at all, in practice. For example, the approach used to extrapolate link ratios into the "tail" of a traditional chain ladder model can introduce model error. An important point to make about model error is that its resulting bias on the actuary's prediction, if any, should be unknown.

4.1 User Error

User error is different from model error. User error occurs when actions, or inactions, of the actuary lead to the **expectation** that the resulting prediction will be biased high or low. Generally accepted actuarial practice is based on the presumption that an actuary's work product is void of significant or

material user error, and hence this type of error should not be incorporated as a component of uncertainty in the actuary's estimate.

4.2 Historical Error

Implicit within most actuarial projection methodologies is the assumption that observations of patterns and trends in the past are indicative of patterns and trends in the future, but future conditions can change and result in materially different processes and outcomes that are often too speculative to estimate. This type of error is a subset of model error and while some changes to future conditions may be reasonably estimable and therefore can be incorporated as an element of uncertainty within the MSE, actuaries oftentimes consider this type of error to be out of scope of their analysis. If so, then the approaches discussed herein will also exclude uncertainty associated with this type of error.

Regardless of the type of error that may exist in a prediction, a goal should be to minimize error within each model to the extent appropriate. Unfortunately, model error often still exists and should therefore be incorporated into the actuary's estimate of uncertainty.

5 Incorporating Model Error

At this point we are ready to introduce a methodology for incorporating model error into an estimate of uncertainty. Various suitable methods exist for estimating the MSE conditional on a single model in isolation so it will be assumed that this analysis has already been performed for each model relied upon by the actuary to derive the central point estimate. This methodology is a simulation-based approach as opposed to a mathematical approach aimed at computing the formulas discussed previously and is perhaps best described through a simplistic example.

5.1 Weighted Sampling

Consider a single actuarial central estimate, $\hat{y}$, to be based on a 50%-50% weighting of estimates produced from two projection methodologies, Model A and Model B, such that:

$$\hat{y} = \sum_{m=A,B} w_m \hat{y}_m$$

Where,

$$\hat{y}_A = the\ prediction\ from\ Model\ A;$$

$$\hat{y}_B = the\ prediction\ from\ Model\ B;$$

$$w_A = 0.5; and$$

$$w_B = 0.5$$

Assume that two distributions of the MSE conditional on Model A and separately for Model B are already estimated and that each distribution is comprised of a series of 10 simulations where each simulation, denoted x_i, is shown in Figure 4.

Figure 4. Single prediction model simulations

Model A Simulations	
Sim	Value
1	3.4
2	2.5
3	1.8
4	3.8
5	4.4
6	3.0
7	2.0
8	6.0
9	3.7
10	6.4

Model B Simulations	
Sim	Value
1	3.6
2	4.6
3	5.2
4	4.4
5	3.4
6	3.6
7	4.4
8	3.9
9	3.4
10	3.0

E.g. simulation x_5 from Model A equals 4.4

A distribution reflecting the inclusion of model error can be estimated by taking a weighted sample without replacement of simulations from Model A and Model B in accordance with their weights. To accomplish this with the example given above, we first create a matrix where we use the weights as the basis for sampling between Model A and Model B for each of the 10 simulations. Because this matrix defines which model to sample for each simulation, we will refer to it as a "Model Matrix," which is shown in Figure 5.

Figure 5. Single prediction Model Matrix

	Wt
Model A	50%
Model B	50%

Model Matrix

Sim	Method
1	B
2	A
3	A
4	B
5	A
6	A
7	B
8	B
9	A
10	A

Once a Model Matrix is created, we select the value corresponding to the simulation number and model to create a series of sampled simulations, which are shown in Figure 6.

Figure 6. Single prediction sampled simulations

Model A Simulations

Sim	Value
1	3.4
2	2.5
3	1.8
4	3.8
5	4.4
6	3.0
7	2.0
8	6.0
9	3.7
10	6.4

Model B Simulations

Sim	Value
1	3.6
2	4.6
3	5.2
4	4.4
5	3.4
6	3.6
7	4.4
8	3.9
9	3.4
10	3.0

	Wt
Model A	50%
Model B	50%

Model Matrix

Sim	Method
1	B
2	A
3	A
4	B
5	A
6	A
7	B
8	B
9	A
10	A

Sampled Simulations

Sim	Value
1	3.6
2	2.5
3	1.8
4	4.4
5	4.4
6	3.0
7	4.4
8	3.9
9	3.7
10	6.4

If we increase the number of simulations in this example to a larger sample size the MSE of the resulting distribution can be estimated by computing the variance of the simulations and the mean of the resulting distribution will be equal to the actuarial central estimate.

Figure 7 shows the results of the distribution before and after incorporating model error when the number of simulations in this example is increased to 10,000.

Figure 7. Single prediction weighted sampling

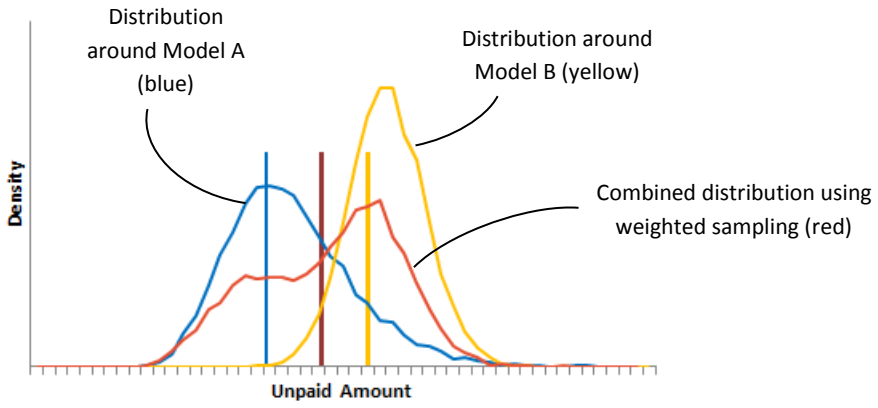
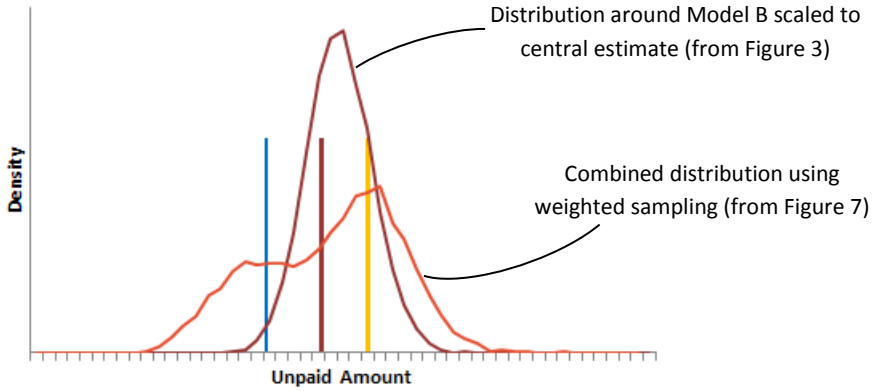


Figure 8 compares weighted sampling in this example to multiplicative scaling Model B's simulations to the central estimate.

Figure 8. Single prediction weighted sampling versus multiplicative scaling



5.2 Considerations

Before we progress the methodology further, it is worth discussing a few points about this approach thus far.

5.2.1 Simulations

It should be noted that in this example, Model B is generated 4 times and Model A is generated 6 times in the Model Matrix. Ideally each model would have been generated an equal number of times since the weighting between the models were equal but the low sample count has led to sample error. For statistically significant sample sizes, we would expect each model in this example to be generated close to 50% of the time.

Sample error must also be considered when evaluating the resulting distribution. Although there is no single number of simulations that is suitable for every circumstance, the user should incorporate a sufficient number to adequately represent the range of potential outcomes, especially if the user is interested in evaluating outcomes generated for extreme tail probabilities.

5.2.2 Individual Model Distributions

Weighted sampling assumes that a distribution of the MSE reflecting the combined effects of process variance and parameter variance is already developed for each model in isolation. Various approaches to estimating the distribution and deriving simulations exist in the literature and example approaches include but are not limited to:

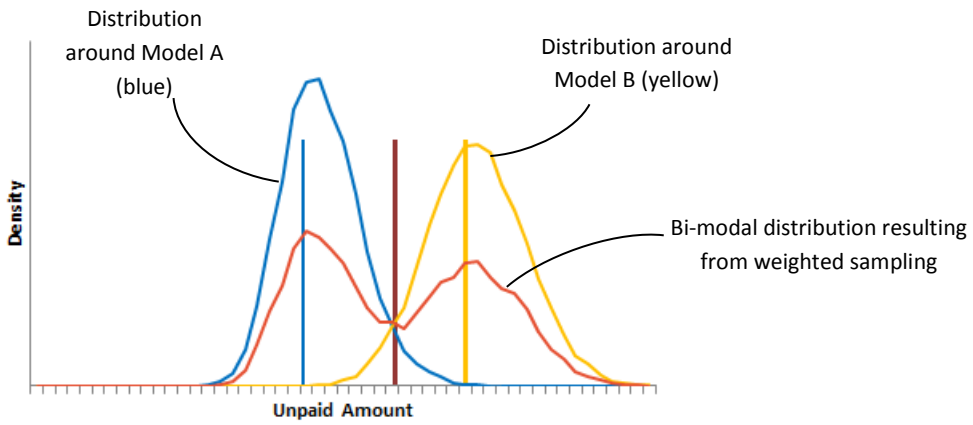
- Simulated approaches – Bootstrapping, Markov-Chain Monte-Carlo simulation or straightforward simulation of outcomes from an assumed distribution using benchmark statistical properties, for example, can be used;
- Analytical approaches – The methodology presented by Thomas Mack is an example of approaches that estimate the statistical properties underlying a model. From this, the user can simulate outcomes once a distributional form is selected; and

- Replicating and scaling – Simulations generated for a particular model can be scaled, either additively or multiplicatively, to the mean of a different model such that an implied distribution of the different model is approximated.

