A NONLINEAR REGRESSION MODEL OF INCURRED BUT NOT REPORTED LOSSES by Scott Stelljes

Discussion by Jeffrey H. Adams, FCAS

The paper by Stelljes [1] the subject of this discussion is a welcome addition to the Casualty Actuarial Society literature on nonlinear regression for loss reserving. This discussion will predominantly concern a key assumption made in [1]. In particular, on page 361:

"Based on the assumption that the incremental pure premiums for different development intervals are independent, the variance of IBNR pure premium is the sum of the variances of the incremental pure premiums for the remaining development intervals."

It may be true that the *historical* incremental pure premiums can be considered independent, but it does not follow that the future *fitted* incremental pure premiums are independent. An analogous situation exists for ordinary linear regression, where the *hat* matrix provides for the covariance of the fitted values. Since the variance of the sum of random variables depends on covariance between the random variables, the variance of the reserve will depend on the covariance of the incremental IBNRs.

After providing a brief review on traditional nonlinear regression in section 2, the bulk of this discussion is concerned with two issues. First, modifying the methods of [1] to reflect covariance among the fitted values and is described in section 3. Second, there are times when a reliable insurance trend factor is not available. In such circumstances the actuary needs to derive the trend as part of the model, as in the model on page 359 of [1]. [1] succinctly describes the problems with such an approach. Section 4 discusses this latter model and shows simulation is not required to calculate confidence intervals. The last section, section 5 will discuss some miscellaneous issues.

2. BRIEF REVIEW OF NON LINEAR REGRESSION BASED ON THE BOOK BY MYERS, MONTGOMERY, VINING [4].

Let y be the dependent variable. Let x be a vector of explanatory variables, and **B** a vector of parameters. We then assume the following function:

(2.1) $y = f(x,\mathbf{B}) + \varepsilon$

ε are the errors and are assumed to be independent normal, with the means zero and constant variance σ^2.

(When fitting the data, this assumption should be checked to see if the error assumption is tenable since insurance claim data is often skewed or the errors may be heteroscedastic. [1] notes the heteroscedasticity and thus modifies the error term).

(2.2) $E(y) = f(x,\mathbf{B})$, denotes the expectation of y.

For example let y = $x_1*B_1/(B_3 + x_2*B_2)$ + error. The expectation of y is f(x,**B**) and is $x_1*B_1/(B_3 + x_2*B_2)$.

Typically, **B** is unknown and replaced with parameter estimates. Based on significance tests (see (2.7) below), it is possible fewer parameters are necessary. Insignificant parameters can be discarded and the function refit.

The parameters may be estimated through nonlinear least squares using the iterative Gauss-Newton method (or other methods).

The (asymptotic) variance covariance matrix of parameter estimators **b** is

(2.3) var(**b**) $\cong \hat{\sigma}^2 (\mathbf{D}^T \mathbf{D})^{-1}$

(2.4) Dij = $\partial f(x_i, \mathbf{B})/\partial B_j)$ is evaluated at final parameter estimates.

In (2.4) i refers to the vector of explanatory variables for observation i, and the j refers to the j'th parameter.

An estimate of the error variance is

(2.5) $\hat{\sigma}^2 = \hat{\varepsilon}^T \hat{\varepsilon} /(n–p)$, n is the number of observations fit, and p the number of parameters in **B**.

(2.6) $\hat{\varepsilon}$ = y- f(x,**b**)

(2.7) A parameter significance test is (**b** ÷ (standard error of the parameter)), which is asymptotically the normal distribution. The denominator is the square root of the appropriate element from the diagonal of the asymptotic variance covariance matrix of the parameters (2.3), or for weighted regression (2.11).

Let g(**b**) be a function of the parameter estimators and observations. Then

(2.8) E(g(**b**)) $\cong$ g(**B**)

The approximate (asymptotic) variance covariance matrix of g(**b**) is

(2.9) var(g(**b**)) $\cong$ $\mathbf{d}^T$ var(**b**) **d** , where

(2.10) $\mathbf{d}^T = [\, \partial g(\mathbf{B})/\partial B_1,..., \partial g(\mathbf{B})/\partial B_p]$ is evaluated at the estimated parameters.

Equations (2.9) emphasizes the discussion in section 1 regarding the non-independence of fitted values. (Take g(**b**) as the predicted values, then (2.9) can be used to derive the covariance of the predicted values).

If weighted non linear regression is used with a diagonal matrix **V** = var(y_i) = diag{$\sigma_1^2,.....,\sigma_n^2$}; $\sigma_i^2 = \sigma^2 / w_i$, and w_i are the weights then

(2.11) $\text{var}(\mathbf{b}) \cong (\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1}$

Weighted non linear regression may be used in the presence of heteroscedasticity.

Let **W** = diag{$w_1,...,w_n$}, then

(2.12) $\hat{\sigma}^2 = \hat{\varepsilon}^T (\mathbf{W}) \hat{\varepsilon} /(n–p)$ is the mean square error, and

(2.13) $\hat{\sigma}_i^2 = \hat{\sigma}^2 / w_i$, provides an estimate for **V**.

After the fit, the model assumptions must be checked. Checks include the usual regression error plots.

For loss reserving, errors should also be checked by accident quarter. The accident quarter fitted values by age, should be plotted against the dependent variable pure premium values. This will appraise the fit and the homogeneity of the accident quarters.

3. THE EQUATIONS APPLIED TO LOSS RESERVING WHEN EXTERNAL TREND IS USED

Let c_i represent the accident quarter exposures for observation i. In [1], the exposures are not inflation sensitive and external inflation factors were utilized to trend the incremental pure premiums. If the exposures are inflation sensitive, no additional inflation adjustment is generally required. (However, you may statistically test whether an additional trend factor is required by fitting (4.1) and (4.2). This will be discussed in section 4). If no additional inflation adjustment is required, the methods in section 4 may be applied, and no simulation is required for confidence intervals.

Start with the basic equation given in [1] for future observation(s) y, the future incremental pure premium(s). There is only one explicit explanatory variable x, the valuation age.

(3.0) $f(x,\mathbf{B}) = B_1 \exp(xB_2) + B_3\exp(xB_4)$

(3.1) $y = f(x,\mathbf{b}) + \varepsilon / (w^{1/2})$

Multiply (3.1) by exposure c gives

(3.2) $cy = cf(x,\mathbf{b}) + c\,\varepsilon /(w^{1/2})$

Taking the variance of (3.2) gives

(3.3) $\text{variance}(cy) = \text{variance}(cf(x,\mathbf{b})) + \text{variance}(c\,\varepsilon/w^{1/2})$

Now take g(**b**) = cf(x,**b**), and then apply (2.9), (2.10), and (2.11) giving,

(3.4) variance(c y) $\cong$ $\mathbf{d}^T$ var(**b**) **d** + $(c^2)\hat{\sigma}^2/w$

For equation (3.4) use equation (2.12) to valuate $\hat{\sigma}^2$.

The second term on right hand side of (3.4) is a diagonal matrix, diag $=\{c_i^2\hat{\sigma}_i^2\}$.

The expectation of (3.2) is

(3.5) E(cy) $\cong$ c f(x,**b**) = g(**b**)

(3.5) provides the vector of means, and (3.4) provides the variance covariance matrix, for a multinormal distribution. It is that distribution that must be sampled to provide an IBNR array. Then, each IBNR value is multiplied by the simulated trend factor, as explained in [1]. Doray [6] page 648 explains a method for simulating the multinormal. The simulations in this discussion were performed in R version 2.4.1 (2006-12-18) (C) 2006 The R Foundation for Statistical Computing.

Exhibit 1 displays a summary and the key results of this discussion. The first four columns are reproduced from Table 3.2.1 of [1]. Columns (7) and (8) are calculated assuming all off diagonal elements of the matrix of (3.4) are set to zero, and then doing 1000 simulations of the multinormal distribution, after which simulated trend factors (using the [1] trending approach) are applied. That is essentially the method in [1]. Columns (5) and (6) are also based on 1000 simulations but incorporate covariance terms of the full matrix (3.4). Although the expected total IBNR are essentially the same in columns (3), (5), (7), and the standard deviations of the total IBNR of (4) and (8) are essentially the same, the standard deviations of the total IBNR in column (6) is significantly higher. Column (6) is the appropriate standard deviation.

Exhibit 2 column (5) and (10) provides a partial listing of the vector of 780 means (3.5) used to simulate the pre- trended IBNRs (these are at calendar quarter 40 level). Exhibit 3 provides a portion of the 780 by 780 variance covariance matrix (3.4).

Accident quarter variances are estimated as a by-product of simulating the entire southeast portion of the loss “triangle”, and should not add up to the variance of total IBNR.

4. THE EQUATIONS APPLIED TO LOSS RESERVING WHEN NO EXTERNAL TREND IS USED

Let y be the incremental losses divided by an inflation or non inflation sensitive exposure base. We use the rejected trend model on page 359 of [1] shown as (4.1) below. (See section 5 paragraph g regarding the extrapolation issue briefly discussed in [1]).

Let B_5 be the trend, u the calendar quarter, and age be the accident quarter valuation age. If an inflation sensitive exposure base is used, B_5 provides for excess trend. (I have assumed the same weights as in [1]. Normally the appropriate weights need to be individually selected for each model).

After the fit, significance levels of the parameters can be checked. If B_5 is not significant then there is no trend other than what is contemplated by the exposure base and age, then $\exp(uB_5)$ may be dropped from equation (4.1) and the model refit.

(4.1) $f(\text{age},u,\mathbf{B}) = (B_1 \exp(B_2\text{age}) + B_3\exp(B_4\text{age}))\exp(uB_5)$

(Denote u and age by the explanatory variable vector x.)

(4.2) $y = f(x,\mathbf{B}) + \varepsilon / (w^{1/2})$

Assume (4.1), (4.2) have been fit to the historical incremental pure premiums. The focus will now be on the future incremental pure premiums.

Using the estimated parameters **b** in (4.2), multiply (4.2) by c to get the future incremental losses:

(4.3) $c\, y = cf(x,\mathbf{b}) + c\, \varepsilon / (w^{1/2})$

Taking the variance of (4.3) gives

(4.4) $\text{variance}(cy) = \text{variance}(cf(x,\mathbf{b})) + \text{variance}(c\, \varepsilon/w^{1/2})$

Now take $g(\mathbf{b}) = cf(x,\mathbf{b})$ and apply (2.9), (2.10), and (2.11) giving

(4.5) $\text{variance}(c\, y) \cong \mathbf{d}^T \text{var}(\mathbf{b})\, \mathbf{d} + (c^2)\hat{\sigma}^2/ w$

For equation (4.5), use equation (2.12) to evaluate $\hat{\sigma}^2$. The second term on the right hand side of (4.5) is a diagonal matrix, $\text{diag} =\{c_i^{\,2}\, \hat{\sigma}_i^{\,2}\}$.

The expectation of (4.3) are the expected future incremental losses
(4.6) $E(cy) \cong g(x,\mathbf{b})$

Now form the sum of the future incremental losses denoted by R for reserve giving

(4.7) $R = \Sigma\, cy$, the sum taken over the southeast portion of the loss "triangle".

The expectation of R is the mean total reserve and is given by

(4.8) $E(R) \cong \Sigma\, g(x,\mathbf{b})$, the sum taken over the southeast portion of the loss "triangle".

The variance of R denoted by var(R) is

(4.9) $\text{var}(R) = \Sigma\ \Sigma\ \text{cov}(c_i y_j,\ c_j,y_j)$

In (4.9), the sum is taken over all future observations (i,j) in the southeast portion of the loss triangle. The covariance terms in (4.9) are from (4.5).

Using the normality assumption, the confidence interval for the reserve becomes

(4.10) $E[R] \pm \ z\cdot var(R)^{1/2}$, z is the appropriate standard normal value.

Applying section 4 equations to Exhibit A data from [1] provides the following:

The estimated parameters for b_1, b_2, b_3 ,b_4, b_5 are 2.364885501 -0.077678377 21.611842502 -0.566532596 0.009735732. The MSE is 2759171.

The parameter variance covariance matrix derived from equation (2.11) is

	b_1	b_2	b_3	b_4	b_5
b_1	0.308171765	-3.248550e-03	2.13645749	-2.257342e-02	-2.046082e-03
b_2	-0.003248550	8.684792e-05	-0.01130499	4.756396e-04	-5.492841e-06
b_3	2.136457489	-1.130499e-02	31.07109676	-2.940411e-01	-1.983273e-02
b_4	-0.022573418	4.756396e-04	-0.29404108	6.488308e-03	-1.960661e-05
b_5	-0.002046082	-5.492841e-06	-0.01983273	-1.960661e-05	3.016662e-05

The parameter standard deviations are the square roots of the diagonal:

0.555132205, 0.009319223, 5.574145384, 0.080550033, 0.005492415 .

The 95% confidence intervals using t(.025,590-5) are

	b_1	b_2	b_3	b_4	b_5
Lower	1.274593134	-0.095981606	10.664076447	-0.724735018	-0.001051543
Upper	3.45518287	-0.05937519	32.55962508	-0.40833007	0.02052296

The trend parameter b_5 is just shy of significance at the 95% level, but will be used.

Exhibit 1, column (9) displays the estimated IBNRs and corresponds to equation (4.6) summed over the accident quarter's IBNRs. The IBNR, by accident quarter and in total, compare favorably with columns (3), (5), and (7), although a bit higher probably due to the higher trend (.0097 versus .005 used by the author). The total IBNR standard deviation calculated using the square root of (4.9) is 3782848, and using (4.10) with z =1.96 provides a 95% reserve confidence interval of : (25254267 , 40083031).

Simulation may also be used to determine confidence intervals. (4.6) provides the vector of means, and (4.5) provides the variance covariance matrix for a multinormal distribution. Exhibit 2 columns (4) and (9) provides a partial listing of the vector of 780 means that may be used to simulate the IBNRs. Exhibit 2 columns (4) and (5) are not comparable, since column (4) already includes trend, while column (5) is still at calendar quarter 40 level. The same applies for columns (9) and (10).

If confidence intervals are desired by accident quarter, the multinormal distribution can be simulated. Accident quarter variances are estimated as a by-product of simulating the entire southeast portion of the triangle, and of course will not add up to the variance of total IBNR. Alternatively, equation (4.9) may be used limiting the summation to the appropriate accident quarter ages. For example, consider accident quarter 4. The portion of the variance covariance matrix (4.5) corresponding to the fourth accident quarter's three IBNR elements is

age	38	39	40
38	605880842	3367957	3291128
39	3367957	582719205	3225207
40	3291128	3225207	560981739

Adding up these nine figures provide the variance for the fourth accident quarter IBNR, which is 1769350371, and a standard deviation of 42064. The diagonal elements are the individual IBNR variances. For example, the variance of the incremental IBNR for accident quarter 4 age 39 is 582719205. Exhibit 1, column (10) displays the standard deviations for the accident quarter IBNRs calculated in such a fashion.

Exhibit 4 displays a partial portion of the variance covariance matrix as calculated in (4.5).

5. MISCELLANEOUS ISSUES

a) On page 354 of [1] "Furthermore, Narayan...remarks that dollar based regression models do not take into account changing levels of exposure. This is a serious flaw because the amount of loss in an accident period is highly correlated to the number of earned exposures." I would concur with this assessment and would suggest incorporating exposure as an explanatory variable in GLM or regression methods, or perhaps an offset in GLM. England and Verrall [2] discuss incorporating exposure in stochastic loss reserving. Incorporating exposure should act to reduce the number of parameters in a GLM or regression type model.

b) Page 231 of [1] formula (2.3.1) should have included the weight function in the minimization since weighted least squares is being performed i.e minimize

$$\sum_{i=1}^{n} w_i(y_i - f(x_i,\mathbf{B}))^2$$

This must have been a typo, and conversations with Stelljes has confirmed this.

c) Page 371 of [1] "Some of the models could be applied to cumulative instead of incremental data." (Page 370 in [1] does note that if autocorrelation occurs other models exist). In my limited experience fitting a single curve to an array of cumulative accident year or report year data results in autocorrelation which violates linear and nonlinear regression assumptions. In addition, heteroscedasticity tends to occur. A plot of the cumulative data for each incurred year versus the fitted curve will help detect autocorrlation as well as detect non-homogeneity of the accident years. A further problem with fitting cumulative data occurs when the estimated ultimate pure premium for a particular incurred year is below the actual emerged pure premium for that year. One way around these problems may be to fit a separate curve to each accident year as in Clark [3] and Kazenski[5]. Kazenski asserts he has detected no autocorrelation using such an approach.

d) Traditional nonlinear regression assumes the error terms are normal which is a symmetric distribution with a range $-\infty$ to $+\infty$. Incremental pure premium data may actually be skewed and can hardly ever be highly negative, therefore, using the normal distribution is approximation at best.

e) Page 358 of [1] formula (2.2.2) should use the square root of the weight, not just the weight. This appears to have been a typo, and conversations with Stelljes has confirmed this. See equation (3.1) above.

f) A note regarding the parameter estimates and the data used for fitting.

[1] excluded the first evaluation of an accident quarter and all evaluations prior to the twenty first calendar quarter when fitting the equation. The same was done in this discussion, both in section 3 and section 4 and section 5 paragraph g. Also, Stelljes [1] has informed me the raw incremental pure premiums (Exhibit A in [1]) are first trended to calendar quarter 40 using a constant trend factor of exp(.005) per calender quarter prior to fitting them. The same was done for the section 3 calculations. Using Exhibit A data (kindly supplied by Stelljes as a computer file), I was able to replicate the following from [1]: parameters on page 362, matrix inversion of $(F^{'}WF)^{-1}$ on page 363, the confidence interval of (-40259,56186) for accident quarter 2 on page 364, and finally, the mean square error of 2987236 on page 364. The parameters in [1] on page 362: 3.1994, -.0754, 29.4446, -.5480 correspond to estimates of B_1, B_2, B_3, B_4 in equation (3.0) of this discussion and are used in section 3.

Keeping within the limited scope of this paper, various diagnostics for the section 4 or section 5 paragraph g fittings have not been performed. Those diagnostic procedures are widely discussed in nonlinear regression texts and should be applied in practice. No claim is made that the fitted parameters are actually the best. Nonlinear regression requires initial starting values, and there is no guarantee the solution will converge, let alone converge to the global minimum mean square error.

g) Extrapolating

In section 4, if B_5 is significant, formula (4.5) extrapolates beyond the fitting space, (in the example for calendar quarters past 40). Discussions with Stelljes, and page 359 in [1] cautions against extrapolating. Pages 86-88 in [4] provides for a confidence interval of a “future observed response”, and seems silent on the issue of extrapolating. Using the approaches in section 4, an alternative model is:

(5.1) $f(\text{age},\text{aqtr},\mathbf{B}) = (B_1 \exp(B_2\text{age}) + B_3\exp(B_4\text{age}))\exp(B_5\ \text{aqtr})$

where aqtr the accident quarter. Using the same data as in section 4, results from (5.1) were very close to those of (4.1), but even (5.1) will also extrapolate beyond the fitting space when B_5 is significant.

If the variances as calculated by (4.5) appear unreasonable in the extrapolated region, perhaps a ceiling or floor may be required after some point. This seems to be an area requiring further research.

h) On the one hand, the approach in [1] (and section 3), assume the availability of an external trend and that the estimates of the parameter in the model are independent of the trend. On the other hand, it's nonlinear regression model is not extrapolated, only the trend needs to be extrapolated. The section 4 model allows for estimation of internal trend and allows for covariance among all the parameters (including trend), but does require extrapolation when B_5 is significant. Neither method is perfect.

REFERENCES

[1] Scott Stelljes, “A Nonlinear Regression Model of Incurred But Not Reported Losses”, Casualty Actuarial Society Forum, Fall 2006 Featuring Reserves Call Papers, pp. 353-377.

[2] Peter D. England and Richard J. Verrall, “A flexible Framework for Stochastic Claims Reserving”, Proceedings of the Casualty Actuarial Society 2001 Volume LXXXVIII, pp. 1-38.

[3] Harold E. Clarke , “Recent Developments in Reserving for Losses in the London Reinsurance Market”, Proceedings of the Casualty Actuarial Society 1988 Volume LXXV, pp. 1-48.

[4] Raymond H. Myers, Douglas C. Montgomery, G. Geoffrey Vining, Generalized Linear Models With Applications in Engineering and the Sciences, 2002 John Wiley and Sons, Inc., pp. 63-92 discuss nonlinear regression. This book also provides accessible explanations of linear regression, GLM, GEE and GAM.

[5] Paul M. Kazenski, “A Nonlinear Modeling Approach to Assessing the Accuracy of Property-Liability Insurer Loss Reserves”, University of Hawaii - Manoa, February 1994.

[6] Louis Doray, “IBNR Reserve Under a Loglinear Location-Scale Regression Model”, Casualty Actuarial Society Forum Spring 1994, Volume Two, pp. 607-652.

EXHIBIT I

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				Discussion	Discussion			Discussion	Discussion
				Paper	Paper			Paper	Paper
		[1]	[1]	section 3	section 3	Check [1]	Check [1]	section 4	section 4
Accident		Expected	Standard	Expected	Standard	Expected	Standard	Expected	Standard
Quarter	Exposure	Value	Deviation	Value	Deviation	Value	Deviation	Value	Deviation
2	50,801	8,190	24,518	7,489	24,719	7,616	24,912	8,010	23,601
3	51,187	16,643	35,835	16,767	36,204	16,816	33,944	16,872	33,922
4	51,146	26,310	44,192	28,985	45,415	24,058	44,909	26,443	42,064
5	51,527	36,541	51,941	33,429	51,328	37,975	52,022	37,157	49,402
6	52,348	49,099	58,839	49,399	59,053	48,470	60,416	49,380	56,446
7	52,480	61,528	65,232	60,100	69,592	60,716	65,327	62,191	62,790
8	53,148	75,340	71,800	75,401	72,824	76,815	72,159	76,954	69,266
9	53,924	91,671	78,552	93,025	79,352	90,003	80,072	93,486	75,738
10	54,403	109,065	85,433	112,127	88,895	108,506	87,839	111,208	81,966
11	54,557	124,874	91,436	126,736	94,084	125,926	91,494	129,920	87,919
12	55,083	144,622	96,258	149,578	100,674	141,166	94,407	151,342	94,221
13	55,292	168,450	103,341	175,839	107,340	166,273	101,628	173,891	100,296
14	55,899	192,189	108,233	189,828	117,084	183,868	108,754	199,906	106,864
15	56,067	215,948	115,108	218,495	119,945	218,185	113,886	226,736	113,100
16	57,025	247,643	123,187	245,486	126,152	249,288	119,610	259,542	120,393
17	57,071	279,736	129,481	277,633	136,171	279,801	129,502	291,148	126,815
18	57,317	311,248	134,933	305,717	133,675	311,388	134,122	326,584	133,667
19	57,907	346,819	143,714	346,509	143,603	336,674	140,549	367,375	141,225
20	58,285	388,878	149,405	383,582	151,327	387,150	152,150	410,598	148,789
21	59,096	433,974	157,772	435,640	164,002	427,185	163,959	461,162	157,349
22	59,193	479,592	165,473	474,623	173,326	478,486	161,765	510,590	165,192
23	59,524	530,342	173,337	524,379	177,440	528,747	169,566	566,470	173,823
24	59,745	583,879	177,894	585,037	183,270	573,480	175,996	626,235	182,747
25	60,427	645,944	188,083	652,774	204,720	639,599	194,014	696,579	193,112
26	60,155	705,701	195,557	709,139	199,170	706,895	193,614	761,641	202,285
27	60,568	776,239	207,953	776,419	222,299	788,439	203,526	841,356	213,588
28	60,708	852,632	215,059	863,905	225,281	844,677	209,276	924,383	225,219
29	60,262	925,896	222,578	921,837	235,006	924,073	229,328	1,005,182	236,460
30	60,606	1,012,197	233,755	1,015,105	247,787	1,016,063	247,362	1,107,100	250,821
31	60,580	1,109,304	251,368	1,099,773	268,201	1,094,682	247,988	1,212,155	265,684
32	60,648	1,213,637	258,802	1,227,733	267,445	1,221,054	254,047	1,330,473	282,513
33	61,159	1,344,114	277,079	1,325,154	281,107	1,348,687	269,254	1,473,989	302,862
34	61,462	1,492,000	292,032	1,470,864	296,064	1,509,526	298,480	1,633,463	325,285
35	61,934	1,660,873	312,021	1,664,619	328,967	1,665,426	304,419	1,826,677	351,853
36	61,716	1,858,275	333,112	1,867,446	348,580	1,863,920	337,684	2,040,965	380,446
37	61,837	2,123,409	361,113	2,128,841	352,122	2,140,963	343,229	2,330,037	417,181
38	62,285	2,514,004	394,000	2,499,739	392,466	2,521,633	404,654	2,738,893	466,097
39	62,728	3,055,695	450,062	3,069,822	465,666	3,061,935	443,104	3,329,815	532,473
40	63,180	3,892,584	522,958	3,892,268	515,975	3,878,801	501,528	4,232,741	633,498
Totals		30,105,085	1,350,093	30,101,242	2,210,162	30,104,966	1,348,733	32,668,649	3,782,848

Exhibit 2

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Section 4	Section 3				Section 4	Section 3
			Incremental	Incremental				Incremental	Incremental
aqtr	age	expos	IBNR	IBNR	aqtr	age	expos	IBNR	IBNR
2	40	50801	8010	7964	40	2	63180	846121	795568
3	39	51187	8723	8653	40	3	63180	553745	520639
3	40	51187	8150	8024	40	4	63180	381677	357292
4	38	51146	9420	9323	40	5	63180	278847	258771
4	39	51146	8801	8646	40	6	63180	215972	198022
4	40	51146	8223	8018	40	7	63180	176251	159387
5	37	51527	10256	10128	40	8	63180	150042	133789
5	38	51527	9583	9392	40	9	63180	131801	115967
5	39	51527	8953	8710	40	10	63180	118339	102858
5	40	51527	8365	8077	40	11	63180	107812	92679
6	36	52348	11261	11095	40	12	63180	99153	84382
6	37	52348	10522	10289	40	13	63180	91736	77348
6	38	52348	9831	9542	40	14	63180	85192	71207
6	39	52348	9185	8849	40	15	63180	79299	65733
6	40	52348	8582	8206	40	16	63180	73920	60784
7	35	52480	12202	11994	40	17	63180	68967	56268
7	36	52480	11400	11123	40	18	63180	64381	52123
7	37	52480	10651	10315	40	19	63180	60120	48304
7	38	52480	9952	9566	40	20	63180	56153	44776
7	39	52480	9298	8871	40	21	63180	52454	41513
7	40	52480	8687	8227	40	22	63180	49002	38491
8	34	53148	13355	13098	40	23	63180	45780	35692
8	35	53148	12478	12147	40	24	63180	42771	33098
8	36	53148	11658	11264	40	25	63180	39960	30693
8	37	53148	10893	10446	40	26	63180	37335	28463
8	38	53148	10177	9688	40	27	63180	34882	26395
8	39	53148	9509	8984	40	28	63180	32591	24478
8	40	53148	8884	8332	40	29	63180	30450	22700
9	33	53924	14645	14330	40	30	63180	28450	21051
9	34	53924	13683	13289	40	31	63180	26581	19522
9	35	53924	12784	12324	40	32	63180	24835	18104
9	36	53924	11944	11429	40	33	63180	23203	16789
9	37	53924	11160	10599	40	34	63180	21679	15570
9	38	53924	10427	9829	40	35	63180	20255	14439
9	39	53924	9742	9115	40	36	63180	18925	13391
9	40	53924	9102	8453	40	37	63180	17682	12418
10	32	54403	15968	15589	40	38	63180	16520	11516
10	33	54403	14919	14457	40	39	63180	15435	10680
10	34	54403	13939	13407	40	40	63180	14421	9904
10	35	54403	13024	12433					
10	36	54403	12168	11530					
10	37	54403	11369	10693					
10	38	54403	10622	9916					
10	39	54403	9924	9196					
10	40	54403	9273	8528					

