Testing the Assumptions of Assumptions Testing

Keith Curley, FCAS, MAAA

Abstract

Motivation. The growing availability and advocacy of stochastic reserving methods is outpacing their critical evaluation, testing, and indeed acceptance. I believe there has not yet been sufficient critical attention given to claims made in favor of stochastic models, and I'll focus here on the particular claims that assumptions can be tested and the models validated.

Method. We'll review some of the statistical background, especially hypothesis testing, needed to understand the issues and see how it applies to reserve modeling with aggregate loss triangles. We'll make use of the concept of statistical power, associated with Type II error, which has been previously absent from reserve modeling discussions. This concept can be used to question the reliability of modeling results and certain common modeling recommendations. A few simplified reserving models and results of simulations that help illuminate the issues are described and reported.

Results. We'll see that significance tests, and testing more generally, might have little power and recommendations based on these tests can be unwise. We'll also see the benefits of a deeper understanding of the claims process and the dangers of relying on statistical methods without that understanding.

Conclusions. This particular argument for stochastic modeling in reserving with aggregate triangles is almost certainly unsound. If this were the only reason to resort to modeling, there are more productive uses of an actuary's time. With or without modeling, better approaches probably rely on simpler methods, hard work, a skeptical and inquisitive attitude, and a deeper knowledge and understanding of the claims generation, reserving and settlement processes.

Keywords Reserving methods; reserve variability; statistical models and methods.

1. INTRODUCTION

Recent technical actuarial literature has been dominated by *advocates* of advanced quantitative techniques. This is perhaps unavoidable, but it has led to a one-sided discussion of stochastic reserve modeling.

Two particularly exaggerated claims are that modeling allows for "assumptions to be tested and models validated." I will show that this is true in only a very limited sense which is often without much practical consequence.

There is a risk that by focusing too narrowly on this one argument, which is perhaps one of the worst arguments made on behalf of modeling, I will fall into the opposite error of the advocates and be unfairly branding the whole stochastic program because of some poorly thought-out claims from its advocates: that I'll be impeding the progress of science by throwing the proverbial baby out with the bath-water while stochastic reserving is still in its infancy. All I can say is that I wish to bring greater clarity to the discussion of the merits of modeling versus traditional actuarial methods.

A fuller discussion would touch on a number of issues of which reasonable people can and

probably always will disagree. All that I ask of a reader is that he or she critically evaluate my arguments and evidence and do the same whenever claims in favor of modeling are encountered.

1.1 Research Context

It is a common, though not universal, modeling practice to screen variables for inclusion within a model by means of significance testing. A number of papers in the actuarial literature have also advocated this practice when selecting variables to model yearly aggregate loss triangles. Though not a necessarily exhaustive list, in this context the usual variables considered are: yearly exposure measures, accident year trends, calendar year trends, and development year trends, which are also called loss development factors when they are a multiplicative factor of the prior development year's losses.

For illustrative purposes and because of limitations of space, we will focus in this paper on loss development factors, but I hope it will be clear that many of the concepts explored here apply equally to significance testing for any variable.

Some authors have reported that the loss development factor is often not found to be significant in their modeling experience. According to some of the authors, in this case, such a factor should then be dropped from a final loss reserve model.

In addition to significance testing, it is common practice, especially among the more thorough modelers, to run a series of additional tests and diagnostics to check that the model assumptions are probabilistically consistent with the data. The possibility of doing such tests is often offered as a distinct advantage of the modeling framework which traditional actuarial methods do not provide. Many modeling advocates make recommendations for how to develop statistical models and methods to project reserves and study reserve variability that lean very heavily on testing results.

Sometimes even bolder claims are made on behalf of the possibility, the advisability, and the effectiveness of assumption and model testing and some have tried to draw implications for which traditional reserving methods to use because of the results of this testing.

1.2 Objective

The CAS Working Party on Quantifying Variability in Reserve Estimates ("CAS Working Party"), "The Analysis and Estimation of Loss & ALAE Variability: A Summary Report," [1] correctly state under the topic of "Model or Specification Uncertainty" (page 35): "In nearly every stochastic model, the modeling begins by making the assumption that the underlying process follows the

model."

Inspired by some of the bold claims in Section 1.1, let me elaborate on the comment immediately above and add a few bold claims of my own:

1. Not only do statistical models rely on various assumptions, it is important to always keep in mind that they are in fact *theories* about the world. These theories have implications for how the world must actually be in order for the assumptions to hold true. If the theories are true, they allow one to predict the future, at least probabilistically. If they are false, they might be a waste of time or even seriously misleading.

2. Although there exist various tests and diagnostics of the assumptions, it's unlikely that they will be *effective*--meaning that they would allow one, with a high probability of success, to correctly draw any conclusion about loss development or to pursue any action, such as using one loss development method rather than another.

3. The strongest and least plausible of modeling assumptions in insurance is that insurance data are observations of random variables that are *independent* and *identically* distributed. This is the main statistical assumption with modeling and if false all of the modeling results are compromised.

Although my claims are predominantly negative, we will also see along the way that any real information, which an actuary can discover about losses and how they are generated or develop, can be highly useful in the reserving process. I believe that traditional actuarial methods, in addition to having adequate statistical properties, bring the actuary in closer contact with the data, without the possibly distorting effects of false assumptions and without time spent on unnecessary tests.

I stake no particular claims to originality here. One can find scattered throughout the actuarial literature[1], comments from actuaries questioning whether standard statistical assumptions apply to insurance data. But in those papers it is usually a caveat which receives little attention while here it's the centerpiece of the paper. So I believe the emphasis and arrangement of ideas in this paper is somewhat unique.

The particular technical issue of statistical power which I'll discuss and which severely qualifies any claims for testing and validation effectiveness, although part of the Actuarial Exam Syllabus[2], has

[1] See, for instance, the discussion of "i.i.d." by David Clark in "LDF Curve-fitting and Stochastic Reserving: A Maximum Likelihood Approach" [2] page 56.
[2] See Stuart Klugman's Study Note *Estimation, Evaluation, and Selection of Actuarial Models* [8]

not been as far as I can tell applied to this issue before.

I owe debts to many previous actuaries, that are too numerous to name or reference here. But I owe a particular debt to the thinking of the late UC Berkeley Statistician David Freedman. The two textbooks, which I reference in this paper, and the numerous papers he has written over the years, are *models* of clear and careful statistical reasoning and how it can be applied to answer real world questions.[3]

1.3 Outline

This discussion will require a review of some basic concepts from statistics, but they are all ideas to which every actuary will have been exposed at one time or another. In particular we will review the meanings of *methods* and *models*, and the assumptions the latter usually rely on.

Unfortunately, we will have to go into some detail about statistical hypothesis testing, and this material is routinely misunderstood by both students and professionals throughout the social sciences. Apparently it is difficult for many to understand. It's possible that I might not do a great job of explaining it either. Here we discuss the importance of the *statistical power* of a test, and the consequential costs, or *loss*, associated with following the results of a test.

All I can ask is that you bear with me and be willing to think a bit. If this material is unfamiliar, I think you'll find that it's well worth learning it, and you may never look at another statistical analysis the same way again.

We will then apply these ideas to insurance reserving to see whether any claims for diagnostic effectiveness are likely to be true.

It's common in the technical actuarial literature to briefly present a model and then to elaborate at length the mathematical implications of that model. I'm not so interested in the math, but in the validity of the very first step. So we will largely travel in the other direction, and starting with models and modeling assumptions we're going to interpret them as assertions about the world and study whether those assertions are true.

We might not be able to reach any absolutely definitive conclusions, but regardless there should be some value in trying to think clearly about the relation of models to our world and the role of

[3] James M. Robins, professor of epidemiology at the Harvard School of Public Health, once wrote about David, that he was "one of the world's leading mathematical statisticians, but he has also assumed the mantle as the skeptical conscience of statistics as it is applied to important scientific, policy and legal issues." See the obituary at: http://berkeley.edu/news/media/releases/2008/10/20_freedman.shtml

those models in our work.

2. BACKGROUND AND METHODS

2.1 Statistical Models and Methods

The Working Party draws a distinction between a "Method" and a "Model." Methods are (page 38) "algorithms or series of steps followed to determine an estimate," with some examples being the "chain-ladder (development factors) method or the Bornhuetter-Ferguson method." They then add that Methods "do not involve the use of any statistical assumptions that could be used to validate reasonableness or to calculate standard error."

On the other hand, a "Model" (page 67) "specifies statistical assumptions about the loss process, usually leaving some parameters to be estimated." They also add: "There are various methods that could be used for estimating the parameters, such as maximum likelihood and various robust estimators, but unless otherwise noted, 'methods' here will refer to algorithms for calculating loss future payments, not methods for estimating model parameters."