5.2.3 Lumpiness

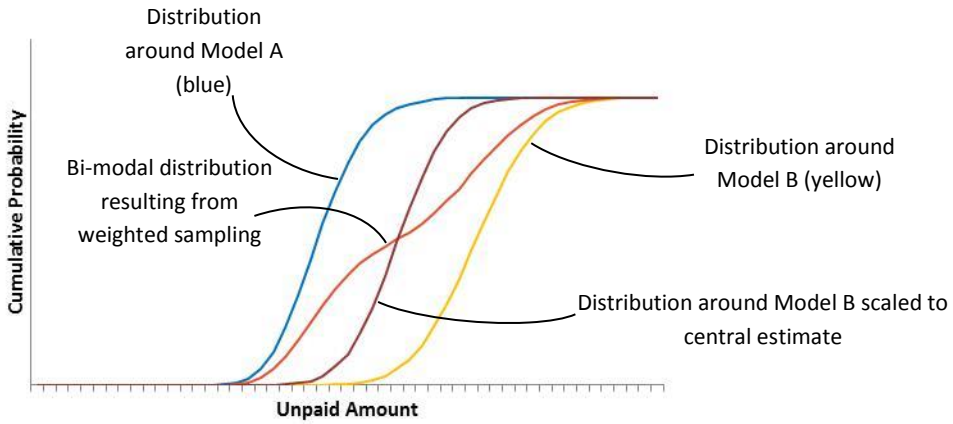
In practice, the user may find the resulting probability density function from weighted sampling to be lumpy, in that there may be multiple modes to the distribution. Figure 9 shows a comparison of weighted sampling from two underlying distributions.

Figure 9. Multi-mode distribution



As a result, it may be challenging to interpret relative probabilities associated with particular outcomes but it is less of an issue when evaluating probabilities associated with a range of outcomes as shown by the corresponding cumulative probability density function for the same example in Figure 9, shown as Figure 10 (also shown in Figure 10 is the distribution around Model B scaled to the selected central estimate).

Figure 10. Multi-mode cumulative probability function



If the shape of the probability density function resulting from weighted sampling is determined to be problematic, the following adjustments could be made:

- Compute the indicated coefficient of variation from the resulting lumpy distribution and re-simulate a newly defined distribution with the same mean and coefficient of variation. Figures 11 and 12 show an example where the lumpy distribution was re-simulated using a Gamma distribution with the same mean and coefficient of variation. It should be noted that a potentially undesired consequence of this adjustment is that probabilities associated with various outcomes within the distribution will be different.
- Probabilities within the range of outcomes where the nodes occur can be re-distributed according to some user-selected smoothed distribution, such as a uniform distribution. An advantage of this adjustment approach is that tail probabilities are unaffected. Figures 13 and 14 show an example of this approach with the probability density graph and the cumulative probability graph, respectively. Note that the actuary should use caution with this approach and be aware that in achieving a more intuitive 'shape' to the distribution, the mean and the coefficient of variation should be maintained.

Figure 11. Re-simulated distribution – probability density

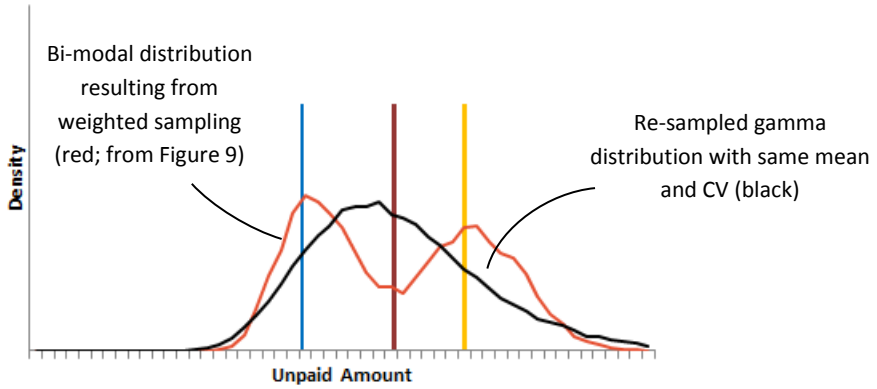


Figure 12. Re-simulated distribution – cumulative probability

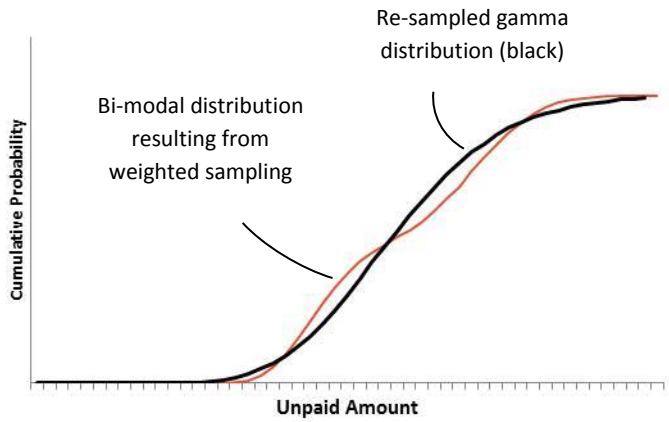


Figure 13. Re-distributed distribution – probability density

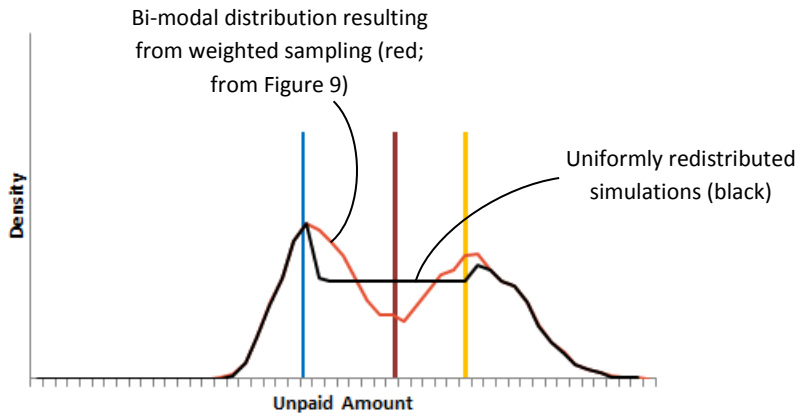
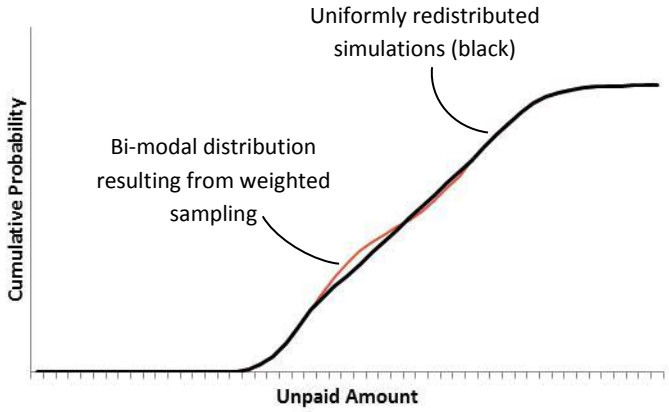


Figure 14. Re-distributed distribution – cumulative probability



5.2.4 Assigning Weights to Models

Assigning weight to a model when using the weighted sampling approach implies that the actuary believes the model is a reliable predictor because otherwise the user may be introducing additional variability that is attributable to user error. Bad practices can exist without harm to deriving a central point estimate, such as having two models that are known to be biased but offset each other so that the average produces a reasonable point estimate (e.g. "two wrongs can make a right" philosophy), but this practice should not be used when estimating uncertainty. In such cases where the models have any known bias, the user may want to consider scaling as a solution instead of weighted sampling.

5.2.5 Effect on MSE

The effect that weighted sampling has on the MSE depends on two factors:

1. The dispersion in the means of the underlying models before weighted sampling; and
2. The MSE of the model distributions before weighted sampling.

As the mean of each model converges to the same point, the resulting MSE using weighted sampling will essentially be an average of the MSE from the various models before weighted sampling. As the mean of each model diverges, the resulting MSE will increase and can be larger than the MSE before weighted sampling of each underlying model.

6 Aggregating Variability

The weighted sampling approach described thus far is an approach to incorporating model error for a ***single*** prediction. Projection methodologies used by actuaries often generate ***multiple*** predictions where each prediction corresponds to a certain subset of claims generally grouped according to a predefined time interval (e.g. accident year, report year, policy quarter, etc.), which we will refer to generically as an origin period. Weighted sampling is suitable for estimating the distribution of any single origin period prediction, however, a separate and more complex approach must be considered when aggregating the variability across multiple origin period predictions.

