Meaningful Intervals

Glen Barnett, Ph.D, David Odell, Ph.D, and Ben Zehnwirth, Ph.D, AIA, AIAA[1]

1 Professorial Visiting Fellow, School of Actuarial Studies, Australian School of Business, University of NSW

Abstract: Reserve ranges and risk capital requirements can be related to statistical interval estimates. While not all sources of uncertainty are readily incorporated into an interval estimate, such intervals give a lower bound on the size of the required interval. We discuss the calculation of interval estimates, for both the estimate of the mean and for the liability process itself, show how to tell if the model is a reasonable description of the data and show that when it is not, the interval estimates may sometimes be disastrously wrong.

Many practitioners are now using probabilistic versions of standard actuarial techniques, sometimes employing quite sophisticated tools in their estimation. However, none of these developments avoid the need for stringent checking of the suitability of model assumptions, a necessity that is often overlooked.

We discuss some of the statistical models underlying a variety of standard methods, construct a number of diagnostics for model assessment for several models and discuss how the underlying ideas carry over to many other methods for the estimation of liabilities. These tools are easy to implement and use. They allow practitioners to use the corresponding models with greater confidence, and gain additional information about the triangle. This information can have important consequences for the insurer.

We illustrate that some popular approaches—the Mack chain ladder, the quasi-Poisson GLM—and consequently predictions based on them (both bootstrapped and otherwise) have structure not present in real triangles, and don't describe some features of the data. Consequently their *associated intervals fail to have the desired properties*.

We point out that many aspects of the reserving problem and the structure of real data lead us to model on the log scale. We briefly describe the Probabilistic Trend Family (PTF) models and its extension to the multivariate case (MPTF) and show that these model families can capture the patterns in real data and produce more reasonable prediction intervals.

INTRODUCTION

It is important to distinguish between variability and uncertainty.

Variability is the effect of chance and a function of the system. Additional data points don't reduce the process variability.

Uncertainty is a lack of knowledge about the parameters that characterize the physical system that is being modeled. This may be reducible with additional information.

While separate concepts, variability and uncertainty are not completely unrelated—in general, the uncertainty of a parameter estimate will be related to the variability.

An interval for the mean is a form of confidence interval, based on the associated parameter uncertainty (and possibly including other sources of uncertainty).

An interval for a future payment (a prediction interval) must incorporate both the variability of the process and the uncertainty in the mean.

A sum or even a linear combination of future payments will similarly incorporate process variability for each term, parameter uncertainty for each term, and parameter covariances between each pair of terms. If the model includes correlations between observations, process covariance will also come into the prediction intervals.

The basic properties of confidence intervals and prediction intervals in regression models are presented in many standard statistical textbooks. See, for example, Wackerly, Mendenhall, and Schaeffer (2002). Many of the same principles encountered in the regression context apply more generally and form a useful basis for extending the discussion to intervals in the context of loss reserving and also calculation of risk capital requirements.

SECTION 1: CONFIDENCE INTERVALS AND PREDICTION INTERVALS

Consider the following simple motivating example.

Suppose a fair coin is tossed 100 times and we count the number of heads (X). To draw a parallel with insurance, imagine you pay a dollar for each head.

The mean number of heads (the mean of X) is 50. The standard deviation of X is 5. The binomial probability of each possible outcome of X (0, 1, 2,…100) is known. There is no uncertainty about the coin's mean, its variability, nor any of the probabilities associated with each outcome.

A 100% confidence interval for the mean is [50, 50]. However, the probability that X is equal to the mean of its distribution, 50, is approximately 0.08. A 95% prediction interval for the outcome X is [40, 60]. This prediction interval cannot be shortened without reducing the coverage probability. The inherent variability in the outcomes is termed process variability.

Suppose now that we do not know the true probability of a head, p, perhaps because the coin, or the method by which it is tossed is in some way not fair. Before the coin is tossed 100 times, a preliminary observation is made: it is tossed 20 times to get an estimate of the probability p.

Let's say that 10 heads are observed. Now the estimate of the probability of a head in one toss (10/20 = 0.5) is just that—an estimate. It is uncertain.

We can create a confidence interval (CI) for p and also for the mean number of heads in 100 tosses ($100p$). The CI is an interval around the estimated mean, 50. This confidence interval is not the same as a prediction interval for the *outcome from 100 tosses*.

The risk you're exposed to is the risk of the process, not the mean of the process (you don't pay the mean). Hence, even when the model is known, adequate risk capital is derived from the process variability. However, estimated process variability is insufficient—because the parameters are unknown. Our estimates are not equal to the true values, so we must *also* account for parameter uncertainty.