Exhibit 3

aqtr		2	3	3	4	4	4
aqtr	age	40	39	40	38	39	40
2	40	602,854,869	3,112,619	3,014,973	3,204,500	3,110,126	3,012,558
3	39	3,112,619	631,054,179	3,136,270	3,334,387	3,235,699	3,133,757
3	40	3,014,973	3,136,270	607,458,435	3,228,848	3,133,757	3,035,448
4	38	3,204,500	3,334,387	3,228,848	655,671,475	3,331,716	3,226,262
4	39	3,110,126	3,235,699	3,133,757	3,331,716	630,546,122	3,131,247
4	40	3,012,558	3,133,757	3,035,448	3,226,262	3,131,247	606,969,439
5	37	3,319,123	3,454,230	3,344,343	3,557,965	3,451,463	3,341,664
5	38	3,228,371	3,359,226	3,252,901	3,459,481	3,356,535	3,250,295
5	39	3,133,294	3,259,802	3,157,102	3,356,535	3,257,191	3,154,573
5	40	3,034,999	3,157,102	3,058,060	3,250,295	3,154,573	3,055,610
6	36	3,458,533	3,599,973	3,484,812	3,708,826	3,597,089	3,482,020
6	37	3,372,008	3,509,268	3,397,630	3,614,656	3,506,457	3,394,908
6	38	3,279,810	3,412,750	3,304,731	3,514,602	3,410,016	3,302,084
6	39	3,183,218	3,311,742	3,207,405	3,410,016	3,309,089	3,204,836
6	40	3,083,357	3,207,405	3,106,785	3,302,084	3,204,836	3,104,296
7	35	3,546,852	3,692,652	3,573,802	3,805,149	3,689,694	3,570,939
7	36	3,467,254	3,609,050	3,493,599	3,718,178	3,606,160	3,490,801
7	37	3,380,511	3,518,116	3,406,197	3,623,770	3,515,299	3,403,469
7	38	3,288,080	3,421,355	3,313,064	3,523,465	3,418,615	3,310,410
7	39	3,191,245	3,320,093	3,215,493	3,418,615	3,317,433	3,212,917
7	40	3,091,132	3,215,493	3,114,619	3,310,410	3,212,917	3,112,124
8	34	3,663,703	3,815,168	3,691,541	3,932,367	3,812,112	3,688,584
8	35	3,591,998	3,739,655	3,619,291	3,853,584	3,736,659	3,616,392
8	36	3,511,387	3,654,989	3,538,068	3,765,505	3,652,061	3,535,234
8	37	3,423,540	3,562,897	3,449,553	3,669,896	3,560,044	3,446,790
8	38	3,329,933	3,464,905	3,355,235	3,568,314	3,462,129	3,352,547
8	39	3,231,865	3,362,353	3,256,422	3,462,129	3,359,660	3,253,813
8	40	3,130,478	3,256,422	3,154,264	3,352,547	3,253,813	3,151,737
9	33	3,778,996	3,936,224	3,807,710	4,058,262	3,933,071	3,804,660
9	34	3,717,196	3,870,872	3,745,440	3,989,782	3,867,772	3,742,440
9	35	3,644,444	3,794,256	3,672,136	3,909,849	3,791,217	3,669,194
9	36	3,562,656	3,708,354	3,589,726	3,820,484	3,705,384	3,586,851
9	37	3,473,527	3,614,918	3,499,919	3,723,479	3,612,023	3,497,116
9	38	3,378,553	3,515,495	3,404,224	3,620,414	3,512,679	3,401,497
9	39	3,279,053	3,411,446	3,303,968	3,512,679	3,408,713	3,301,321
9	40	3,176,185	3,303,968	3,200,318	3,401,497	3,301,321	3,197,755
10	32	3,861,688	4,023,507	3,891,031	4,149,543	4,020,284	3,887,914
10	33	3,812,564	3,971,189	3,841,533	4,094,311	3,968,008	3,838,456

Exhibit 4

aqtr		2	3	3	4	4	4
aqtr	age	40	39	40	38	39	40
2	40	557,010,999	3,076,909	3,007,116	3,181,310	3,115,232	3,044,109
3	39	3,076,909	583,099,003	3,141,809	3,324,615	3,255,594	3,181,296
3	40	3,007,116	3,141,809	561,346,864	3,249,349	3,183,066	3,111,471
4	38	3,181,310	3,324,615	3,249,349	605,880,842	3,367,957	3,291,128
4	39	3,115,232	3,255,594	3,183,066	3,367,957	582,719,205	3,225,207
4	40	3,044,109	3,181,296	3,111,471	3,291,128	3,225,207	560,981,739
5	37	3,309,854	3,459,950	3,381,707	3,580,453	3,506,202	3,426,258
5	38	3,248,025	3,395,306	3,319,801	3,513,545	3,442,061	3,364,808
5	39	3,180,038	3,324,227	3,251,437	3,439,980	3,371,218	3,296,648
5	40	3,106,968	3,247,836	3,177,730	3,360,919	3,294,831	3,222,924
6	36	3,465,072	3,623,357	3,541,518	3,750,844	3,673,110	3,589,404
6	37	3,408,292	3,563,915	3,484,814	3,689,234	3,614,287	3,533,269
6	38	3,344,022	3,496,650	3,420,281	3,619,536	3,547,345	3,469,022
6	39	3,273,491	3,422,845	3,349,190	3,543,076	3,473,604	3,397,971
6	40	3,197,796	3,343,649	3,272,681	3,461,044	3,394,244	3,321,287
7	35	3,571,152	3,735,590	3,651,328	3,868,496	3,788,379	3,702,096
7	36	3,521,668	3,683,685	3,602,101	3,814,585	3,737,218	3,653,557
7	37	3,463,277	3,622,482	3,543,587	3,751,066	3,676,435	3,595,423
7	38	3,397,363	3,553,427	3,477,220	3,679,434	3,607,512	3,529,163
7	39	3,325,166	3,477,814	3,404,282	3,601,029	3,531,782	3,456,095
7	40	3,247,795	3,396,803	3,325,924	3,517,048	3,450,436	3,377,400
8	34	3,708,164	3,880,415	3,793,016	4,020,164	3,936,970	3,847,360
8	35	3,667,175	3,837,284	3,752,499	3,975,213	3,894,734	3,807,677
8	36	3,615,573	3,783,079	3,700,941	3,918,824	3,841,063	3,756,609
8	37	3,554,928	3,719,439	3,639,966	3,852,693	3,777,639	3,695,822
8	38	3,486,652	3,647,838	3,571,034	3,778,341	3,705,975	3,626,814
8	39	3,412,007	3,569,596	3,495,454	3,697,135	3,627,427	3,550,925
8	40	3,332,125	3,485,894	3,414,395	3,610,295	3,543,208	3,469,358
9	33	3,846,268	4,026,672	3,936,130	4,173,643	4,087,347	3,994,380
9	34	3,815,764	3,994,377	3,906,352	4,139,764	4,056,121	3,965,605
9	35	3,772,676	3,948,954	3,863,507	4,092,331	4,011,365	3,923,380
9	36	3,718,786	3,892,266	3,809,438	4,033,270	3,954,991	3,869,594
9	37	3,655,700	3,825,989	3,745,804	3,964,313	3,888,716	3,805,947
9	38	3,584,857	3,751,625	3,674,094	3,887,013	3,814,085	3,733,966
9	39	3,507,548	3,670,523	3,595,643	3,802,764	3,732,479	3,655,021
9	40	3,424,928	3,583,889	3,511,646	3,712,812	3,645,138	3,570,336

Chain Ladder Reserve Risk Estimators

Daniel M. Murphy, FCAS, MAAA

Abstract

Mack (1993) [2] and Murphy (1994) [4] derived analytic formulas for the reserve risk of the chain ladder method. In 1999, Mack [3] gave a recursive version of his formula for total risk. This paper provides the recursive versions of Mack's formulas for process risk and parameter risk and shows that they agree with the formulas in Murphy [4] except for a parameter risk cross-product term. MSE is decomposed into variance and bias components. For the unbiased all-year weighted average link ratios in Mack [2] and Murphy [4] the MSE decomposition in this paper yields formulas that agree with Murphy [4]. For well-behaved triangles the difference between Mack and Murphy parameter risk estimates should be negligible. The concepts are illustrated with an example using data from Taylor and Ashe [5].

Keywords: chain ladder; reserve risk; Mack; mean square error; parameter risk; bias; benchmarks.

Introduction

Mack [1] derived formulas for the chain ladder reserve risk when the age-to-age factors are based on the all-year weighted average. Murphy [4] derived recursive formulas for the chain ladder reserve risk under assumptions that are equivalent to Mack's. The authors' formulas yield different results, for reasons to be discussed herein.

Mack [3] presented a recursive version of the total risk formula. In Section 1 we show recursive formulas for process risk and parameter risk not shown in [3]. We compare them with Murphy's recursive formulas using Mack's notation and note that the difference between the Mack and Murphy reserve risk estimates lies in the parameter risk component.

Mack's reserve risk is measured by the mean square error (MSE). Murphy's reserve risk is measured by total variance. Although MSE is employed in many authors' actuarial research, a mathematically precise definition, particularly as regards reserve risk, is not readily found in the literature. In Section 2 we present a definition of mean square error using the calculus of probability density functions. We will see that MSE can be decomposed into three terms: process risk, parameter risk, and bias. Since total variance is the sum of process variance and parameter variance, the difference between the Mack and Murphy reserve risk measures is bias. A separate mathematical manipulation, this time of parameter risk, yields a recursive formula that agrees with Murphy's. Most of the mathematics will be relegated to the appendix.

Bias is ubiquitous in actuarial practice. When an actuary employs benchmark or industry factors in reserving, there arises a very real potential for bias. Yet biased development factors can yield estimated ultimates with smaller MSE than ultimates based solely on a company's own experience, especially when that experience lacks sufficient credibility. The role that bias plays in estimating reserves and reserve risk has received little attention in the literature.

In Section 3 we illustrate the above with an example using the Taylor/Ashe data analyzed by Mack [2] and elsewhere in the literature. We expand on the discussion by exploring the data a bit more with the regression perspective of [1]. We show how a simple graphical diagnostic leads to a different deterministic method with a not insignificantly smaller MSE.

1 Recursive Reserve Risk Formulas

We start with the model of loss development presented in [2] and [4], employing Mack's notation.

Suppose we are given a triangle of cumulative loss amounts C_{ij} by accident year *i* and development age *j*, $1 \leq i,j \leq I$. The triangle is assumed to be sufficiently large that age *I* can be considered "ultimate." Note that for a given accident year *i* the triangle's current diagonal observation has column index $j = I + 1 - i$, a useful fact to keep in mind when reading Mack's formulas. The triangle in hand can be considered a sample from a theoretical set of random variables $D = \{C_{ij} \mid 1 \leq i \leq I, 1 \leq j \leq I + 1 - i\}$.

Under the assumptions[1]

(CL1) $E(C_{i,k+1} \mid D) = C_{ik} f_k$,

(CL2) $Var(C_{i,k+1} \mid D) = C_{ik}\sigma_k^2$ for unknown parameters σ_k^2, $1 \leq i \leq I$, $1 \leq k \leq I-1$,

and (CL3) accident years are independent,

Mack derived the following closed-form formula for the estimate of the mean square error (MSE) of the chain ladder estimated ultimate losses:

$$\hat{mse}(\hat{C}_{iI}) = \hat{C}_{iI}^2 \sum_{k=I+1-i}^{I-1} \frac{\hat{\sigma}_k^2}{\hat{f}_k^2} \left(\frac{1}{\hat{C}_{ik}} + \frac{1}{\sum_{j=1}^{I-k} C_{jk}} \right) \tag{1}$$

where

[1] Assumptions from Mack [2], pp. 214-217, which agree with those of Model IV in [4]; labeling from Mack [3].

- $\hat{\sigma}_k^2 = \dfrac{1}{I-k-1}\sum_{i=1}^{I-k} C_{ik}\left(\dfrac{C_{i,k+1}}{C_{ik}} - \hat{f}_k\right)^2$ for $1 \le k \le I-2$; (2)
- $\hat{\sigma}_{I-1}^2$ is judgmentally selected[2];
- the link ratio estimates are calculated using the all-year weighted averages

$$\hat{f}_k = \frac{\sum_{j=1}^{I-k} C_{j,k+1}}{\sum_{j=1}^{I-k} C_{jk}} ;$$

- accident year losses for future ages ($k > I + 1 - i$) are predicted using the chain ladder method

$$\hat{C}_{ik} = C_{i,I+1-i}\hat{f}_{I+1-i}\cdots\hat{f}_{k-1} ;$$

- and, despite being scalars and not estimates, the current diagonal elements are granted "hats" ($C_{i,I+1-i} = \hat{C}_{i,I+1-i}$), which makes the formula more concise.

Formula (1) is a combination of process risk and parameter risk (a.k.a., "estimation error," but more about that later).

We next look at recursive versions of the process and parameter risk components of equation (1). In the remainder of this paper unless otherwise noted it is understood that all expectations are conditional expectations, conditional on the triangle *D*. Also, depending on the context, sometimes it will be convenient to refer to "risk" in terms of variance and sometimes in terms of standard deviation.

1.1 Process Risk

It can be seen in [2] that Mack's closed-form estimator[3] for the process risk component of equation (1) is

$$V\hat{a}r(C_{iI}) = \hat{C}_{iI}^2 \sum_{k=I+1-i}^{I-1} \frac{\hat{\sigma}_k^2 \big/ \hat{f}_k^2}{\hat{C}_{ik}} . \tag{3}$$

Mack based the derivation of equation (3) on the recursive property[4] of process risk

$$\mathrm{Var}(C_{ik}) = E(C_{i,k-1})\sigma_{k-1}^2 + \mathrm{Var}(C_{i,k-1})\, f_{k-1}^2 \tag{4}$$

[2] Mack suggests $\hat{\sigma}_{I-1}^2 = \min(\hat{\sigma}_{I-2}^4 \big/ \hat{\sigma}_{I-3}^2, \min(\hat{\sigma}_{I-3}^2, \hat{\sigma}_{I-2}^2))$.

[3] p. 218; the hat notation in (3) shows that $\mathrm{V\hat{a}r}(C_{iI})$ is an estimator of the variance Var(C_{iI}).

[4] Ibid.

for ages *k* beyond the first future diagonal for the given accident year *i*. For the first future diagonal, (4) reduces to

$$\mathrm{Var}(C_{ik}) = E(C_{i,k-1})\sigma_{k-1}^2 = C_{i,I+1-i}\sigma_{k-1}^2,$$

which is assumption CL2 above.

We obtain a recursive version of Mack's estimator for process risk by substituting estimators of the unknowns in (4):

$$ProcessRisk\hat{}_{ik} = \begin{cases} \hat{f}_{k-1}^2 ProcessRisk\hat{}_{i,k-1} + \hat{C}_{i,k-1}\hat{\sigma}_{k-1}^2 & \text{for } k > I+2-i \\ C_{i,I+1-i}\hat{\sigma}_{k-1}^2 & \text{for } k = I+2-i\,. \end{cases} \tag{5}$$

The process risk estimator in (5) has the same form as Murphy's recursive estimator[5]. To demonstrate that the authors' formulas are identical in substance as well as form, it remains to be shown that Mack and Murphy have the same formula for the variance estimator $\hat{\sigma}_k^2$ (both authors' models yield weighted average link ratios).

Mack's formula (2) for the variance estimator[6] can be rewritten as

$$\hat{\sigma}_k^2 = \frac{1}{I-k-1}\sum_{i=1}^{I-k}\left(C_{i,k+1} - \hat{f}_k C_{ik}\right)^2.$$

So $\hat{\sigma}_k^2$ is the sum of the squared deviations of losses at the end of the development period from the chain ladder predictions given the losses at the beginning of the period, all divided by *n*-1, where *n* is the number of terms in the summation. This is the formula for residual variance when the regression line (the paradigm in Murphy [4]) is determined by only a slope parameter, no intercept. Thus, the Mack and Murphy formulas for the variance estimator, and in turn for process risk, are equivalent.

1.2 Parameter Risk

It can be seen in Mack [2] that the author's closed-form estimator for parameter risk[7] is

$$ParameterRisk\hat{}_{ik} = \hat{C}_{ik}^2 \sum_{j=I+1-i}^{k-1} \frac{\hat{\sigma}_j^2}{\hat{f}_j^2} \frac{1}{\sum_{r=1}^{I-j} C_{rj}}. \tag{6}$$

This can be reformulated recursively as follows:

[5] Murphy [4], p. 168, under the weighted average development model.

[6] Mack [2], p. 217.

[7] In Mack's derivation of equation (1).

$$
\begin{aligned}
Parameter\hat{R}isk_{ik} &= \hat{C}_{i,k}^{2} \sum_{j=I+1-i}^{k-1} \frac{\hat{\sigma}_j^2}{\hat{f}_j^2} \frac{1}{\sum_{r=1}^{I-j} C_{rj}} \\
&= \hat{f}_{k-1}^{2} \hat{C}_{i,k-1}^{2} \left(\sum_{j=I+1-i}^{k-2} \frac{\hat{\sigma}_j^2}{\hat{f}_j^2} \frac{1}{\sum_{r=1}^{I-j} C_{rj}} + \frac{\hat{\sigma}_{k-1}^2}{\hat{f}_{k-1}^2} \frac{1}{\sum_{r=1}^{I-k-1} C_{r,k-1}} \right) \\
&= \hat{f}_{k-1}^{2} \hat{C}_{i,k-1}^{2} \sum_{j=I+1-i}^{k-2} \frac{\hat{\sigma}_j^2}{\hat{f}_j^2} \frac{1}{\sum_{r=1}^{I-j} C_{rj}} + \hat{C}_{i,k-1}^{2} \frac{\hat{\sigma}_{k-1}^2}{\sum_{r=1}^{I-k-1} C_{r,k-1}} \\
&= \hat{f}_{k-1}^{2} Parameter\hat{R}isk_{i,k-1} + \hat{C}_{i,k-1}^{2} V\hat{a}r(\hat{f}_{k-1}) \quad .
\end{aligned}
$$

For *k* equal to the first future diagonal, the prior parameter risk is zero, and Mack's estimator above reduces to simply the second term.

Murphy's recursive estimator for parameter risk in Mack's notation is[8]

$$
Parame\hat{t}erRisk_{ik} = \begin{cases} \hat{f}_{k-1}^{2} Parame\hat{t}erRisk_{i,k-1} + \hat{C}_{i,k-1}^{2} V\hat{a}r(\hat{f}_{k-1}) + \\ \qquad\qquad V\hat{a}r(\hat{f}_{k-1}) Parame\hat{t}erRisk_{i,k-1} & \text{for } k > I+2-i \\ C_{I+1-i}^{2} V\hat{a}r(\hat{f}_{k-1}) & \text{for } k = I+2-i \ . \end{cases} \tag{7}
$$

Thus, the Mack and Murphy formulas differ only by the third, cross-product term in (7).[9] The derivation in theorem 2 in the appendix also yields the recursive formula (7).

2 Decomposition of the Mean Square Error

2.1 MSE Defined

Dispensing with the subscripts for accident year *i* and ultimate development age *I*, the mean square error (MSE) of the predictor $\hat{C}$ is defined[10] as the expected squared deviation of the predictor $\hat{C}$, a random variable, from the value of the random variable *C* being predicted; in operator notation

$$
mse(\hat{C}) = E(\hat{C} - C)^2
$$

where the expectation is taken with respect to the joint probability distribution of $\hat{C}$ and *C*.

[8] Mack [1] p. 167, assuming no constant term in the loss development model.

[9] The missing cross-product term has been noted elsewhere. See Buchwalder [1] for an example.

[10] For an example, see Mack [2], p. 216.

2.1 MSE Decomposed

Theorem 1 in the appendix shows that the MSE can be decomposed into variance and bias terms:

$$mse(\hat{C}) = Var(C) + Var(\hat{C}) + Bias^2(\hat{C}) \ . \tag{8}$$

The bias of the estimator is the difference between its mean and the mean of its target:

$$Bias(\hat{C}) = E(\hat{C}) - E(C) \ .$$

Thus, the MSE is the sum of process risk, parameter risk, and the squared bias of the estimator.

As can be seen from equation (8), it is possible for the MSE of a biased estimator to be smaller than the MSE of an unbiased estimator. For example, when a company's triangle is small or "thin" the resulting link ratios can bounce around too much from one reserve review to the next – high parameter risk. To stabilize the indications between reserve reviews, actuaries often supplement unstable company factors with more stable industry benchmarks. Do those benchmark factors introduce bias? Perhaps. If so, what might be the magnitude of that bias, and how does it compare with the corresponding reduction in MSE? Those questions are beyond the scope of this paper.

The all-year weighted averages in Mack [2] and Murphy [4] are unbiased.

2.2 Estimation Error Decomposed

Equation (12) in Theorem 1 in the appendix shows that an intermediate decomposition of the MSE has two terms, process risk and estimation error:

$$mse(\hat{C}) = Var(C) + E_{\hat{C}}(\hat{C} - \mu_C)^2 \ .$$

Estimation error $E_{\hat{C}}(\hat{C} - \mu_C)^2$ is the expected squared deviation of the estimator, not from its own mean, but from the mean of its target.[11] That expectation can be decomposed into the squared deviation of the estimator from its own mean plus the squared difference between the two means:

$$\begin{aligned} E_{\hat{C}}(\hat{C} - \mu_C)^2 &= E_{\hat{C}}(\hat{C} - \mu_{\hat{C}})^2 + (\mu_{\hat{C}} + \mu_C)^2 \\ &= Var(\hat{C}) + Bias^2(\hat{C}) \ . \end{aligned}$$

Thus, for unbiased estimators, estimation error and parameter risk are synonymous. For biased estimators, they are not.

[11] Contrast this with Mack's formulation of estimation error (Mack [2], p. 217), $(\hat{C} - \mu_C)^2$, a random variable.

2.3 The Magnitude of the Cross-Product Parameter Risk Term

Theorem 2 in the appendix proves (in parameter notation) that an estimator of the parameter risk of losses projected to age *k* is

$$\hat{\sigma}^2_{\hat{C}_k} = \hat{f}^2_{k-1}\hat{\sigma}^2_{\hat{C}_{k-1}} + \hat{C}^2_{k-1}\hat{\sigma}^2_{\hat{f}_{k-1}} + \hat{\sigma}^2_{\hat{f}_{k-1}}\hat{\sigma}^2_{\hat{C}_{k-1}} \, .$$

The ratio of the cross product term to the parameter risk estimator gives an idea of the relative magnitude of its contribution to the parameter risk estimate:

$$\begin{aligned}\frac{\hat{\sigma}^2_{\hat{f}_{k-1}}\hat{\sigma}^2_{\hat{C}_{k-1}}}{\hat{\sigma}^2_{\hat{C}_k}} &= \frac{\hat{\sigma}^2_{\hat{f}_{k-1}}\hat{\sigma}^2_{\hat{C}_{k-1}}}{\hat{f}^2_{k-1}\hat{\sigma}^2_{\hat{C}_{k-1}} + \hat{C}^2_{k-1}\hat{\sigma}^2_{\hat{f}_{k-1}} + \hat{\sigma}^2_{\hat{f}_{k-1}}\hat{\sigma}^2_{\hat{C}_{k-1}}} \\ &= \frac{1}{\dfrac{\hat{f}^2_{k-1}}{\hat{\sigma}^2_{\hat{f}_{k-1}}} + \dfrac{\hat{C}^2_{k-1}}{\hat{\sigma}^2_{\hat{C}_{k-1}}} + 1} \, . \end{aligned} \tag{9}$$

As can be seen from equation (9) the contribution of the cross-product term to the parameter risk estimate will be large when the denominator in (9) is small, which can occur when the link ratio variation is large relative to the square of link ratio. So for small triangles or triangles with wildly varying development, it would behoove the actuary not to ignore the cross-product term. In our experience, with reasonably stable triangles the impact of the cross-product term has been negligible.