Consider a situation where each model used by the actuary generates a prediction, $\hat{y}_m$, for multiple different origin periods, t, such that:

$$\hat{y}_{m,t} = \left[\hat{y}_{m,t=1}, \hat{y}_{m,t=2}, \hat{y}_{m,t=3}, \dots\right]$$

and the actuary's selected central estimate for each origin period, t, is

$$\hat{y}_t = \sum_{m=A,B,\dots} w_{m,t}\hat{y}_{m,t}$$

where $w_{m,t}$ corresponds to the weight assigned to model m and origin period t and

$$\sum_{m=A,B,\dots} w_{m,t} = 1$$

Then we wish to derive an approach for aggregating the Mean Squared Error of predictions across all origin periods,

$$MSE = E\left[\left(\sum_{t=1}^{N}\left(y_t - \sum_{m=A,B,\ldots} w_{m,t}\hat{y}_{m,t}\right)\right)^2\right] = ?$$

6.1 Weighted Sampling Revisited

Expanding on the previous example in Section 5.1, consider actuarial central estimates for three separate origin periods, $\hat{y}_{t=1}, \hat{y}_{t=2}\ and\ \hat{y}_{t=3}$, to be based on a 50%-50% weighting of predictions produced from two projection methodologies, Model A and Model B, such that:

$$For\ origin\ periods\ t = 1, 2\ and\ 3$$

$$\hat{y}_t = \sum_{m=A,B,\ldots} w_{m,t}\hat{y}_{m,t}$$

Where,

$$\hat{y}_{A,t} = the\ prediction\ from\ Model\ A\ for\ origin\ period\ t;$$

$$\hat{y}_{B,t} = the\ prediction\ from\ Model\ B\ for\ origin\ period\ t;$$

$$w_{m,t} = \begin{bmatrix} w_{A,1} & w_{A,2} & w_{A,3} \\ w_{B,1} & w_{B,2} & w_{B,3} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{bmatrix}$$

Assume that distributions of the MSE for each origin period conditional on Model A and separately for Model B are already estimated and that each origin period distribution is comprised of a series of 10 simulations where each simulation, denoted x_i, is shown in Figure 15.

Figure 15. Multiple prediction model simulations

Model A Simulations

Sim	t = 1	t = 2	t = 3
1	3.4	5.8	28.8
2	2.5	12.5	28.0
3	1.8	6.5	24.0
4	3.8	8.8	20.0
5	4.4	8.7	14.5
6	3.0	10.7	14.0
7	2.0	9.4	16.9
8	6.0	7.6	24.9
9	3.7	9.7	25.0
10	6.4	8.6	29.0

Model B Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.0	19.9
2	4.6	13.3	26.9
3	5.2	16.1	27.2
4	4.4	11.3	22.7
5	3.4	17.2	26.9
6	3.6	11.3	15.7
7	4.4	10.7	22.9
8	3.9	13.3	22.6
9	3.4	13.5	20.4
10	3.0	13.2	15.0

Once again, a distribution incorporating model error can be estimated for each origin period by taking a weighted sample without replacement of simulations from the distributions of Model A and Model B for each origin period independently in accordance with their weights. As before, this is accomplished by

creating a Model Matrix, shown in Figure 16, where the weights are used as the basis for sampling between Model A and Model B for each set of origin period simulations.

Figure 16. Multiple prediction Model Matrix

Weighting Selections

	t = 1	t = 2	t = 3
Model A	50%	50%	50%
Model B	50%	50%	50%

Model Matrix

Sim	t = 1	t = 2	t = 3
1	B	B	B
2	A	B	A
3	A	B	A
4	B	B	A
5	A	A	B
6	A	A	A
7	B	B	A
8	B	A	B
9	A	B	A
10	A	A	B

Then based on the Model Matrix, we select the value corresponding to the simulation number, model and origin period to create a series of sampled simulations, which can be used as a distribution incorporating model error for each origin period's actuarial central estimate as shown in Figure 17.

Figure 17. Multiple prediction sampled simulations

Model Matrix

Sim	t = 1	t = 2	t = 3
1	B	B	B
2	A	B	A
3	A	B	A
4	B	B	A
5	A	A	B
6	A	A	A
7	B	B	A
8	B	A	B
9	A	B	A
10	A	A	B

Sampled Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.0	19.9
2	2.5	13.3	28.0
3	1.8	16.1	24.0
4	4.4	11.3	20.0
5	4.4	8.7	26.9
6	3.0	10.7	14.0
7	4.4	10.7	16.9
8	3.9	7.6	22.6
9	3.7	13.5	25.0
10	6.4	8.6	15.0

The weighted sampling approach works for multiple separate estimates much in the same way it works for a single estimate; however, dependencies need to be considered before aggregating uncertainty across multiple origin periods. In this example, a total distribution of the three origin periods remains unanswered as depicted in Figure 18.

Figure 18. Multiple prediction weighted sampling

Sampled Simulations

Sim	t = 1	t = 2	t = 3	Total
1	3.6	12.0	19.9	?
2	2.5	13.3	28.0	?
3	1.8	16.1	24.0	?
4	4.4	11.3	20.0	?
5	4.4	8.7	26.9	?
6	3.0	10.7	14.0	?
7	4.4	10.7	16.9	?
8	3.9	7.6	22.6	?
9	3.7	13.5	25.0	?
10	6.4	8.6	15.0	?

6.2 Dependencies

If it can be assumed that within each model the predictions for each origin period are independent then an aggregate distribution representing the total of the three origin periods above can be created quite easily by summing across the values generated above for each simulation (assuming the weighted sampling used to derive the Model Matrix was generated randomly).

Unfortunately, the assumption of independence among different origin periods within a particular model is generally not true. Instead, origin period dependencies are generally inherent within the structure of a model and the process of weighted sampling among various different models for each origin period independently (as described in this example thus far) will break these origin period dependencies. Before discussing an approach to establishing a dependency, if any, among origin periods, it is useful to consider how origin period dependencies may exist within the components that make up uncertainty.

6.2.1 Origin Period Dependency – Process Error

Given that the actual outcome, y, is assumed to be a random variable, we would not expect there to be any dependency in the order in which actual outcomes occur. Therefore, it is usually assumed that the outcome of any given origin period is independent of the outcomes in any other origin period.

6.2.2 Origin Period Dependency - Parameter Error

Parameter variance measures the uncertainty in the actuary's estimate of the model's parameters used to generate a prediction. For many actuarial models, the same parameters and assumptions are used to generate predictions for all origin periods, and as such, any change to a parameter estimate or assumption will permeate through some or all of the origin periods and result in a dependency. Approaches, such as Bootstrapping, produce results which enable the user to measure this dependency.

6.2.3 Origin Period Dependency - Model Error

The model we use to predict $\hat{y}$ is likely an imperfect representation of the true model that defines the actual outcome, y, and as such may result in an unknown tendency to overestimate or underestimate the intended measure. The degree to which a model's error, if any, is dependent across different origin periods is debatable and may depend on the circumstances.

In certain circumstances, it may be argued that a model's error will be consistent across all origin periods. Consider a hypothetical example where the only difference between two chain-ladder models is the approach used to select the tail factor, which results in different values being chosen. Because the tail factor affects the predictions for all origin periods, any error may affect all origin periods.

In other circumstances, it may be argued that error, if any, in any given model may not be consistent across origin periods. For example, chain ladder models tend to be sensitive to the magnitude of cumulative amounts to which the link-ratios are applied and it may be that the cumulative amounts across origin periods exhibit an amount of reasonable volatility with respect to their size relative to historical experience simply because the volume of business being analyzed is not statistically voluminous. If the volatility observed is somewhat random across the origin periods, then the corresponding error in the model, if any, may also be random across origin periods as a result of this attribute.

Because it can be argued that model error dependency may or may not exist across origin periods, we discuss two different approaches to aggregating the weighted sampling distributions across origin periods so that a range of model error dependency assumptions can be used.

6.3 Rank Tying

One approach to aggregating the weighted sampling results across origin periods is to borrow a dependency structure from one of the underlying sampled models. Since process variance does not usually create a dependency across origin periods, any dependency observed is wholly attributable to parameter variance in standard models.

Continuing with the example discussed in Section 6.1, we can create another type of matrix, called a Rank Matrix, that identifies the "rank order" of each simulation within a given model and origin period where the largest value of all simulated values is assigned a rank value of 1. Then, the second largest value of all simulated values within that same model and origin period is assigned a rank value of 2. This process is repeated until all simulations are assigned a rank order value. The Rank Matrix for Model A and Model B are shown in Figure 19.

Figure 19 – Rank Matrix for Model A and Model B

Model A Simulations

Sim	t = 1	t = 2	t = 3
1	3.4	5.8	28.8
2	2.5	12.5	28.0
3	1.8	6.5	24.0
4	3.8	8.8	20.0
5	4.4	8.7	14.5
6	3.0	10.7	14.0
7	2.0	9.4	16.9
8	6.0	7.6	24.9
9	3.7	9.7	25.0
10	6.4	8.6	29.0

Model A Rank Matrix

Sim	t = 1	t = 2	t = 3
1	6	10	2
2	8	1	3
3	10	9	6
4	4	5	7
5	3	6	9
6	7	2	10
7	9	4	8
8	2	8	5
9	5	3	4
10	1	7	1

Model B Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.0	19.9
2	4.6	13.3	26.9
3	5.2	16.1	27.2
4	4.4	11.3	22.7
5	3.4	17.2	26.9
6	3.6	11.3	15.7
7	4.4	10.7	22.9
8	3.9	13.3	22.6
9	3.4	13.5	20.4
10	3.0	13.2	15.0

Model B Rank Matrix

Sim	t = 1	t = 2	t = 3
1	7	7	8
2	2	4	3
3	1	2	1
4	4	9	5
5	8	1	2
6	6	8	9
7	3	10	4
8	5	5	6
9	9	3	7
10	10	6	10

Currently, the weighted sample results for each origin period in Figure 18 produces a different Rank Matrix from the Rank Matrix of Model A and Model B because the underlying Model Matrix was generated randomly in accordance with the weights and therefore broke the origin period links intrinsic to the underlying models. Figure 20 shows the implied Rank Matrix from Figure 18 which is crossed out to denote that the origin period dependencies may not be appropriate.