A prediction interval includes process variability and is therefore wider than a confidence interval. A confidence interval is an interval for a parameter, which is a constant (though unknown), while a prediction interval is for a random variable. The liability on a line of business is a random variable, not a parameter.

Predictive variance (the average variation between the predicted value and the actual outcome) is the sum of process variance and the variance of the parameter estimate (parameter uncertainty).

A 95% prediction interval for the number of heads in 100 tosses will be wider than [40, 60] (which accounts for process variability alone). For example, it might be, say [35, 65]. This interval could only be reduced to at best [40, 60] through additional sampling to reduce the parameter uncertainty. But you cannot make a 95% prediction interval narrower than [40, 60] in this circumstance—only the parameter uncertainty can be reduced.

Consider another situation with the same mean. A fair roulette wheel, numbered 0, 1, 2, 3,....., 100 is spun only once, and let R be the random variable that represents the outcome. The mean of R is 50, and its standard deviation is about 29. There is no uncertainty about the variability in the outcome. The probability that $R=50$ is 1/101.

A 100% *confidence interval* for the mean is [50, 50], as it was for the fair coin (no parameter uncertainty). A 95% *prediction* interval for the outcome R is [3, 98] (there are several such intervals of equal width).

Each process (fair coin and balanced roulette wheel) has the same mean, or if you like the same "best estimate." But the wheel requires more risk capital.

Reserve Ranges

In some countries, a "range" for reserving relates to uncertainty in the mean. As we have seen, a

confidence interval can be produced by considering the uncertainty in the mean arising from parameter uncertainty. In some other countries, reserve calculations will incorporate process uncertainty and consider predictive distributions. In some cases a one-sided interval may be required. More sophisticated approaches can formally incorporate several additional sources of uncertainty into either kind of interval, but consideration of these is beyond the scope of this paper. When some important components of uncertainty are not formally introduced into the calculations, the upper ends of ranges implied by the statistical intervals (whether confidence intervals or prediction intervals) would be lower bounds on the required endpoints (and conversely where lower ends of ranges are required).

Risk Capital

It is important to recognize that the insurer is exposed to the loss process itself, not its mean. That risk includes process risk, and in order to hold risk capital adequate to cover the risk faced, process risk must be included in calculations. This implies that prediction intervals, rather than confidence intervals, are the appropriate starting point.

Reliance on Assumptions

When the assumptions of the model are reasonable, the derived interval estimates will be suitable inputs to reserving and risk-capital calculations. Conversely, when model assumptions are not met, derived intervals may have nothing like the desired properties. It is important to see that the model for the past is a reasonable description, and that the model for the future contains any relevant information about how that may change.

It is unfortunately the case that many models that have been used in loss reserving frequently fail to describe the data. We illustrate this problem by predicting the final calendar year (without using it in the estimation) and showing that the prediction intervals don't behave as they should if the model was adequate. It would seem that being able to predict the past is at least a minimum requirement for a model. If we cannot predict the past, on what basis can we assert we are able to predict the future?

SECTION 2: A BASIC PREDICTION PROBLEM

The simplest prediction problem illustrates many of the issues. Consider the following example.

We have n observations $Y_1, \ldots Y_n \sim F(\mu,\sigma^2)$.

Equivalently, $Y_i = \mu + \varepsilon_i \quad \varepsilon_i \sim F(0, \sigma^2)$.

While for sufficiently large samples, the distribution may not be critical for a confidence interval, it becomes important in the case of prediction intervals, since the observation being predicted will come from that distribution. In order to deal with prediction, we either need to make a distributional assumption (and check it), or use some methodology—such as the bootstrap—that allows us to deal (at least approximately) with the distribution.

Now we want to forecast another observation, Y_{n+1} ($= \mu + \varepsilon_{n+1}$).

So we have:

$\hat{Y}_{n+1} = \hat{\mu} + \hat{\varepsilon}_{n+1}$ ($\hat{\varepsilon}_{n+1}$ is the forecast of the error term)

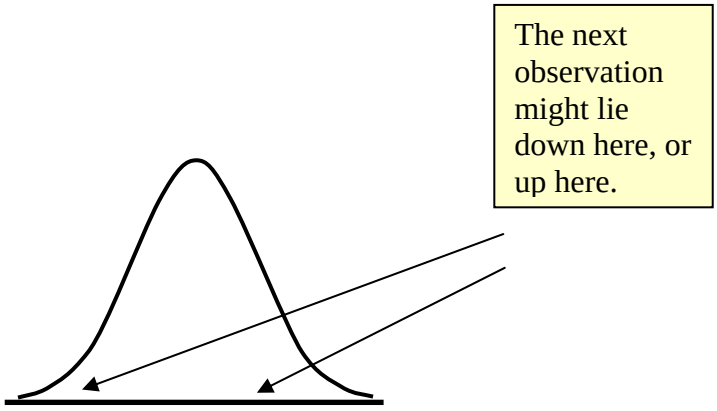
$= \hat{\mu} + 0.$

That is, our best estimate of the next observation is exactly equal to our current estimate of the mean.