In this paper, I will not follow that convention about *methods* only applying to reserving algorithms and not parameter estimation. In the statistical context, both are functions of random variables and hence *estimators*—random variables themselves, which one hopes will take on values close to what one is estimating. Reserving methods might not explicitly rely on stochastic assumptions, but once those assumptions are introduced into the discussion, those methods become estimators.

2.1.1 An Illuminating Example of a Method and a Model

Because there is often much confusion on this score and because it will be useful throughout the paper, we should compare the ordinary least squares (OLS) *method* to the ordinary least squares (OLS) regression *model*.

The OLS *method* is merely a way of solving a system of linear equations where there are more equations than unknowns, say n equations in p unknowns with $n>p$. In this situation, such a system is usually inconsistent and has no solution. But by taking weighted averages of all the equations with weights given by the unknowns' coefficients in the equations, one reduces the number of equations to just p equations in p unknowns, and this usually has a unique solution. The *method* is just linear

algebra.[4]

The OLS *model* is a set of statistical assumptions, for which the OLS *method* becomes a well-suited estimator of the unknowns. We will only need two unknowns in our discussion, in which case the OLS *model* assumes that (adapted from David Freedman's, *Statistical Models: Theory and Practice* [5])[5]:

1. There are two *observable* random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$; they are $n \times 1$ random vectors; there is also an $n \times 1$ random vector $\boldsymbol{\varepsilon}$ that is *not* observed and is called the *random error* or *disturbance* term; $\boldsymbol{Y}$ is a linear function of $\boldsymbol{\varepsilon}$ and $\boldsymbol{X}$ via unknowns *a* and *b*, which usually have to be estimated from the data.

2. The vector relationship above unpacks into *n* ordinary equations:

$$Y_i = a + bX_i + \varepsilon_i \tag{2.1}$$

3. A fundamental assumption is that "the data on Y_i are observed values of $a + bX_i + \varepsilon_i$." As David Freedman points out "[w]e have observed values for $\boldsymbol{X}$ and $\boldsymbol{Y}$, not the random variables themselves. We do not know [*a* and *b*] and do not observe $\boldsymbol{\varepsilon}$." Recall: *random variables* have distributions, means, standard deviations[6], etc.; *observed values* of random variables, aka *data*, are just numbers.

4. "The ε_i are independent and identically distributed, with mean 0 and variance σ^2."

5. "If $\boldsymbol{X}$ is random, we assume that $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$."

   Warning: Many applications of regression and many of the standard theorems assume that $\boldsymbol{X}$ is fixed, as it could be, for instance, in an experiment where the experimenter is able to control the value of $\boldsymbol{X}$. For us, since we will be using $\boldsymbol{X}$ to represent losses during exposure periods, $\boldsymbol{X}$ is random, not fixed, and many of the standard theorems do not apply or apply only "conditionally on $\boldsymbol{X}$ being given."[7]

As mentioned previously the OLS *method* with 2 unknowns will reduce a system of *n* equations to just 2 equations, by taking a weighted average of all the equations with weights equal to the coefficients of the unknowns in the model. So let's suppose that we have *n* pairs of observed values (y_1,x_1), (y_2,x_2),…, (y_n,x_n).

This gives us *n* linear equations:

---

[4] See, for instance, Gilbert Strang, *Linear Algebra and Its Applications,* 3rd Edition.
[5] Since we're focusing on only two unknowns we don't present the whole matrix formulation; and we leave out mention of certain niceties like the rank of our system of equations being at least 2 which will almost always be the case for us.
[6] Theoretically not all random variables have moments; but any which appear in insurance probably will.
[7] The parameter estimators tend to be *unconditionally* unbiased as well, but not necessarily the standard error estimators.

$$\begin{aligned} y_1 &= a + bx_1 \\ y_2 &= a + bx_2 \\ &\dots \\ y_n &= a + bx_n \end{aligned} \tag{2.2}$$

The coefficient of the first unknown *a* is always 1, so the first reduced equation will simply be an average of all *n* equations, where we now also employ the hat "^" over *a* and *b*, to represent that they are not the same *a* and *b* as above (in fact they cannot be the same because the above is inconsistent):

$$\bar{y} = \hat{a} + \hat{b}\bar{x}, \tag{2.3}$$

Barred variables, for instance, $\bar{y}$ and $\bar{x}$, will just indicate the averages of the data series.

We get our second equation by multiplying through on both sides of the *i*th equation by x_i and averaging those equations:

$$\begin{aligned} x_1 y_1 &= a x_1 + b x_1 x_1 \\ x_2 y_2 &= a x_2 + b x_2 x_2 \\ &\dots \\ x_n y_n &= a x_n + b x_n x_n \end{aligned} \tag{2.4}$$

$$\overline{xy} = \hat{a}\bar{x} + \hat{b}\overline{x^2} \tag{2.5}$$

Now (2.3) and (2.5) gives us two equations in two unknowns, $\hat{a}$ and $\hat{b}$, and those equations almost always have unique solutions:

$$\hat{a} = \frac{\overline{x^2}\bar{y} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2} \tag{2.6}$$

$$\hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \tag{2.7}$$

If we're going to use our estimate to, for instance, project the *n+1*-th observation, a little creative algebra and we can write:

$$y_{n+1} = \hat{a} + \hat{b}x_{n+1} \tag{2.8}$$

As:

$$y_{n+1} = (1 - Z)\bar{y} + Z\frac{\bar{y}}{\bar{x}}x_{n+1}, \tag{2.9}$$

$$\text{where } Z = \frac{\bar{x}}{\bar{y}}\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2},$$

which can also be rewritten as

$$Z = \rho \frac{CV_y}{CV_x},$$

where ρ is the correlation between the two data series x and y, and the CV′s are their coefficients of variation, i. e. , their standard deviations over their means.

Or, in words, as many actuaries before have noticed: if x_i represents losses from the *i*th exposure period, say accident year, at some evaluation age, and y_i represents the incremental losses at the next evaluation age, then the OLS projection y_{n+1} for the losses from the latest diagonal x_{n+1}, is just a weighted average of the standard chain ladder estimate, and the overall mean $\bar{y}$. One might replace incremental loss with cumulative loss or ultimate loss. But, regardless, actuarial methods which are now standard, such as Bornhuetter-Ferguson, Stanard-Bühlmann, and Benktander, can be viewed as methods for estimating the various parameters such as Z and $\bar{y}$ which appear in equation 2.9.

Now, in order to connect this OLS *method* with the OLS *model,* we have to invoke assumption 3, which was that "the data on Y_i are observed values of$a + bX_i + \varepsilon_i$." If we make this connection, replacing our equations in $\boldsymbol{x}$ and $\boldsymbol{y}$ above with equations in $\boldsymbol{X}$ and $\boldsymbol{Y}$, then $\hat{a}$ and $\hat{b}$ are now *estimators*, i.e. functions of random variables, and hence random variables themselves with their own distributions.

If all the other assumptions on page 6 are true as well, then we are allowed to conclude that $\hat{a}$ and $\hat{b}$ are conditionally *unbiased* estimators, meaning their expected values, conditional on a given $\boldsymbol{X}$, are equal to the unknown parameters *a* and *b*. We can also calculate their variances and correlations with each other, conditional on a given $\boldsymbol{X}$, if we know the variance of the error term σ^2; and if that's unknown we have a conditionally unbiased estimator available for it as well. Finally, conditional on $\boldsymbol{X}$, we can show that the squared difference between $\boldsymbol{Y}$ and the linear combination of $\boldsymbol{X}$ is minimized when $a = \hat{a}$ and $b = \hat{b}$, which is of course the origin of the *ordinary least squares* method and model's name.

I apologize if all of the above is old "hat," but I think it's important that we keep in mind the difference between a method which can be applied regardless of whether any statistical assumptions are true, and a model for which that method might have some nice properties when regarded as an estimator.

For instance, if in the equations (2.2) we had dropped the *a*'s and only considered *y* as a function of *x*, i.e., taken only:

$$y_i = bx_i \tag{2.10}$$

Then there are many ways we might estimate *b*. We can simply average all the equations and divide, which would give us the standard chain-ladder over all the years, or we can take the last *m* equations for any $1 \leq m$ and average them, as we might do in practice if those were more representative years.[8] More generally, we could apply any weighted average whatsoever to the *n* equations to reduce them to 1 equation in order to estimate *b*.