3 An Example

Mack [1] applied his formulas to the following triangular array of data from Taylor and Ashe [5]:

357848	1124788	1735330	2218270	2745596	3319994	3466336	3606286	3833515	3901463
352118	1236139	2170033	3353322	3799067	4120063	4647867	4914039	5339085	
290507	1292306	2218525	3235179	3985995	4132918	4628910	4909315		
310608	1418858	2195047	3757447	4029929	4381982	4588268			
443160	1136350	2128333	2897821	3402672	3873311				
396132	1333217	2180715	2985752	3691712					
440832	1288463	2419861	3483130						
359480	1421128	2864498							
376686	1363294								
344014									

Given the all-year weighted average link ratios below and the cumulative loss development factors (LDFs)

	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	tail
Link Ratio	3.491	1.747	1.457	1.174	1.104	1.086	1.054	1.077	1.018	1.000
LDF	14.447	4.139	2.369	1.625	1.384	1.254	1.155	1.096	1.018	1.000

the completed triangle is

i/*k*	*k*=1	*k*=2	*k*=3	*k*=4	*k*=5	*k*=6	*k*=7	*k*=8	*k*=9	*k*=10
i=1	357,848	1,124,788	1,735,330	2,218,270	2,745,596	3,319,994	3,466,336	3,606,286	3,833,515	3,901,463
i=2	352,118	1,236,139	2,170,033	3,353,322	3,799,067	4,120,063	4,647,867	4,914,039	5,339,085	5,433,719
i=3	290,507	1,292,306	2,218,525	3,235,179	3,985,995	4,132,918	4,628,910	4,909,315	5,285,148	5,378,826
i=4	310,608	1,418,858	2,195,047	3,757,447	4,029,929	4,381,982	4,588,268	4,835,458	5,205,637	5,297,906
i=5	443,160	1,136,350	2,128,333	2,897,821	3,402,672	3,873,311	4,207,459	4,434,133	4,773,589	4,858,200
i=6	396,132	1,333,217	2,180,715	2,985,752	3,691,712	4,074,999	4,426,546	4,665,023	5,022,155	5,111,171
i=7	440,832	1,288,463	2,419,861	3,483,130	4,088,678	4,513,179	4,902,528	5,166,649	5,562,182	5,660,771
i=8	359,480	1,421,128	2,864,498	4,174,756	4,900,545	5,409,337	5,875,997	6,192,562	6,666,635	6,784,799
i=9	376,686	1,363,294	2,382,128	3,471,744	4,075,313	4,498,426	4,886,502	5,149,760	5,544,000	5,642,266
i=10	344,014	1,200,818	2,098,228	3,057,984	3,589,620	3,962,307	4,304,132	4,536,015	4,883,270	4,969,825

The variance estimates are

	k=1	*k*=2	*k*=3	*k*=4	*k*=5	*k*=6	*k*=7	*k*=8	*k*=9
$\hat{\sigma}_k^2$	160,280	37,737	41,965	15,183	13,731	8,186	447	1,147	447
$\hat{\sigma}_{f_k}^2$	0.048170	0.003681	0.002789	0.000823	0.000764	0.00051	0.00004	0.00013	0.00012

Using formula (5) the process risk (variance) estimates of the future losses displayed above are calculated recursively left to right. The variance of the sum is the sum of the variances because years *i*=1…10 are independent.

	k=1	*k*=2	*k*=3	*k*=4	*k*=5	*k*=6	*k*=7	*k*=8	*k*=9	*k*=10
i=1										
i=2										2.38E+09
i=3									5.63E+09	8.19E+09
i=4								2.05E+09	7.92E+09	1.05E+10
i=5							3.17E+10	3.71E+10	4.81E+10	5.19E+10
i=6						5.07E+10	9.32E+10	1.05E+11	1.28E+11	1.34E+11
i=7					5.29E+10	1.21E+11	1.79E+11	2.01E+11	2.39E+11	2.50E+11
i=8				1.20E+11	2.29E+11	3.46E+11	4.53E+11	5.06E+11	5.93E+11	6.17E+11
i=9			5.14E+10	2.09E+11	3.41E+11	4.71E+11	5.93E+11	6.61E+11	7.72E+11	8.02E+11
i=10		5.51E+10	2.14E+11	5.42E+11	7.93E+11	1.02E+12	1.23E+12	1.37E+12	1.59E+12	1.65E+12
Sum		5.51E+10	2.65E+11	8.71E+11	1.42E+12	2.00E+12	2.58E+12	2.88E+12	3.39E+12	3.53E+12

For example, for *i*=8, *k*=6, $3.46\cdot 10^{11} = 1.104^2 \cdot 2.29\cdot 10^{11} + 4900545\cdot 13731$.

Using formula (6) the parameter risk (variance) estimates of the future losses are also calculated recursively left to right. The variance of the sum is calculated using formulas in Murphy [4].

	k=1	*k*=2	*k*=3	*k*=4	*k*=5	*k*=6	*k*=7	*k*=8	*k*=9	*k*=10
i=1										
i=2										3.32E+09
i=3									3.25E+09	6.62E+09
i=4								7.38E+08	4.00E+09	7.30E+09
i=5							7.70E+09	9.17E+09	1.33E+10	1.64E+10
i=6						1.04E+10	2.08E+10	2.38E+10	3.05E+10	3.46E+10
i=7					9.99E+09	2.50E+10	3.99E+10	4.52E+10	5.59E+10	6.16E+10
i=8				2.29E+10	4.59E+10	7.43E+10	1.03E+11	1.15E+11	1.39E+11	1.49E+11
i=9			6.84E+09	3.04E+10	5.18E+10	7.59E+10	9.99E+10	1.12E+11	1.33E+11	1.42E+11
i=10		5.70E+09	2.27E+10	6.06E+10	9.13E+10	1.21E+11	1.51E+11	1.68E+11	1.98E+11	2.08E+11
Sum		5.70E+09	4.16E+10	2.39E+11	4.95E+11	9.20E+11	1.44E+12	1.64E+12	2.12E+12	2.46E+12

For example, for *i*=8, *k*=6,

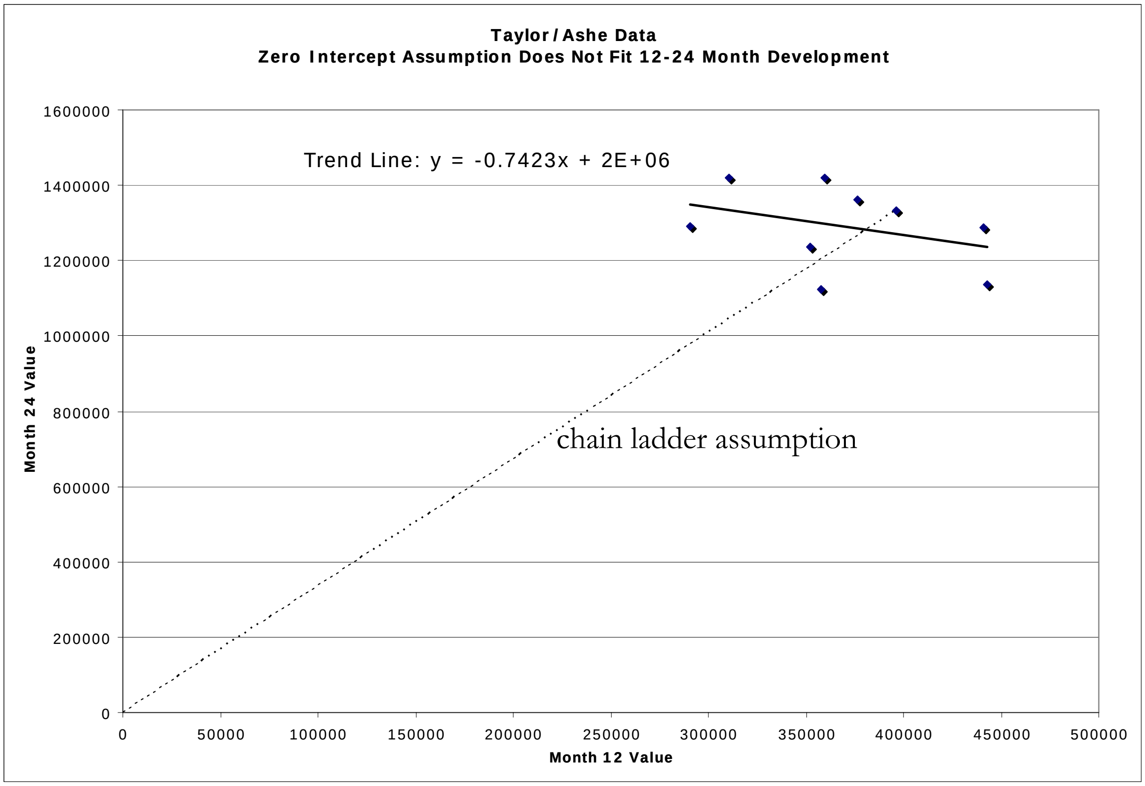
$$7.43 \cdot 10^{10} = 1.104^2 \cdot 4.59 \cdot 10^{10} + 4900545^2 \cdot 0.000764 + 0.000764 \cdot 4.59 \cdot 10^{10} .$$

Comparisons of these Murphy-formula results with the Mack-formula results from Mack [2] are displayed in row detail, and in total, in the following table:

	Reserve Risk Estimates					
Origination Year	**Mack Formula**			**Murphy Formula**		
	Process	**Parameter**	**Total**	**Process**	**Parameter**	**Total**
i=2	48,832	57,628	75,535	48,832	57,628	75,535
i=3	90,524	81,338	121,699	90,524	81,340	121,700
i=4	102,622	85,464	133,549	102,622	85,467	133,551
i=5	227,880	128,078	261,406	227,880	128,091	261,412
i=6	366,582	185,867	411,010	366,582	185,907	411,028
i=7	500,202	248,023	558,317	500,202	248,110	558,356
i=8	785,741	385,759	875,328	785,741	385,991	875,430
i=9	895,570	375,893	971,258	895,570	376,222	971,385
i=10	1,284,882	455,270	1,363,155	1,284,882	455,957	1,363,385
Total:	1,878,292	1,568,532	2,447,095	1,878,292	1,569,349	2,447,618

The Mack and Murphy process risk estimates are identical. Differences in parameter risk occur, at most, only in the 3rd or 4th significant digit.

Continuing with this example, the regression perspective of Murphy [4] provides additional insight into the Taylor/Ashe data. The graphical display below of the historical relationship between 12- and 24-month losses clearly shows that the data violate the first chain ladder assumption (Mack's CL1), i.e., that the expected relationship is a line through the origin.



Although the indicated slope of the trend line is negative, the regression statistics support the statement that it is not significantly different from zero, implying that the 12- and 24-month losses are actually uncorrelated. Therefore, a reasonable estimate of the 24-month losses for year 10 would simply be the average of all of the previous years' 24-month losses, 1,290,505. This estimate would be reasonable not just from a statistical standpoint but from a business standpoint if we knew, for instance, that all losses are on-level and of equal exposure. The standard deviation of those losses is 108,885 = process risk, and the standard deviation of the mean is 38497 = sqrt(108895^2/(9-1)) = parameter risk.

This demonstrates one of the advantages of recursive formulas: flexibility. The recursive formulas (5) and (7) do not know how the predictions and variances are estimated, nor do they care (e.g., see Theorem 2). One need only substitute these two new process risk and parameter risk estimates for year 10 into the corresponding (i=10,k=2) cells in the tables above and the recursive calculations for k>2 carry on as before. The new comparison table is

	Reserve Risk Estimates					
Origination Year	**Mack Formula**			**Murphy Formula**		
	Process	**Parameter**	**Total**	**Process**	**Parameter**	**Total**
i=2	48,832	57,628	75,535	48,832	57,628	75,535
i=3	90,524	81,338	121,699	90,524	81,340	121,700
i=4	102,622	85,464	133,549	102,622	85,467	133,551
i=5	227,880	128,078	261,406	227,880	128,091	261,412
i=6	366,582	185,867	411,010	366,582	185,907	411,028
i=7	500,202	248,023	558,317	500,202	248,110	558,356
i=8	785,741	385,759	875,328	785,741	385,991	875,430
i=9	895,570	375,893	971,258	895,570	376,222	971,385
i=10	1,284,882	455,270	1,363,155	980,971	390,295	1,055,762
Total:	1,878,292	1,568,532	2,447,095	1,685,041	1,568,504	2,302,079

Thus, after a simple diagnostic of the underlying data and an appropriate adjustment in the actuarial projection, process risk for year 10 is reduced by 22.5%, parameter risk by 14.3%, and total risk by 21.5%, and the total risk estimate for all years combined is 6% lower than that produced by the Mack method. This example also points out how it is not necessary – or even advisable – to use a single reserving method for the entire future development of a given year. In some instances it is beneficial to "change methods in the middle of the development stream."

4 Conclusion

Although Mack's reserve risk formulas omit a parameter risk cross-product term, the understatement should be negligible for reasonably behaved triangles. The advantage of closed-form formulas as in Mack [2] is that they are concise. Recursive formulas by Murphy [4], by Mack [3], and in this paper are not as concise but are more flexible, e.g., allowing for projections based on a shift in model from one development period to the next.

Mean square error is comprised of process risk, parameter risk, and bias. Estimation error and parameter risk are equivalent when the link ratios are unbiased. Within the context of the chain ladder method, utilization of industry benchmark factors might introduce bias into the projections, but in the actuary's judgment the resulting stabilization may outweigh whatever bias might occur. Estimating the magnitude of the potential for bias and reduction in MSE are areas of further actuarial research.

Appendix

The definition of the mean square error (MSE) of the predictor $\hat{C}$ is the expected squared deviation of the (random variable) predictor $\hat{C}$ from the value of the random variable C being predicted:

$$mse(\hat{C}) = E(\hat{C} - C)^2 \tag{10}$$

where the expectation is taken with respect to the joint probability distribution of $\hat{C}$ and *C*.

Theorem 1: The MSE Decomposition Theorem

$$\text{mse}(\hat{C}) = \text{Var}(C) + \text{Var}(\hat{C}) + \text{Bias}^2(\hat{C}).$$

Proof: Let $f(c,\hat{c})$ represent the joint density of *C* and $\hat{C}$. Then the MSE is the integral

$$\text{mse}(\hat{C}) = \iint (\hat{c} - c)^2 f(c,\hat{c}) dc d\hat{c}$$

taken over the joint sample space.

To decompose the MSE into variance and bias components, we will use the fact that the joint density of the two random variables can be factored into a conditional density and a marginal density:

$$f(c,\hat{c}) = f(c \mid \hat{c}) f(\hat{c}).$$

This fact allows us to write equation (10) as

$$\text{mse}(\hat{C}) = E_{\hat{C}}(E((\hat{C} - C)^2 \mid \hat{C})) \tag{11}$$

where the inner expectation is taken with respect to *C* conditional on the value of $\hat{C}$. We will manipulate the inner expectation first, taking advantage of the "scalar" nature of $\hat{C}$ with respect to that conditional expectation.

We add and subtract the mean μ_C of the predicted random variable inside the quadratic, group the result into two terms, square the binomial, and observe that the cross-product term disappears. To wit

$$\begin{aligned} E_C((\hat{C} - C)^2 \mid \hat{C}) &= E_C[(\hat{C} - \mu_C + \mu_C - C)^2 \mid \hat{C}] \\ &= E_C[((\hat{C} - \mu_C) + (\mu_C - C))^2 \mid \hat{C}] \\ &= E_C[(\hat{C} - \mu_C)^2 + 2(\hat{C} - \mu_C)(\mu_C - C) + (\mu_C - C)^2 \mid \hat{C}] \\ &= E_C[(\hat{C} - \mu_C)^2 \mid \hat{C}] + 2E_C[(\hat{C} - \mu_C)(\mu_C - C) \mid \hat{C}] + E_C[(\mu_C - C)^2 \mid \hat{C}]. \end{aligned}$$

The third term above is just Var(*C*), the first term (conditional on $\hat{C}$) is simply $(\hat{C} - \mu_C)^2$, and the middle term disappears because

$$\begin{aligned} E_C[(\hat{C} - \mu_C)(\mu_C - C) \mid \hat{C}] &= E_C[(\hat{C}\mu_C - \hat{C}C - \mu_C^2 + \mu_C C) \mid \hat{C}] \\ &= \hat{C}\mu_C - \hat{C}E_C[C \mid \hat{C}] - \mu_C^2 + \mu_C E_C[C \mid \hat{C}] \\ &= \hat{C}\mu_C - \hat{C}\mu_C - \mu_C^2 + \mu_C^2 \\ &= 0\ . \end{aligned}$$

Substituting these expressions into (11), we have that

$$\text{mse}(\hat{C}) = Var(C) + E_{\hat{C}}(\hat{C} - \mu_C)^2 \tag{12}$$

which shows that the MSE equals the process variance plus the expected squared deviation between the predictor and the mean of its target.[12] The second term on the right in (12) is called "estimation error."

To continue the decomposition, we address the estimation error term in (12) by adding and subtracting the mean $\mu_{\hat{C}}$ inside the quadratic and proceeding as above:

$$\begin{aligned}
E_{\hat{C}}(\hat{C} - \mu_C)^2 &= E_{\hat{C}}(\hat{C} - \mu_{\hat{C}} + \mu_{\hat{C}} - \mu_C)^2 \\
&= E_{\hat{C}}[(\hat{C} - \mu_{\hat{C}}) + (\mu_{\hat{C}} - \mu_C)]^2 \\
&= E_{\hat{C}}[(\hat{C} - \mu_{\hat{C}})^2] + 2E_{\hat{C}}[(\hat{C} - \mu_{\hat{C}})(\mu_{\hat{C}} - \mu_C)] + E_{\hat{C}}[(\mu_{\hat{C}} - \mu_C)^2] \\
&= Var(\hat{C}) + 2E_{\hat{C}}[(\hat{C}\mu_{\hat{C}} - \hat{C}\mu_C - \mu_{\hat{C}}^2 + \mu_{\hat{C}}\mu_C)] + (\mu_{\hat{C}} - \mu_C)^2 \\
&= Var(\hat{C}) + 2[\mu_{\hat{C}}^2 - \mu_{\hat{C}}\mu_C - \mu_{\hat{C}}^2 + \mu_{\hat{C}}\mu_C] + (\mu_{\hat{C}} - \mu_C)^2 \\
&= Var(\hat{C}) + (Bias(\hat{C}))^2 \ .
\end{aligned}$$

Substituting this expression for $E_{\hat{C}}(\hat{C} - \mu_C)^2$ into (12), we have

$$mse(\hat{C}) = Var(C) + Var(\hat{C}) + Bias^2(\hat{C})$$

which proves the theorem.

Theorem 2: The Parameter Risk Recursion Theorem

$$V\hat{a}r(\hat{C}_{ik}) = \hat{f}_{k-1}^2 V\hat{a}r(\hat{C}_{i,k-1}) + \hat{C}_{i,k-1}^2 V\hat{a}r(\hat{f}_{k-1}) + V\hat{a}r(\hat{f}_{k-1}) V\hat{a}r(\hat{C}_{i,k-1})$$

Proof: Following a similar path as in equation (4) in Section 1 above:

$$\begin{aligned}
Var(\hat{C}_{ik}) &= E_{\hat{C}_{i,k-1}}(Var(\hat{C}_{ik} \mid \hat{C}_{i,k-1})) + Var_{\hat{C}_{i,k-1}}(E(\hat{C}_{ik} \mid \hat{C}_{i,k-1})) \\
&= E_{\hat{C}_{i,k-1}}(Var(\hat{f}_{k-1}\hat{C}_{i,k-1} \mid \hat{C}_{i,k-1})) + Var_{\hat{C}_{i,k-1}}(E(\hat{f}_{k-1}\hat{C}_{i,k-1} \mid \hat{C}_{i,k-1})) \\
&= E_{\hat{C}_{i,k-1}}(\hat{C}_{i,k-1}^2 Var(\hat{f}_{k-1})) + Var_{\hat{C}_{i,k-1}}(\hat{C}_{i,k-1} E(\hat{f}_{k-1})) \\
&= Var(\hat{f}_{k-1}) E(\hat{C}_{i,k-1}^2) + Var_{\hat{C}_{i,k-1}}(\hat{C}_{i,k-1} f_{k-1}) \\
&= Var(\hat{f}_{k-1})(Var(\hat{C}_{i,k-1}) + E^2(\hat{C}_{i,k-1})) + f_{k-1}^2 Var(\hat{C}_{i,k-1}) \\
&= Var(\hat{f}_{k-1}) Var(\hat{C}_{i,k-1}) + Var(\hat{f}_{k-1}) C_{i,k-1}^2 + f_{k-1}^2 Var(\hat{C}_{i,k-1}) \ .
\end{aligned}$$

Substituting estimates for the unknown parameters yields the desired result.

[12] Contrast this with Mack's expression for the MSE in [2]: $\text{mse}(\hat{C}) = \text{Var}(C) + (\mu_C - \hat{C})^2$.

Acknowledgment

The author wishes to acknowledge Ali Majidi, Doug Collins, and Emmanuel Bardis for their clarifying comments, direction, and support.

References

[1] Buchwalder, M., et al. 2005. "Legal Valuation Portfolio in Non-Life Insurance." Lecture, ASTIN Colloquium, Zurich, Switzerland, September 5-7, 2005.

[2] Mack, Thomas. 1993. "Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates." *ASTIN Bulletin* 23, no. 2:213-225.

[3] Mack, Thomas. 1999. "The Standard Error of Chain Ladder Reserve Estimates: Recursive Calculation and Inclusion of a Tail Factor." *ASTIN Bulletin* 29, no. 2:361-366.

[4] Murphy, Daniel. 1994. "Unbiased Loss Development Factors." *PCAS* 81:154-222.

[5] Taylor, G., and F. Ashe. 1983. "Second Moments of Estimates of Outstanding Claims." *Journal of Econometrics* 23:37-61.

Biography

Daniel Murphy is a consulting actuary with the Tillinghast business of Towers Perrin. He is a Fellow of the CAS, a Member of the American Academy of Actuaries, and a member of the CAS program planning committee.

Generalized Linear Models Beyond the Exponential Family with Loss Reserve Applications

Gary G. Venter, FCAS, MAAA

Abstract

The formulation of generalized linear models in *Loss Models* by Klugman, Panjer, and Willmot [5] is a bit more general than is often seen, in that the residuals are not restricted to following a member of the exponential family. Some of the distributions this allows have potentially useful applications. The cost is that there is no longer a single form for the likelihood function, so each has to be fit directly. Here the use of loss distributions (frequency, severity, and aggregate) in generalized linear models is addressed, along with a few other possibilities.

Keywords. Loss reserving; regression modeling; generalized linear models.

1 INTRODUCTION

The paradigm of a linear model is multiple regression, where the dependent variables are linear combinations of independent variables plus a residual term, which is from a single mean-zero normal distribution. Generalized linear models, denoted here as GLZ[1], allow nonlinear transforms of the regression mean as well as other forms for the distribution of residuals.

Since many actuarial applications of GLZ are to cross-classified data, such as in a loss development triangle or classification rating plan, a two-dimensional array of independent observations will be assumed, with a typical cell's data denoted as $q_{w,d}$. That is easy to generalize to more dimensions or to a single one.

Klugman, Panjer, and Willmot (2004) [5] provide a fairly general definition of GLZs. To start with, let $\mathbf{z_{w,d}}$ be the row vector of covariate observations for the w, d cell and β the column vector of coefficients. Then a GLZ with that distribution models the mean of $q_{w,d}$ as a function η of the linear combination $\mathbf{z_{w,d}}\beta$, where all the other parameters, including β, are constant across the cells.

It appears that their intention is that η does not take any of the parameters of the

[1] Often GLM is used but with more restrictions on distributional form, typically the exponential family.

distribution as arguments, although this is not explicitly stated. An interesting special case is where η is the identity function, so the mean of $q_{w,d}$ is $\mathbf{z_{w,d}}\boldsymbol{\beta}$. Another key case is where η is exp, so $E[q_w] = \exp(\mathbf{z_{w,d}}\boldsymbol{\beta})$. This is a multiplicative model in which the mean is the product of the exponentiated summands in $\mathbf{z_{w,d}}\boldsymbol{\beta}$.

Standard regression sets the mean $\mathbf{z_{w,d}}\boldsymbol{\beta}$ to the μ of a normal distribution, which has another parameter σ that is constant across the cells. But almost any distribution that has a mean could be reparameterized so that the mean is one of the parameters. This allows virtually any distribution to be used for the residuals. The mean-parameter will be referred to as μ hereafter.

Usual GLM requires the distribution to be from the exponential family. Mildenhall (1999) [7] defines this as a distribution that can be written in the form $f(x;\mu,\phi) = c(x,\phi)/\exp[d(x;\mu)/(2\phi)]$ where $d(x;\mu) = 2w\int_{\mu}^{x}\frac{x-\mu}{V(t)}dt$ for a strictly positive function $V(t)$ and weighting constant w. The tricky part is that μ appears only in the exponent and is constrained in how it combines with ϕ. For any μ, c has to make the integral unity. While quite a few models are possible with this family and various η functions, expanding the universe of distributions leads to other interesting models. Some of the simplicity of exponential models is lost, however.

Standard theory shows the mean of an exponential model is μ and the variance is $\phi V(\mu)/w$. The V function defines the exponential model uniquely. Using w=1 and $V = \mu^j$ with j = 0, 1, 2, 3 gives the normal, Poisson, gamma, and inverse Gaussian distributions, respectively. The ratio of the coefficient of skewness to the coefficient of variation (or CV, which is the standard deviation divided by mean) for these distributions is also 0, 1, 2, 3, respectively. Renshaw (1994) [10] has a formula that implies more generally that skewness/CV is $\mu\partial lnV/\partial\mu$ whenever w=1.

The relationship of variance to mean is one of the issues in selecting a distribution for GLZs. The relationship no longer uniquely defines the distribution, however. For the normal and *t*-distributions[2] the mean and variance are not related, which could be expressed as the variance being proportional to μ^0. The Poisson has variance proportional to μ^1, and quite a few distributions have variance proportional to μ^2. Other relationships of mean and variance will be discussed below. One advantage of GLZs is that distributions with the same relationship of variance to mean might have different tail properties, including different skewnesses and higher moments, giving more flexibility in fitting models to data.

In linear regression the failure of the observations to match the predictions of constant variance is called heteroscedasticity. Often this occurs because the variance is smaller for smaller observations. In such a case using a distribution with variance proportional to a power of the mean might solve the heteroscedasticity problem. A simple example is the Poisson, where μ is the λ parameter, which then gets set to $\eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$ for each cell and then is the mean and variance of the cell.

Virtually any distribution can be used in a GLZ. Specific examples of frequency, severity, and aggregate loss distributions in GLZs are discussed next, followed by estimation issues and examples from modeling loss development triangles.

2 FREQUENCY DISTRIBUTIONS IN GLZ

For the Poisson in λ, the mean and variance are both $\lambda = \eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$. The negative binomial is more interesting. In the usual parameterization, the variance is a fixed multiple of, but greater than, the mean. Negative binomial distributions are in the $(a,b,0)$ class, which means that for $k>0$, there are values a and b so that probabilities follow the recursive relationship $p_k = (a+b/k)p_{k\text{-}1}$. The negative binomial has two positive parameters, r and β, with mean $= r\beta$ and variance $= r\beta(1+\beta)$. Skewness/CV is $1+\beta/(1+\beta)$, which is between 1 and 2. Probabilities start with $p_0 = (1+\beta)^{-r}$ and in the recursion $a = \beta/(1+\beta)$ and $b = (r\text{-}1)a$.

There are two simple ways to express the negative binomial mean as a parameter. First, keeping the parameter β, replace r by μ/β, so there are two parameters β and μ and the mean

[2] Having *t*-distributed residuals is one of the many possibilities this formulation of GLZ allows. Also the Laplace, which has exponential tails in both directions from the origin, or the logistic, which is like a heavy-tailed normal, could be used for symmetric residuals.

is μ. The variance $r\beta(1+\beta)$ becomes $\mu(1+\beta)$. In a GLZ the mean is $\mu = \eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$ and the variance is $\eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})(1+\beta)$, which is proportional to the mean. On the other hand if you keep r and replace β by μ/r, the parameters are r and μ, and the mean is again μ, but the variance $r\beta(1+\beta)$ is $\mu(1+\mu/r)$, which is quadratic in μ. This form is in the exponential family. Thus depending on how you parameterize the negative binomial, its variance can be either linear or quadratic in the mean.