Figure 20. Rank Matrix from weighted sampling

Sampled Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.0	19.9
2	2.5	13.3	28.0
3	1.8	16.1	24.0
4	4.4	11.3	20.0
5	4.4	8.7	26.9
6	3.0	10.7	14.0
7	4.4	10.7	16.9
8	3.9	7.6	22.6
9	3.7	13.5	25.0
10	6.4	8.6	15.0

Sampled Implied Rank Matrix

Sim	t = 1	t = 2	t = 3
1	7	4	7
2	9	3	1
3	10	1	4
4	3	2	6
5	2	8	2
6	8	5	10
7	5	6	8
8	4	10	5
9	6	2	3
10	1	9	9

If we select Model B as the model to use as the basis for dependency in aggregating simulations across all origin periods, then all we have to do is reorder our sampled simulation values in Figure 20 within

each origin period separately so that the Rank Matrix of Model B is replicated. Then we can aggregate across each simulation as shown in Figure 21 (differences in the total occur because of rounding).

Figure 21. Reordered simulations using Model B Rank Matrix

Sampled Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.0	19.9
2	2.5	13.3	28.0
3	1.8	16.1	24.0
4	4.4	11.3	20.0
5	4.4	8.7	26.9
6	3.0	10.7	14.0
7	4.4	10.7	16.9
8	3.9	7.6	22.6
9	3.7	13.5	25.0
10	6.4	8.6	15.0

Model B Simulation Ranks

Sim	t = 1	t = 2	t = 3
1	7	7	8
2	2	4	3
3	1	2	1
4	4	9	5
5	8	1	2
6	6	8	9
7	3	10	4
8	5	5	6
9	9	3	7
10	10	6	10

Sampled Simulations: Re-ordered by Model B

Sim	t = 1	t = 2	t = 3	Total
1	3.6	10.7	16.9	31.2
2	4.4	12.0	25.0	41.4
3	6.4	13.5	28.0	47.8
4	4.4	8.6	22.6	35.6
5	3.0	16.1	26.9	46.1
6	3.7	8.7	15.0	27.5
7	4.4	7.6	24.0	35.9
8	3.9	11.3	20.0	35.2
9	2.5	13.3	19.9	35.7
10	1.8	10.7	14.0	26.5

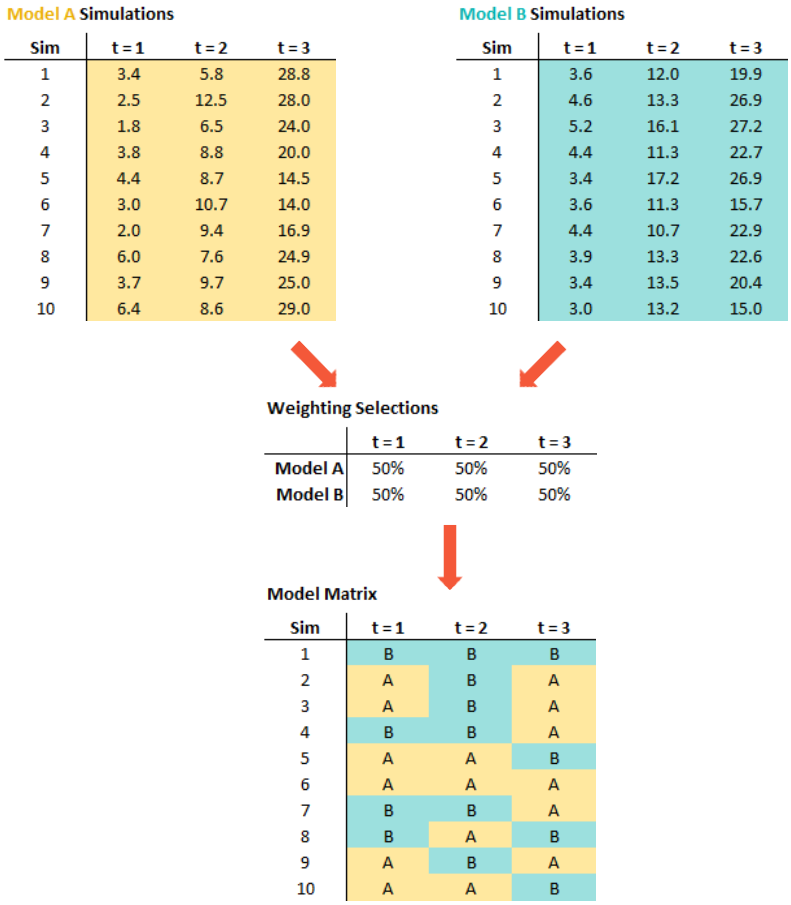
Note that the resulting reordered simulations are not color-coded because the link to the Model Matrix no longer exists.

The Rank Tying approach is a means to combine the simulations across origin periods while maintaining the same parameter variance dependency structure associated with one of the underlying projection models. In essence, this approach assumes that the introduction of model uncertainty does not produce any additional dependency across origin periods.

6.4 Model Tying

The Model Tying approach attempts to incorporate dependencies associated with model error into the aggregate estimate. In order to accomplish this, we will need to revisit the case study in Section 6.1 and revert to the step where the Model Matrix was created in Figure 16. The Model Matrix in Figure 16 and underlying model simulations in Figure 15 are summarized in Figure 22.

Figure 22. Multiple prediction Model Matrix



Model A Simulations

Sim	t = 1	t = 2	t = 3
1	3.4	5.8	28.8
2	2.5	12.5	28.0
3	1.8	6.5	24.0
4	3.8	8.8	20.0
5	4.4	8.7	14.5
6	3.0	10.7	14.0
7	2.0	9.4	16.9
8	6.0	7.6	24.9
9	3.7	9.7	25.0
10	6.4	8.6	29.0

Model B Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.0	19.9
2	4.6	13.3	26.9
3	5.2	16.1	27.2
4	4.4	11.3	22.7
5	3.4	17.2	26.9
6	3.6	11.3	15.7
7	4.4	10.7	22.9
8	3.9	13.3	22.6
9	3.4	13.5	20.4
10	3.0	13.2	15.0

Weighting Selections

	t = 1	t = 2	t = 3
Model A	50%	50%	50%
Model B	50%	50%	50%

Model Matrix

Sim	t = 1	t = 2	t = 3
1	B	B	B
2	A	B	A
3	A	B	A
4	B	B	A
5	A	A	B
6	A	A	A
7	B	B	A
8	B	A	B
9	A	B	A
10	A	A	B

Under the Model Tying approach, we will rearrange the Model Matrix with the goal of maximizing the degree to which the same model is selected across as many origin periods as possible within a given simulation. In this specific example, we want to maximize the degree to which 'A's in one origin period are grouped with 'A's in other origin periods, and the degree to which 'B's are grouped with 'B's. The resulting reordered Model Matrix might look like the example in Figure 23.

Figure 23. Reordered Model Matrix

Model Matrix

Sim	t = 1	t = 2	t = 3
1	B	B	B
2	A	B	A
3	A	B	A
4	B	B	A
5	A	A	B
6	A	A	A
7	B	B	A
8	B	A	B
9	A	B	A
10	A	A	B

Model Matrix: Reordered

Sim	t = 1	t = 2	t = 3
1	A	B	A
2	A	A	A
3	B	B	B
4	A	A	A
5	A	A	A
6	B	B	B
7	A	A	A
8	B	B	B
9	A	B	A
10	B	B	B

Note that sampling error in this example means that we do not achieve an exact 50/50 split reflecting the weights chosen in each year between Model A and Model B so 'perfect strings' are not possible for all simulations.

With the reordered Model Matrix, we are now ready to select the value corresponding to the simulation number, model and origin period to derive our values for each origin period as shown in Figure 24. Also, the total can be derived by aggregating across each simulation (differences in the total occur because of rounding). It should be noted that the resulting distributions for each origin period from this approach should produce similar results to the distributions derived from weighted sampling because the reordered Model Matrix maintains the exact same weighting between the models.

Figure 24. Model Tying simulations

Model Matrix: Reordered

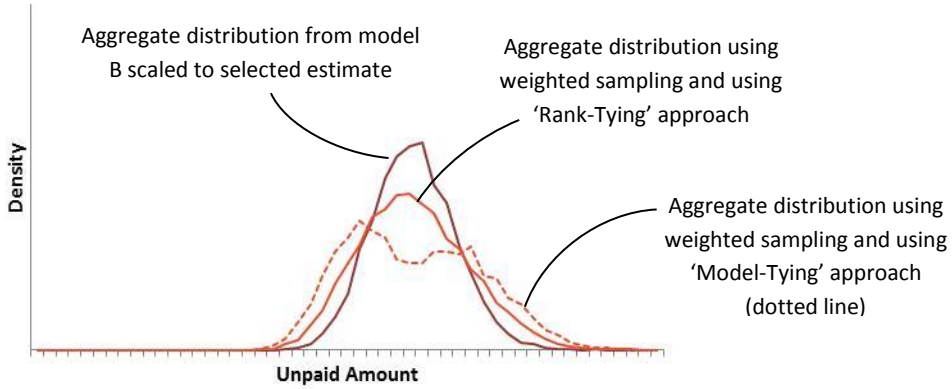
Sim	t = 1	t = 2	t = 3
1	A	B	A
2	A	A	A
3	B	B	B
4	A	A	A
5	A	A	A
6	B	B	B
7	A	A	A
8	B	B	B
9	A	B	A
10	B	B	B

Sampled Simulations: Using reordered Model Matrix

Sim	t = 1	t = 2	t = 3	Total
1	3.4	12.0	28.8	44.2
2	2.5	12.5	28.0	43.0
3	5.2	16.1	27.2	48.6
4	3.8	8.8	20.0	32.7
5	4.4	8.7	14.5	27.6
6	3.6	11.3	15.7	30.6
7	2.0	9.4	16.9	28.2
8	3.9	13.3	22.6	39.7
9	3.7	13.5	25.0	42.2
10	3.0	13.2	15.0	31.3

Figure 25 shows the resulting aggregate distribution for all three origin periods combined resulting from Model Tying, Rank Tying to Model B's dependency structure and scaling the distribution (multiplicatively) around Model B to the selected central estimate when the number of simulations in this example is increased to 10,000. All three approaches have the same mean value, which is equal to the actuarial selected central estimate for all three origin periods combined.