And referring back to the model assumptions, where now there is no *a*:

$$Y_i = bX_i + \varepsilon_i \tag{2.11}$$

As long as the ε_i have mean 0 and the weights in combination with the X_i are independent of ε_i then the resulting estimator will be an unbiased estimator of *b*.[9]

2.2 Models and Their Assumptions

If one reviews any application of statistical models, such as the example we give above, one sees that model assumptions come in five flavors[10]:

1. There is always the assumption, though usually implicit, connecting the mathematical formulation to the world: that data are *observed values* of random variables.
2. There are *functional* assumptions such as $\boldsymbol{Y}$ is a linear function of $\boldsymbol{X}$.
3. There are *parameter* assumptions, such as that the ε_i*'s* have means of 0 and variances of σ^2, or that there are non-zero *a* and *b*, even if they usually have to be estimated from the data.
4. There are *independence (*and *identically distributed)* assumptions, such as that the ε_i are *independent (*and *identically distributed)* and that if $\boldsymbol{X}$ is random, $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$.
5. Though *not* an OLS model assumption, in practice there are usually more specific *distributional* assumptions as well. For instance, in OLS one often assumes that the ε_i are *normally* distributed; this allows one to calculate t-tests, p-values, *F* tests, etc. and draw *inferences* about how well the model fits the data. Full distributional assumptions allow one to discuss a reserve distribution as well.

---

[8] See Stigler, *History of Statistics* [13]. Before the advent of least squares, the Method of Averages was to use simple unweighted averages of subsets of equations to reduce a system to solvable form.

[9] One minor caveat is to avoid dividing by anything with a non-zero probability of being zero.

[10] There is some regrettable but inconsequential overlap in my classification system: with for instance functional relationships creating dependency relationships. Also, it can be very hard conceptually to separate the two i's in i.i.d..

I fear that, through over-use, most of us have become deadened to the true force and meaning of *assumptions*.

Assumptions are *theories* about the world.

In *pure* mathematics, the fundamental concepts are *undefined*, since one has to start the chain of definitions somewhere; and all concepts are *un-interpreted*, meaning no particular meaning in the real world is ascribed to them.[11] One merely studies how statements (*axioms*) postulated about these concepts in terms of each other imply other statements (*theorems*) made up of them or of concepts freshly defined in terms of them.[12]

In *applied* mathematics, on the other hand, one must first *interpret* some of these undefined and un-interpreted terms, so that they refer to something in the world. The hardest term to interpret in our context is *probability* itself, and I will not attempt to do so here.[13]

But once *some* interpretation is given for that term, we can connect the assumptions and results of our statistical modeling to the real world. Assumptions, which were just conditions of theorems in pure mathematics, become in this way declarative statements about the world and how it functions. That is, they are *theories*.

In addition to *probability* itself, the hardest modeling assumption to comprehend is the assumption of *independent* and *identically distributed* (i.i.d.) random variables. We will not make much use of this, but I think an extremely useful tool in trying to understand the real meaning of *i.i.d.* is given by a *conceptual* model of the loss generation process as a *box* model and presented, for instance, in the textbook *Statistics,* by Freedman, Pisani, Purves ([6] page 389.) The idea is that the situation to be described by a statistical model must generate data like draws of lottery tickets from a box with fixed numbers of different tickets, where each ticket is equally likely to be drawn. There can be multiple levels of boxes and selections required, and there need not literally be a *box*, of course, but one has to be able to conceptualize the process in such a manner.

In insurance we have a ready-made box model for us in the form of Collective Risk Theory. Recall, in this model, there is a box for the claim count during a period, and a box for the claim

---

[11] This is not to say that most mathematicians don't have some interpretation in mind, merely that there is no *official* interpretation for a fundamental undefined concept such as *set membership*.

[12] This is the source of mathematical logician (later philosopher) Bertrand Russell's popular definition of mathematics as: "the science in which we do not know what we are talking about, and do not care whether what we say about it is true."

[13] See Don Gillies, *Philosophical Interpretations of Probability* [7], for a highly readable and sympathetic account of the main interpretations of probability, from a philosopher who has also done applied work in statistics.

severities. For a single period, one selects a ticket from the claim count box. Then, based on the number shown, one selects that many tickets from the severity box, making sure to replace each severity ticket after recording its value, and shaking the box thoroughly before selecting the next severity ticket. Finally, after one has drawn the requisite number of severity tickets for that period, one vigorously shakes the claim count box, selects from it for the next year, and draws again from the severity box.

We will return to the assumption of i.i.d. draws in the results section and examine a little more how well it fits to insurance. For now note that rather than establishing by means of facts, theory, or argument that their data is really from i.i.d. random variables, most modelers merely *assume* it, for among other things the enormous computational convenience it provides. Then, if they have any doubts about these assumptions and the other modeling assumptions, they rely on tests and diagnostics to indicate if they might possibly be in error. So, we had better discuss now, tests and diagnostics.

2.3 Statistical Hypothesis Testing

Tests, *diagnostics*, *validation*, *reasonability checks, goodness-of-fit*—they all mean roughly the same and can all be treated in the same general framework, which is *Statistical Hypothesis Testing*.[14]

The CAS Working Party says (page 47) "[b]y overall model reasonability checks, we mean 'what measures can we use to judge the overall quality of the model?'" and then on page 49 "[b]y goodness-of-fit and prediction error evaluation, we mean 'what measures can we use to judge whether a model is capturing the statistical features in the data?'" They go on to list various tests, such as, "Coefficient of Variation by Year," "Validity of Link Ratios," and specific "Goodness-of-Fit Measures"...in all over a dozen criteria.

In all of these tests, the same abstract framework pertains: a *measure* is a *function* of the data being modeled and the model specification (all the various assumptions and specific parameter values,) and the measure is used to reach some *decision* about whether the current model specification is adequate or not. Sometimes it won't be the result of a single test, but a combination of tests will be examined, but in this case as well the final outcome is usually the same: a *yes* or *no* decision is reached

[14] A fairly clear and concise treatment of the elements of hypothesis testing is available in Klugman's exam study note [8]. A classic graduate-level text is Lehmann and Romano's *Testing Statistical Hypotheses* [12]. As of this writing in 2013, the Wikipedia article for *Statistical Hypothesis Testing* is an informative introduction.

about the whole model or some features in the model.[15]

So, we can represent our measure (or combination of measures) as a function φ which takes on the value 1, whenever we would reach a *yes* decision and the value 0, whenever we would reach a *no* decision.

Two essential concepts in hypothesis testing are: *power* and *loss. Power* gives us the probability that φ=1, that a *yes* decision is made, as a function of the models under consideration. *Loss* is a function of both the *decision* we make and the *models* under consideration, and measures the consequences to us of making some decision when a particular model is true.

This is all very abstract, so we will look in detail at the example which most concerns us, which is significance testing of a variable. In this type of testing we assume that, except for the *parameter* assumptions, *all* other assumptions are known and fixed: the *functional* assumptions (such as linearity,) the *distributional* assumptions (such as normality,) the *independence* assumptions (such as i.i.d. errors.)

2.3.1 A Fully Worked Example

Consider the following simple model:

$$Y_i = bX_i + \varepsilon_i \tag{2.12}$$

Where,
X_i are *i.i.d.* lognormal with mean 1 and standard deviation of 1
ε_i are *i.i.d.* normal with mean of 0 and standard deviation of 1
All X_i*'s* and ε_i*'s* are independent of each other
$b = .5$

X_i might represent the cumulative losses for accident year i at a certain development age; b is an incremental multiplicative loss development factor that applies between that development age and the next; ε_i is random variation in the development; and Y_i is the resulting incremental losses for accident year i at that next development age.

Say we had nine full observations of X_i and Y_i, and 1 more of just X_{10}; i.e., assume X_{10} is on the latest diagonal of a triangle.

---

[15] I said "usually" because occasionally one will stop with just a probability, such as a *p-value*, and draw no particular decision or action as a result. That need not concern us. Note also: Standard Bayesian methods do not usually rely much on classical hypothesis testing for parameters, but rather rely on prior knowledge encapsulated within a prior distribution(s) which is updated with data. Nonetheless, if a Bayesian modeling exercise ever needs to reach a *yes/no* decision about variable inclusion or whether any assumptions are *true* or *not*, all of the comments in this section apply; in a Bayesian analysis one has to of course include the prior distribution(s) in all such calculations.

In order to test our model specification we consider whether the model isn't really

$$Y_i = a + bX_i + \varepsilon_i, \text{ where } a \neq 0 \tag{2.13}$$

To test this we calculate the OLS parameter estimates for that model from our data.

What we are testing in particular is called a simple *null hypothesis* that $a=0$. We also state an *alternative hypothesis* which is that $a \neq 0$. In a significance test, if we find that our estimate of *a* is "significantly different" from 0, then we have some possible evidence that our null hypothesis, which corresponds to model 2.12, is incorrect. And this is possibly some evidence in favor of the alternative which corresponds to 2.13.