The parameterization chosen does not make any difference for a single distribution. Suppose for example that X has $r = 3$ and $\beta = 10$ and so mean $\mu=30$ and variance 330. The variance is $\mu(1+\beta)$ in the first formulation and $\mu(1+\mu/r)$ in the second, both of which are 330. A difference comes when modeling other variables while keeping parameters other than μ constant. Suppose Y has mean 100. If β is kept at 10, $\mu(1+\beta) = 1100$, while if r is kept at 3, $\mu(1+\mu/r) = 3433$. The parameterization to choose would be the one that best captures the way the variance grows as the risk size increases. This same idea is applied to severity distributions next.

3 SEVERITY DISTRIBUTIONS IN GLZ

A parameter θ of a distribution of X is a scale parameter if the distribution of a multiple of X is obtained by substituting that multiple of θ into the original distribution. The k^{th} moment of the distribution is then proportional to θ^k. Thus if the mean μ is a scale parameter, the variance is proportional to μ^2.

3.1 Inverse Gaussian

As an example, consider the inverse Gaussian distribution with density

$$ig_1(x;\mu,\alpha) = \sqrt{\frac{\mu}{2\pi\alpha x^3}}\,e^{\frac{2-x/\mu-\mu/x}{2\alpha}}.$$

Here μ is a scale parameter, with EX = μ and VarX = $\alpha\mu^2$. However it is more usual to parameterize the inverse Gaussian with $\lambda = \mu/\alpha$, so α is replaced by μ/λ:

$$ig_2(x;\mu,\lambda) = e^{\frac{2-x/\mu-\mu/x}{2\mu/\lambda}}\sqrt{\frac{\lambda}{2\pi x^3}}.$$

Now μ is no longer a scale parameter, even though it is still the mean. The variance is μ^3/λ, and so is proportional to μ^3 instead of μ^2. This is in the exponential family as μ is just in the exponent. Both forms meet the requirements to be GLZs, so either variance assumption can be accommodated. The choice would depend on how the squared deviations from the cell means tend to vary with the means $\eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$. If they seem to grow proportionally to the square of the mean, ig_1 would be indicated, but if they grow with the mean cubed, ig_2 would be preferred.

How the variance relates to the mean is thus not a fundamental feature of the inverse Gaussian, but is a result of how it is parameterized. A characteristic constant of this distribution, not dependent on parameterization, is the ratio of the skewness to the CV. In ig_1, with μ a scale parameter, the third central moment is $3\mu^3\alpha^2$ while it is $3\mu^5/\lambda^2$ in ig_2. Thus in ig_1 the CV is $\alpha^{½}$ and the skewness is $3\alpha^{½}$, so the ratio is 3. In ig_2 these coefficients are $(\mu/\lambda)^{½}$ and $3(\mu/\lambda)^{½}$, so the ratio is again 3.

3.2 Gamma

Substituting alternative parameters can be done for other distributions as well. For instance the gamma distribution is usually parameterized $F(x;\ \theta,\alpha) = \Gamma(x/\theta;\ \alpha)$ with the incomplete gamma function Γ. This has mean $\alpha\theta$ and variance $\alpha\theta^2$. To get the mean to be a parameter, set $F(x;\ \mu,\alpha) = \Gamma(x\alpha/\mu;\ \alpha)$. Then the variance is μ^2/α and μ is still a scale parameter. But other parameterizations are possible. Similarly to the inverse Gaussian, setting $F(x;\ \mu,\lambda) = \Gamma(x\lambda/\mu^2;\ \lambda/\mu)$ still gives mean μ but now the variance is μ^3/λ. Other variance functions can be reached by this method. For instance $F(x;\ \mu,\lambda) = \Gamma[x/(\lambda\mu^p);\ \mu^{1-p}/\lambda]$ has mean μ and variance $\mu^{1+p}\lambda$. This works for any real p, so the gamma variance can be made to be proportional to any power of the mean, including zero. This will be called the gamma p.

Hewitt (1966) [3] noted that if larger risks were independent sums of small risks, the variance would grow in proportion to the mean. He found in fact that aggregate loss distributions for some insurance risks can be modeled by gamma distributions, and that the gamma variance grows by about $\mu^{1.227}$. This relationship could be modeled by the gamma p with $p = 0.227$.

As with the inverse Gaussian, the ratio of skewness to CV is a characteristic constant of the gamma distribution. With power *p*, the third central moment is $2\lambda^2\mu^{1+2p}$. This gives skewness of $2\lambda^{0.5}\mu^{0.5p-0.5}$, which is twice the CV, so the ratio is 2 for the gamma regardless of *p*. Thus an inverse Gaussian is 50% more skewed than the gamma with the same mean and variance.

3.3 Lognormal

The lognormal density can be parameterized as:

$$f(x;\theta,\tau)=\frac{e^{-[\log(x/\theta)]^2/(2\tau)}}{x\sqrt{2\pi\tau}}.$$

Here θ is a scale parameter. The mean is $\theta e^{\tau/2}$ and the variance is $\theta^2 e^{\tau}(e^{\tau}\text{-}1)$. Taking $\alpha = e^{\tau/2}$ and $\mu = \alpha\theta$, the mean and variance are μ and $\mu^2(\alpha^2-1)$ and

$$f(x;\mu,\alpha)=\frac{e^{-[\log(\alpha x/\mu)]^2/(4\log\alpha)}}{2x\sqrt{\pi\log\alpha}}.$$

For the lognormal a characteristic constant is the ratio of skewness to CV minus the CV-squared. This is always 3, regardless of parameterization.

The usual parameterization of the lognormal is: $F(x;\mu,\sigma)=N\left(\frac{\ln(x)-\mu}{\sigma}\right)$. This has mean $e^{\mu+\sigma^2/2}$ and variance $e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$. Now reparameterize with two parameters *m* and *s*:

$$F(x;m,s)=N\left(\frac{\ln\left((x/m)\sqrt{1+s^2/m^2}\right)}{\sqrt{\ln\left(1+s^2/m^2\right)}}\right)$$

It is not hard to see that μ has been replaced by $\ln\left(\frac{m^2}{\sqrt{s^2+m^2}}\right)$ and σ^2 has been replaced by $\ln\left(\frac{s^2+m^2}{m^2}\right)$. Thus e^{μ} is $\frac{m^2}{\sqrt{s^2+m^2}}$ and e^{σ^2} is $\frac{s^2+m^2}{m^2}$. From this it follows that the mean is *m* and the variance is s^2. This parameterization makes the mean and variance completely unrelated. By the way, skewness is then also a fairly simple function of the parameters: skewness = $3\frac{s}{m}+\frac{s^3}{m^3}$. As with the gamma, other reparameterizations of the lognormal are possible, and can give any relationship of variance and mean. In fact,

$$F(x;m,s,p)=N\left(\frac{\ln\left((x/m)\sqrt{1+s^2m^{p-2}}\right)}{\sqrt{\ln\left(1+s^2m^{p-2}\right)}}\right)$$

has mean *m*, variance s^2m^p, and skewness $3t+t^3$, where $t = sm^{p/2\text{-}1}$. Here μ has been replaced by $\ln\left(\frac{m}{\sqrt{1+s^2m^{p-2}}}\right)$ and σ^2 by $\ln(1+s^2m^{p-2})$.

3.4 Pareto

The Pareto is another interesting case. Consider $F(x;\ \theta,\alpha) = 1 - (1+x/\theta)^{-\alpha}$. This has mean $\theta/(\alpha-1)$. Taking $\mu = (\alpha-1)\theta$ gives $F(x;\ \mu,\alpha) = 1 - (1+x/(\mu\ \alpha- \mu))^{-\alpha}$. This has mean μ and variance $\mu^2/(\alpha-2)$ if $\alpha > 2$. But if $\alpha \leq 1$ this does not work, as the mean does not exist. There does not seem to be any reason not to extend the GLZs to this case. Perhaps the easiest way to do this is to model $\theta_{w,d}$ as $\eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$ for each cell. Or the median $m = \theta(2^{1/\alpha} - 1)$ could be the parameter modeled, by setting $\mathrm{F}(x;\ m,\alpha) = 1 - (1+x(2^{1/\alpha}-1)/m)^{-\alpha}$, with $m = \eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$. This is median regression in the GLZ framework.

The skewness for the gamma, inverse Gaussian and lognormal distributions can be expressed as 2CV, 3CV, and $3\mathrm{CV}+\mathrm{CV}^3$, respectively. For the Pareto, if the skewness exists, CV^2 is in the range (1,3). Then the skewness is $\frac{2}{CV}\frac{\alpha+1}{\alpha-3} = 2CV\frac{3-CV^{-2}}{3-CV^2}$. This is less than the lognormal skewness when $CV^2<2$ and less than the inverse Gaussian skewness when $CV^2<0.5+\sqrt{11/12}\approx 1.4574$. This illustrates the different tail possibilities for GLZs with the same means and variances.

3.5 Origin Shifting

Severity distributions have their support on the positive reals, so all fitted values have to be positive. Frequency and aggregate distributions extend the support to include zero, but not negative values. However, any of the positive distributions can be location shifted to allow the possibility of negative values or even negative means. For instance, the shifted gamma has $F(x) = \Gamma[(x–b)/\theta, \alpha]$, with mean $b+\alpha\theta$ and variance $\alpha\theta^2$. Making the mean a parameter gives the distribution $F(x) = \Gamma[(\alpha(x–b)/(\mu–b), \alpha]$. The variance is then $(\mu–b)^2/\alpha$, which is still quadratic in μ.

4 AGGREGATE DISTRIBUTIONS IN GLZ

Aggregate distributions can be especially useful for residuals that are continuous on the positive reals but also could take a positive probability at zero. This is often seen out in late lags of a development triangle, for example.

4.1 Poisson-Gamma Aggregates

An example of an aggregate loss model in the exponential family is the Tweedie distribution. This starts by combining a gamma severity in α and θ that has mean $\alpha\theta$ and variance $\alpha\theta^2$ with a Poisson frequency in λ. Then the aggregate distribution has mean $\mu =\lambda\alpha\theta$ and variance $= \lambda\alpha\theta^2(\alpha+1) = \mu\theta(\alpha+1)$. Since this can also be written as $\lambda(\alpha\theta)^2(1/\alpha+1)$, it is clear that the variance is linear in the frequency mean and quadratic in the severity mean.

If the restriction $\lambda = k(\alpha\theta)^{\alpha}$ is imposed, then $\mu = k(\alpha\theta)^{\alpha+1}$, and the variance is $k\alpha^{\alpha+1}\theta^{\alpha+2}(1+\alpha)$, or $\mu^{1+1/(\alpha+1)}(1+1/\alpha)k^{-1/(\alpha+1)}$. This is the Tweedie distribution. The variance is proportional to a power of the mean between 1 and 2, which is often realistic for sums of claims. The link between frequency and severity is problematic, however. It would seem unusual for the observations with the smaller number of claims to also have the smaller claim sizes.

Kaas (2005) [4] expresses the Tweedie by replacing the three parameters λ, α, θ of the Poisson-Gamma with three others μ, ψ, and p by the formulas:

$$\lambda = \mu^{2-p}/[\psi(2-p)] \qquad \alpha = (2-p)/(p-1) \qquad \theta = \psi(p-1)\mu^{p-1}$$

This looks like a 3 for 3 swap of parameters, so it is not clear that a relationship between the frequency and severity means has been imposed. But $(\alpha\theta)^{\alpha}$ in this notation is:

$$(\alpha\theta)^{\alpha} = \lambda[\psi(2-p)]^{1/(p-1)}.$$

Thus taking $k = [\psi(2 - p)]^{1/(1 - p)}$ gives $\lambda = k(\alpha\theta)^{\alpha}$, which is the restriction originally imposed above. This k is not a function of μ and can also replace ψ by $\psi = k^{1 - p}/(2 - p)$. This gives a parameterization of the Tweedie in terms of k, p, and μ:

$$\lambda = \mu(\mu/k)^{1-p} \qquad \alpha = (2 - p)/(p - 1) \qquad \theta = (\mu/k)^{p-1}/\alpha$$

The mean is still μ, the frequency mean is k times the severity mean raised to the power $(2 - p)/(p - 1)$, and the aggregate variance is now $\mu^p k^{1-p}/(2 - p)$. Since p is $(\alpha+2)/(\alpha+1)$, it is between 1 and 2. The parameters are a bit simpler than Kaas' but the variance is more complicated than his $\psi\mu^p$. In any case skewness/CV is p, consistent with Renshaw's formula.

Not requiring the exponential family form gives other possibilities. Without imposing any relationship between frequency and severity, as noted above, the Poisson-gamma can be parameterized with mean μ and variance $\mu\theta(\alpha+1)$. This has replaced λ with $\mu/(\alpha\theta)$. A somewhat different relationship between frequency and severity can be established by setting $\lambda = (\alpha\theta)^k$. This gives mean $\mu = (\alpha\theta)^{k+1}$ and variance $(\alpha\theta)^{k+2}(1+1/\alpha) = \mu^{(k+2)/(k+1)}(1+1/\alpha)$, which is again proportional to a power of the mean between 1 and 2.

4.2 Poisson-Normal

A limiting case is the Poisson-normal. This has a point mass at zero but could have some negative observations. For the normal in m and s^2 it has mean $\mu = \lambda\mathrm{m}$, variance $\lambda(m^2+s^2) = \mu m[1+(s/m)^2]$ and skewness $(1+3\mathrm{CV}^2)\lambda^{-1/2}(1+\mathrm{CV}^2)^{-1.5}$. Fixing m and s and setting $\lambda_{w,d}$ to $\mu_{w,d}/m$ makes the variance proportional to the mean. Another possibility is to make λ and s constant and set $m_{w,d}$ to $\mu_{w,d}/\lambda$. Then the variance of each cell is $\mu_{w,d}{}^2/\lambda + \lambda s^2$. This is quadratic in $\mu_{w,d}$ and any $\mu_{w,d}$ can be negative. This is possible for the normal regression as well, but for the Poisson-normal, homoscedasticity is not required (or possible).

4.3 Poisson-Constant Severity Aggregates

The simplest aggregate loss distribution is probably Poisson frequency with a constant severity, called the PCS distribution. If θ is the severity, a cell with frequency λ has mean $\theta\lambda = \eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$ and variance $\theta^2\lambda = \theta\eta(\mathbf{z}_{\mathbf{w,d}}\boldsymbol{\beta})$. This is sometimes called the over-dispersed Poisson distribution, but PCS may be more descriptive, especially if $\theta < 1$. Some authors define the over-dispersed Poisson more broadly as any distribution in the exponential family for which the variance is proportional to the mean. But by uniqueness properties of the exponential family the PCS is the only such distribution, and so is the unique over-dispersed Poisson.

If X is the total loss random variable, X/θ is Poisson in $\lambda = EX/\theta = \mu/\theta$. Thus $\Pr(X/\theta = n) = e^{-\mu/\theta}(\mu/\theta)^n/n!$. For $x = \theta n$, $\Pr(X=x) = e^{-\mu/\theta}(\mu/\theta)^{x/\theta}/(x/\theta)!$. If x is not an integer multiple of θ, $\Pr(X=x) = 0$. If μ is modeled by covariates and parameters, say $\mu_{w,d} = U_ng_d$, with θ fixed, then an observation of $X_{w,d}$, say $q_{w,d}$, with $q_{w,d}/\theta$ a non-negative integer, has $\Pr(X_{w,d} = q_{w,d}) = p(q_{w,d}) = e^{-\mu \mathrm{w},d/\theta}(\mu_{w,d}/\theta)^{q_{w,d}/\theta}/(q_{w,d}/\theta)!$, and $p(q_{w,d})$ is zero otherwise. The PCS is a discrete distribution with positive probability only at integer multiples of θ. By its uniqueness, there is no continuous over-dispersed Poisson distribution in the exponential family. Thus over-dispersed Poisson probabilities are always zero except at integer multiples of θ.

A continuous analogue of the PCS is discussed in Mack (2002)[3] [6]. This can be described as a zero-modified continuous scaled Poisson, or ZMCSP. To specify it, start by using $p(x)/\theta$ as a density on the positive reals, extending the factorial by the gamma function, i.e., defining $a! \equiv \Gamma(1+a)$. But this density gives total probability above unity. Mack's solution is to reduce the probability mass at zero.

The ZMCSP is defined by the density $f(x;\mu,\theta) = e^{-\mu/\theta}(\mu/\theta)^{x/\theta}/[\theta(x/\theta)!]$ for $x > 0$ and by setting the point mass at $x = 0$ enough to make the total probability 1. To see how much probability is needed at 0, define the function $\mathrm{pois}(x,\lambda) = \lambda^x e^{-\lambda}/x!$ and the function $\mathrm{zm}(\lambda) = 1 - \int_{0+}^{\infty} \mathrm{pois}(x,\lambda)dx$. Then with a change of variable in $f(x)$ to $y = x/\theta$ and defining $\lambda = \mu/\theta$, it is easy to see that $\int_0^{\infty} f(x;\mu,\theta)dx$ is $1 - \mathrm{zm}(\lambda)$. Thus the point mass needed at zero is $\mathrm{zm}(\mu/\theta)$. The function $\mathrm{zm}(\lambda)$ is less than the Poisson's point mass of $e^{-\lambda}$ but is strictly positive.

There is an extra θ in the denominator of f that is not in p, but that will not affect the MLE of μ or the components of μ if μ is a function of covariates. This is interesting because setting $\mu_{w,d} = U_ng_d$ in the PCS and estimating by MLE is known to give the chain-ladder

[3] Chapter 1.3.7 [6].

reserve estimates. Since the estimates of U_w and g_d for Mack's ZMCSP will be the same as for the PCS (as long as there are not any zero observations), this looks like it extends the match of the chain ladder to the continuous case - no longer requiring that all cells in the triangle are integer multiples of θ. It turns out however that this is approximately but not exactly so.

The divergence arises from the fact that the ZMCSP expected value is not exactly μ. Integrating $xf(x)$ shows that the mean is actually:

$$EX = \mu[1 - \mathrm{zm}(\mu/\theta) + \int_{-1}^{0} \mathrm{pois}(x, \mu/\theta)dx].$$

This is greater than μ, but not by much, unless λ is small, as Table 1 shows. Since the function of μ needed to produce the mean depends on the parameters of the distribution, the ZMCSP is probably not a GLZ. As with the Pareto with infinite mean, extending the definition of GLZ a bit to include linear modeling of a parameter that is not the mean may make sense. Whether or not this is considered a GLZ, it is still a useful model.

The variance is a bit less than θμ for small values of λ. Integrating $x^2f(x)$ shows that $EX^2 = \theta^2\lambda \int_{0+}^{\infty} \mathrm{pois}(x-1, \lambda)xdx$. For large values of λ the integral is λ+1, but it is different for smaller λ.

Table 1: Point mass and moment adjustment by λ

λ = μ/θ	**zm(μ/θ)**	***EX*/μ – 1**	***EX*²/[θ²λ(λ+1)] – 1**	**Var/θ²λ– 1**
0.2	.48628	.33861	.03976	-0.11066
1	.16619	.03291	–8.73e-04	-0.06865
5	.00216	9.43e-05	-3.75e-06	-0.00097
25	3.19e-12	1.96e-14	-7.00e-13	-1.9E-11

In a recent study of a fairly noisy runoff triangle, μ/θ was less than two for just one observation and less than five for five observations, out of 55. Thus, a few small observations would have fitted means a percent or two different from the chain ladder's. While the noted match of the PCS and chain-ladder reserve estimates holds exactly only when all probability is concentrated on integer multiples of θ, the ZMCSP comes close to having this relationship in the continuous case.

4.4 Geometric – Exponential

The geometric frequency distribution can be described with a parameter α by $p_k = \alpha(1 - \alpha)^k$ for $k \geq 0$. This has mean $(1 - \alpha)/\alpha$ and variance $(1 - \alpha)/\alpha^2$, which is higher than the mean. With an exponential severity in mean θ, the aggregate distribution has mean $\theta(1 - \alpha)/\alpha$ and variance $\theta^2(1 - \alpha^2)/\alpha^2$. The aggregate survival function is known[4] to be $S(x) = (1 - \alpha)e^{-x\alpha/\theta}$. Both the frequency and aggregate distributions have a point mass of α at 0.

Either α or θ can be replaced by the mean μ, but probably keeping a constant θ would be useful more often. This replaces α by $\theta/(\mu+\theta)$. Thus when μ is higher, the probability α of an observation of zero is lower, which would make sense in many cases. The aggregate mean and variance become μ and $\mu(\mu+2\theta)$ with survival function $S(x) = \mu/(\mu+\theta)e^{-x/[\mu+\theta]}$. The variance is quadratic in the mean but with the linear term it increases more slowly than μ^2. For MLE the aggregate density is $f(x) = \mu/(\mu+\theta)^2e^{-x/[\mu+\theta]}$ for $x > 0$ and $p_0 = \theta/(\mu+\theta)$.

5 ESTIMATION ISSUES

Key to estimation is having an efficient optimizer to estimate the likelihood function including the covariates. Advances in computing power and the availability of optimization algorithms, even as spreadsheet add-ins, is what makes it possible to go beyond the exponential family and to use full MLE estimation.

The modified distributions like gamma p and lognormal basically substitute formulas for the usual parameters. For example in the gamma p, $F(x; \mu,\lambda) = \Gamma[x/(\lambda\mu^p); \mu^{1-p}/\lambda]$ can be written as $F(x) = \Gamma(x/\theta; \alpha)$ with $\theta = \lambda\mu^p$ and $\alpha = \mu^{1-p}/\lambda$. Thus a routine that searches for optimal gamma parameters can be used to estimate the gamma p by first expressing the gamma parameters in terms of λ, μ, and p and then searching for the best values for these three parameters. Since μ will be a function of covariates involving several parameters, this is the part where efficient algorithms comes in.

As long as there are no zero observations the ZMCSP loglikelihood function is

[4] See *Loss Models* [5], page 154.

$l = \sum \left(\frac{q_{w,d}}{\theta} \ln \frac{\mu_{w,d}}{\theta} - \frac{\mu_{w,d}}{\theta} - \ln\left(\frac{q_{w,d}}{\theta} ! \right) - \ln(\theta) \right)$. The last two terms in the sum can be omitted when maximizing for μ. In fact the function to maximize can be reduced to $l* = \sum \left(q_{w,d} \ln \mu_{w,d} - \mu_{w,d} \right)$. Taking the derivative shows that this is maximized when $0 = \sum \left(\frac{q_{w,d}}{\mu_{w,d}} - 1 \right)$. Thus the average relative error should be zero. If μ is a function of covariates and the vector **β** is being estimated, the derivative of *l** wrt the *j*th element of **β**, β*j*, gives the *n* equations $0 = \sum \frac{\partial \mu_{w,d}}{\partial \beta_j} \left(\frac{q_{w,d}}{\mu_{w,d}} - 1 \right)$. This could be considered a series of weighted average relative errors, all of which should be 0. After finding the estimates of the β*j*, the likelihood can be maximized for θ. The Poisson is analogous to the normal distribution case where the loglikelihood reduces to minimizing $\sum \left(q_{w,d} - \mu_{w,d} \right)^2$. This gives the *n* equations $0 = \sum \frac{\partial \mu_{w,d}}{\partial \beta_j} \left(q_{w,d} - \mu_{w,d} \right)$. Here the weighted average errors should be 0.

In non-parametric estimation, it is common to adopt the criterion of minimizing the sum of the squared errors, regardless of distribution. This treats a fixed squared error in any observation as equally bad – basically incorporating a constant variance assumption. This reduces to the normal distribution when in the exponential family, so minimizing squared error is a normal non-parametric approach. It sets the sum of weighted errors to 0. This is called unbiased, which sounds like something you want to be, but is not always that important.

If the same weighted relative error is equally bad across observations this is more of a Poisson assumption. This could also be used in a non-parametric context, where the weighted sums of relative errors are set to 0. This could be done without assuming the form of the distribution, so could be a Poisson non-parametric approach. The reasoning above shows that this results from finding the parameters that minimize $\sum \left(fitted - actual \ln \ fitted \right)$. This forces the actual/fitted toward 1.

For the Poisson-gamma aggregate and its special cases (Tweedie, etc.) the density for the likelihood function can be calculated by inverting the characteristic function $\varphi(t) = \exp[-1+\lambda(1 - it\theta)^{-\alpha}]$. Mong (1980) [9] worked out a purely real integral for this in terms of λ, α, θ and the aggregate standard deviation σ:

$$f(x) = \frac{1}{\sigma\pi}\int_0^\infty e^{\lambda j(t)} \cos\left[\frac{xt}{\sigma} - \lambda k(t)\right] dt\text{, where } j(t) = \delta(t)\cos[\rho(t)]-1\text{ ; } k(t) = \delta(t)\sin[\rho(t)]$$

$\delta(t) = [1+(t\theta/\sigma)^2]^{-\alpha/2}$; $\rho(t) = \alpha\tan^{-1}(t\theta/\sigma)$. The scaling by σ is probably done for numerical efficiency. With covariates, $\mu/\theta\alpha$ could replace λ in the characteristic function and its inversion. For the Tweedie it is also possible to express the density using an infinite sum, as in Clark and Thayer (2004) [1].

The gamma characteristic function is $\varphi_\Gamma(t) = (1 - it\theta)^{-\alpha}$, and $\varphi_\Gamma(t/\sigma) - 1 = j(t) + ik(t)$. For the normal distribution in m and s^2 the characteristic function is $\varphi_N(t) = \exp(itm - 0.5(st)^2)$. Scaling by s instead of σ gives $\varphi_N(t/s) - 1 = j(t) + ik(t)$ where $j(t) = \exp(-0.5t^2)\cos(tm/s) - 1$ and $k(t) = \exp(-0.5t^2)\sin(tm/s)$. These can be used in the integral above to give the Poisson-normal density if σ is replaced by s.

Mong's comments are: "(The) formula and its consequent computations may seem complex in the form shown above. However, the implementation is quite simple. Any standard numerical integration technique would handle the computation effectively; for example, the extended Simpson's rule is adequate to calculate the integration and is easy to code in any scientific programming language."

The extended Simpson's rule breaks down a finite range of integration into $2n$ intervals of length h, with $2n+1$ endpoints $x_0, \ldots, x_{2n}$. The function to be integrated is evaluated at each of the $2n+1$ points and multiplied by h. Then these are weighted by the following factors and summed: x_0 and x_{2n} get weight 1/3; odd points $x_1, x_3, \ldots, x_{2n-1}$ get weight 4/3; even points $x_2, \ldots, x_{2n-2}$ get weight 2/3.