Case (i): The mean, μ, is known:

$\hat{Y}_{n+1} = \mu + 0.$

This is equivalent to the case where we knew the coin was fair. Even though we know everything about the process, in predicting Y_{n+1}, we are predicting a random quantity.



The variance of the prediction of Y_{n+1} is:

$\text{Var}(\mu) + \text{Var}(\varepsilon_{n+1}) = 0 + \sigma^2.$

It is important to remember that the *risk to the business* is not simply from the uncertainty in the mean—for example, the value at risk is related to the amount you will actually pay, not its mean.

Even when the mean is known exactly, there is still underlying process uncertainty (with 100 tosses of a fair coin, might get 46 heads or 57 heads).

It doesn't really make sense to talk about a mean (or any other aspect of the distribution) in the absence of a probabilistic model—otherwise what distribution is it the mean *of*? Without some kind of model we cannot even be clear what we're talking about.

With loss modeling, you design a model to describe what's going on with the data. Assumptions need to be explicit so that you can check that the distribution is consonant with the data.

In our motivating example, you would not want to use the coin model if your data was actually coming from the roulette wheel.

Given a suitable model, in practice we simply won't know the mean or other relevant parameters—we only have a sample to tell us about them. (For simplicity we confine our attention to the mean in this discussion.)

Case (ii): The mean, μ, is unknown.

The best estimate for the next observation is still the mean, but we now have to estimate it. This estimate is based on past values and is not exactly equal to the actual underlying population mean — even with a perfect model due to the effects of random variation, the estimate will differ from the underlying value.

While the estimate is uncertain, we can obtain an estimate of the uncertainty, if the model is a good description. So we will have an estimate of the mean and we'll also be able to get a confidence interval for the mean.

This interval is designed so that if we were able to re-run history (re-toss our coin, respin our roulette wheel), many times, the intervals we generate will include the unknown mean a given fraction of the time.

If the model doesn't describe the data, however, the confidence interval may not have anything close to required probability coverage.

Confidence interval for the mean of the coin model

Again, we count how many heads in 100 tosses, but we have a small sample (20 tosses) with which to assess the probability of a head. As mentioned before, let's assume that we observe 10 heads (to keep the mean prediction unchanged).

Obtaining 10 heads in 20 tosses yields $\hat{p} = ½$. $\hat{\mu} = 100.\hat{p} = 50$.

$\text{Var}(\hat{\mu}) = 100^2 \text{Var}(\hat{p}) = 100^2 ½. ½. /20 = 11.18^2$.

An approximate 95% CI for the mean, μ is

$\hat{\mu} \pm 1.96$ s.e.$(\hat{\mu}) \approx (29, 71)$.

Note that the interval here can be based on a normal approximation, due to the central limit theorem. (If the distribution is sufficiently skewed or heavy tailed, the sample may need to be larger for the normal approximation to be reasonable, but in the case of the binomial with p not too far from 0.5, a sample of 20 should be plenty).

Now we again want to look at a *prediction interval*, but here with an estimated mean.

We want to predict a random outcome where we don't know the mean. (In this example we assume the variance is known. In many practical cases uncertainty in the variance does not greatly alter the limits of intervals.)

To understand your business you need to understand the actual risk of the process, not just the risk in the estimate of the mean.

Let's revisit the simple model $Y_{n+1} = \mu + \varepsilon_{n+1}$:

$\hat{Y}_{n+1} = \hat{\mu} + \hat{\varepsilon}_{n+1}$.

Now μ is unknown. So $\hat{Y}_{n+1} = \hat{\mu} + 0$.