The significance test itself is often a *t-test* via a *t-stat*[16] which is the ratio of the parameter estimate for *a* divided by the estimate for the standard error of our estimator for *a*. Recall $\hat{a}$ is an *estimator* for *a* and a random variable, so it has a mean, standard deviation etc. When we have data we are modeling, we end up calculating one particular *estimate* which we assume is an observed value for that estimator. We also have an estimator for the standard deviation (aka standard error) of that random variable, which when applied to our data gives us an *estimate* of $\hat{a}'s$ standard error.

In order to use the t-stat for any inferences, we must decide a *critical region* of values which we will regard as *significant*. Any cut-off is arbitrary, but many modelers suggest using the value of 2 for the t-stat to judge significance as an easy rule-of-thumb. If another critical region were chosen, our numbers below would of course change, but the issues would remain the same.

So, now we have our full test function φ: it will equal 1 whenever the t-stat has an absolute value of 2 or more, which corresponds to finding the parameter estimate *significant*, and otherwise it will equal 0. Our *power* will give us the probability of φ equaling 1, which is also the expected value of φ, which we can calculate because we have specified our model in (2.12.)

Estimator	Mean	Mean Standard Error	Power
$\hat{a}$	(0.00)	0.58	0.08
$\hat{b}$	0.50	0.54	0.28

With respect to $\hat{a}$, our significance test worked *exactly* the way it's supposed to: our estimate for *a* was only found to be "significantly different from 0" 8% of the time, which is good, because it is

[16] So called because when $\boldsymbol{X}$ is fixed and $\boldsymbol{\varepsilon}$ normally distributed, the *t-stat* will follow Student's *t* distribution. Because our $\boldsymbol{X}$ is random and not fixed, and lognormal besides, we chose to simulate the results here rather than try to find any closed-form solutions.

actually 0. But keep in mind that this is a probabilistic result: in 8% of the cases, one would conclude that *a* was *not 0* when in fact it was. This is the un-memorably named *Type I Error* rate.

But notice also that for $\hat{b}$, it was only found to be significant, 28% of the time. This despite the fact that *b* is not 0, it is .5. So, if anyone were to try to conclude whether *b* was 0, because of its significance test, he or she would be right only 28% of the time and wrong 72% of the time. The latter is the un-memorably named *Type II Error* rate. So, before one can really judge if a non-significant finding is very good evidence that a parameter is 0, one should attend to the power of the test to judge whether the test can even tell with much probability whether the parameter is "significantly different from 0."

But all the tests in the world are for naught unless one takes some action as a result of the test. The authors cited at the start of the paper suggested that one should drop a variable if its parameter estimate is found insignificant. In order to see whether this is a wise decision, we must specify some *loss* function that will measure the costs to us of taking an action: our action will be choosing one estimator over another and our loss will be a reduction in accuracy. There is infinite freedom available in choosing loss functions, but we will simply take a common measure of predictive accuracy for our loss. The one we will take is the squared difference between an estimator's prediction and the variable it is trying to predict. We will also look at just the (un-squared) difference between the two.[17] The expected value of the former is the *mean square error of prediction* (MSEP) and of the latter is the *bias*. We take the square *root* of the latter, which we call the *root mean square error of prediction* (RMSEP), to have the same units as the variable we are trying to predict.

What we are trying to predict is the incremental loss Y_{10} based on the previous 9 accident years and X_{10}.

As mentioned in section 2.1 we have infinite flexibility about what estimators and resulting predictors to use. But let's look at just the following:

[17] In the standard theory, loss functions are non-negative; so un-squared difference would not usually be considered a loss function.

1. The **Full OLS**[18] estimator, where we estimate both *a* and *b* by OLS and predict Y_{10} as $\hat{a} + \hat{b}X_{10}$.
2. The **Average** of the Y_i 's over the 9 accident years.
3. The OLS estimator for *b* only, with only a single variable in the regression; the **LDF Only** predictor, $\hat{b}X_{10}$.
4. The **Modeler** recommendation that we use predictor 2, the **Average**, unless the estimate of *b* is significant, in which case predictor 3, the **LDF Only**, is used. One should note that this is in defiance of standard actuarial practice, which would be to down-weight any unusually high LDFs.
5. So we can look at the **Anti-Modeler** (or **Actuary**), which does the exact opposite of the **Modeler** estimator, and uses predictor 3, the **LDF Only**, unless it's found to be significant, in which case it uses 2, the **Average**.

	Estimators				
Loss Criteria	**Full OLS**	**Average**	**LDF Only**	**Modeler**	**Anti-Modeler**
Bias	(0.00)	0.00	(0.00)	(0.01)	0.01
RMSEP	1.25	1.18	1.10	1.15	1.13

The negative signs in the Bias row mean the estimator is biased high; positive that it's biased low. It's not too difficult to understand this table. The **Full OLS**, **Average**, and **LDF Only** are all unbiased estimators of the mean of Y_{10}. The **Modeler** predictor *randomly* switches between the **Average** and the **LDF Only**, based on the significance test of *b*. Since *b* is positive, including it in the model only when it's high relative to its standard error, means that the expected value of $\hat{b}$, conditional on finding it significant, will be biased high. So, the **Modeler** predictor is biased high. The **Anti-Modeler** predictor does the opposite, so is biased low.

The RMSEP of **LDF Only** is the lowest, which is nice because that reflects the true model of the data. The MSEP (before the square root) of the **Modeler** is basically the *power*-weighted MSEPs of

[18] A reviewer of parts of an earlier draft of this paper asked whether it wouldn't be more appropriate to use *weighted* regressions in some contexts. It might be, but would introduce many technical complications which would take us too far afield to address. See note 9 for instance. Suffice to say, certain numbers would change in that case, but the conceptual issues would remain the same.

the **LDF Only** and the **Average**. The MSEP of the **Anti-Modeler** has just the converse weights. Since *b* is found significant only 28% of the time the **Modeler** estimator spends only 28% of its time as the **LDF Only,** which has the lower RMSEP**,** and spends 72% of its time as the **Average,** which has the higher RMSEP.

Since the biases are the same, but the RMSEP of the **Anti-Modeler** is lower, in this case we would say that sometimes it is better to do the *exact opposite* of what the modelers recommend.

To generalize this example, if one *knows* the correct model form, one should design an estimator and predictor for it; significance testing has nothing to do with it. If one does *not* know the correct model form, it's at least possible, as this example shows, that significance testing will lead one to make a suboptimal choice. If *b* had been a little larger, its estimate would have been found to be significant more often; but including the variable in the model only then, would still bias the result high. As *b* grows larger, the bias due to truncating the estimator via significance testing would go to 0 as the probability of finding significance goes to 100%. But we'll see in the results section that this special case is unlikely to occur.

I bring this up as an illustration of the dangers of relying blindly on significance testing, but I will leave it as an area of future research to delineate precisely the situations where one should or shouldn't run significance tests and what actions should be taken as a result. I have seen no such delineation in any of the actuarial literature which takes into account the above issues.

2.3.2 Estimator Selection

The issue of variable selection for an estimator, as far as I can tell, is far more complicated than simply running significance tests, and may have nothing really to do with it.[19] An optimal estimator, by *whatever* loss criteria, is likely to depend on the precise distributional and parameter assumptions underlying the *true* model; an estimator might be optimal in some regions of the parameter space and less optimal in others, optimal for some distributions and not for others. Since the true model in any real application is more or less unknown, it's unclear what relevance such results would have unless the estimators are very robust to a wide range of potential models and parameter combinations. Many of the theorems available to help find estimators are restricted to *unbiased* estimators, and when *biased* estimators are included, the problem becomes so much more extensive. Finally, many of those theorems are restricted to *fixed* variables, and in insurance our potential variables are mostly

[19] See Lehmann, *Theory of Point Estimation* [11], for a graduate-level treatment of estimator theory.

random variables.[20]

2.3.3 Statistical Power More Generally

We mentioned previously that *power* is a *function*, but then only gave the single values of power for our regression estimators. We now complete the discussion.

So, consider a simple test function φ. φ will take on the value of 1 when the criteria we're testing for is met (or conversely not met,) or 0 if it is not (or the converse.) Given all the other model assumptions as fixed, let's let only the parameter assumptions vary. We can describe our model of the data as some joint distribution P_θ of the random variables, while θ is a possible parameter assumption.

The *power* function[21], often denoted $\beta(\Theta)$, is then a function of θ as θ varies over its possible values, and equals the probability that φ=1, calculated on the assumption that P_θ is the true model for the data. One can write this

$$\theta \to \beta(\Theta) = P_\Theta(\varphi(M) = 1), \tag{2.14}$$

where M potentially encompasses all of the random variables of the model.