Figure 1: Integrand for Poisson-gamma density

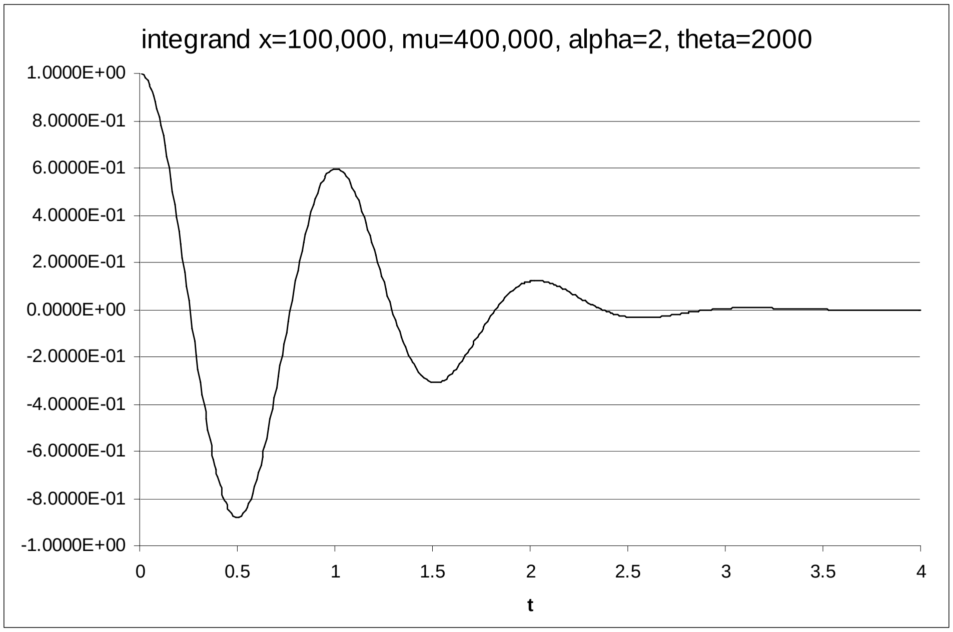


Figure 1 shows an example of the integrand for the Poisson-gamma density. This is for an x that is more than six standard deviations below the mean for a positively skewed distribution, so the integrated probability is low (7.5e-19). This makes the integration a bit more difficult as the dampening cycles have to get quite small before it stabilizes. However this occurred by about t = 10. Less remote probabilities have cycles that damp out more quickly.

There is a problem with this integral, however. The integration for $f(x)$ does not converge[5]! For both the gamma and normal severities, as t gets large $j(t)\rightarrow -1$ and $k(t) \rightarrow 0$. Thus the integrand becomes $e^{-\lambda}\cos(xt/\sigma)/(x\sigma)$, which fluctuates and does not go to 0. If λ is sufficiently large, this fluctuation is well beyond any reasonable degree of accuracy, and so is not a problem. Otherwise an alternative is to use the inversion formula for the distribution function to calculate $[F(x+\varepsilon) - F(x-\varepsilon)]/2\varepsilon$ for some appropriate ε, perhaps ½. According to Mong that inversion is: $F(x) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{e^{\lambda j(t)}}{t}\sin\left[\frac{xt}{\sigma} - \lambda k(t)\right]dt$, which does converge.

[5] My colleague John Major pointed this out.

6 DEVELOPMENT FACTOR EXAMPLE

Venter (2007) [12] fit the development triangle in Table 2 by a regression model for the incremental losses at lags 1 and above. The independent variables were the cumulative losses at lags 0 through 4, a dummy variable equal to 1 for the 3rd diagonal and 0 elsewhere, a dummy variable equal to 1 on the 4th, 7th, and 9th diagonals, -1 on the 10th diagonal, and 0 elsewhere, and a constant term. The diagonals are numbered starting at 0, so the 3rd is the one beginning with 7,888 and the 10th starts with 19,373. The cumulative loss independent variables are set to 0 for incremental losses that are not in the immediately following column.

Table 2: Cumulative Loss Development Triangle

Lag0	**Lag1**	**Lag2**	**Lag3**	**Lag4**	**Lag5**	**Lag6**	**Lag7**	**Lag8**	**Lag9**	**Lag10**	**Lag11**
11,305	30,210	47,683	57,904	61,235	63,907	64,599	65,744	66,488	66,599	66,640	66,652
8,828	22,781	34,286	41,954	44,897	45,981	46,670	46,849	47,864	48,090	48,105	48,721
8,271	23,595	32,968	44,684	50,318	52,940	53,791	54,172	54,188	54,216	54,775	
7,888	19,830	31,629	38,444	43,287	46,032	47,411	47,677	48,486	48,498		
8,529	23,835	35,778	45,238	51,336	53,574	54,067	54,203	54,214			
10,459	27,331	39,999	49,198	52,723	53,750	54,674	55,864				
8,178	20,205	32,354	38,592	43,223	44,142	44,577					
10,364	27,878	40,943	53,394	59,559	60,940						
11,855	32,505	55,758	64,933	75,244							
17,133	45,893	66,077	78,951								
19,373	50,464	75,584									
18,433	47,564										
20,640											

This is a loss development model with a constant term and calendar-year adjustments up through lag 5, but for lags 6 and beyond the constant term and the calendar-year adjustments operate but there are no development factors. The late development appears to be random in time and not dependent on the level of the accident year. There are heteroscedasticity issues, however. The smaller incremental losses at the end tend to have lower residuals – which actually seems desirable. Also the 0 to 1 development factor fits unreasonably well, so the residuals are also lower for the large increments at lag 1.

To address these issues, the same model was fit using Mack's ZMCSP distribution and the gamma *p*, where *p* was -0.29. The other parameters, negative loglikelihood, and AICc/2 are shown in Table 3.

Table 3: Parameters and fit statistics

	lag0	lag1	lag2	lag3	lag4	diag3	4+-10	const	θ,λ,σ	-lnL	AICc/2
ZMCPS	1.618	0.508	0.223	0.103	0.026	-2072	107.1	487.9	306.1	637.8	646.9
Gamma *p*	1.624	0.504	0.217	0.102	0.027	-1922	132.0	499.8	3,969.0	630.3	642.0
Normal	1.601	0.499	0.211	0.102	0.021	-1832	801.6	527.8	1,387.7	662.2	671.2

For N observations and p parameters, taking half of the small sample AIC, denoted AICc, penalizes the negative loglikelihood by $Np/(N - p - 1)$. For small samples ($N < 40p$) this is growing in popularity as the best way to penalize for extra parameters. Usually all parameters are penalized but for comparing fits maybe parameters that do not affect the fit can be ignored. Here for the normal and ZMCSP, p was set to 8, as θ and σ do not affect the fit. However for gamma p it was set to 10, as λ and p do. Still it gave the best AICc. N is 77 for this data.

The fit is clearly worse for the normal regression, reflecting the heteroscedasticity issue. The variance for the gamma p is $\mu^{0.71}$. Usually a power less than 1 is not anticipated, thinking of losses coming from a compound frequency-severity distribution. The abnormally good fit for the 0 to 1 factor, which has the largest observations, may be pushing the power down. The regression coefficients are quite similar for all the distributions, reflecting the common wisdom that heteroscedasticity does not greatly distort regression estimates. The distribution of possible results will vary among the distributions, however.

Figures 2 and 3 compare PP Plots for the normal and gamma p fits. The gamma p looks more reasonable. Figures 4 and 5 look at standardized residuals vs. fitted for the two distributions. They both look positively skewed, which they should be for gamma p but not for normal. Also the normal extremes are more extreme. The small fitted values have standardized residuals more like the other values for the gamma p, but not for the normal. Overall the gamma p seems to fit better.

Figure 2

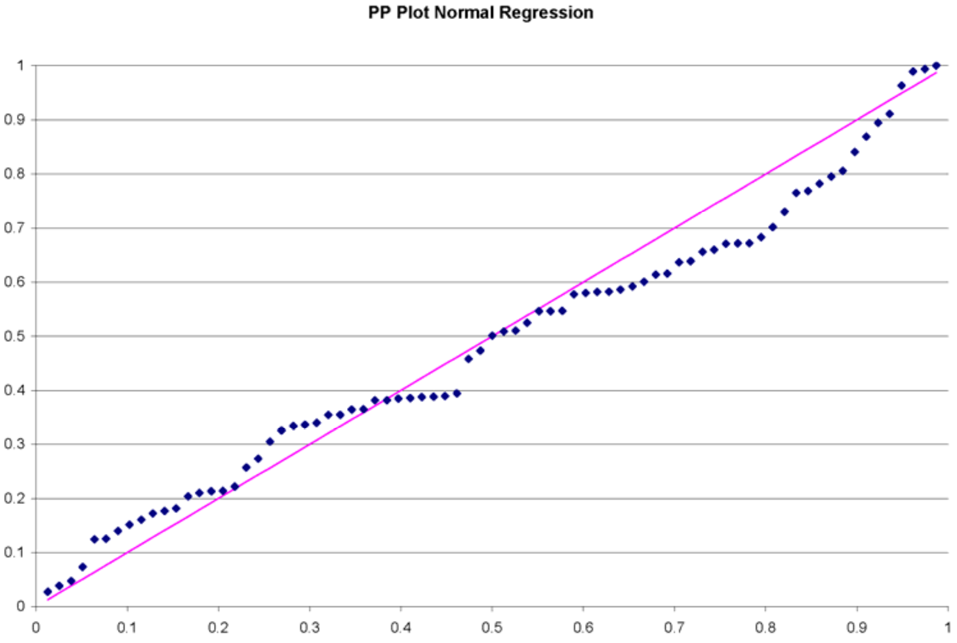


Figure 3

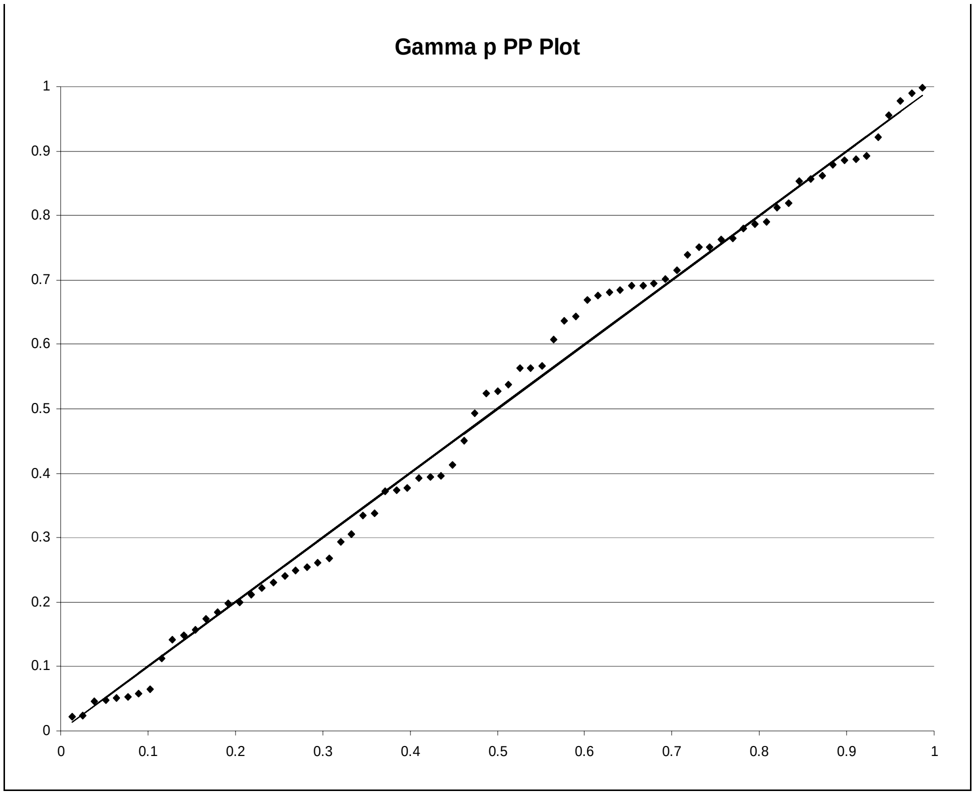


Figure 4

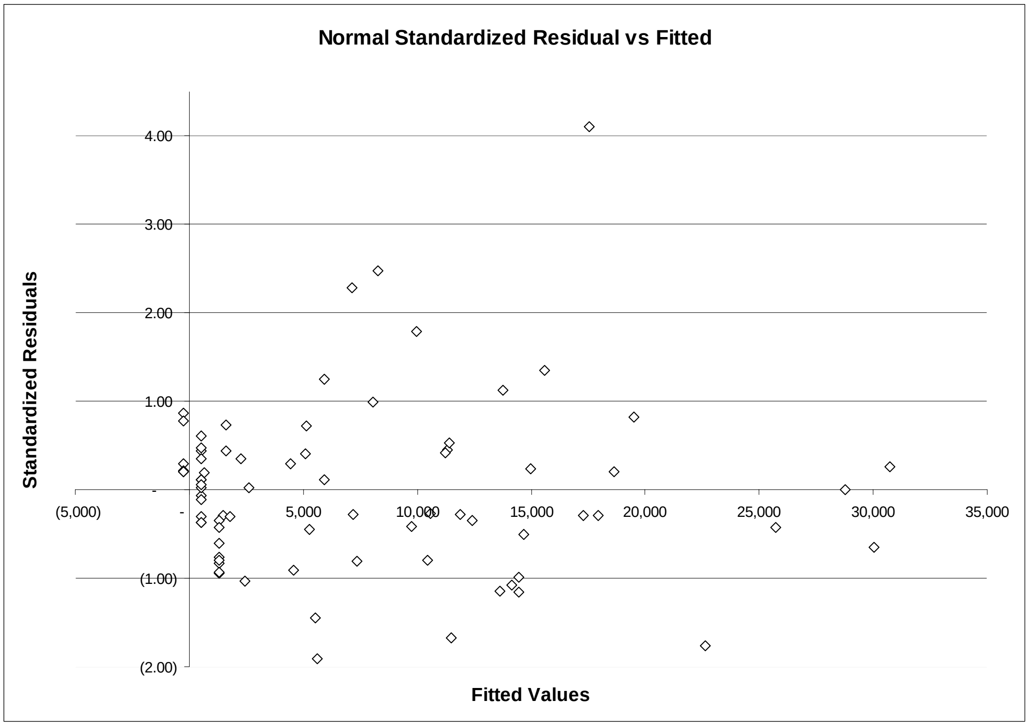
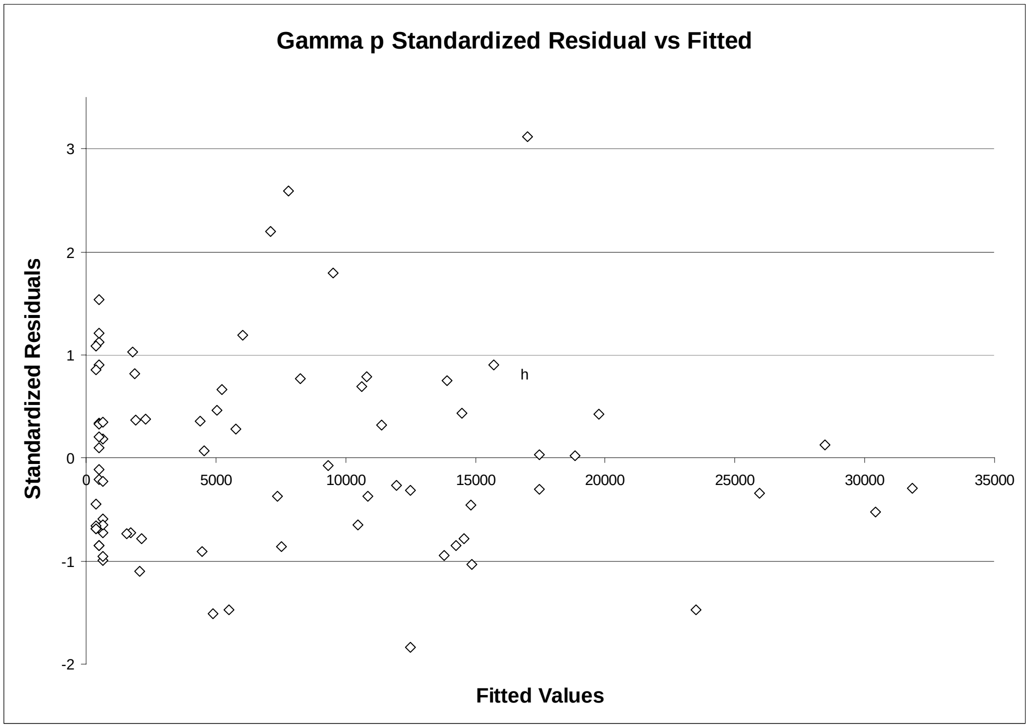


Figure 5



7 MULTIPLICATIVE MODEL ISSUES

Multiplicative fixed-effects models can be treated in the GLZ framework. Take the case where $\mu_{w,d} = Eq_{w,d} = U_w g_d h_{w+d}$. The covariates are 0, 1 dummies picking out which factors apply to each cell, and the vector of coefficients $\boldsymbol{\beta}$ is the log of all the accident year factors U_w followed by the log of all the delay factors g_d followed by the log of all the calendar year factors h_{w+d} in the model. Let $\mathbf{z_{w,d}}$ be the vector that has zero in all positions except for ones for the positions of the w^{th} row, d^{th} column and $w+d^{\text{th}}$ diagonal. Then $\eta(\mathbf{z_{w,d}}\boldsymbol{\beta}) = \exp(\mathbf{z_{w,d}}\boldsymbol{\beta})$ is $E\mu_{w,d}$. This can be used in any of the distributions discussed above. However the factors all have to be positive to take the logs, even though some observations can be negative with the right choice of distribution around the mean. However, if negative means are needed for some columns, $\mu_{w,d} = Eq_{w,d} = U_w g_d h_{w+d}$ with some negative g's can be used directly as the mean of any of the distributions discussed. This could be fit by MLE, but it would not really be considered a linear model any more, unless $\boldsymbol{\beta}$ is allowed to have complex coefficients that become negative reals when exponentiated. The line between GLZ and truly non-linear models is thus a bit imprecise, but the labeling is not really very important anyway.

Fu and Wu (2005) [2] provide an iterative scheme, using constants labeled here as r and s, that can in some cases help in the estimation of multiplicative models. The Fu-Wu iteration for the row-column model can be expressed as[6]:

$$g_d = \left\{ \sum_{w=0}^{n-d} U_w^{r-s} q_{w,d}^s \Big/ \sum_{w=0}^{n-d} U_w^r \right\}^{1/s} \text{ and } U_w = \left\{ \sum_{d=0}^{n-w} g_d^{r-s} q_{w,d}^s \Big/ \sum_{d=0}^{n-w} g_d^r \right\}^{1/s}.$$

The idea is to start with reasonable guesses for the U's and then alternatively apply the two formulas to get new g's and U's until they converge. Often this iteration gives the MLE for some model. For instance, taking $r = 2$ and $s = 1$ gives the normal regression. The case $r=s=1$ gives the estimate where $q_{w,d}$ is Poisson in $U_w g_d$. Both of these cases work fine if some column of q's tends to be negative and so its mean g is as well. Mildenhall (2005) [8] shows that there is a model for each r and s for which this iteration gives a reasonable estimate. The cases s=1, r = -1, 0, 1, and 2 are the inverse Gaussian, gamma, Poisson, and normal

[6] They also include weighting factors that here are set to unity.

distributions, respectively, and the estimates are MLE for the β's if the other parameters are known or do not affect the estimates of the β's.

With arbitrary *s* the power transforms of these distributions are realized. Taking *r*=0 gives the transformed gamma or inverse transformed gamma, depending on the sign of s, and so a wide range of distribution shapes. If 1 < *r* < 2 and *s* = 1, the Tweedie with *p* = *r* is produced. If *p* and ψ are known, the iteration gives the MLE for the β's. This could be done within an optimization routine that is looking for the MLE values for *p* and ψ, so would only require a routine that works for two variables.

For the multiplicative models with diagonal factors $E[q_{w,d}] = U_w g_d h_{w+d}$, the Fu-Wu iterative estimates become:

$$g_d = \left[\sum_{w=0}^{n-d} \left(U_w h_{w+d}\right)^{r-s} q_{w,d}^{s} \Big/ \sum_{w=0}^{n-d} \left(U_w h_{w+d}\right)^{r} \right]^{1/s},$$

$$U_w = \left[\sum_{d=0}^{n-w} \left(g_d h_{w+d}\right)^{r-s} q_{w,d}^{s} \Big/ \sum_{d=0}^{n-w} \left(g_d h_{w+d}\right)^{r} \right]^{1/s}, \text{ and}$$

$$h_j = \left[\sum_{w+d=j} \left(U_w g_d\right)^{r-s} q_{w,d}^{s} \Big/ \sum_{w+d=j} \left(U_w g_d\right)^{r} \right]^{1/s}.$$

8 MULTIPLICATIVE MODEL EXAMPLE

Table 4 is a development triangle from Taylor-Ashe (1983) [11]. Venter (2007) [12] fit a form of the PCS multiplicative effects model to this data. Each cell $\mu_{w,d}$ was set to the product of row, column, and diagonal effects, but some parameters are used more than once. Accident year 0, a low year, gets its own parameter U_0. Accident year 7 also gets its own parameter U_7 as it is high. All the other years get the same parameter U_a, except year 6 which is a transition and gets the average of U_a and U_7. Thus, there are three accident-year parameters.

The years are divided into high and low payment years with parameters g_a and g_b for fraction of total loss paid in the year. Delay 0 is a low year as payments start slowly. Delays 1, 2, and 3 are the higher payment lags and all get g_b. Delays 5, 6, 7, and 8 are again low getting

g_a. Delay 4 is a transition and gets the average of g_a and g_b. Finally delay 9 gets the rest, i.e., 1 – 5.5g_a – 3.5g_b. Thus there are only two delay parameters. Three of the diagonals were modeled as high or low, getting factors 1+c or 1–c. The 7th diagonal is low and the 4th and 6th are high. Thus, only one diagonal parameter c is used. The diagonals are numbered from 0, so the 7th starts with 359,480.

Table 4: Incremental triangle Taylor-Ashe (1983) [11]

Lag 0	**Lag 1**	**Lag 2**	**Lag 3**	**Lag 4**	**Lag 5**	**Lag 6**	**Lag 7**	**Lag 8**	**Lag 9**
357,848	766,940	610,542	482,940	527,326	574,398	146,342	139,950	227,229	67,948
352,118	884,021	933,894	1,183,289	445,745	320,996	527,804	266,172	425,046	
290,507	1,001,799	926,219	1,016,654	750,816	146,923	495,992	280,405		
310,608	1,108,250	776,189	1,562,400	272,482	352,053	206,286			
443,160	693,190	991,983	769,488	504,851	470,639				
396,132	937,085	847,498	805,037	705,960					
440,832	847,631	1,131,398	1,063,269						
359,480	1,061,648	1,443,370							
376,686	986,608								
344,014									

Fitting the PCS is done by maximizing $l^* = \sum \left(q_{w,d} \ln \mu_{w,d} - \mu_{w,d}\right)$, where $\mu_{w,d}$ = $U_w g_d h_{w+d}$. This pretends that every observation $q_{w,d}$ is a multiple of θ, as in fact the PCS probability is zero otherwise. This is the same function to be maximized for fitting the ZMCSP, which does not require observations to be multiples of θ. Thus, the row, column, and diagonal parameters are the same for both models. The difference is that θ is fit by an ad hoc method for the PCS and by MLE for ZMCSP. The likelihood function is $l = \sum \left(\frac{q_{w,d}}{\theta} \ln \frac{\mu_{w,d}}{\theta} - \frac{\mu_{w,d}}{\theta} - \ln\left(\frac{q_{w,d}}{\theta}!\right) - \ln(\theta) \right)$, and now θ is the only parameter needed to maximize over. The MLE estimate of θ is 30,892. Estimating it by a moments method $\hat{\theta} = \frac{1}{N-p} \sum_{w,d} \frac{\left(q_{w,d} - U_w g_d\right)^2}{U_w g_d}$ gives 37,184.

Just changing θ makes a difference in the estimated runoff distribution and parameter errors. The estimated parameters and their PCS standard errors from the information matrix with the moment and MLE θ's are in Table 5. The runoff variance separated into process and parameter is in Table 6.

Table 5: Parameter se's with two estimates of θ

Parameter	U_0	U_7	U_a	g_a	g_b	c
Estimate	3,810,000	7,113,775	5,151,180	0.067875	0.173958	0.198533
se 37,184	372,849	698,091	220,508	0.003431	0.005641	0.056896
se 30,892	339,846	636,298	200,989	0.003127	0.005142	0.051860

Table 6: Runoff Variance with two estimates of θ

Model	Moment 37,184	MLE 30,892
Parameter Variance	1,103,569,529,544	916,846,252,340
Process Variance	718,924,545,072	597,282,959,722
Total Variance	1,822,494,074,616	1,514,129,212,061
Parameter Std Dev	1,050,509	957,521
Process Std Dev	847,894	772,841
Standard Deviation	1,349,998	1,230,500

So far this is all from keeping the PCS framework and replacing the estimated θ from the moment method by that from MLE from ZMCSP. The ability to estimate θ by MLE is actually the main difference between the two distributional assumptions. In this case the MLE θ is quite a bit lower, which gives a lower variance. It is also useful to have an optimized negative loglikelihood to compare to other models, as in the development factor example. Here that is 725.

Recall that the mean and variance of each cell differs a little from μ and θμ in the ZMCSP model for the smaller cells. In this case only the last projected column has low values of λ = μ/θ and these are around 3. This has only a very slight effect on the projected mean and variance. The estimated reserve of 19,334,000 increases by about 1,000 and the standard deviation of 1,230,500 decreases by about 100. Thus in this case that is a very minor impact. Only the change in the estimated θ has any significant influence on the projections.

A good starting point for investigating other possible distributions for the same models structure is fitting the gamma *p*. Aggregate losses are often approximately gamma distributed, and the value of *p* gives an indication of how the variance can be expressed as a multiple of the mean.

For this data the MLE of p is -0.136, which gives the variance as proportional to the mean raised to 0.864. This is not suggestive of any other popular models, however. The negative loglikelihood is 723.06 compared to 725.00 for the ZMCSP. With 8 parameters compared to 7 for the ZMCSP, the AICc's come out as 732.6 and 733.2, so the gamma p is a little lower. However, if only 6 parameters are counted for the ZMCSP under the view that θ does not affect the fit, its AICc reduces to 731.9. Thus, there is some ambiguity as to which is the best fit. Better ways of counting the degrees of freedom a model uses up would be helpful. In any case the variance is close to proportional to the mean in either model.

Another model with that property is the Poisson-normal. MLE using Mong's formula for $f(x)$ gives m = 35,242 and s = 3,081, with λ's ranging from 2 to 35. The negative loglikelihood is 722.4, which is the best so far. The resulting AICc for 8 parameters is 732.0, which is still ambiguous in comparison to the ZMCSP. The integral for $f(x)$ for the one cell with $\lambda = 2$ is of limited accuracy, so there is a slight degree of additional ambiguity in the value of the AICc.

9 CONCLUSIONS

GLMs provide a powerful modeling tool, but the exponential family has some limitations. By not requiring this form, even familiar distributions can be reparameterized to provide different relationships between the mean and variance of the instances of the fitted dependent variable. When fitting aggregate loss distributions, the gamma is often a good starting point for the shape of the distribution, and so fitting the gamma p, which is a gamma but allows for the variance to be any desired power of the mean, is often a good way to get an indication of the form of the variance to mean relationship. Other distributions can then be tried which have approximately that relationship.