Figure 25. Aggregating multiple predictions: Model Tying versus Rank Tying to Model B



The difference between Model Tying and Rank Tying occurs only in the aggregate results. Rank Tying uses the parameter variance dependency attributable to only one of the models whereas Model Tying incorporates parameter variance dependencies from all models in accordance with their weights. Rank Tying excludes origin period dependencies associated with model error whereas Model Tying incorporates origin period dependency associated with model error.

6.5 Aggregation Considerations

A few points about using the Rank Tying or Model Tying approaches are noteworthy.

6.5.1 Broken Strings

With respect to the Model Tying approach, a broken string refers to a Model Matrix simulation where the same model is not identified for all origin periods. Examples of broken strings and perfect strings are shown in Figure 26.

Figure 26. Broken strings versus perfect strings

Model Matrix: Reordered

Sim	t = 1	t = 2	t = 3	
1	A	B	A	← 'Broken' string
2	A	A	A	
3	B	B	B	
4	A	A	A	
5	A	A	A	
6	B	B	B	
7	A	A	A	
8	B	B	B	
9	A	B	A	← 'Broken' string
10	B	B	B	

Broken strings can occur because of sample error as demonstrated in the previous example or because of the particular weighting attributed to the various models by origin period. A broken string is noteworthy for two reasons. First, a broken string raises the question of how to address parameter

variance dependency since values are being pulled from different models within that particular simulation. One solution is to pre-sort the simulations within each model in ascending order by some measure, such as the total unpaid claim estimate across all origin periods, before applying the Model Matrix. The result will be an approximate Rank Tying of parameter variance dependency between models.

Second, a broken string implies that a dependency associated with model error does not run throughout all origin periods in that particular simulation. This should be considered a desirable effect if the broken string was caused by the particular weighting chosen for each model and origin period.

6.5.2 Increasing Complexity

The example used for Rank Tying and Model Tying was simplistic in that it used only two models, three origin periods and equal weights across all origin periods. The Rank Tying and Model Tying approaches are scalable to multiple models, an increased number of origin periods and varying weights across origin periods, however, some considerations are worth noting.

As mentioned previously, Rank Tying superimposes the parameter variance dependency structure from a single model. As the number of models is increased the relevance of any single parameter variance dependency structure is diminished accordingly. If Rank Tying is used, preference for the selected parameter variance dependency structure should be given to one of the models that contribute to the largest proportion of the total unpaid claim estimate.

Increasing the number of models and origin periods and varying the weights with Model Tying may result in broken strings and a situation where there are multiple solutions for the Model Matrix. Weightings among models should be sensible such that broken strings produce a desirable effect on the resulting distribution. An example of a desirable effect is if the actuary believes that a particular model is appropriate and hence given weight in the actuarial central estimate for only a subset of origin periods. As a result, a perfect string will not exist across all origin periods if the weight for some origin periods is zero.

With regards to multiple solutions for the Model Matrix, consider the following example in Figure 27 where we have three models used to estimate three origin periods:

Figure 27. Multiple prediction model simulations

Model A Simulations

Sim	t = 1	t = 2	t = 3
1	3.4	5.8	28.8
2	2.5	12.5	28.0
3	1.8	6.5	24.0
4	3.8	8.8	20.0
5	4.4	8.7	14.5
6	3.0	10.7	14.0
7	2.0	9.4	16.9
8	6.0	7.6	24.9
9	3.7	9.7	25.0
10	6.4	8.6	29.0

Model B Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.0	19.9
2	4.6	13.3	26.9
3	5.2	16.1	27.2
4	4.4	11.3	22.7
5	3.4	17.2	26.9
6	3.6	11.3	15.7
7	4.4	10.7	22.9
8	3.9	13.3	22.6
9	3.4	13.5	20.4
10	3.0	13.2	15.0

Model C Simulations

Sim	t = 1	t = 2	t = 3
1	3.6	12.5	19.4
2	4.6	14.1	26.2
3	5.3	17.3	26.5
4	4.4	11.7	22.1
5	3.4	18.5	26.2
6	3.6	11.7	15.4
7	4.5	11.1	22.3
8	3.9	14.0	22.0
9	3.4	14.2	19.9
10	3.0	13.9	14.7

We can, again, create a Model Matrix, shown in figure 28, based on the selected weights from each of the Models A, B and C across 10 simulations:

Figure 28. Multiple predictions Model Matrix

Weighting Selections

	t = 1	t = 2	t = 3
Model A	33%	33%	33%
Model B	33%	33%	33%
Model C	33%	33%	33%

Model Matrix

Sim	t = 1	t = 2	t = 3
1	B	C	B
2	C	B	A
3	A	A	A
4	C	C	C
5	A	A	A
6	B	C	B
7	A	A	B
8	B	B	C
9	A	C	B
10	C	B	C

Under the Model Tying approach, we rearrange the Model Matrix with the goal of maximizing the degree to which the same model is selected across as many origin periods as possible within a given simulation. Two unique solutions exist and are shown in Figure 29:

Figure 29. Multiple solutions

Model Matrix

Sim	t = 1	t = 2	t = 3
1	B	C	B
2	C	B	A
3	A	A	A
4	C	C	C
5	A	A	A
6	B	C	B
7	A	A	B
8	B	B	C
9	A	C	B
10	C	B	C

Model Matrix: Re-ordered (solution 1)

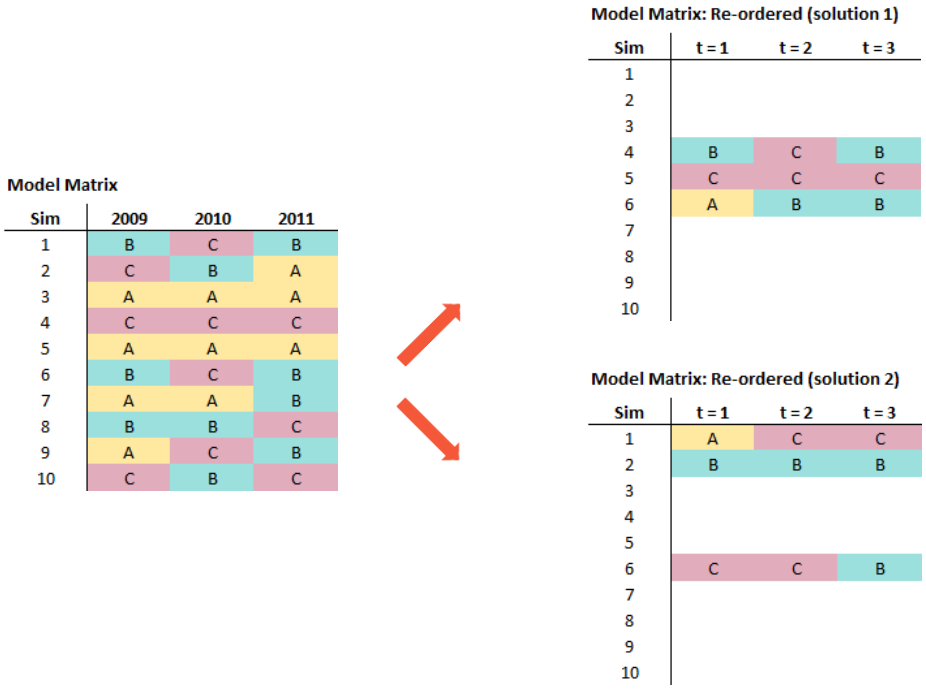
Sim	t = 1	t = 2	t = 3
1	C	C	C
2	C	C	C
3	A	A	A
4	B	C	B
5	C	C	C
6	A	B	B
7	B	B	B
8	A	A	A
9	A	A	A
10	B	B	B

Model Matrix: Re-ordered (solution 2)

Sim	t = 1	t = 2	t = 3
1	A	C	C
2	B	B	B
3	A	A	A
4	C	C	C
5	A	A	A
6	C	C	B
7	C	C	C
8	A	A	A
9	B	B	B
10	B	B	B

Removing common strings in Figure 30 helps identify the differences:

Figure 30. Isolated differences



Although both solutions maximize origin period dependency as measured on the Model Matrix, the origin period dependency measured on the sampled simulations (i.e. values) between both solutions may differ and the preferred solution may depend on the circumstances.

6.5.3 Effects on MSE

It is difficult to make blanket statements about the impact between Rank Tying and Model Tying approaches on the overall variance of aggregate origin period predictions because it will depend on each unique situation. With regards to model error, the dependency assumed in Model Tying will generally increase the aggregate variance as compared to Rank Tying in situations where the predictions of the underlying models diverge in the same direction relative to the actuarial central estimate across origin periods. However, model error dependency assumed in Model Tying can reduce the aggregate variance in situations where the predictions of the underlying models fluctuate between being greater and less than the actuarial central estimate across origin periods.

With regards to parameter variance, the dependency assumed in Rank Tying is unaffected by the complexity in the number of models, origin periods and weights, and the dependency structure selected may be different from the dependency structures observed in other models. On the other hand, parameter variance dependency structures across models will be averaged under Model Tying and their effect may be diminished as the complexity of the approach increases.