This is all a little abstract again, but the general formulation is useful to keep in mind, whenever one comes across any statistical analysis whatsoever:

1. One should ask what the model assumptions are and why; in other words what is P_θ the joint probability distribution of the random variables which correspond to the data; and then why: a useful heuristic might be, rather than contemplating "reasonableness" of assumptions, to start from the premise that the assumptions are false and see what reason there is to believe otherwise;
2. Even granting the model assumptions, if drawing any conclusion from the analysis, one should ask what the power of the analysis was to come to an opposite conclusion. If the test had low power to detect an alternative, there might be little reason to believe the results.

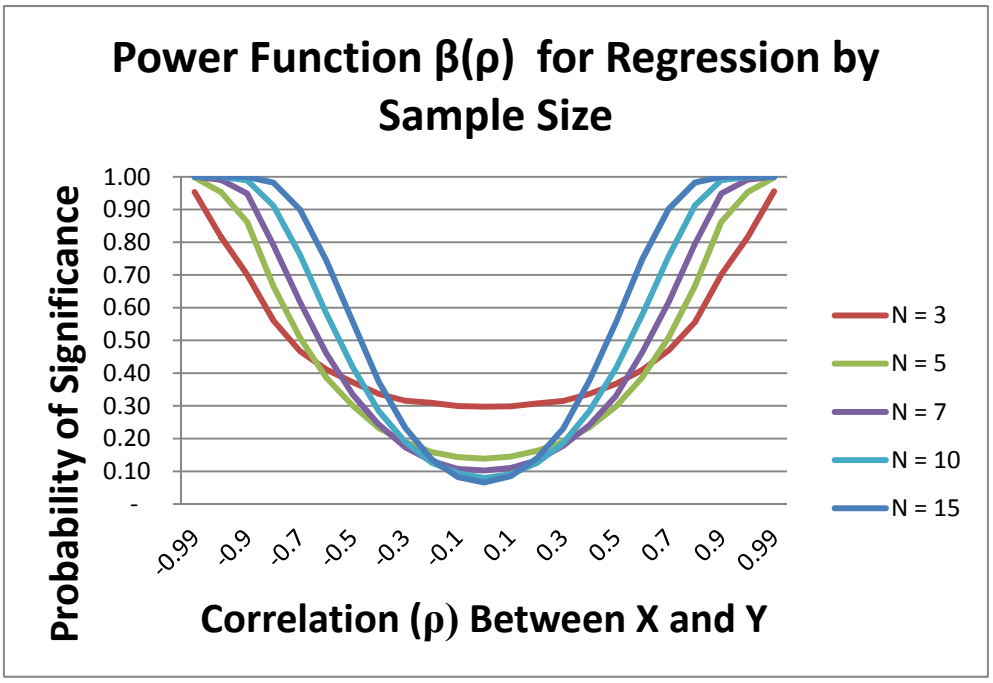
---

[20] For those who wish to pursue the topic further, the correct topic heading here appears to be *errors-in-variables* models and also *latent variable* models. Please note the early pioneering work that James Stanard did on estimator properties for certain types of loss development, and I understand Hans Bühlmann and others have continued some of that work as well.

[21] This definition is adapted from Lucien LeCam's useful comparison of frequentist and subjectivist approaches to statistics [10]. According to Wikipedia, he "was the major figure during the period 1950 – 1990 in the development of abstract general asymptotic theory in mathematical statistics."

Returning to the specific test we're most interested in, significance, we can easily find the power function for any OLS regression as a function of the correlation ρ of two variables X and Y, where they are both normally distributed with means of 0 and SDs of 1. One can then generalize this result by adding constants for their means and scaling by different standard deviations. (Recall from formula 2.9 that the OLS coefficient in front of x_{n+1} is the product of the chain ladder estimate, the ratio of the y's mean to the x's, and the correlation and CV's, $\rho \frac{CV_y}{CV_x}$.)

Using the t-test, with the critical value of 2 discussed in the earlier example, the power function of the significance test for small sample sizes which might be relevant for yearly reserving triangles is the following:



Even with a sample of size 10, *real* correlations between X and Y that are between about (-55% and 55%) have a 50-50 chance or less of being detected. Whether in a particular modeling exercise a relationship of this size between two variables (such as the cumulative losses in one development age and the incremental losses in the next) will be declared significant is just a coin-flip. For a sample of size 5, a correlation of more than +/-70% is needed to have a better than even chance of detecting it.

Another way of looking at the same issue is to consider the number of years needed in order to

have, say, a 50-50 chance of detecting a correlation of a given size. This could be an involved simulation exercise, so we will make the simplifying assumption that the sample correlation coefficient $\hat{\rho}$ is symmetric around its mean value ρ. This is probably ok unless ρ is close to +/-1, but then power isn't much of an issue anyway. There is a standard formula available for this[22]:

$$2 = \sqrt{\frac{\rho^2(n-2)}{(1-\rho^2)}} \tag{2.15}$$

where the first 2 is our critical value for the significance test,
the second 2 is for the number of variables, and n is the number of years.

We easily rearrange this to express *n* as a function of ρ.

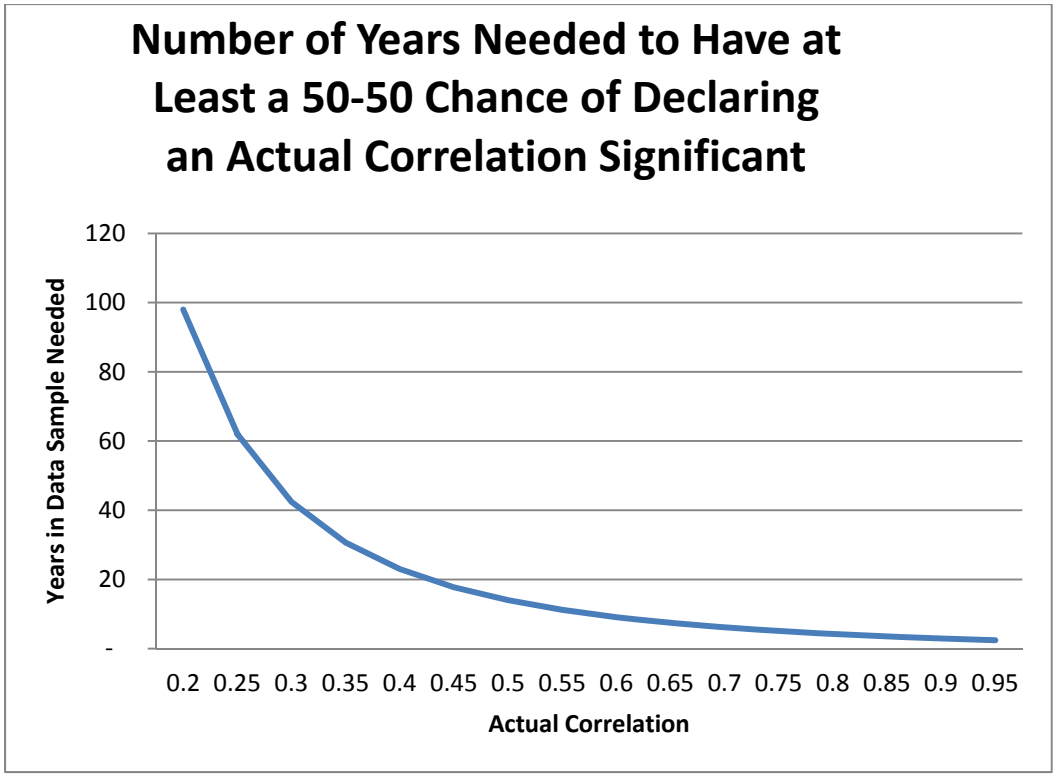
$$n = \frac{4}{\rho^2} - 2 \tag{2.16}$$

See Graph 2.1 on next page. I don't show the number of years needed for correlations of size .15, .10, or .05 because those are 176, 398, and 1598 respectively.

Keep in mind also that a finding of significance for an effect in a real-world modeling application is a function of: the model specification, the size of the effect, the size of the sample, and random chance. One can't in practice conclude from a significant or an insignificant finding alone which of those causes are responsible for the finding.

The most natural question to ask at this point is, what correlations would we expect to find in insurance? And we will address that issue in the next section, but first it might be worthwhile to discuss an amazing paper by David Freedman and generalize our discussion a little to other tests besides significance.

[22] See the Wikipedia article on "Statistical Hypothesis Testing," for instance.



Graph 2.1

2.3.4 Diagnostic Power or the Lack Thereof

Statistical diagnostics, checks, tests, etc. usually address themselves to limited breakdowns of model assumptions. As we saw with significance, one fixes all of the distributional, independence, and functional assumptions and looks at whether or not the data is consistent with some parameters being 0. That's all the t-test is used for.