Even when using exponential family distributions, computing power is usually sufficient to calculate the full likelihood function, instead of approximations sometimes used in GLMs. GLZs thus expand the limitations of GLMs, yet there are still situations where it may be useful to use strictly nonlinear models.

REFERENCES

[1] Clark, David R. and Charles A. Thayer. 2004. "A Primer on the Exponential Family of Distributions." *CAS Discussion Paper Program*, 117-148.

[2] Fu, Luyang and Cheng-sheng Peter Wu. 2005. "Generalized Minimum Bias Models." *CAS Forum*,(Winter): 73-121.

[3] Hewitt, Charles C. 1966. "Distribution by Size of Risk-A Model." *PCAS* 53:106-114.

[4] Kaas, Rob. 2005. "Compound Poisson Distributions And GLM's — Tweedie's Distribution."
Lecture, Royal Flemish Academy of Belgium for Science and the Arts,
http://www.kuleuven.be/ucs/seminars_events/other/files/3afmd/Kaas.PDF.

[5] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. 2004. *Loss Models: From Data to Decisions*, 2nd Ed. Hoboken, NJ: Wiley.

[6] Mack, Thomas. 2002. *Schadenversicherungsmathematik*, 2nd Ed. Karlsrühe, Ger.:Verlag Versicherungs-wirtshaft.

[7] Mildenhall, Stephen J. 1999. "A Systematic Relationship between Minimum Bias and Generalized Linear Models." *PCAS* 76:393-487.

[8] Mildenhall, Stephen J. 2005. "Discussion of Generalized Minimum Bias Models." *CAS Forum* (Winter): 122-124.

[9] Mong, Shaw. 1980. "Estimating Aggregate Loss Probability and Increased Limit Factor." *CAS Discussion Paper Program:* 358-393.

[10] Renshaw, Arthur E. 1994. "Modeling the Claims Process in the Presence of Covariates." *ASTIN Bulletin* 24, no.:2:265–286.

[11] Taylor, Greg C. and Frank R. Ashe. 1983. "Second Moments of Estimates of Outstanding Claims." *Journal of Econometrics* 23:37-61.

[12] Venter, Gary G. 2007. "Refining Reserve Runoff Ranges." *CAS E-Forum,* Summer (forthcoming)..

Biography of the Author

Gary Venter is managing director at Guy Carpenter, LLC. He has an undergraduate degree in philosophy and mathematics from the University of California and an MS in mathematics from Stanford University. He has previously worked at Fireman's Fund, Prudential Reinsurance, NCCI, Workers Compensation Reinsurance Bureau and Sedgwick Re, some of which still exist in one form or another. At Guy Carpenter, Gary develops risk management and risk modeling methodology for application to insurance companies. He also teaches a graduate course in loss modeling at Columbia University.

917.937.3277 **gary.g.venter@guycarp.com**

Refining Reserve Runoff Ranges

Gary G. Venter, FCAS, MAAA

Abstract

Reserve runoff ranges are often wider than they need to be. This paper applies some practical tools used by regression modelers to find ways to reduce the ranges. Four approaches are explored: finding better-fitting models; getting rid of insignificant parameters; using exposure information; and considering whether some part of the triangle should be ignored.

Keywords. Loss reserving; regression modeling; range estimates; parameter reduction.

1. INTRODUCTION

Techniques that can reduce the runoff variance and reserve ranges are outlined and illustrated through three examples of fitting models to development triangles. Two basic paradigms for development models are used:

[1] Future development is a proportion of losses emerged to date, plus a random error.

[1] Future development is a proportion of the as yet unknown ultimate, plus a random error.

The chain ladder method is the paragon of the first paradigm, and the Bornhuetter-Ferguson (BF) method is an early example of the second. Multiplicative effects models, where the mean of each cell is a product of a row and column parameter, are also of the second type, as the row parameters can be scaled to be expected ultimate.

The factors estimated for both model types can be distorted if there are diagonal (calendar-year) influences in the data. It is possible to identify and take into account such influences in either of the modeling paradigms. This is investigated in all of the examples.

Exposure information, if available, can also improve the model fit and reduce the variance and ranges. Also there are situations where the common models fit better to a portion of the triangle than to the whole triangle, and this is explored as well.

The first example is a triangle whose development pattern is much better explained as a factor times ultimate than a factor times already emerged, but the multiplicative effects model has so many parameters that the estimated variance and runoff ranges are higher than for the chain ladder, despite the better fit, due to greater parameter uncertainty. Ways to maintain a good fit while eliminating insignificant parameters are explored, and lead to a lower variance. These are somewhat ad hoc methods out of the regression modelers' tool bag. Their application is more of an art than a science but they can produce better models in many cases. The multiplicative effects model can easily handle calendar-year influences by including row, column, and diagonal factors.

The second example is one in which development factors appear to provide a reasonable fit to the data, at least at the early lags. The chain ladder is often presented in a regression context, where factors are calculated using some form of regression on the previous cumulative losses. That gives a separate variance for each factor. It is possible to include diagonal effects in the chain ladder, but the factors have to be computed in a single overall regression. This can get into problems with heteroscedasticity, where a single variance is assumed for each cell but latter lags in fact have lower variances. This does not usually affect the parameter estimates very much, but it does distort the estimated runoff variance. A heteroscedasticity adjustment is introduced and applied to this case. Further use of parameter-reduction techniques are also illustrated.

The third example is of a triangle that exhibits a good deal of change in development patterns over time, and ways to test for that are explored. It also has exposure information available, and using that improves the model. Parameter reduction by fitting a distribution to the emergence lag pattern is applied to this triangle as well.

Section 2 reviews some details of the two modeling paradigms and provides a common notation to discuss them. Section 3 addresses how to compare fits of alternative models. Sections 4, 5, and 6 are the three examples. Section 7 concludes. Standard assumptions, discussed in each case, are used for the distributions of the error terms, but other distributions could be used. These are beyond the scope of this paper, but should not be ignored in application.

2. BACKGROUND ON DEVELOPMENT TRIANGLE MODELS

Mack (1993) [13] presents statistical assumptions and criteria under which the chain-ladder estimate is optimal, and shows how to calculate the implied variance. Mack's assumptions are intuitive from the viewpoint of what actuaries might imagine development factors are doing. Basically they postulate that the incremental losses at a given lag are a factor times the previous cumulative, plus a random innovation.

Having a model like Mack's allows for testing how well the chain-ladder assumptions apply to specific triangles[1]. Which model works best for a given data set is an empirical matter, but when the chain-ladder assumptions fail it is often because incremental losses are not fit well as a factor times previous cumulative. Then the losses at each lag might be modeled as a fraction of the yet-unknown ultimate losses. This is an element of the Bornheutter-Ferguson approach, so all such models can be regarded as formalized versions of BF. Typically these take the form of multiplicative fixed-effects models (MFE), where each cell's expected loss is a product of row and column (and perhaps diagonal) factors.

2.1 Variants of Chain Ladder

Murphy (1994) [16] gives several versions of the chain ladder in a regression setting. Losses at one age are expressed as a factor times the cumulative losses at a previous age plus a random error, plus possibly a constant term. For each age the variance of the random error could be constant, or it could be proportional to the level of the previous cumulative losses, or to the square of the previous cumulative. Murphy shows that for the model with no additive term and a constant variance, standard regression theory gives the estimator $\Sigma xy/\Sigma x^2$, where y represents the current losses and x the previous. He calls this the LSM model, for least-squares multiplicative. Using transformed regressions Murphy shows that the factor estimators when the variance is proportional to losses or losses squared are $\Sigma y/\Sigma x$ and average(y/x), respectively. $\Sigma y/\Sigma x$ is typical in actuarial applications and is the same estimator as Mack's. It is the regression estimator for a no-constant regression of $y/x^{1/2}$ on $x^{1/2}$ that converts the constant variance to a variance proportional to x. Unfortunately it is difficult to tell which behavior of the variance best holds for a single column, so judgment is often needed.

2.2 Multiplicative Fixed Effects Models

These models express the losses in a cell in a triangle as a product of a row constant and a column constant, which are the fixed effects plus a random innovation. Some notation is needed to discuss this.

The n+1 columns of a triangle are numbered 0, 1, … n and denoted by the subscript d. The rows are also numbered from 0 and denoted by w. The last observation in each row of a full triangle will then have $w+d=n$. The cumulative losses in cell w,d are denoted $c_{w,d}$ and the incrementals by $q_{w,d}$.

For the MFE model, $E[q_{w,d}]$ is $U_w g_d$, where U_w and g_d are the row and column parameters,

[1] See for example Mack (1994) [14], Venter (1998) [21], Barnett and Zehnwirth (2000) [3].

respectively. Note that increasing each *g* by the same factor and dividing each *U* by that factor does not change the mean for any cell. To have specificity, it is often convenient to have the *g*'s sum to 1. Then U_w can be interpreted as the ultimate loss for year *w* and g_d the fraction that are at lag *d*.

Assuming that the distribution around the cell mean is lognormal, each cell's observation is log [$q_{w,d}$] =log U_w+ log g_d + $\varepsilon_{w,d}$, which is a linear model with a normal error term, and so estimable by regression. This was already studied by Kremer (1982) [9]. On the other hand, if the distribution is normal, so $q_{\mathrm{w},d} = U_w g_d + \varepsilon_{w,d}$, the model is non-linear. Mack (1991) [12] linked this model of development triangles to MFE models in classification ratemaking, such as those in Bailey (1963) [1], Bailey-Simon (1960) [2], etc. These models can be estimated by a generalization of fixed-point iteration called Jacobi iteration, using $g_d = \sum_{w=0}^{n-d} U_w q_{w,d} \Big/ \sum_{w=0}^{n-d} U_w^2$ and $U_w = \sum_{d=0}^{n-w} g_d q_{w,d} \Big/ \sum_{d=0}^{n-w} g_d^2$. This is just the result of alternatively treating the *g*'s and the *U*'s as known constants, so the model temporarily becomes a simple factor model in the other parameter.

2.3 Poisson – Constant Severity Distribution

A convenient starting point for multiplicative fixed-effects models is to assume the error terms follow the Poisson – constant severity (PCS) distribution. This is the aggregate loss distribution consisting of a Poisson frequency and a constant severity. In this context that assumes all claims or payments in all cells are the same size, call it *b*. This of course is rarely the case, but the model has some advantages. First, it is a distribution of aggregate claims, which most triangles consist of. However its historical appeal is that an PCS model estimated by MLE gives the same reserve estimate as the chain ladder.

In the pure Poisson case, the agreement of methods was shown by Hachemeister and Stanard (1975) [6] although that finding was not published formally until Kremer (1985) [10] in German (translated into Russian as well) and Mack (1991) [12] in English. Renshaw and Verrall (1998) [17] extend this to the over-dispersed Poisson, which in generalized linear model terminology is defined as any member of the exponential family whose variance is proportional to its mean. However the only distribution meeting this criterion is the PCS. A good presentation is Clark (2003) [4], who in addition uses a parameterized distribution for the payout pattern. None of the cited papers compare the MFE – PCS variance to the chain ladder's, however.

Giving the same answer as the chain ladder is not a particularly useful criterion for evaluating models, but it starts from a familiar base. Thus the error terms will be assumed approximately PCS distributed for MFE models here.

For the PCS model, a cell with frequency λ has mean $b\lambda$ and variance $b^2\lambda$. For the MFE implementation then $b\lambda_{w,d} = U_wg_d$. This model is applied here to incremental losses, so that the observation $q_{w,d}/b$ is Poisson with mean U_wg_d/b. The loglikelihood function[2] can be shown to be:

$l = C + \sum\left(\frac{q_{w,d}}{b}\ln\frac{U_wg_d}{b} - \frac{U_wg_d}{b}\right)$, where $C = -\Sigma \ln \Gamma(1 + q_{w,d}/b) \equiv -\Sigma \ln [(q_{w,d}/b)!]$. Taking derivatives, the MLE estimates can be expressed as: $g_d = \sum_{w=0}^{n-d} q_{w,d} \Big/ \sum_{w=0}^{n-d} U_w$ and $U_w = \sum_{d=0}^{n-w} q_{w,d} \Big/ \sum_{d=0}^{n-w} g_d$, which do not depend on b. Technically, the Poisson probabilities are zero unless $q_{w,d}$ is an integral multiple of b. However Mack (2002) [15], chapter 1.3.7, shows that there is a continuous analogue of the Poisson that can be scaled by b and gives estimates close to the PCS. When the PCS is applied in a continuous setting it can be thought of as using this distribution. For more details see Venter (2007) [22].

The MLE formulas can be solved by iteration, starting with some values then solving alternatively for the *g*'s and *U*'s until the results converge. If then the *g*'s do not sum to 1, just divide each by their sum and multiply each *U* by the same sum. Starting at the upper right corner of the triangle and working back can show that these estimates correspond to the chain-ladder calculation. Essentially the *U*'s are the last diagonal grossed up to ultimate by the development factors and the *g*'s are the factors converted to a distribution of ultimate. The fitted incrementals are then the *g*'s applied to the *U*'s, and can be calculated by using the development factors to back cumulatives down from the last diagonal.

From the chain-ladder viewpoint these use future information to predict the past, but this is not the chain-ladder paradigm. Sometimes incremental losses are better fit as a fraction of ultimate (MFE model) than as a factor times previous cumulative (chain-ladder model). The drawback is that

[2] Note that this requires not fitting just one Poisson distribution but $(n/2+1)(n+1)$ of them, defined by $2n+1$ row-column parameters plus b. But MLE applies to fitting multiple distributions with the same parameters. This is noted in the *Loss Models* textbook [8], for instance.

there are more parameters needed for MFE. The chain ladder models each column conditionally on the previous column and does not estimate the first column of the triangle. It requires the calculation of *n* factors. The PCS model does estimate the first column but uses 2*n*+1 parameters. Comparing the fits of the two models is thus awkward. Perhaps comparing the estimated variances is the best way to do this. The process variances can be thought of as measuring the accuracy of the models, and the parameter variance is the parameter penalty.

Clark (2003) [4] discusses calculating the MFE – PCS variance. First an estimate of *b* is needed. Since the variance of each cell is *b* times its mean, he suggests estimating *b* by:

$$\widehat{b} = \frac{1}{N-p} \sum_{w,d} \frac{\left(q_{w,d} - U_w g_d\right)^2}{U_w g_d}.$$

This is a kind of moment matching, but it is not clear how good an estimate of *b* this might be. The estimated variance of each projected incremental cell is the cell's mean times this *b*, and so the reserve variance is the reserve times *b*. This is the process variance, assuming all the parameters are known. Since in fact they are estimated, another element of reserve variance is the parameter variance. Clark suggests estimating this by the delta method. The delta method (see *Loss Models*) starts with the usual covariance matrix of the parameters, calculated as the inverse of the MLE information matrix (matrix of 2nd derivatives of the negative loglikelihood wrt the parameters). The delta method calculates the parameter variance of a function of the parameters by the covariance matrix left and right multiplied by the vector of the derivatives of the function wrt the parameters. In this case the function of the parameters is the reserve. For the PCS model, the 2nd derivatives of the loglikelihood function wrt the parameters are:

$$\frac{\partial^2 l}{\partial U_w^2} = -\sum_{d=0}^{n-w} \frac{q_{w,d}}{bU_w^2} \quad ; \quad \frac{\partial^2 l}{\partial g_d^2} = -\sum_{w=0}^{n-d} \frac{q_{w,d}}{bg_d^2} \quad ; \quad \frac{\partial^2 l}{\partial U_w \partial g_d} = -\frac{1}{b}, \text{ otherwise } 0.$$

The derivative of the reserve wrt g_d is $\Sigma_{w>n-\mathrm{d}} U_w$ and wrt U_w is $\Sigma_{d>n-w} g_d$. But with g_n set to $1 - \Sigma_{d<n} g_d$, these have to be adjusted. First $\frac{\partial^2 l}{\partial U_0 \partial g_d} = 0$. Also now $\frac{\partial^2 l}{\partial g_d^2} = -\frac{q_{0,n}}{bg_n^2} - \sum_{w=0}^{n-d} \frac{q_{w,d}}{bg_d^2}$ and for $d \neq j$, $\frac{\partial^2 l}{\partial g_d \partial g_j} = -\frac{q_{0,n}}{bg_n^2}$. The derivative of the reserve wrt U_w is not affected by this adjustment, but wrt

g_d it is $-\sum_{w=1}^{n-d} U_w$.

2.4 Adding in Calendar-Year Effects

Diagonal effects can be a result of accelerated or stalled claim department activity in a calendar year. Such a departure would often be made up for in a later year or years, so more than one diagonal can be affected. A similar pattern can arise from inflation operating on calendar years. Inflation operating on accident year is in the factor approach, as each year gets its own level. But there can appear to be inflation by accident year that is actually generated by calendar year. If that inflation varies by year, high and low residuals can show up by diagonal. Large differences in residuals among diagonals would suggest that either calendar-year inflation or claim department variation is operating. In many cases there are diagonal effects in triangles, and modeling them can provide better fits. Not accounting for such effects when they are present can lead to misestimating row and column parameters.

Taylor (1977) [18], following Verbeek (1972) [23], discusses a method for estimating calendar-year effects, which he calls the separation method. For some decades after that, models of calendar-year effects were informally called separation models, even when they did not use that technique.

In the lognormal MFE model given by $q_{w,d} = U_w g_d h_{w+d}(1+\eta_{w,d})$, taking logs gives log $q_{w,d}$ = log U_w + log g_d + log h_{w+d} + $\varepsilon_{w,d}$, which is a linear multiple regression model.

Barnett and Zehnwirth (2000) [3] set up a model framework of this type, but in a way that facilitates parameter reduction. They denote log U_w by α_w and express log g_d = $\sum_{k=1}^{d}\gamma_k$ and log h_{w+d} = $\sum_{t=1}^{w+d}\iota_t$. This makes $\gamma_d = \log[g_d/g_{d\text{-}1}]$, for instance. Thus it makes sense to call γ a trend. If the g's are trending upwards or downwards by a power curve for several columns, the same γ can be used for those columns, reducing the number of parameters in the model. Similarly the ι's are trends over calendar years and may be constant for a few years, reducing the number of diagonal parameters.

3. COMPARING MODELS

This paper's goal is finding ways to increase the accuracy and reduce the variance and ranges of reserve estimates. A lower predictive variance is suggestive but not absolutely definitive for having the best model. Calculating variances can also be tedious. Thus, when searching for models, variances are calculated only for a few models and comparison of fits are based on other criteria from information theory. The original information criterion, Akaike's information criterion, or AIC, can be interpreted as imposing a penalty of 1 to the maximized loglikelihood for each parameter in the model. This is often regarded as too low a penalty, however. The Hannan-Quinn information criterion (HQIC) has a per-parameter penalty of the log of the log of the number of observations N. For instance for a 10×10 triangle with 55 observations, this gives a penalty of 1.388 for each parameter. The Schwartz-Bayesian information criterion penalty is higher, at the log of the square root of N, which is per-parameter penalty of 2 for 55 observations. This may be a bit high, however. An alternative is the small sample AIC, denoted by AIC_c. Its per-parameter penalty with p parameters is $N/[N - p - 1]$, which increases with the number of parameters. The penalty is a bit less than that of the HQIC when there are not too many parameters, but is higher with over-parameterized models. A typical standard for what is a small sample is anything less than 40 times the number of parameters, so would include most loss-development triangles.

Here the AIC_c is favored but the HQIC also used. The formal criteria are actually double what are stated above, but dividing by 2 is convenient in that it directly penalizes the loglikelihood. Since the MFE – PCS loglikelihood increases with b, as does the variance, worse fitting models with a higher variance can have a higher loglikelihood. Thus, comparing likelihoods across PCS models requires fixing a value of b and using it for different models. The choice of b affects the scale of the loglikelihood and, thus, the meaning of the parameter penalties. Therefore, these can only be regarded as general guidelines and not strict cutoffs for this model.

4. EXAMPLE 1

In this example the MFE – PCS model is fit to a triangle that has often been used as an example and for which the Mack estimates are known. This is first fit by the MFE – PCS model, then some diagonal parameters are added in, and then ways to reduce the number of parameters used are explored. The starting point in Table 1 is the incremental development triangle for years 1972 - 81 from Taylor and Ashe (1983) [20] that has been used by Mack, Clark, and many other authors. The first column is estimated ultimate counts.

Often dividing the losses by exposure information like counts produces a more stable triangle, but preliminary analysis suggests that in this case it does not. The source of the data has not been identified, but it is consistent with excess losses with an increasing retention, which with inflation can make the losses more stable than average claim size. Exposure information is not useful in every case, and will not be used here, but is included for reference.

Table 1 – Taylor Ashe triangle with ultimate claim counts (#)

#	Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lg 9
40	357,848	766,940	610,542	482,940	527,326	574,398	146,342	139,950	227,229	67,948
37	352,118	884,021	933,894	1,183,289	445,745	320,996	527,804	266,172	425,046	
35	290,507	1,001,799	926,219	1,016,654	750,816	146,923	495,992	280,405		
41	310,608	1,108,250	776,189	1,562,400	272,482	352,053	206,286			
30	443,160	693,190	991,983	769,488	504,851	470,639				
33	396,132	937,085	847,498	805,037	705,960					
32	440,832	847,631	1,131,398	1,063,269						
43	359,480	1,061,648	1,443,370							
17	376,686	986,608								
22	344,014									

Mack's methods lead to a reserve estimate of 18,681,000 to the end of the triangle and a prediction standard error of 2,447,000. The MFE – PCS model calculated as outlined above gives the same reserve estimate but a prediction standard error of 2,827,000. The difference is due to the combination of a much better fit from the MFE – PCS model, with an almost 50% reduction in process standard deviation, and a parameter standard deviation greater by almost 70% due to the greater number of parameters.

Figure 1

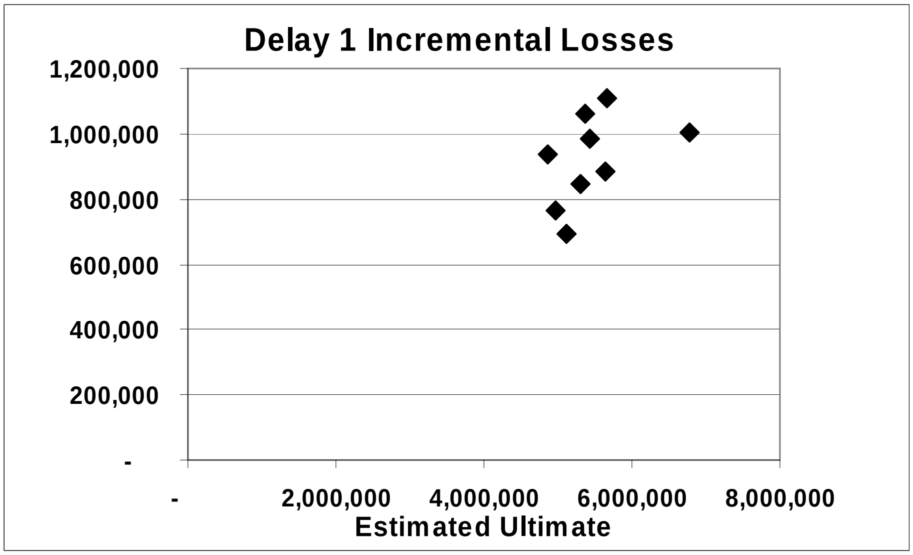
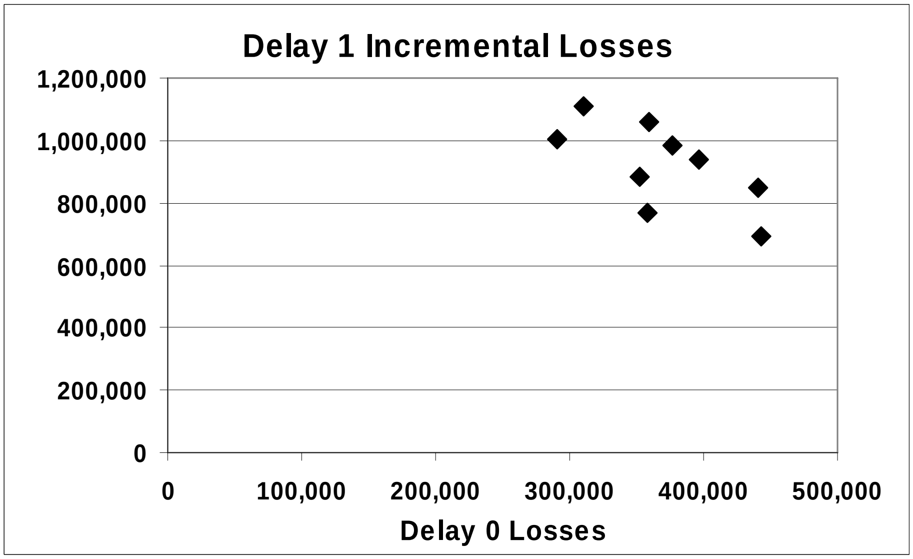


Figure 2



To illustrate the difference in fits, Figures 1 and 2 graph the delay 1 incremental losses as a function of the delay 0 losses and as a function of the estimated ultimate losses. A factor times ultimate losses looks like a much better explanation of the incremental losses than does a factor times losses at 0.

There are of course assumptions that need to be verified for either model. For instance in MFE all of the observations are assumed independent, while for Mack at least the rows should be independent. Both assumptions are violated when there are strong calendar-year (diagonal) effects, as in this triangle.

Table 2 shows the residuals by diagonal for the MFE – PCS model. Diagonals 2, 3, 4, 6, and 7 are all suspicious, with 7 being the most problematic. A related issue is correlation of residuals among columns. This can be a result of diagonal effects that have not been modeled. Table 3 shows the correlation of the MFE – PCS residuals from one column to the next for the first four columns. All the correlations are negative and two are quite significant.

Table 2

Diagonal	Average Residual	Fraction Positive
0	87,787	1 of 1
1	35,158	1 of 2
2	(76,176)	0 of 3
3	(74,853)	1 of 4
4	100,127	4 of 5
5	(26,379)	2 of 6
6	103,695	5 of 7
7	(115,163)	1 of 8
8	(17,945)	3 of 9
9	38,442	6 of 10

Table 3

Columns	0-1	1-2	2-3	3-4
Correlation	-21.5%	-89.5%	-48.9%	-85.4%
Significance	0.289	0.001	0.133	0.015

4.1 Incorporating Diagonal Effects

Factors can be put into the model for diagonal effects. Denoting the factor for the j^{th} diagonal as h_j, then the cell expected loss is not given by $b\lambda_{w,d} = U_w g_d$, but by $b\lambda_{w,d} = U_w g_d h_{w+d}$. Still assuming that the λ's are Poisson means, the likelihood function is:

$$l = C + \sum \left(\frac{q_{w,d}}{b} \ln \frac{U_w g_d h_{w+d}}{b} - \frac{U_w g_d h_{w+d}}{b} \right)$$

The unconstrained parameter estimates still have an iterative formulation:

$$g_d = \sum_{w=0}^{n-d} q_{w,d} \Big/ \sum_{w=0}^{n-d} U_w h_{w+d} \,, \quad U_w = \sum_{d=0}^{n-w} q_{w,d} \Big/ \sum_{d=0}^{n-w} g_d h_{w+d} \,, \quad \text{and} \quad h_j = \sum_{w+d=j} q_{w,d} \Big/ \sum_{w+d=j} U_w g_d \,.$$

These converge a bit slowly, but 50 or so iterations often suffice. This can be done in a spreadsheet without programming any functions. Again the *g*'s can be made to sum to 1, and so represent a payout pattern, but with the calendar-year factors the *U*'s are then no longer the ultimate losses.