7 Summary

It has been shown that the uncertainty in a prediction, as defined by the mean squared error, is comprised of the sum of three components: process variance, parameter variance and squared bias. Suitable approaches exist in the literature to measure these components and its corresponding distribution when a single model is considered in isolation. When multiple models are considered reasonable indicators of unpaid claims, it may be appropriate to incorporate model uncertainty into the actuary's distribution of uncertainty. Various approaches for incorporating model uncertainty were introduced. The first approach, called weighted sampling, is an approach that can be used to incorporate model uncertainty into a single prediction. Rank Tying and Model Tying are approaches that can be used to incorporate model uncertainty into an aggregation of multiple predictions that exhibit dependencies in either parameter or model uncertainty. These approaches are somewhat more complex to apply but are nevertheless important to consider when measuring the aggregate uncertainty of multiple predictions.

References

Actuarial Standards Board, *Actuarial Standard of Practice No. 43 Property/Casualty Unpaid Claim Estimates*, June 2007, Updated for Deviation Language Effective May, 1, 2011, Doc. No. 159, http://www.actuarialstandardsboard.org/pdf/asops/asop043_159.pdf

England, P. D., and Verrall, R. J., *Analytic and bootstrap estimates of prediction errors in claims reserving*, Insurance: Mathematics and Economics, 1999, 25, pp. 281-293

England, P. D., *Addendum to Analytic and bootstrap estimates of prediction errors in claims reserving*, 2001, Actuarial Research Paper No. 138a, Department of Actuarial Science and Statistics, City University, London, EC1V 0HB

England, P. D., and Verrall, R. J., *Stochastic Claims Reserving in General Insurance*, British Actuarial Journal, 8, III, 2002, pp. 443-544, http://www.actuaries.org.uk/system/files/documents/pdf/sm0201.pdf

England, P.D., and Verrall, R. J., *Predictive Distributions of Outstanding Liabilities in General Insurance*, Annals of Actuarial Science, 2006, Vol. 1, No. 2, pp. 221-270, http://cassknowledge.co.uk/sites/default/files/article-attachments/371~~richardverrall_-_predictive_distributions_of_general_insurance_outstanding_liabilities.pdf

Mack, T, *Measuring the Variability of Chain Ladder Reserve Estimates*, Casualty Actuarial Society Forum, Spring 1994, pp. 102-182, http://www.casact.org/pubs/forum/94spforum/94spf101.pdf

Appendix A

Excerpts of the following case study are used throughout this paper. In this appendix we will discuss the complete case study and will highlight relevant sections corresponding to the Figures displayed in the body of the paper.

Overview of data and selections

- This case study is based on data spanning a nine year history of origin periods, where an origin period represents an accident year.
- Development factor models (i.e. chain ladder models) were applied to each of the paid ('model A') and incurred ('model B') data in order to project to ultimate.
- A 'central estimate' was selected based on a simple average of the two development factor models for each accident year.
- Distributions reflecting process and parameter variance for each model were achieved using stochastic methods. The type of stochastic methods used is irrelevant for this illustration, but in this instance a 'practical stochastic' method was applied to Model A and a Bootstrapping approach to Model B. 'Practical stochastic' in this instance is used to describe a process whereby the analyst generates samples from a selected distribution with a user-defined mean and coefficient of variation.

For the purpose of this case study we are going to concentrate on results for just the three most recent accident years, however, any totals shown will represent the cumulative results of the full nine years of accident period history (rounding may occur with totals).

Central Estimate

The table in Figure A.1 summarizes the point estimates produced by each model for 'prior' years (1997 – 2008), 2009, 2010 and 2011 accident years, alongside the weighting used to determine the selected central estimate and the resulting amount of that estimate.

Figure A.1. Selected central estimates

	Model A	Weight	Model B	Weight	Selected Central Estimate
Prior	$2,784	50%	$8,783	50%	$5,783
2009	2,774	50%	3,838	50%	3,306
2010	8,275	50%	12,871	50%	10,573
2011	19,114	50%	23,534	50%	21,324
Total	$32,947		$49,026		$40,987

Implicit in the equal weightings used in this case study is the assumption that each model is an equally reliable predictor of the final outcome. The challenge is to estimate the corresponding uncertainty around this prediction that adequately reflects this inherent assumption.

Distributions conditional on each model

We begin the process of estimating uncertainty by developing distributions around each of the underlying models that reflect both process and parameter variance. The table in Figure A.2 summarizes the results of the stochastic uncertainty analyses performed around each of the underlying models in terms of the prediction error ("Pred. Error", $000s) of the resulting distributions as well as the coefficient of variation ("CV", prediction error as a percentage of the mean), for the most recent three accident years and in total.

Figure A.2. Summary of uncertainty conditional on each model

Model A

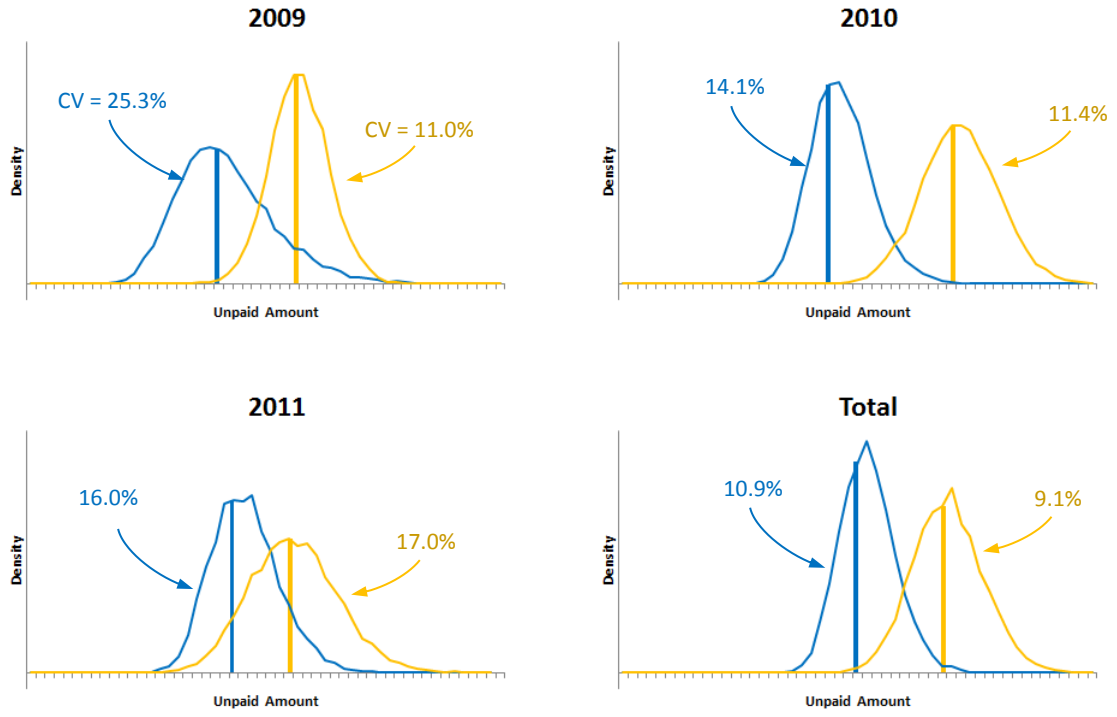
	Mean	Pred. Error	CV
2009	$2,774	$702	25.3%
2010	8,275	1,167	14.1%
2011	19,114	3,058	16.0%
Total	$32,947	$3,595	10.9%

Model B

	Mean	Pred. Error	CV
2009	$3,838	$423	11.0%
2010	12,871	1,465	11.4%
2011	23,534	3,995	17.0%
Total	$49,026	$4,441	9.1%

These distributions are also shown graphically in Figure A.3 along with the means (represented by the vertical bar) and corresponding CV's from each model (blue line is Model A, yellow line is Model B).

Figure A.3. Distributions around Model A and Model B



It should be noted that the distributions for each origin period and in total are not generated independently but rather collectively as a single process defined by the stochastic methods. As a result, origin period dependencies exist and can be measured. As a precursor for what is to come, each origin period can be treated as a 'single period prediction' as discussed in the paper through weighted sampling, however, the intrinsic origin period dependencies created by these stochastic methods will be broken. Rank Tying and Model Tying are options to restoring some sort of origin period dependency in order to recreate a 'total' aggregate distribution.

Distribution around selected central estimate using scaling

Once we have generated our distributions reflecting process and parameter uncertainty for each of the underlying models, we are faced with the challenge of producing a distribution around our selected central estimate.

One commonly-used approach is to select an underlying model and scale the associated simulated output from that model in an appropriate manner (see Section 2, Scaling).

In this example, we might select underlying Model B as our preferred model and choose multiplicative scaling to generate a distribution of simulated outcomes with a mean equal to our selected central estimate.

Figure A.4 summarizes the statistical properties of our distribution around our selected central estimate derived by multiplicatively scaling the simulations from Model B. Again, we show the prediction error ("Pred. Error", $000s) of the resulting distribution as well as the coefficient of variation ("CV", prediction error as a percentage of the mean), for the most recent three accident years and in total.