So, diagnostics are not usually direct tests of *all* of the assumptions of a model, but continue to assume some parts of that model. One should always keep in mind that a diagnostic outcome might be the result of some other model breakdown than what one is explicitly testing; this is too often forgotten in rushes to find some result, such as significance for instance.

All diagnostics have probabilistic results and all the issues with chance occurrence and power highlighted above apply. Even with non-parametric tests or robust inference procedures, once one has a probability model to work with, one can calculate the probabilities of passing or failing any test and the issue of power comes into play.

It's a common practice among modelers to also "teach to the test," meaning if they know that

they'll be running some test, they will include additional features in their model, so that it will pass that particular test. A common such addition is to include some sort of time-series model in order to pass an autocorrelation test of the residuals. The *residuals* are the differences between the observed data and model estimates for that data. An autocorrelation (or serial correlation) test looks at the sample correlation between residuals in near-by years.

In some sense this is perfectly legitimate: if the residuals would otherwise show autocorrelation, there is trend within the data, and the model should perhaps be adjusted. It will definitely improve the model fit to the data. But the question is whether it will improve the model's predictions as well. The big--usually unasked--question is whether the particular model that is used to adjust for this trend is really true and why. Including such a sub-model within a model makes it harder to analyze the whole model, and may just be sweeping model misspecification issues under the rug.

If one opens up one's imagination to more general statistical assumptions than usually contemplated, then there is even more reason to question modeling results. David Freedman in a paper entitled "Diagnostics Can Not Have Much Power Against General Alternatives" [3] took one-by-one the standard assumptions underlying most statistical models, and showed that given any test, diagnostic or combination of them, which the standard assumptions would pass with a certain probability, there are alternative assumptions that are very different from the standard ones, but which would pass with the same or greater probability.

Although Freedman advises that all diagnostics should be viewed with a healthy dose of skepticism, he does *not* conclude that diagnostics should be ignored, quite the opposite, and he recommends they be employed and published more often. This is because diagnostics can still occasionally detect gross violations of model assumptions. But his results also clearly imply that one cannot simply rely on diagnostics to determine whether a model is true. One needs prior theory and experience to convincingly narrow down the model possibilities before diagnostics can be of much use.

Even then, as we saw above, with small samples or volatile processes, we need to be realistic about what we can and cannot accomplish from statistical analyses and only the data at hand.

3. RESULTS AND DISCUSSION

3.1 Some Examples of Loss Development Correlation in Insurance

3.1.1 Claim Count Development on Claims-Made Business

Consider a claims-made book of business. Suppose that X_i are the claims still open at the end of the first development period for accident year i and Y_i are the incremental changes in open claims during the next development period. Let's assume, as is often done, that X_i is Poisson distributed with mean μ. Now suppose that for each of the X_i claims, independently of the other claims, there is a certain probability p that it will close during the development period. Then Y_i is -1 times a Binomial random variable with parameters X_i and p.

Then it is possible to show that the correlation between X and Y is:

$$-\sqrt{p} \tag{3.1}$$

So, for instance, referring to Table 2.1 for a quick approximation, for a $|p|$ below .25 ($\sqrt{p}$<.5), we are unlikely to find the correlation significant with less than about 14 years. For p's over .5 however, we need only about 6 years or less.

3.1.2 Completely *Dependent* Development on Claims-Made Business

The above is a special case of a more general situation where the development on each claim during the next period is a multiple of the claim itself (in the above case, either -1 or 0.) We can generalize this example, by means of Collective Risk Theory, to include reported claims simply, rather than just open claims, reported severity distributions in the first period and then a distribution of incremental reported development factors which will be multiplied to each claim severity for the next. Let's call the development random variable λ and say it has a coefficient of variation (its standard deviation over its mean) of CV_λ.

Then the correlation is (where the -1 is if there is expected negative development):

$$\frac{\pm 1}{(1 + (CV_\lambda)^2)^{1/2}} \tag{3.2}$$

Intuitively, if λ were a constant, $\boldsymbol{X}$ and $\boldsymbol{Y}$ would be perfectly correlated; but there is random variation in the development (CV_λ) which is clouding the relationship. If the expected development were 0 the CV would be infinite and there would be no correlation, though there would still be

dependence via the claim count.[23]

Now, the question is what magnitudes of CV's should we expect in insurance? Well, that's going to vary by line-of-business, the specifics of a company's reserving practice and the details of the claims.

On one extreme, one could imagine that the claims-adjusters were nearly always right, at least on average, about the ultimate values of the claims they adjusted during the first period. In this case, the mean of λ would be close to 0, in which case, because the mean is in the denominator, CV_λ could be enormous. One would then have very low correlation and find it almost impossible to detect it, even though the development was still dependent on the losses from the prior period.

On another extreme, one could imagine that the claims department is stair-stepping their reserves, using claims signals for instance, or some other reserving practice which doesn't match averages, and the average development might be quite large in the next period. Then it would depend on the spread of that development, which might still be quite wide if there's a diverse set of claims.

I certainly don't know what ranges this parameter might take for different books, but the one book I did look at, which was a not too volatile professional liability account, had CV's in the second period of about 15, which means a correlation around 7%, and very little chance whatsoever of detecting that in a significance test (one would need 800 years.)

3.1.3 Completely *Independent* Development on Claims-Made Business

At another extreme, we can continue with our model for reported losses, but this time we assume that the development in the next period is completely *independent* of the severities in the prior period: they are just additive amounts that emerge for each claim independently of what the severity on the claim was previously. Because the claim count is common to both periods, the losses are still correlated however. Let's still call the development random variable λ and with a CV_λ. But now we must include the reported severity in the first period, which we'll assume has CV_S.

Then the correlation is:

$$\frac{\pm 1}{(1+(CV_S)^2)^{1/2}(1+(CV_\lambda)^2)^{1/2}} \tag{3.3}$$

[23] Zero correlation does not imply independence except with normally distributed variables for instance.

Intuitively, ***X*** and ***Y*** are correlated only by having the same claim count in common, but there is random variation in the development (CV_λ) and random variation in the severity (CV_S) which do nothing but weaken the relationship.

Again, any CV's will depend upon the particulars of a book of business, but for my one account above, I found CV's of about 3 for the reported severity in the first period and about 8 for the incremental development in the second. (Note this is the CV of λ viewed as an additive amount rather than the multiplicative amount from before.) I find a correlation here of about 4%, again undetectable for all practical purposes.

So far, I've only considered the first two development periods. As we move along the triangle, taking ***X*** as the cumulative development and ***Y*** as the incremental development, I might expect the CV's associated with **X** to grow as more information became available to precisely determine claim values. We might also expect the CV_λ to perhaps explode while the prior estimates are getting more and more accurate so the incremental development averages are close to 0, while whatever development there is might be highly volatile, and perhaps volatile enough to overcome the large number of claims undergoing no more development. But this is just speculation, and there could be all sorts of patterns of CV development. It might not even make much sense to think of these CV's as immutable parameters that could be meaningfully estimated, though the one account I looked at had much more stable CV's by report and development year than I would have expected.

3.1.4 Adding Independent IBNR

If we generalize to occurrence business and add pure IBNR claims going into the second period, which are completely independent of the claims and losses from the first period, then one can show that the correlation gets scaled down by a factor which includes the ratio of the additional variance of the new losses to the original variance.

3.1.5 Adding *Dependent* IBNR

If one adds pure IBNR claims going into the second period, which are dependent in any way on the claims and losses from the first period, as they would be if they were the result of common exposure for instance, then one can show that once again the correlation gets scaled down by the additional variance of the new losses, but a new additive term enters into the equation as well for correlation as a result of the common exposure.

3.1.6 Discussion

So, in all of the above cases, except where there is 0 expected development, there is non-zero

correlation between the aggregate losses in one period and the incremental losses in the next. So if modelers have failed to find the loss development factor significant, it might very likely be due to the lack of statistical power of their analyses.

3.2 So What Development Methods Should We Use?

As I mentioned even with the simple example of 2.3.1, that is actually a very involved question. For what it's worth, based on a few tests with even tamer parameters than the ones I found for the account discussed above, I could not find much practical difference between using an average incremental development, chain-ladder, an average of the two, or the modeler or anti-modeler estimators. OLS seemed to perform slightly worse than those, but a 5-10% difference in RMSEP hardly seems to matter much.

There is a certain amount of *irreducible* uncertainty[24] to development that cannot be decreased by any estimator no matter how clever. I suspect almost any standard estimators or actuarial methods would be about as good (or bad) as any other, and as long as a number of methods are applied, there is just no practical benefit from worrying about estimator optimality.

What *does* have a real practical benefit is if an actuary can determine the parameters themselves independently of this data. Or, barring a definitive determination, a Bayesian method, as long as the prior concentrated close to the true answer, could make a practical difference as well.