Using this method, two models with calendar-year effects were fit to the Taylor-Ashe data, adding diagonal parameters for the 7th diagonal, and for the 6th and 7th. The other *h*'s in the iteration were kept at 1. To compare the loglikelihoods, *b* was fixed at 37,183.5. This is the estimated value for another MFE – PCS model, discussed below. With this value, the maximum loglikelihood values for zero, one, and two diagonal factors are:

-149.11, -145.92, -145.03.

With 55 observations, the HQIC penalty for an additional parameter is 1.388. According to this, the model with both diagonals is better than the one with no diagonal parameters, but not as good as the one with only the 7th diagonal. The AIC$_c$ strongly penalizes having so many parameters (up to 21) with only 55 observations, and charges the first diagonal parameter 2.5 and the second 2.65. This makes no diagonal parameters better than two but worse than one. The factors for the 6th and (in both models) 7th diagonal are 1.136 and 0.809.

Including these parameters corrects for potential random errors in the row and column parameter estimates from ignoring diagonal effects. The chain ladder and original PCS reserves were 18,681,000. Adding one diagonal parameter increases this to 19,468,000 and having them both increases it further to 19,754,000. Thus it appears that the original reserve estimates were low.

4.2 Reducing the Number of Parameters

Regression modelers use various techniques to get rid of insignificant parameters without hurting the fit. Parameters that are not significantly different from 0 or 1 are sometimes defaulted to those values. Also parameters that are not significantly different from each other can be set equal. Also, when changes are systematic, a parameter for a year or delay could be set to the average of the parameters before and after it. More generally, several parameters in a row could be expressed as a linear or geometric trend, which can reduce the number of parameters further. Reducing the parameters in these ways can eliminate distinctions that are not supported by the data. This can be done for row, column, or diagonal parameters. For instance, up to random effects, the upward and downward diagonal deviations could be indistinguishable. This could hold for many of the late small lag factors and some accident-year mean levels as well.

Several of these methods were tried for the Taylor-Ashe data. A fairly extreme example gets the MFE model down to just six parameters without significantly degrading the fit. In this model, accident year 0 is low so gets its own parameter U_0. Accident year 7 is high and also gets its own parameter U_7. All the other years get the same parameter U_a, except year 6 which is a transition and gets the average of U_a and U_7. Thus there are three accident year parameters.

The fraction paid is divided into high and low payment years with parameters g_a and g_b. Delay 0 is a low year as payments start slowly. Delays 1, 2, and 3 are the higher payment lags and all get g_b. Delays 5, 6, 7, and 8 are low getting g_a, but delay 4 is a transition and gets the average of g_a and g_b. Finally delay 9 gets the rest, i.e., $1 – 5.5g_a – 3.5g_b$. This leaves only two delay parameters. Three of the diagonals were specified as high or low diagonals, getting factors $1+c$ or $1–c$. The 7th diagonal is low and the 4th and 6th are high. Thus only one diagonal parameter c is used.

This model uses the techniques of setting parameters equal if they are not significantly different and putting other parameters on trend lines – in this case averages – of other parameters. The loglikelihood for this six-parameter model is -146.66. This is not as good as the twenty-parameter model above, with a loglikelihood of -145.92, but it gets an HQIC penalty that is less by 19.4 and an AIC_c penalty that is lower by 25.5. These clearly overwhelm the difference in loglikelihood of 0.74. The resulting parameters and their standard errors are:

Parameter	U_0	U_7	U_a	g_a	g_b	c
Estimate	3,810,000	7,113,775	5,151,180	0.0678751	0.1739580	0.1985333
StdError	372,849	698,091	220,508	0.0034311	0.0056414	0.0568957

Table 4

Estimating the parameters was done by an add-in spreadsheet optimizer on the loglikelihood. Most of the build-in spreadsheet optimizers have trouble estimating this many parameters. The parameter variances came from the information matrix. The 2nd derivatives of the unconstrained loglikelihood wrt U_w and g_d do not change with the inclusion of diagonal parameters. The other 2nd partials are:

$$\frac{\partial^2 l}{\partial h_j^2} = -\sum_{w+d=j} \frac{q_{w,d}}{bh_j^2}, \quad \frac{\partial^2 l}{\partial U_w \partial g_d} = -\frac{h_{w+d}}{b}, \quad \frac{\partial^2 l}{\partial U_w \partial h_j} = -\frac{g_{j-w}}{b}, \quad \frac{\partial^2 l}{\partial g_d \partial h_j} = -\frac{U_{j-d}}{b}.$$

The derivatives of the loglikelihood wrt U_a, g_a, g_b, and c, use the chain rule on the sum of the derivatives of the loglikelihood wrt the parameters above. However U_a and U_7 are now not independent, as they go into estimation of some of the same cells, and similarly for g_a and g_b. The correlations of adjacent residuals improve a good deal with the diagonal parameters, as shown in Table 5. This is still somewhat problematic, however, as the correlations are all negative and some are weakly significant. These correlations are still there after accounting for diagonal effects, so might indicate some degree of actual serial correlation in accident year payments. Perhaps ARIMA models could have a role in this modeling. The logic is that high development in one year would be followed by low development the next, which is possible. But forcing the column factors to sum to one would induce some degree of negative correlation across columns, so the extent of this would have to be established before any firm conclusions about auto-correlated development could be made.

Table 5

Columns	0-1	1-2	2-3	3-4
Correlation	-0.9%	-58.1%	-50.7%	-74.1%
Significance	0.491	0.066	0.123	0.046

Table 6

Model	Original 19 Parameter	6 Parameter
Parameter Variance	7,009,527,908,811	1,103,569,529,544
Process Variance	982,638,439,386	718,924,545,072
Total Variance	7,992,166,348,198	1,822,494,074,616
Parameter Std Dev	2,647,551	1,050,509
Process Std Dev	991,281	847,894
Standard Deviation	2,827,042	1,349,998

The reserve estimate from this model is 19,334,000, which is quite close to that of the twenty-parameter model. The prediction standard error (with *b* = 37,183.5) is down to 1,350,000, compared to 2,827,000 for the full MFE – PCS and 2,447,000 for the chain ladder. The better fit from including calendar-year effects and the reduced number of parameters has decreased the standard error appreciably. The breakdown of the variance into parameter and process is in Table 6. There is a decrease in the process standard deviation of 15%, probably coming from recognizing the diagonal effects, and a 60% reduction in the parameter standard deviation in going from 19 to 6 parameters, for a total decrease in the prediction standard error of over 50%.

4.3 Testing the Variance Assumption

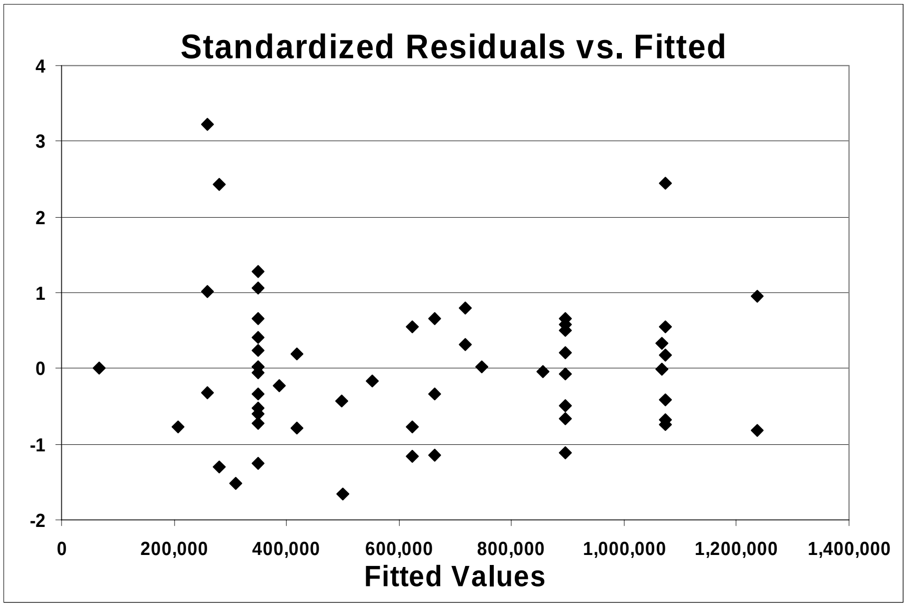
In the PCS model the variance of each cell is *b* times its mean. If the variance is proportional to a higher power of the mean, then the PCS standardized residuals (residuals divided by modeled standard deviation) would tend to be larger in absolute value for the cells with the larger means. A plot of standardized residuals vs. fitted values would be a way to show this. These are graphed in

Figure 3 for the six-parameter model. This effect does not appear. However, the positive residuals have more extreme values than do the negative residuals, which could be indicative of a more highly skewed model.

There is a possible analogue to the PP-plot as well. A PP-plot for a probability distribution fitted to data compares the empirical cumulative probability to the fitted cumulative probability at each sample point. Here there are 55 Poisson distributions, each of which has a sample of 1, namely $q_{w,d}/b$. The typical empirical probability for the p^{th} observation out of a sample of N is $p/(N+1)$, so this would be ½ for each of our 55 observations. But you could start with the fitted probability at each point, rank these 55 fitted values from 1 to N and then assign the empirical probability = rank/(N+1) to each. This gives something like a PP-plot, and is shown in Figure 4 for the six-parameter model.

The fit is not too bad, but is better below the median than above. Above there are more observations below most of the probability levels than the Poissons would predict, as shown by the empirical probabilities being higher than the Poisson probabilities. That is a bit surprising, in that usually you would expect observed data to have more large observations than the Poisson. Probably overall this graph is supportive of the distributional assumption, but Figures 3 and 4 both weakly suggest a lighter tail than the Poisson.

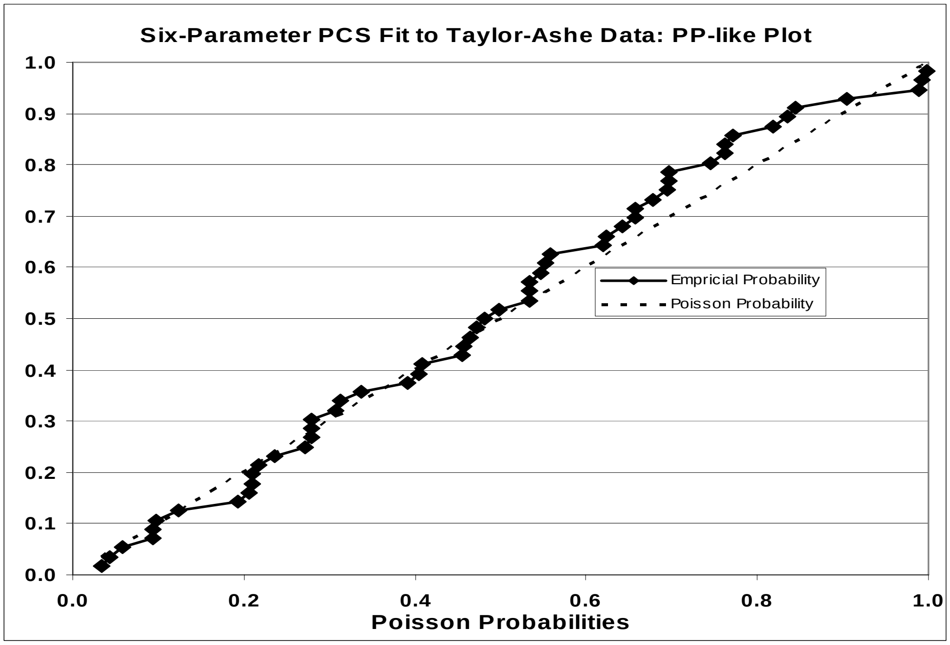
Figure 3



4.4 Example 1 Conclusions

The MFE – PCS model with one parameter for each row and column matches the chain-ladder reserve calculation but can have very different fitted values in the triangle. It has more parameters so a better fit would be expected, but the variance calculation reflects the parameter uncertainty, so the chain ladder can easily give a lower variance. The fit and assumptions of both models can be strained by calendar-year effects, but these can be modeled with their own parameters in either model. As in this example, it should usually be possible to reduce the number of parameters in the models through the use of trends, combination of similar parameters, etc. The MFE models also allow for eliminating some accident year parameters, which can be reduced even to a single parameter in the Cape Cod case. In the example here, three levels sufficed for 10 years. Other possible models, including MFE with different distributional assumptions, have not been considered and may give better fits to this data. Negative correlations between adjacent columns might also be real, and these could be modeled by time-series techniques. Taylor (2000) [19] and de Jong (2006) [5] explore time-series modeling for development triangles. In summary, getting a better fit by recognizing calendar-year effects and then reducing the number of parameters in the model can decrease the both the process and parameter variances of the reserve estimate. The MFE paradigm is appealing when incremental losses are not well explained as a factor times previous cumulative.

Figure 4



5. EXAMPLE 2

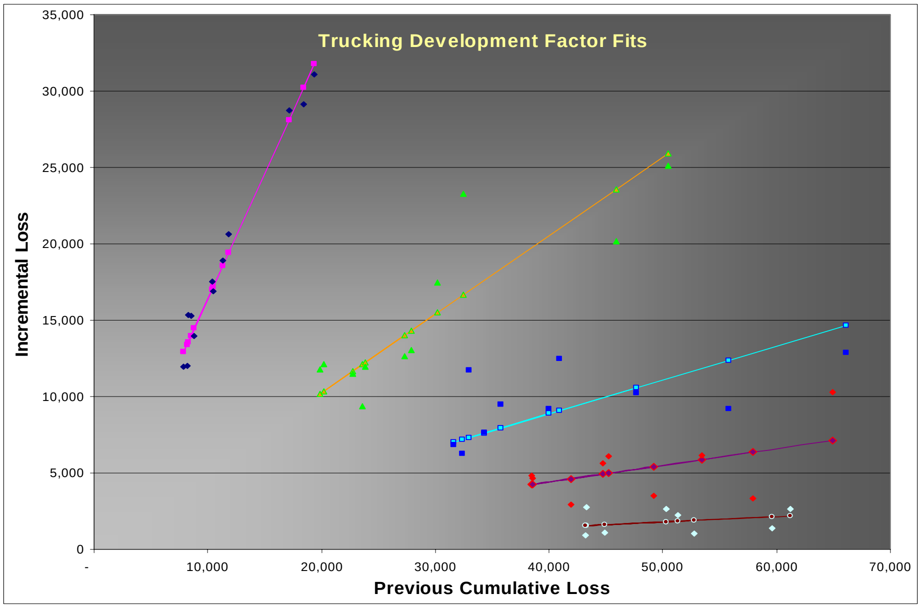
For those who like development factors, it is possible to do many of the steps of Example 1 in a factor setting. Calendar-year effects can be modeled, and parameter-reduction techniques can be applied. These can lead to better-fitting models with fewer parameters. Such ideas are illustrated in this example, using a triangle of long-haul trucking liability losses.

Table 7 Long-Haul Trucking Development Triangle and Murphy LSM Factors

Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Lag 11
11,305	30,210	47,683	57,904	61,235	63,907	64,599	65,744	66,488	66,599	66,640	66,652
8,828	22,781	34,286	41,954	44,897	45,981	46,670	46,849	47,864	48,090	48,105	48,721
8,271	23,595	32,968	44,684	50,318	52,940	53,791	54,172	54,188	54,216	54,775	
7,888	19,830	31,629	38,444	43,287	46,032	47,411	47,677	48,486	48,498		
8,529	23,835	35,778	45,238	51,336	53,574	54,067	54,203	54,214			
10,459	27,331	39,999	49,198	52,723	53,750	54,674	55,864				
8,178	20,205	32,354	38,592	43,223	44,142	44,577					
10,364	27,878	40,943	53,394	59,559	60,940						
11,855	32,505	55,758	64,933	75,244							
17,133	45,893	66,077	78,951								
19,373	50,464	75,584									
18,433	47,564										
20,640											
Factors	2.640	1.5132	1.2220	1.1102	1.0359	1.0149	1.0108	1.0093	1.0017	1.0035	1.0045

The data is for 1984 to 1995. Recall that the LSM model calculates each factor by a least-squares regression. For this data the factors provide a believable representation of the development process for the first five lags. The actual and fitted incremental losses at these lags are graphed as a function of the previous cumulative losses in Figure 5. Some of the deviations from the lines are fairly substantial, but the factors do seem to capture the general pattern of development. This is not to say that factors give the best model for this data – in fact no other models were tested. The goal is just to show how to apply the methods above to factor models.

Figure 5



5.1 Multiple Regression Format

To add in diagonal elements, these regressions can be converted to a single multiple regression, and dummy variables added in for the diagonals. Table 8 shows part of the design matrix for such a regression. The incremental losses at lags 1 to 5 (partial) are strung out into the first column, then the subsequent columns are the cumulative losses at lags 0 to 4 that are to predict the next incremental losses.

The last column is a dummy variable that picks out the incremental losses that are on the 4th diagonal, which are highlighted. Before looking at diagonals, a standard normal-residual regression routine provided the output in Table 9 on the 11 development factors estimated by a single no-constant multiple regression.

Table 8

Incremental	L0	L1	L2	L3	L4	D4
18,904	11,305	-	-	-	-	-
13,953	8,828	-	-	-	-	-
15,324	8,271	-	-	-	-	-
11,942	7,888	-	-	-	-	1
15,306	8,529	-	-	-	-	-
16,873	10,459	-	-	-	-	-
12,027	8,178	-	-	-	-	-
17,515	10,364	-	-	-	-	-
20,650	11,855	-	-	-	-	-
28,759	17,133	-	-	-	-	-
31,091	19,373	-	-	-	-	-
29,131	18,433	-	-	-	-	-
17,474	-	30,210	-	-	-	-
11,505	-	22,781	-	-	-	-
9,373	-	23,595	-	-	-	1
11,799	-	19,830	-	-	-	-
11,943	-	23,835	-	-	-	-
12,668	-	27,331	-	-	-	-
12,150	-	20,205	-	-	-	-
13,065	-	27,878	-	-	-	-
23,253	-	32,505	-	-	-	-
20,184	-	45,893	-	-	-	-
25,120	-	50,464	-	-	-	-
10,221	-	-	47,683	-	-	-
7,668	-	-	34,286	-	-	1
11,716	-	-	32,968	-	-	-
6,815	-	-	31,629	-	-	-
9,460	-	-	35,778	-	-	-
9,199	-	-	39,999	-	-	-
6,238	-	-	32,354	-	-	-
12,451	-	-	40,943	-	-	-
9,175	-	-	55,758	-	-	-
12,874	-	-	66,077	-	-	-
3,331	-	-	-	57,904	-	1
2,943	-	-	-	41,954	-	-
5,634	-	-	-	44,684	-	-
4,843	-	-	-	38,444	-	-
6,097	-	-	-	45,238	-	-
3,524	-	-	-	49,198	-	-
4,631	-	-	-	38,592	-	-
6,165	-	-	-	53,394	-	-
10,312	-	-	-	64,933	-	-
2,671	-	-	-	-	61,235	-
1,084	-	-	-	-	44,897	-
2,623	-	-	-	-	50,318	-
2,745	-	-	-	-	43,287	-
2,238	-	-	-	-	51,336	-
1,027	-	-	-	-	52,723	-

The first five factors are all highly significant, but none of the others are. Yet they are all positive, so some recognition of development beyond 5th is clearly needed. Since the differences among the factors is small compared to their standard deviations, one possibility is combining some, like 6th through 8th and 9th through 11th, or trending them, or replacing them by a constant or constants. For this example a constant term was included in the regression and factors f6 to f11 dropped. That reduced the number of parameters by five while still recognizing late development.

Table 9

Parameter	**Est value**	**St dev**	***t* student**	**Prob(>|*t*|)**
f1	1.64042	0.03751	43.7337	6.2E-50
f2	0.5132	0.01564	32.8085	3.6E-42
f3	0.22199	0.0118	18.8143	5.3E-28
f4	0.11017	0.01095	10.061	7E-15
f5	0.0359	0.01111	3.23205	0.00193
f6	0.01486	0.01173	1.26635	0.20991
f7	0.01079	0.0122	0.88452	0.37968
f8	0.00931	0.01329	0.69999	0.48643
f9	0.0017	0.0147	0.1155	0.90841
f10	0.00348	0.01636	0.21279	0.83216
f11	0.00451	0.01959	0.23034	0.81855

5.2 Modeling Diagonal Effects

Table 10 shows the average residual from the all-factors model and the percent positive for each diagonal. The *j*th diagonal has *j*+1 fitted values in it except for the 11th, which has 11 values. The 3rd, 4th, 7th, 9th and 10th diagonals are suspicious. Adding them all to the regression gives the results in Table 11. The same factors are significant but with slightly different values. The 3rd diagonal is significant at the 5% level, and the 4th and 9th at a bit weaker levels. Some combination of the diagonal adjustments might be more significant.

Table 10

Diagonal	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Avg Residual	359	721	402	(1,681)	1,226	(142)	93	599	(157)	902	(734)	(63)
% Positive	100	50	33	25	80	17	71	88	44	50	27	36

This model gives separate parameters to all the development factors and the suspicious diagonals. Trying parameter reduction, a fairly minimalist model is to keep the first five factors, add a constant to the regression for the later development, keep the 3rd diagonal, and have a common factor for the 4th, 7th, 9th, and 10th diagonals, but with the 10th subtracted. The constant for all development after 5th works well enough because this development is highly random and does not seem to depend on the level of previous cumulative. The late development could be due to lawsuits coming to a conclusion late in the process, with the timing being highly random. There is still a possibility of improving the model by differentiating stages of the late development, however, which is not explored here. The regression results are in Table 12. All the parameters are significant enough to keep in the model. Table 11

Parameter	**Est value**	**St dev**	***t* student**	**Prob(>|*t*|)**
f1	1.6345	0.0364	44.947	6.58E-48
f2	0.5127	0.0151	33.988	6.72E-41
f3	0.2208	0.0115	19.274	2.18E-27
f4	0.1103	0.0108	10.236	8.76E-15
f5	0.0293	0.0108	2.7165	0.0086
f6	0.0117	0.0112	1.0430	0.3011
f7	0.0080	0.0117	0.6902	0.4927
f8	0.0043	0.0130	0.3344	0.7392
f9	0.0005	0.0140	0.0359	0.9715
f10	-0.0004	0.0158	-0.0270	0.9786
f11	0.0110	0.0187	0.5855	0.5604
D3	-1657.7	779.5	-2.1266	0.0376
D4	1325.9	700.0	1.8941	0.0630
D9	1041.5	535.1	1.9463	0.0563
D10	-655.2	528.3	-1.2403	0.2197
D7	726.5	573.2	1.2675	0.2099

Table 12

Parameter	**Est value**	**St dev**	***t* student**	**Prob(>|*t*|)**
Constant	527.81	255.77	2.0636	0.0428
f1	1.601	0.03767	42.4984	3.23E-51
f2	0.499	0.01558	32.0293	3.77E-43
f3	0.211	0.01167	18.0798	7.01E-28
f4	0.102	0.01083	9.4008	5.59E-14
f5	0.021	0.01076	1.9818	0.0515
D3	-1832	724.59	-2.5284	0.0138
D4+D7+D9-D10	801.61	245.88	3.2601	0.0017

5.3 Comparing Fits

The loglikelihood at the maximum for a regression with normal residuals on *n* observations can be expressed as a function of the SSE:

$$\log L = (n/2)\log[2\pi e\mathrm{SSE}/n]$$

Using this, with *p* parameters, $AIC_c/2 = (n/2)\log[2\pi eSSE/n] + np/[n - p - 1]$. The se of the regression is also a function of goodness of fit and number of parameters, so it is a related comparative measure. The models discussed above are compared on this basis in Table 13.

The minimalist model is not a special case of the 16 parameter model because it has a constant term. This appears to provide a somewhat better explanation of the development than does the combination of factors even before adjusting for number of parameters.

Model	*p*	SSE	se	$AIC_c/2$
All Development Factors	11	171,040,478	1609.821	684.913
All Factors and Five Diagonals	16	133,609,815	1479.975	682.907
Minimalist	8	132,867,569	1387.666	671.218

Table 13

5.4 Analysis of Residuals

Figure 6 is a QQ plot of the residuals of the minimalist model vs. the normal distribution regression assumption. The QQ plot graphs the residuals, whereas the PP plot graphs the probabilities of the residuals. In the right tail the last few residuals are much higher than the normal percentiles, while most of the positive residuals are lower than the normal would suggest. This is not very supportive of the normal assumption.

Figure 7 plots the residuals by delay. Regression assumes that all the residuals have the same distribution, but delays 2 through 4 or 5 appear to have a higher variance. Failure to have the same residual distribution is a regression problem known as heteroscedasticity. It does not necessarily affect the estimates of the coefficients, but it does require a different variance calculation.

There is a formal test for heteroscedasticity known as White's test, which when applied to this model is ambiguous about the presence of heteroscedasticity. However White's test is not regarded as definitive. In this model heteroscedasticity would be suspected and even preferred in the sense that the smaller observed increments at later stages of development should have lower error variances than the larger increments earlier on. A preference for equalizing relative errors actually would suggest a lognormal model, which is not explored here. However there are correction methods available for adjusting the variance for heteroscedasticity in the model, and these come at little cost, because they do not change the estimate much when the variances are in fact constant. Thus such an adjustment would be appropriate for calculating the variance for this model.