Figure A.4. Summary of uncertainty for selected central estimate using scaling

Uncertainty Summary: Selected - Scaling ($000s)

	Mean	Pred. Error	CV
2009	$3,306	$364	11.0%
2010	10,573	1,203	11.4%
2011	21,324	3,620	17.0%
Total	$40,987	$3,958	9.7%

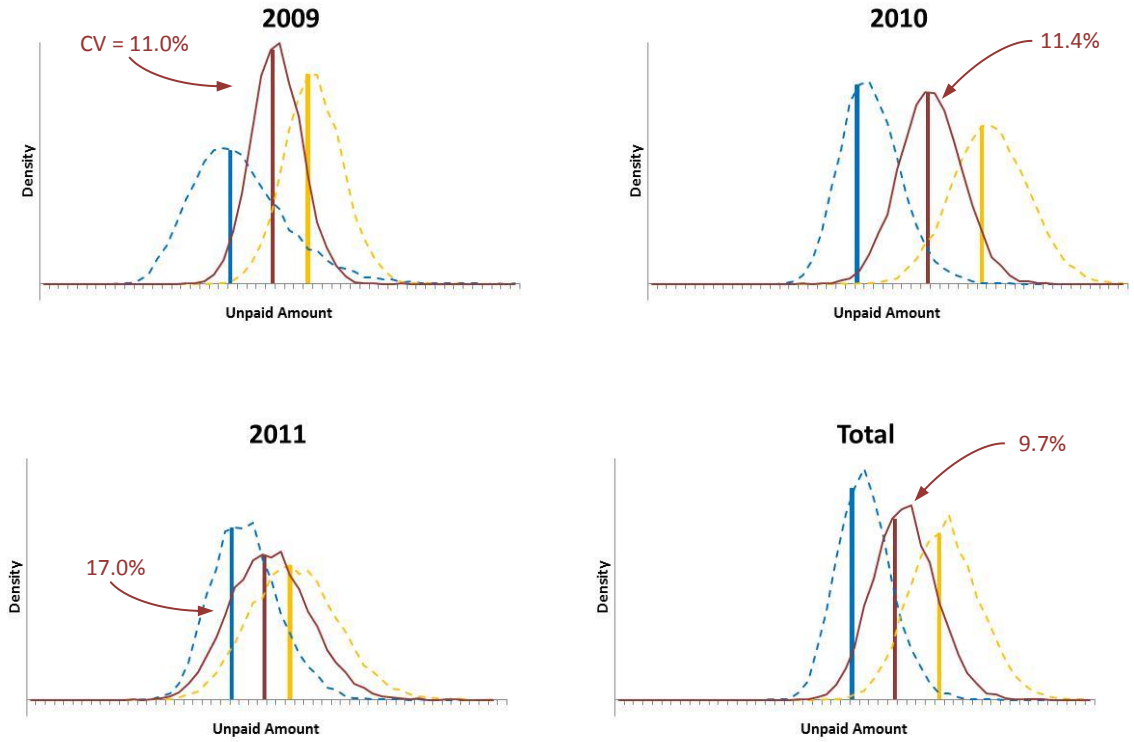
Note that, because we selected to use multiplicative scaling, the mean of the distribution is equal to our selected central point estimate and the coefficients of variation for each accident year are equivalent to the corresponding measure from the distribution developed around Model B.

Had we selected to scale additively, the mean of our distributions would still align with our selected central estimate but the coefficient of variation for each accident year would change when compared to Model B. Under additive scaling, the prediction error for each accident year remains equivalent instead of the CV.

Note also that the 'Total' coefficient of variation from multiplicative scaling is not equivalent to the 'Total' coefficient of variation from Model B. This is due to differences in the magnitude of scaling for each year.

Figure A.5 shows these scaled distributions for each accident year and in total. The selected mean and the scaled distributions are shown as solid green lines, and the distributions and means from our underlying models are shown as blue (Model A) and yellow (Model b) broken lines.

Figure A.5. Distributions using scaling

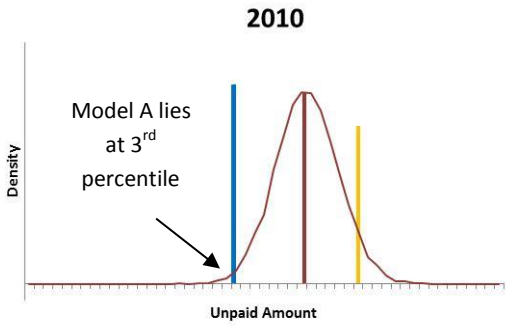


It should be noted that the graph shown in Figure A.5 for 2009 is similar to the graph shown in Figure 3 in the main text.

With regards to scaling, we are simply *borrowing* a distribution from one of our underlying models, which the actuary is forced to select. This may not adequately reflect the assumption that both models are considered to be equally valid as implied by the equal weighting used in the selection of the central estimate.

Furthermore, we may end up in a situation where our selected scaled distribution around our central estimate implies that the prediction from one of our underlying models is a relatively unlikely outcome. If we consider the 2010 accident year, our scaling approach suggests that the point estimate projected by Model A, as shown as the blue bar in Figure A.6, lies at the 3rd percentile of our range of probable outcomes.

Figure A.6. Distribution using scaling for 2010



Distribution around selected estimate using weighted sampling

We can instead employ weighted sampling for each accident year in a manner that reflects the weights selected for the determination of our selected central estimate that perhaps better represents the full distribution of possible outcomes suggested by the underlying models (see Section 5.1, Weighted Sampling).

For each accident year, we sample randomly and without replacement from each of the underlying distributions – in this case, we select 50% of the sample from the distribution around Model A and 50% from the distribution around Model B.

The table in Figure A.7 summarizes the statistical properties of our distribution around our selected central estimate derived by weighted sampling from each of the underlying models. Again, we show the prediction error ("Pred. Error", $000s) of the resulting distribution as well as the coefficient of variation ("CV", prediction error as a percentage of the mean), for the last three accident years.

Figure A.7. Summary of uncertainty using weighted sampling

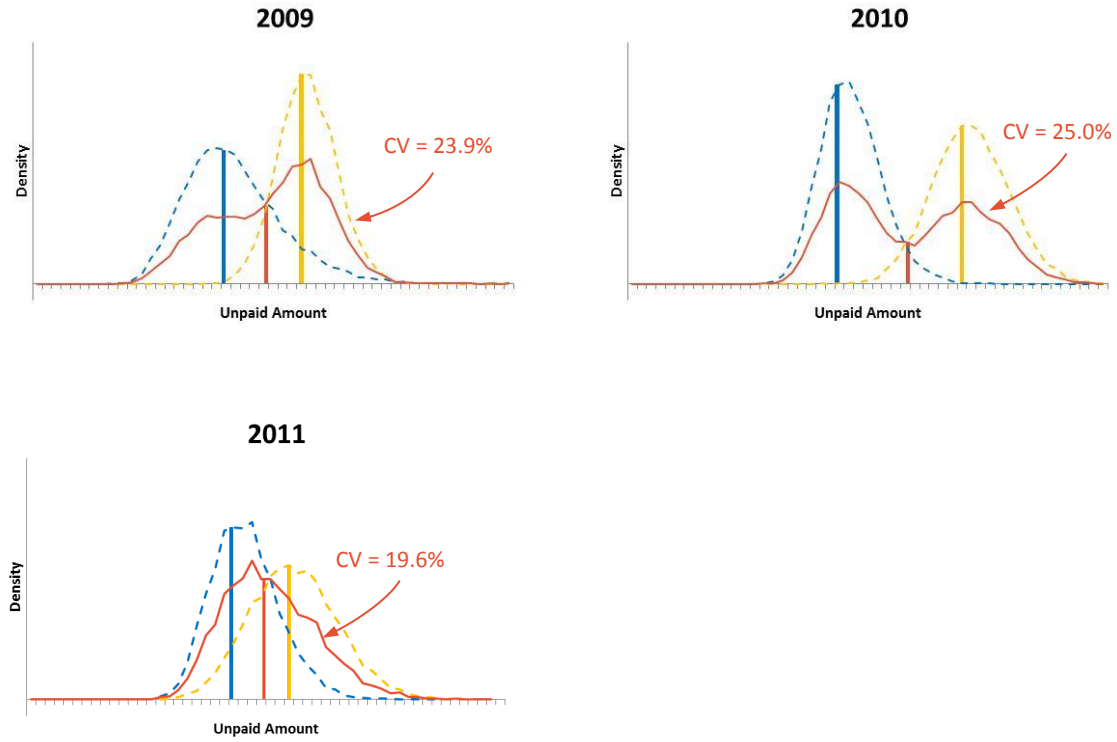
Uncertainty Summary: Selected - Wtd Sampling ($000s)

	Mean	Pred. Error	CV
2009	$3,306	$790	23.9%
2010	10,573	2,646	25.0%
2011	21,324	4,174	19.6%
Total	$40,987	?	?

As noted previously, the origin period dependencies intrinsic in the stochastic methods have been broken as a result of weighted sampling so the total aggregate distribution is no longer discernible.

The graphs in Figure A.8 show these distributions for each of the last three accident years. The selected mean and the weighted sampling distributions are shown as solid red lines, and the distributions and the means from our underlying models are again shown as broken lines.

Figure A.8. Distributions using weighted sampling



It should be noted that the graphs shown in Figure A.8 for 2009 and 2010 are similar to the graphs shown in the main text as Figures 7 and 9, respectively.

Weighted sampling will produce distributions for each accident year in isolation (as discussed for single period predictions in Section 5.1). In order to create a distribution around the selected total central estimate of unpaid claims across multiple accident years we must decide how to reintroduce an origin period dependency.

As suggested by this paper, we have the options of using either:

- Rank Tying, which reorders the year-by-year simulations such that a pre-defined accident-year correlation is targeted (as discussed in Section 6.3); or
- Model Tying, which uses a Model Matrix designed in such a manner to maximize the degree to which the same model is selected across as many different accident years as possible within a given simulation (as discussed in Section 6.4)

If using Rank Tying, the analyst should produce the Rank Matrix that is to be used to reorder the simulation. In this example, we have selected to use the Rank Matrix from the simulated distribution around Model B.

The tables in Figure A.9 summarize the point estimates and statistical properties of our distribution around each of:

- Model A;
- Model B;
- Selected central estimate using multiplicative scaled simulations from Model B;
- Selected central estimate using weighted sampling and Rank Tying accident years according to the correlation matrix suggested by Model B; and
- Selected central estimate using weighted sampling and optimized Model Tying.