3.2.1 Example of Meaningful Estimation Improvement

The real advantage of traditional actuarial methods is not their optimality properties, but that they bring the actuary quickly into close contact with the data. If there is any information the actuary can discover which will *reduce* uncertainty, then *that* could have a large effect on estimation accuracy.

Let's suppose we're at a primary medical malpractice writer. Fortunately we only write claims-made policies so except for the occasional DD&R policy, which we can always reserve separately, there's no pure IBNR, and all loss development is from claims reported in the first period. We buy reinsurance to put a cap on our maximum loss as well.

For our model let's assume that we have 10 years of data and we're trying to project the 11th. For each report year,

[24] There is an unfortunate ambiguity is many uses of the term *uncertainty*. Sometimes it refers to a psychological state, something akin to doubt. And it sometimes refers to random, or apparently random, variation in the world which is merely one potential *cause* of that psychological state.

1. There's a fixed number of "nuisance" claims, let's say 10, of negligible value (like clearly illegitimate claims.)
2. One in every 10 years there a "catastrophic" claim (like a fetal brain injury with negligence from the OB) that will hit the reinsurance. This is a Bernoulli variable with p=0.1
3. Otherwise there are "regular" claims each year that are Poisson distributed with a mean of 10.
4. During the first development period the claims department assigns them all 1 unit (think $100,000 maybe) until it can complete an initial investigation which won't finish till at least the next year.
5. During the next incremental development period, they will discover that the 10 nuisance claims were just that, and drop all their reserves to 0, for an incremental change of -10.
6. Any catastrophic claim will be discovered and its reserve increased by 10 ($1M.)
7. Of the remaining regular claims, there's a 30% chance that each can have its reserve increased by 2 ($200K.) The others will remain as is.

Since we've specified the model we can calculate explicitly anything we can imagine.

For instance, from the above the expected losses in the first year are 20.1: 10 nuisance claims of 1, .1 catastrophic claims, and 10 regular claims.

The expected incremental losses reported in the next year are: 10 decline by 1 each, so -10; .1 increase by 10, for 1 on average; and 30% of 10, or 3, increase by 2, for 6. So the expected incremental development is -3.[25]

We can explicitly derive a number of other values such as conditional expectations based on various levels of knowledge and detail known or knowable to an actuary. But for now, let's simulate our model with 10 report years of data and apply our methods, starting with OLS regression.

[25] Negative development is not uncommon in medical malpractice.

		Mean		
	Mean	Standard	Mean	Power
	Value	Error	+/-2 SE	(β)
$\hat{a}$	(16.56)	8.93	(-34.43,1.31)	0.51
$\hat{b}$	0.67	0.44	(-0.21,1.55)	0.39

So, despite the effect of the number of claims from the prior period going into the next, and even $\hat{b}$'s relatively high expected value, it's only found significant 39% of the time.

One of the first things we should perhaps have done as actuaries is of course look at the age-to-age (ATA) factors in a triangle. Here is a single iteration, which represents for us something like the position we're actually in when trying to reserve: we don't have the possibility of simulating 200,000 separate draws, nor do we know the full distributions of potential outcomes with certainty. Here's one iteration where the LDF was found to be insignificant:

Report Year	Cumulative Reported Loss By Development Age 1	2	ATA LDF
1	21	15	0.71
2	22	16	0.73
3	21	33	1.57
4	22	26	1.18
5	22	14	0.64
6	29	29	1.00
7	18	14	0.78
8	21	21	1.00
9	16	12	0.75
10	24	30	1.25
Average:	21.60	21.00	0.96
Incremental Development:		(0.60)	

It's a fairly typical sample year: the average of the losses at Age 1 is close to the mean of 20.1, and the incremental development is a little low compared to the mean, but nothing shocking. The one ATA LDF which stands out is the 1.57 for Report Year 3, maybe we should investigate? It turns out that is the one year that had a catastrophic claim. Maybe we should reserve that separately?

But we do not stop there. Our job is to project the 11th year from the data available. But the *real* data available is not just the sheet of numbers above; it's all of our experience, and the experience of those we can learn from in the claims department and elsewhere. Based on our level of curiosity,

knowledge, and energy we might assume that we:

1. Take the lazy method and simply add an average. Let's call this **Average**.
2. Run a two variable OLS regression, call it **Full OLS**. Recall this is an average of an LDF estimator and the **Average**.
3. Apply the **Modeler** routine: test for significance first and assume a constant **Average** amount of development in the next year unless the LDF is found significant.
4. Having been around awhile and figured out that there's always 10 nuisance claims, and about 30% of claims increase by 2, while 1 in 10 years have the catastrophic claim, we parameterize a conditional expected value that tells us the expected number of claims based on the number of aggregate claims in development year 0 and apply that to report year 10. Let's call this the **Parameterized.**
5. Finally, let's suppose we are energetic and experienced enough to actually determine to which class each one of the claims belong and then apply the parameters from 4. Let's call this **Energetic**, and basically the only variability left in the reserve forecast comes from the 30% probability of a regular claim developing upward.

Running our simulation with these loss development methods we find, with % reduction in RMSEP measured relative to the **Average** method:

Improvement in Prediction Error From Different Approaches to Reserving

Approach	Bias	RMSEP	% Reduction
Average	0.00	4.82	
Full OLS	0.00	4.54	-6%
Modeler	(0.00)	4.73	-2%
Parameterized	(0.01)	4.04	-16%
Energetic	(0.00)	2.90	-40%

Please note that only the first three are standard estimators, and the latter are estimators where additional knowledge has been brought to bear on the parameters as described in 4 and 5 above.

Once again, the **Modeler** recommendations are not the best and do worse than just OLS regression, and show those recommendations can be unwise; but the practical difference in RMSEP is hardly important. But look on the other hand at how it could pay to understand and know something about the data.

The **Energetic** does best because he or she is able to *reduce* the uncertainty around the claim type, even though the uncertainty around the development of regular claims is still irreducible.

Of course if one has 20,000 claims a year rather than 20, it's not practical to personally read every claim file to determine its underlying allegations and what its claim type is. But even as claims databases get bigger it's not at all a given that the most effective computer algorithms would not be doing essentially the same that a person would do if equipped with the same patience and computational ability as a machine. I suspect every situation must be examined on a case-by-case basis to determine the most appropriate approaches. I also strongly suspect that in many cases where some advanced quantitative technique can be shown to be effective, there are simpler and more direct methods which are just as effective; but I'll have to leave the exploration of this topic as an area for future research.

The point of this example is not that modeling could not do what a traditional actuarial method could do nor vice versa. The point was merely that following a traditional actuarial method, the actuary was alerted quickly to an issue that when investigated further yielded a very large pay-off in terms of accuracy. Once an actuary *knows* that, for instance, certain claim types are an important variable, there are always ways to design programs to capture claim type or to model for it. The question is *how* is the actuary going to *first* discover this? And here I would think that the simpler the exploratory method the better.

In reality, we'll never know the parameters underlying the *true* model, though some believe that Bayesian estimation might get us close. There's a risk as well that we will be *fooled by randomness* to use Nassim Taleb's felicitous phrase, and by digging into the data and "learning" more we are just fooling ourselves that something is more predictable than it really is. There's no way to tell ahead of time, but if one doesn't look, one doesn't find.

It's also highly judgmental how much time one should spend looking for information, especially given the risk of self-deception, rather than just making a selection and moving on. I doubt any hard and fast rules can ever be given, and different actuaries will choose to spend their time differently.

Nonetheless, I believe we often face a situation in which there *is* reducible uncertainty, and with

enough hard work and looking beyond the mere numbers in an aggregate loss triangle we might discover it.

Finally, in this example, the large losses came from truly horrific claims: fetal brain injuries. An insurer has a social obligation to learn as much as it can about the circumstances and possible prevention of such events, given the insurer's other social responsibilities. If as a result of improved risk management the probability of such claims dropped from .1 to .05, it might invalidate the model assumptions, but so be it. There are more important things than model validity.

3.3 The Miraculous Assumption of I.I.D.[26]

I.i.d. is the workhorse of statistical assumptions and without it little gets done.

By definition, random variables are *independent* if the probability that any one takes on a value (or set of values) remains the same no matter what values the others take on. They're *identically* distributed if they have the same distributions.

Mathematically, independence means that the *joint* distribution of the variables factors into the product of the *marginal* distributions of each variable. Identically distributed implies that all means, variances, and higher moments are the same (though not conversely.)

Statistically, if one has a series of i.i.d. random variables, the sample mean is often an excellent estimator of the random variables' means and the sample variance is unbiased and allows one to even calculate the error in one's mean estimate. In short, the knowledge of the values of any subset of the series of random variables will allow one to predict all of the others, at least probabilistically.