Figure 6

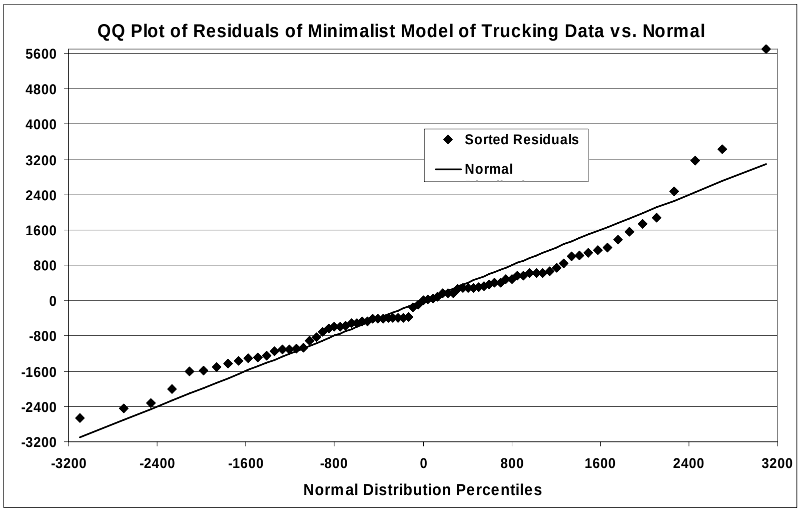
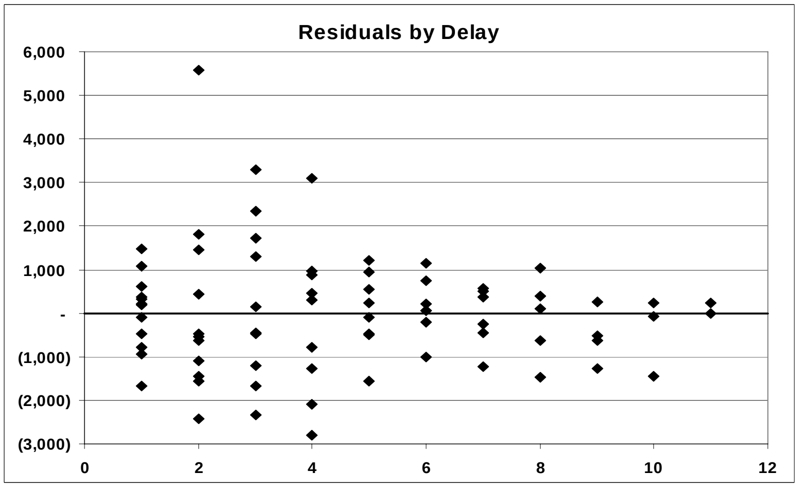


Figure 7



5.5 Estimating the Variance

Again the parameter variance can be estimated by the delta method, and the process variance by using the standard error. The covariance matrix of the parameters needed for the delta method is a standard output of multiple regression software. However when heteroscedasticity is suspected, an adjusted covariance matrix is appropriate.

This discussion is based on Long and Ervin (2000) [11]. They recommend a heteroscedasticity consistent covariance matrix they call HC3 whenever there is any chance of heteroscedasticity. Explaining this requires getting into the calculations underlying multiple regression. The starting point is the matrix **X** of independent variables, which is an $n \times p$ matrix with a row for each observation and a column for each variable. The $p \times p$ matrix $\mathbf{Z} = (\mathbf{X'X})^{-1}$ is widely used in regression.

The $p \times p$ covariance matrix for the parameter estimates can be expressed in terms of $\mathbf{Z}$ and the $n \times n$ covariance matrix $\mathbf{\Phi}$ of the observations of the dependent variable as $\mathbf{ZX'\Phi XZ}$. When the error variances of the observations are constant and independent, i.e., $\mathbf{\Phi} = \sigma^2\mathbf{I}$, the parameter covariance matrix simplifies to $\sigma^2\mathbf{Z}$. This is the usual parameter covariance matrix put out by regression programs. A convenient calculation of $\mathbf{Z}$ is thus to simply divide this matrix by σ^2.

To correct for possible heteroscedasticity, let e_i be the residual for the i^{th} observation and define s_i = $\mathbf{x_i Z x_i}'$, where $\mathbf{x_i}$ is the row vector of the i^{th} observations of the independent variables. Then $e_i/(1 – s_i)$ is an adjusted residual. The adjusted parameter covariance matrix uses the diagonal matrix of squared adjusted residuals as the estimate of $\mathbf{\Phi}$. Thus:

$$\mathbf{HC3 = ZX'diag[e_i^2/(1 - s_i)^2]XZ}$$

is the adjusted covariance matrix of the parameters.

Since the heteroscedasticity is expected to come from differences among column variances, it would be reasonable to extend this approach to estimating adjusted column variances as well. The average of the squared adjusted residuals down a column of the triangle could be used as the estimate of the variance of the residuals for that column.

For the minimalist model this methodology was applied. The original and revised t-statistics for each parameter are in Table 14. The adjusted standard deviations σ_j by column are in Table 15. Using these standard deviations, the actual residuals standardized are graphed against standard normal percentiles in Figure 8. While light in the left tail, this adjustment makes the residuals look more normal.

Table 14

h	Constant	f1	f2	f3	f4	f5	D3	D4+D7+
								D9-D10
Original	2.064	42.498	32.029	18.080	9.401	1.982	(2.528)	3.260
Adjusted	3.501	72.264	17.985	12.838	6.035	3.206	(1.926)	2.574

Table 15

927	2,46	2,13	2,01	831	713	800	919	697	808	228

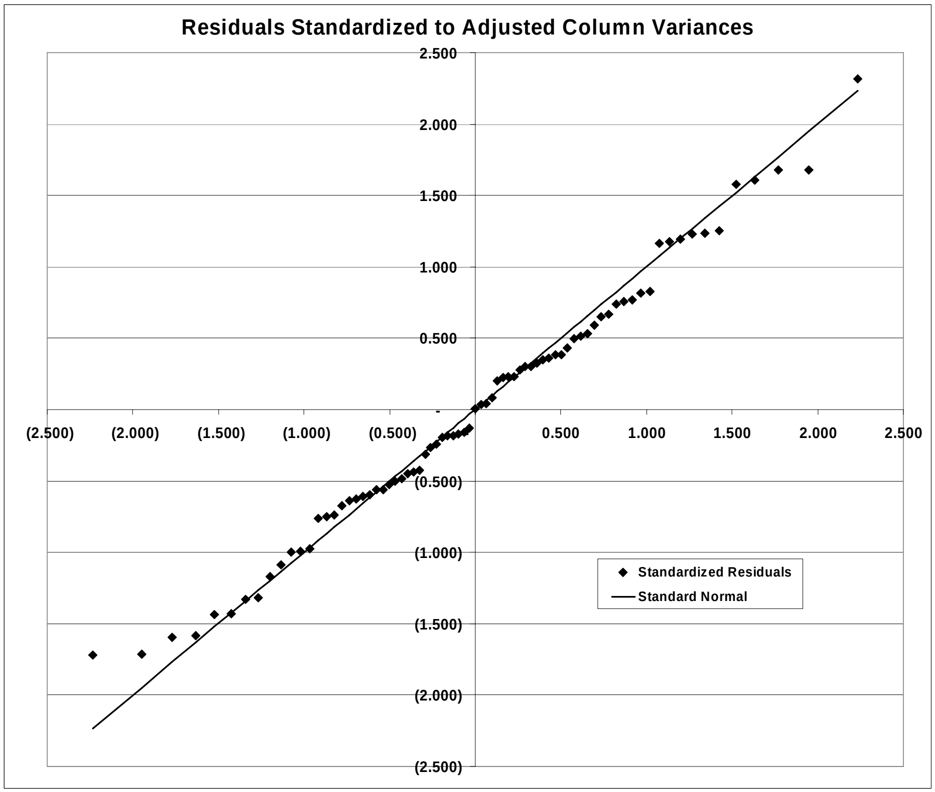


Figure 8

To calculate the variance of the projection, the recursive scheme of Murphy can be applied. First denote by S_j the cumulative losses up through lag j for all accident years in the triangle not already observed through j. The recursion begins:

$$ES_1 = c_{n,0}(1+f_1)+b$$

$$ES_j = (\mathrm{c}_{n-j+1,j-1}+ES_{j-1})(1+f_j)+jb, \text{ where } f_j = 0 \text{ for } j > 5.$$

For the process variance given that the parameters are known:

$$\mathrm{Var}(S_1) = \sigma_1^2$$

$$\mathrm{Var}(S_j) = \mathrm{EVar}(S_j \,|\, S_{j-1}) + \mathrm{VarE}(S_j \,|\, S_{j-1}) = j\sigma_j^2 + \mathrm{Var}[(1+f_j)S_{j-1}] = j\sigma_j^2 + (1+f_j)^2\mathrm{Var}(S_{j-1})$$

For the delta method the derivatives of S_n can be calculated by recursion as well:

$$\partial ES_1/\partial b = 1;\ \partial ES_j/\partial b = j + (1+f_j)\partial ES_{j-1}/\partial b$$

$$\partial ES_j/\partial f_j = c_{n-j+1,j-1}+ES_{j-1}$$

$$\partial ES_j/\partial f_i = 0 \text{ if } i > j \text{ and } \partial \mathrm{ES}_j/\partial f_i = (1+f_j)\partial ES_{j-1}/\partial f_i\ 0 \text{ if } i < j.$$

The results are in Table 16.

Table 16

	Minimal	Original Murphy LSM
Reserve estimate	213,553	221,800
Process variance	89,501,787	92,565,591
Parameter variance	86,856,827	138,084,020
Variance	176,358,614	230,649,611
Standard deviation	13,280	15,187

The reserves corrected for calendar-year effects are lower in this case, the process variance is lower due to a bit better fit, and the parameter variance is lower because of 8 parameters vs. 11.

5.6 Variants of the Chain Ladder

Murphy considered three calculations of chain ladder factors, namely regression, ratio of sums, and average of ratios. As mentioned above, the ratio of sums is a regression for each column where the incremental losses for the column and the cumulative losses for the previous column are both divided by the square root of the previous cumulative, and the average of ratios is the regression divided by the previous cumulative itself.

These adjustments can be done for multiple regression as well. There is only one previous cumulative in each row of the design matrix, so the entire row, including the dummy variables and the 1 for the constant term if included, can be divided by the previous cumulative or its square root. Thus calendar-year effects can be modeled with any variant of the chain ladder. This adjustment is not likely to remove heteroscedasticity from the regressions, however, as the smallest incremental losses are still going to be factors times the largest previous cumulatives.

Further variants of the chain ladder using generalized linear models are also possible. Generalized linear models replace the normal distribution assumption of the residuals with other distributions. The PCS could be used, for example, which would have variance proportional to mean for the entire multiple regression. This could in itself eliminate the problem of heteroscedasticity.

6. EXAMPLE 3

This example looks at using exposure data, distributions instead of lag factors, and leaving out data. Factors are sometimes based on the last five diagonals, or even last five diagonals excluding the high and low observations in each column. This example illustrates that it can sometimes be appropriate to leave out some data. This is when it is clear that there has been a change in the development process. Otherwise leaving out data will increase the variance of the estimate.

Excluding high and low observations is particularly problematic in that if factors are from a skewed distribution this will bias the estimated factors downward.

The triangle in Table 17 is cumulative claim counts with exposures for 1978 - 1995 from Taylor (2000) [19]. Exposures are growing over time. The usual assumption is that this consists of more units from the same population. That is not necessarily the case, however, and may not be so here. The development factors are grouped by selected accident-year ranges in Table 18. The 0 to 1 factors for the four groups are 1.52, 1.37, 1.47, and 1.32, and the factors are fairly consistent within each group. Most of the development occurs from 0 to 1, so it is critical to get a good estimate for this factor.

Table 17 Cumulative Claim Count Triangle with Exposures

Exposure	**Lag 0**	**Lag 1**	**Lag 2**	**Lag 3**	**Lag 4**	**Lag 5**	**Lag 6**	**Lag 7**	**Lag 8**	**Lag 9**	**Lg 10**
71,543	368	559	587	595	601	606	609	610	610	610	611
75,681	393	544	569	575	579	584	588	589	591	592	592
98,960	517	702	731	748	759	769	777	778	778	778	779
102,974	578	832	881	903	920	926	929	929	930	930	930
106,810	622	828	867	883	886	893	893	894	894	894	894
110,779	660	903	931	943	955	959	963	964	964	964	964
114,307	666	900	953	963	971	975	981	982	982	982	982
117,306	573	839	901	913	918	925	931	936	937	937	938
123,304	582	863	895	922	934	947	953	955	956	956	
125,533	545	765	808	826	838	847	852	854	854		
131,265	509	775	824	846	861	865	873	873			
139,661	589	799	828	845	857	861	870				
152,895	564	760	783	795	804	809					
160,331	607	810	839	848	855						
162,900	674	843	863	875							
170,045	619	809	850								
173,248	660	821									
175,941	660										

One approach to verifying that there actually has been a change in development is to compare the variance of the estimate using the full data and using only the more recent data that appears to be from a different population. In this case the claims through lag 6 (7th column) were developed from all accident years and for the last seven years. Using the Mack formulas, estimating the factors from all the years combined gives an expected future claim count for the last seven years of 481, of which 68% are from the last accident year, and a standard deviation of 62. From just the last seven years alone these estimates are 417 claims with a standard deviation of 42, and 65% are from the last accident year. The estimated standard deviation is much lower with the last seven years alone, which supports the idea that there has been a change in development patterns.

Table 18 – Development Factors for Claim Count Triangle

1.519	1.050	1.014	1.010	1.008	1.005	1.002	1.000	1.000	1.002
1.384	1.046	1.011	1.007	1.009	1.007	1.002	1.003	1.002	1.000
1.358	1.041	1.023	1.015	1.013	1.010	1.001	1.000	1.000	1.001
1.439	1.059	1.025	1.019	1.007	1.003	1.000	1.001	1.000	1.000
1.331	1.047	1.018	1.003	1.008	1.000	1.001	1.000	1.000	1.000
1.368	1.031	1.013	1.013	1.004	1.004	1.001	1.000	1.000	1.000
1.351	1.059	1.010	1.008	1.004	1.006	1.001	1.000	1.000	1.000
1.464	1.074	1.013	1.005	1.008	1.006	1.005	1.001	1.000	1.001
1.483	1.037	1.030	1.013	1.014	1.006	1.002	1.001	1.000	
1.404	1.056	1.022	1.015	1.011	1.006	1.002	1.000		
1.523	1.063	1.027	1.018	1.005	1.009	1.000			
1.357	1.036	1.021	1.014	1.005	1.010				
1.348	1.030	1.015	1.011	1.006					
1.334	1.036	1.011	1.008						
1.251	1.024	1.014							
1.307	1.051								
1.244									

Figure 9 graphs the 0 to 1 factors, with the groupings indicated. The last group is subdivided into two sub-groups of three years each. It appears that there have been different eras of internally consistent development factors, and that the last six factors tend to be lower than the others. This supports ignoring most of the older data, especially for the 0 to 1 factor. It raises the question of a possible continuing downward trend, however.

The exposure data is helpful in resolving the question of homogeneity of the last seven years. Table 19 shows the claims per 10,000 exposures for the 0 and 1 lags. The grouping of years is a bit different here. For cumulative claims, the last six years appear homogeneous and different from the years before them. This supports the idea that either the new exposures are from a different population or there has been a change in risk conditions. The claims through lag 1 have gone down from about 80 per 10,000 exposures to less than 50.

The last six years show what actuaries would like to see from using exposures: all the years seem to be about the same level after dividing by exposures. This allows for application of an additive model, where each column has its own expected increment. There may still be a downward trend within these years for incremental claims at lag 1, but this will be ignored for now. Figure 9

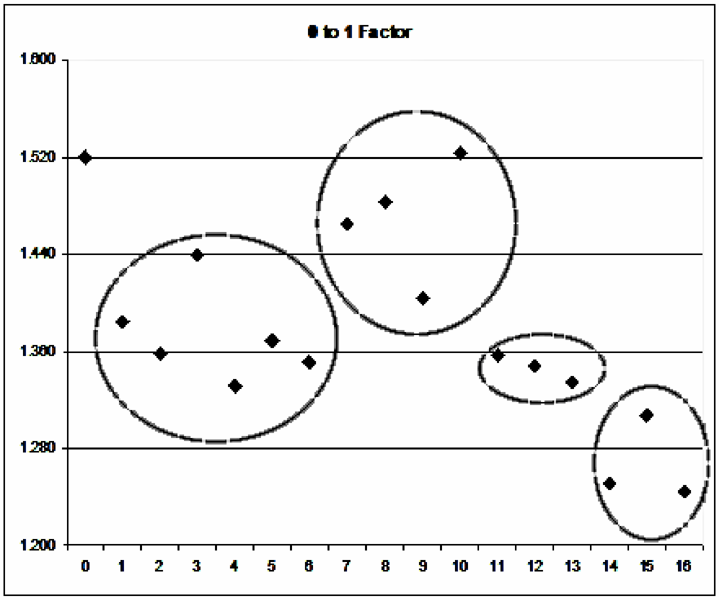


Table 19 – Cumulative and incremental claims per 10,000 exposures

Lag 0	Lag 1 cum	Lag 1 incr
51.4	78.1	26.7
51.9	71.9	20.0
52.2	70.9	18.7
56.1	80.8	24.7
58.2	77.5	19.3
59.6	81.5	21.9
58.3	78.7	20.5
48.8	71.5	22.7
47.2	70.0	22.8
43.4	60.9	17.5
38.8	59.0	20.3
42.2	57.2	15.0
36.9	49.7	12.8
37.9	50.5	12.7
41.4	51.7	10.4
36.4	47.6	11.2
38.1	47.4	9.3
37.5		

Additive development of claims per exposure for the last six years through lag five gives an outstanding reserve of 357 claims. These years can be developed through the end of the triangle using data from earlier accident years. Comparing claims per exposure at lags 0 to 5 for the first 11 years to the last 6 shows an average ratio around 1.945. Dividing the average claims per exposure by this for the older years at each lag for lags 6 and on gives a projection of the future claims for the last 6 years. This adds 35 claims to the expected emergence. Finally doing an additive development for the 4 incomplete older years adds 6 more claims, for a total estimated outstanding of 398 claims.

This is considerably less than the 500 projected from the whole triangle, and can be considered an improved estimate due to the use of exposures and the changes that have occurred in the data. This shows that ignoring data can give a better and possibly significantly different estimate when there are demonstrable changes in the process. However ignoring data otherwise can degrade the estimate. It may be possible to find ways to use the older data with time-series methods instead of discarding it for the first several lags. The apparent continuing downward trend in the claims per exposure at lag 1 gives incentive for following up on this. Taylor (2000) [19] explores some alternatives with this data.

The last 6-year triangle with exposures provides an opportunity to apply a parametric model suggested by Clark (2003) [4]. Denoting the exposures for year w by P_w and the probability of claims appearing by lag d as G_d, assume that $q_{w,d}$ is Poisson in $P_w r(G_d - G_{d\text{-}1})$, where r is an overall ratio of claims to exposures. Any distribution can be used for G, but here Weibull was assumed, with $G_d = 1 - \exp[-(d/\theta)^\omega]$ for $d = 1, 2, \ldots 5$. Weissner (1978) [24] suggests fitting a truncated version of the Weibull, which is technically correct, but for simplicity that was not done here, although it does not seem to make a lot of difference in this case since claims have almost finished their development by lag 5. By starting at $d = 1$ the Weibull is fit for claim appearance after lag 0.

Clark provides the likelihood function and its first two derivatives. MLE for this triangle gives $r = 0.001525$, $\theta = 0.5637$ and $\omega = 0.4980$. The resulting outstanding through lag 5 is 354 claims, which is similar to the 357 from the additive development. However this model has only 3 parameters, while additive development has 5, so there may be a lower variance.

The sample variance for each column of claims per exposure is the sum of the squares of the deviation from the average divided by $n - 1$. This variance would apply to each projected incremental cell. In addition there is the variance of the estimated mean, which is the column variance divided by n. This all results in a factor of $(n+1)/[n(n-1)]$ applied to the sum of squares of the column deviations. For the last column with only one observation an ad hoc variance is typically imputed, and here that was the ratio of the squares of the means applied to the previous variance. This procedure gives the variance of the ratios to exposure for each column of the triangle. In the projection period these are multiplied by the square of the exposures to give the variance of each projected cell. The sum of these through lag 5 is 1087.5, so the standard deviation is near 33.

For the Poisson-Weibull model the process variance of each cell is its mean, by the Poisson assumption. The parameter variance for each projected cell can be calculated by the delta method, using the derivatives of the loglikelihood from Clark. The covariance matrix of the parameters is in Table 20.

r	ω	θ
6.230E-09	-4.717E-06	3.605E-06
-4.717E-06	6.643E-03	-2.336E-03
3.605E-06	-2.336E-03	5.950E-03

Table 20 – Covariance matrix of Poisson-Weibull fit

The *w*, *d* projected cell has mean $rP_w(G_d - G_{d\text{-}1})$ and its derivatives wrt the 3 parameters are as in Clark. Summing over the projected cells gives the derivatives of the reserve wrt *r*, ω, θ as 231,931.82, 95.74 and 65.36. Multiplying the covariance matrix on the left and right by this as a vector gives the delta method estimate of parameter uncertainty of 292. When added to the mean this gives a total variance of 646, or standard deviation of 25.4. Going from 5 to 3 parameters is a 40% reduction in the number of parameters and not much goodness-of-fit was lost, so the standard deviation of the estimated outstanding decreased.

7. CONCLUSIONS

Two paradigms dominate loss development triangle modeling. The conditional approach models each incremental cell's expected value as a linear function of the previous cumulative losses. The unconditional approach models the cell expected losses as a portion of an unobserved level parameter for the year. The chain ladder and BF methods are the original examples of these two paradigms. The unconditional model often fits better but since it uses more parameters (for the accident-year levels), it can have higher variances and wider runoff ranges.

Alternative unconditional or conditional models can be compared on parameter-penalized maximized loglikelihood, but it is difficult to compare across the two paradigms by this method. Perhaps the variance of the estimate is the best common comparison. How to compare models is not a settled issue, however.

Through three examples, ways of improving the estimate were explored. First it is critical to identify calendar-year effects. If these are significant, ignoring them biases the estimates of the other factors. Including them can improve the fit. After that, improving the model primarily consists of getting rid of insignificant parameters. This is not a matter of simply dropping such parameters. It instead involves finding models with fewer parameters that nonetheless account for the observable features of the data.

Replacing level parameters by trends has considerable potential for reducing the number of parameters without sacrificing the fit of the model. In the examples here only linear trends were used and even then just for short periods, but non-linear trends and longer trend periods can be helpful in many cases. A related approach that helped in Example 3 is to use probability distributions for the lag factors. Exposure data when available may improve the modeling as well. When the data has undergone clearly demonstrable changes in structure, using only part of the data can improve the estimates, but otherwise ignoring data will usually increase the variance of the projection. Time series models that account for the changes in structure may be a useful alternative. These could apply vertically, to account for changes in level, horizontally, if high and low development seem to alternate, or by diagonal for evolving cost trends.

Both the conditional and unconditional models can be framed in the notation of multiple regression and put into generalized linear models for alternative residual distributions. The examples only touched on those possibilities, and many more distributions could be tried. If the normal distribution is used, a heteroscedasticity adjustment is needed. A major issue not explored is using calendar-year trends that are projected into the future instead of constants for the diagonal effects. Changing cost trends can strongly affect the projections, and could be considered a key contributor to model risk, also not addressed.

REFERENCES

[1] Bailey, Robert A. 1963. "Insurance Rates with Minimum Bias." *PCAS* 50:4-11.

[2] Bailey, Robert A. and Leroy J. Simon. 1960. "Two Studies In Automobile Insurance Rate Making." *ASTIN Bulletin* 1:192-217.

[3] Barnett, Glen and Ben Zehnwirth. 2000. "Best Estimates for Reserves," *PCAS* 87:245-321.

[4] Clark, David R. 2003. "LDF Curve-Fitting and Stochastic Reserving: A Maximum Likelihood Approach." *CAS Forum*, Fall:41-92.

[5] de Jong, Piet. 2006. "Forecasting Runoff Triangles." *NAAJ* 10, no. 2:28-38.

[6] Hachemeister, Charles A., and James N. Stanard. 1975. "IBNR Claims Count Estimation With Static Lag Functions." (Lecture, ASTIN Colloquium, Portimão, Portugal, September 30-October 3, 1975.

[7] Hewitt, Charles C. 1966 "Distribution by Size of Risk-A Model," *PCAS* 53:106-114.

[8] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. 2004 *Loss Models: From Data to Decisions*, 2nd Ed. Hoboken, NJ: Wiley.

[9] Kremer, Erhard.1982. "IBNR Claims and the Two Way Model of ANOVA. *Scandinavian Actuarial Journal* 1:47-55.

[10] Kremer, Erhard. 1985. *Einfuhrung in die Versicherungsmathematik*. Göttingen, Ger.: Vandenhoek & Ruprecht.

[11] Long, J. Scott and Laurie H. Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *The American Statistician* 54:217-224.

[12] Mack, Thomas. 1991. "A simple parametric model for rating automobile insurance or estimating IBNR claims reserves." *ASTIN Bulletin* 21, no. 1:93-109.

[13] Mack, Thomas. 1993. "Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates." *ASTIN Bulletin* 23, no. 2:213-225.

[14] Mack, Thomas. 1994. "Measuring the Variability of Chain Ladder Reserve Estimates." *CAS Forum* (Spring):101-182.

[15] Mack, Thomas. 2002. *Schadenversicherungsmathematik*. 2nd Edition. Karlsruhe, Ger.: Verlag Versicherungs-wirtshaft.

[16] Murphy, Daniel M. 1994. "Unbiased Loss Development Factors," *PCAS* 81:154-222.

[17] Renshaw, Arthur E. and Richard J. Verrall. 1998. "A Stochastic Model Underlying the Chain Ladder Technique." *British Actuarial Journal* 4:903-923.

[18] Taylor, Greg C. 1977. "Separation of Inflation and Other Effects From the Distribution of Non-Life Insurance Claim Delays." *ASTIN Bulletin* 9, No. 1-2:219-230.

[19] Taylor, Greg C. 2000. *Loss Reserving: An Actuarial Perspective.* Boston: Kluwer Academic Publishers.

[20] Taylor, Greg C. and Frank R. Ashe. 1983. "Second Moments of Estimates of Outstanding Claims." *Journal of Econometrics* 23:37-61.

[21] Venter, Gary G. 1998. "Testing the Assumptions of Age-to-Age Factors." *PCAS* 85:807-847.

[22] Venter, Gary G. 2007. "Generalized Linear Models beyond the Exponential Family with Loss Reserve Applications," *CAS E-Forum,* Summer (forthcoming).

[23] Verbeek, Harry G. 1972. "An approach to the analysis of claims experience in motor liability excess of loss reassurance." *ASTIN Bulletin* 6, no. 3:195-202.

[24] Weissner, Edward W. 1978. "Estimation of the Distribution of Report Lags by the Method of Maximum Likelihood." *PCAS* 65:1-9.

Biography of the Author

Gary Venter is managing director at Guy Carpenter, LLC. He has an undergraduate degree in philosophy and mathematics from the University of California and an MS in mathematics from Stanford University. He has previously worked at Fireman's Fund, Prudential Reinsurance, NCCI, Workers Compensation Reinsurance Bureau and Sedgwick Re, some of which still exist in one form or another. At Guy Carpenter, Gary develops risk management and risk modeling methodology for application to insurance companies. He also teaches a graduate course in loss modeling at Columbia University.

917.937.3277 **gary.g.venter@guycarp.com**