Figure A.9. Summary comparing uncertainty from various models

Point Estimate Selection Summary ($000s)

	Model A	Model B	Selected: Scaled	Selected: Wtd Sample (Rank Tying)	Selected: Wtd Sample (Model Tying)
2009	$2,774	$3,838	$3,306	$3,306	$3,306
2010	8,275	12,871	10,573	10,573	10,573
2011	19,114	23,534	21,324	21,324	21,324
Total	$32,947	$49,026	$40,987	$40,987	$40,987

Uncertainty Summary: Comparison (Prediction Error)

	Model A	Model B	Selected: Scaled	Selected: Wtd Sample (Rank Tying)	Selected: Wtd Sample (Model Tying)
2009	$702	$423	$364	$790	$785
2010	1,167	1,465	1,203	2,646	2,664
2011	3,058	3,995	3,620	4,174	4,187
Total	$3,595	$4,441	$3,958	$5,854	$8,973

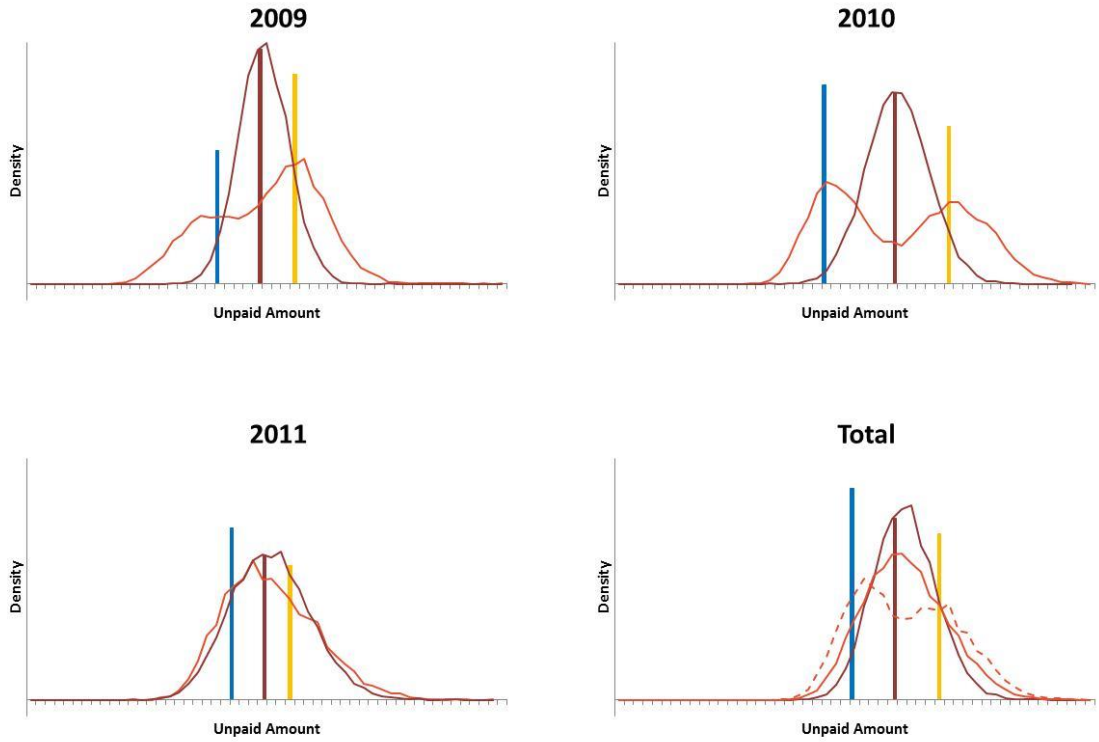
Uncertainty Summary: Comparison (CVs)

	Model A	Model B	Selected: Scaled	Selected: Wtd Sample (Rank Tying)	Selected: Wtd Sample (Model Tying)
2009	25.3%	11.0%	11.0%	23.9%	23.9%
2010	14.1%	11.4%	11.4%	25.0%	25.0%
2011	16.0%	17.0%	17.0%	19.6%	19.6%
Total	10.9%	9.1%	9.7%	14.3%	21.9%

As before, graphs assist in the interpretation and comparison of these results and the associated distributions. Such graphs corresponding to Figure A.9 can be viewed in Figure A.10. Please note:

- The blue and yellow columns represents the point estimate prediction from Models A and B
- The red column represents the selected central estimate
- The burgundy line represents the distribution around the selected central estimate using multiplicative scaling
- The red line represents the distribution around the selected central estimate using weighted sampling
- In the 'Total' graph, the distribution is shown around the total aggregate point estimate using:
    - Rank Tying (solid red line)
    - Model Tying (broken red line)
    - Scaling (solid burgundy line)

Figure A.10. Comparison of distributions using weighted sampling and scaling



Appendix B

In statistics, the Mean Squared Error (MSE) measures the difference between an estimate and what the true value is. Consider a random variable, y and a predicted variable, $\hat{y}$. The mean squared error (MSE) is:

$$E[(y - \hat{y})^2]$$

Expanding this term through additive properties yields:

$$E[(y - \hat{y})^2] = E[(y - \hat{y} + (E[y] - E[y]) + (E[\hat{y}] - E[\hat{y}]))^2]$$

Reordering yields

$$= E\left[\left((y - E[y]) - (\hat{y} - E[\hat{y}]) + E[y] - E[\hat{y}]\right)^2\right]$$

A series of expanding terms and subsequent simplification yields,

$$\begin{aligned} = E\big[&(y - E[y])^2 - (y - E[y])(\hat{y} - E[\hat{y}]) + E[y](y - E[y]) - E[\hat{y}](y - E[y]) + (\hat{y} - E[\hat{y}])^2 \\ &- (y - E[y])(\hat{y} - E[\hat{y}]) - E[y](\hat{y} - E[\hat{y}]) + E[\hat{y}](\hat{y} - E[\hat{y}]) + E[y]^2 \\ &+ E[y](y - E[y]) - E[y](\hat{y} - E[\hat{y}]) - E[y]E[\hat{y}] + E[\hat{y}]^2 - E[\hat{y}](y - E[y]) \\ &+ E[\hat{y}](\hat{y} - E[\hat{y}]) - E[y]E[\hat{y}]\big] \end{aligned}$$

$$\begin{aligned} = E[&(y - E[y])^2 - 2(y - E[y])(\hat{y} - E[\hat{y}]) + 2E[y](y - E[y]) - 2E[\hat{y}](y - E[y]) + (\hat{y} - E[\hat{y}])^2 \\ &- 2E[y](\hat{y} - E[\hat{y}]) + 2E[\hat{y}](\hat{y} - E[\hat{y}]) + E[y]^2 - 2E[y]E[\hat{y}] + E[\hat{y}]^2] \end{aligned}$$

$$\begin{aligned} = E[&(y - E[y])^2 - 2y\hat{y} + 2yE[\hat{y}] + 2\hat{y}E[y] - 2E[y]E[\hat{y}] + 2yE[y] - 2E[y]^2 - 2yE[\hat{y}] \\ &+ 2E[\hat{y}]E[y] + (\hat{y} - E[\hat{y}])^2 - 2\hat{y}E[y] + 2E[y]E[\hat{y}] + 2\hat{y}E[\hat{y}] - 2E[\hat{y}]^2 + E[y]^2 \\ &- 2E[y]E[\hat{y}] + E[\hat{y}]^2] \end{aligned}$$

$$= E[(y - E[y])^2 - 2y\hat{y} + 2yE[y] - E[y]^2 + (\hat{y} - E[\hat{y}])^2 + 2\hat{y}E[\hat{y}] - E[\hat{y}]^2]$$

$$= E[(y - E[y])^2] - E[2y\hat{y}] + E\big[2yE[y]\big] - E[E[y]^2] + E[(\hat{y} - E[\hat{y}])^2] + E\big[2\hat{y}E[\hat{y}]\big] - E[E[\hat{y}]^2]$$

$$= E[(y - E[y])^2] - 2E[y\hat{y}] + 2E\big[yE[y]\big] - E[E[y]^2] + E[(\hat{y} - E[\hat{y}])^2] + 2E\big[\hat{y}E[\hat{y}]\big] - E[E[\hat{y}]^2]$$

$$= E[(y - E[y])^2] - 2E[y\hat{y}] + 2E[y]E[y] - E[y]^2 + E[(\hat{y} - E[\hat{y}])^2] + 2E[\hat{y}]E[\hat{y}] - E[\hat{y}]^2$$

$$= E[(y - E[y])^2] - 2E[y\hat{y}] + 2E[y]^2 - E[y]^2 + E[(\hat{y} - E[\hat{y}])^2] + 2E[\hat{y}]^2 - E[\hat{y}]^2$$

$$= E[(y - E[y])^2] - 2E[y\hat{y}] + E[y]^2 + E[(\hat{y} - E[\hat{y}])^2] + E[\hat{y}]^2$$

If we assume y and $\hat{y}$ are independent, then $E[y\hat{y}] = E[y]E[\hat{y}]$ and

$$= E[(y - E[y])^2] - 2E[y]E[\hat{y}] + E[y]^2 + E[(\hat{y} - E[\hat{y}])^2] + E[\hat{y}]^2$$

Reordering yields,

$$= E[(y - E[y])^2] + E[(\hat{y} - E[\hat{y}])^2] + E[y]^2 - 2E[y]E[\hat{y}] + E[\hat{y}]^2$$

which simplifies to,

$$= E[(y - E[y])^2] + E[(\hat{y} - E[\hat{y}])^2] + (E[y] - E[\hat{y}])^2$$