If one's data series comes in the form of, for instance, a loss triangle where the lower right triangle is still to come, assuming i.i.d. is no less than **assuming one can predict the future**.

If random variables are *not* i.i.d., the sample mean need not be a very good estimator of the random variables' means, the standard errors in significance tests may be wrong, and the sample variance may be very biased. In the last case, one can even think one has a much better estimate of the mean than one really has.

The Collective Risk Theory assumes that the claim counts in every year are i.i.d., the severities are

[26] See William Kruskal's highly regarded American Statistical Association presidential address: "Statistics and Miracles: The Casual Assumption of Independence" [9] for a discussion of the importance of the independence assumption in the evaluation of testimony for miracles, among other things.
Many of the remarks in this section apply equally to certain similar concepts such as *exchangeability* in Bayesian analyses.

independent of the claim counts, and they are i.i.d. within a particular year and across the years. [27]

The easiest way to appreciate the implausibility of the independence assumption is to recognize that any common cause that is neither certain to happen nor to not happen and that could effect, say, the means of two variables, even with different effects, would give them a non-zero correlation, and hence they'd no longer be independent.

So, any underwriting changes, marketing changes, settlement changes, inflationary changes, weather changes, social changes, etc….anything which *could* serve to change the occurrences of claims or their settlement amounts will either act to change the distributions or to create dependence or to do both. The assumption of i.i.d. would fail.

In classical applications of statistics that are widely regarded as successful, such as for example, casino games, medical testing, population surveying, and general experimentation, i.i.d. is not simply *assumed*, it requires hard work to achieve. And even then, it's usually not perfect.

Now is it possible that all of the different dependency effects will somehow negate each other? Or that we will somehow be able to adjust accurately for all of them? Sure it's possible, but it would be little short of a miracle.

If someone is simply presenting a theorem in pure mathematics, then one can assume whatever one likes. But if one is presenting any real world conclusion, one should pay attention to the validity of one's assumptions. Since the assumers of i.i.d. are arrogating to themselves the ability to predict the future, the onus should be on them to establish that the assumption is true or at least cannot be very far from the truth. I would think that they should at least study and present the sensitivity of all of their conclusions to this assumption, but this appears to be a still largely unexplored area of actuarial research.

4. CONCLUSIONS

We saw immediately above that the fundamental assumption of statistical analysis is almost certainly false when applied to insurance.

We saw earlier that even making this assumption, one is unlikely, because of limitations to statistical power, to be able to discover by statistical means anything very useful about the claims

[27] One often adjusts for some theorized trend and portfolio changes (like deductible, etc.) first, but *then* the variables are assumed i.i.d..

generation process or which reserving method to use.

We saw for a few examples that some common modeling recommendations can be unwise, though we did not show that they would *always* be unwise.

We only focused on modeling recommendations which have been applied to aggregate loss triangles; the recommendations might make more sense for individual claims modeling. I believe that many of the results in this paper generalize to contexts outside of modeling just yearly aggregate loss triangles, but that has to remain an area of future research for now.

I believe that we've *practically* refuted two particular claims made on behalf of stochastic reserving of yearly aggregate triangles: the claims that assumptions can be tested and models validated *effectively*. But we've only showed this for a few examples and provided the conceptual tools to help study the issue. Others might want to research the situation and determine a precise delineation of when significance testing might lead to an optimal model and when the model diagnostics are most effective.

We saw that one might be able to make greater progress in estimating reserves by focusing on fact-finding rather than model-checking.

Outside of the aggregate triangle modeling we've considered, as datasets become larger, there are computer algorithms (such as text searches, clustering, etc.) that might prove extremely useful in the data exploratory process. Many of these methods don't rely on statistical assumptions at all, though no doubt they have their own issues. Regardless, my criticism in this paper was leveled at some careless applications of statistical assumptions and modeling, and not at all "advanced" techniques whatsoever.

There are also other arguments made on behalf of stochastic modeling. Some of these are more plausible, and, regardless, some of them are persuasive in certain situations.

Generally, I believe that actuaries need to become much more skeptical and critical of the claims made on behalf of statistical modeling. For many, the technology and the imagined power of statistical analyses are just too seductive. The result can be a lot of wasted effort and misleading models.

David Freedman in "As Others See Us: A Case Study in Path Analysis" [4], which carefully analyzes an application of path models in social science, notes about social scientists in general who apply advanced quantitative techniques that: "nobody pays much attention to the assumptions, and

the technology tends to overwhelm common sense. "

Freedman also cites in that paper studies showing that major econometric forecasting models do very poorly unless frequently revised and unless some of their parameters are re-estimated *subjectively* by modelers; and even then they do no better than forecasters without models. (page 123)

The solution is simple, in fact probably too simple for many to accept. Freedman:

"My opinion is that investigators need to think more about the underlying process, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models. Estimating nonexistent parameters cannot be very fruitful. And it must be equally a waste of time to test theories on the basis of statistical hypotheses that are rooted neither in prior theory nor in fact, even if the algorithms are recited in every statistics text without caveat."

Acknowledgment

I received many recommendations and much helpful support, instruction, and encouragement from Michael Rozema. Anthony Manzitto provided interesting perspectives that both here and more generally over the years have helped to shape the views expressed here. I have benefited greatly from many conversations with James Guszcza and from the reading of many of his papers.

I received helpful direction and comments from the CAS reviewers Denise Ambrogio, Lynne Bloom, and one anonymous reviewer.

Last but not least, Nicolas Annoni gave a very detailed and close reading of an earlier draft of this paper, and provided many good ideas for improving it. He also proofread one of the latest drafts.

The opinions expressed here are not necessarily those of any but the author. And anything good about the paper is due to the above reviewers, while any errors or obnoxious opinions that remain are entirely those of the author.

Supplementary Material

Most of the articles referred to in notes of this paper are available on the internet free, for purchase, or through a library. Note in particular that many of the articles by David Freedman, including ones now published in *Statistical Models and Causal Inference* are freely available at http://www.stat.berkeley.edu/~freedman/

Klugman's *Estimation, Evaluation, and Selection of Actuarial Models* is available here: www.casact.org/library/studynotes/**klugman**4.pdf

5. REFERENCES

[1] CAS Working Party on Quantifying Variability in Reserve Estimates, "The Analysis and Estimation of Loss & ALAE Variability: A Summary Report," *Casualty Actuarial Society* Forum, Fall 2005

[2] Clark, David R., "LDF Curve-Fitting and Stochastic Reserving: A Maximum Likelihood Approach," *Casualty Actuarial Society* Forum, Fall 2003

[3] Freedman, David A. "Diagnostics Can Not Have Much Power Against General Alternatives," in *Statistical Models and Causal Inference*, Cambridge University Press, 2010

[4] Freedman, David A., "As others see us (with discussion)."*J.Educ.Statist.* 12, 101-223. Reprinted in J. Shaffer, ed. *The Role of Models in Nonexperimental Social Science* AERA/ASA Washington D.C. (1997)

[5] Freedman, David A., *Statistical Models: Theory and Practice*, Cambridge University Press, 2005

[6] Freedman, Pisani, Purves, *Statistics,* W.W. Norton & Company, 1978

[7] Gillies, Donald, *Philosophical Theories of Probability*, Routledge, 2000

[8] Klugman, Stuart, *Estimation, Evaluation, and Selection of Actuarial Models*, Klugman, 2002
[9] Kruskal, William, "Miracles and Statistics: The Casual Assumption of Independence (ASA Presidential Address)," *Journal of the American Statistical Association*, December 1988, Vol 83, No 409, 929-940, American Statistical Association
[10] LeCam, Lucien, "A Note on Metastatistics or 'An Essay Toward Stating a Problem in the Doctrine of Chances,'" *Synthese,* Vol. 36, No.1, Foundations of Probability and Statistics, Part I, 1977, pp133-160
[11] Lehmann, E.L. and George Casella, *Theory of Point Estimation,* 2nd Edition, Springer, 1998
[12] Lehmann, E.L. and Joseph P. Romano, *Testing Statistical Hypotheses,* 3rd Edition, Springer, 2005
[13] Stigler, Stephen, *The History of Statistics*, Belknap, 1986

Abbreviations and notations

CV, coefficient of variation
DD&R, death disability and retirement
MSEP, mean square error of prediction
OLS, ordinary least squares
OB, obstetrician
RMSEP, root mean square error of prediction

Biography of the Author

Keith Curley has done reserving for Deloitte & Touche, predictive modeling for Farmers Insurance, and casualty reinsurance pricing and underwriting for Swiss Reinsurance Company. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. In his work, he's been responsible for more bad models being built than anyone he knows.

The only statistics course he ever took was fortunately for him taught by David Freedman. Keith didn't appreciate it nearly as much as he should have at the time.

He can be reached at: keith_curley@hotmail.com