Variance of forecast

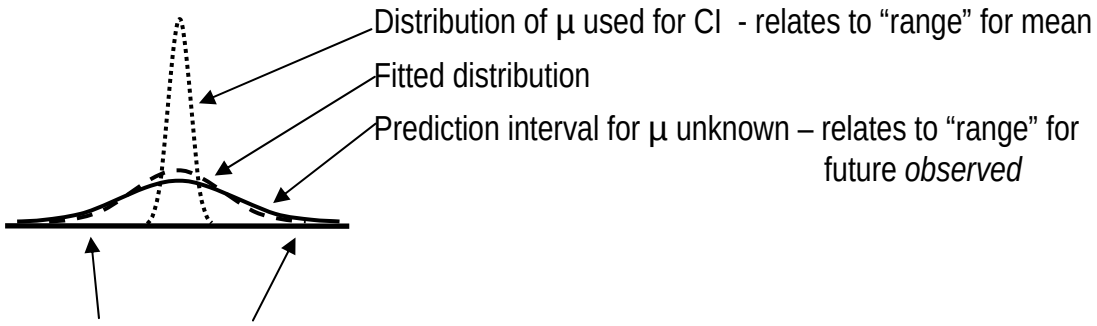
$= \text{Var}(\hat{\mu}) + \text{Var}(\varepsilon_{n+1})$

$= \sigma^2/n + \sigma^2$.

(In practice, σ^2 is replaced by its estimate, of course. We are ignoring the parameter uncertainty in

the variance for the present discussion.)

Now imagine that the distribution, F, is normal. The next observation might lie



down here, or up here. (While we've assumed normality here, the issues in the diagram apply more widely.)

In the coin experiment, the predictive distribution is approximately normal.

Y_{n+1} is the number of heads on our next run of 100 tosses. Its predictive variance is

$\text{Var}(Y_{n+1} \mid \mu = \hat{\mu}) + \text{Var}(\hat{\mu}) = 100 \ \tfrac{1}{2}.\tfrac{1}{2} + 100^2 \tfrac{1}{2}.\tfrac{1}{2}./20 = 12.25^2$

So an approximate 95% CI for the forecast Y_{n+1} is

$\hat{Y}_{n+1} \pm 1.96 \text{ s.e.}(\hat{Y}_{n+1}) \approx (26,74)$.

An alternative way to look at the predictive variance

Here we derive the predictive variance as the variance of the prediction error.

Prediction error = $Y_{n+1} - \hat{Y}_{n+1}$.

Predictive variance = Var(prediction error)

$= \text{Var}(Y_{n+1} - \hat{Y}_{n+1})$

$= \text{Var}(Y_{n+1}) + \text{Var}(\hat{Y}_{n+1}) - 2 \text{ Cov}(Y_{n+1}, \hat{Y}_{n+1})$

$= \text{Var}(Y_{n+1}) + \text{Var}(\hat{Y}_{n+1}) - 0$

$= \text{Var}(Y_{n+1}) + \text{Var}(\hat{Y}_{n+1})$

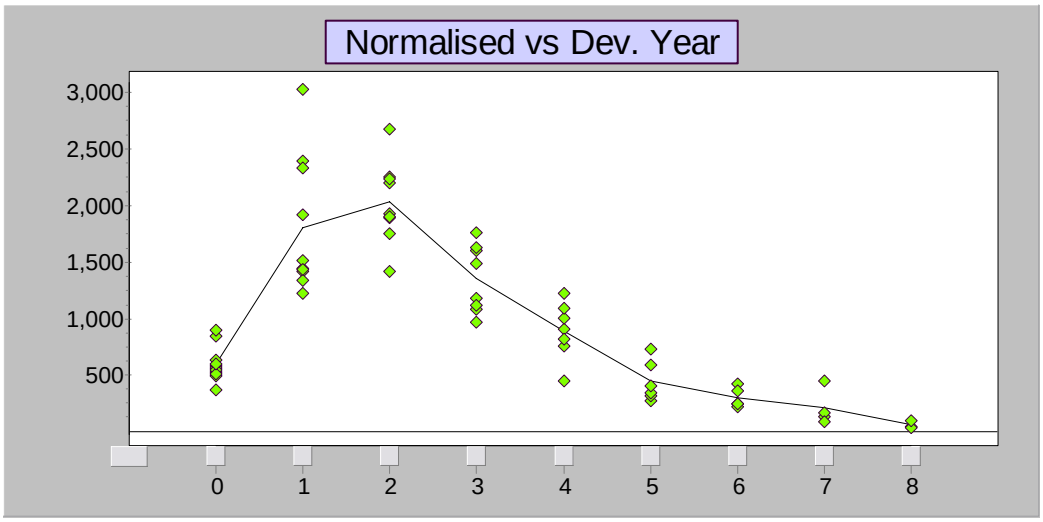
= process variability + parameter uncertainty.

Note that this result only relies on the fact that Cov $(Y_{n+1}, \hat{Y}_{n+1}) = 0$, which, for example, occurs when observations are independent (since then the forecast $\hat{Y}_{n+1}$ is a function of only past observations, while Y_{n+1} is a future observation, which will be independent of it) and the result follows.

With more complex models the calculation of the variance of the estimate of the mean is more complicated, but the principle remains the same.

Confidence intervals vs. Prediction intervals—a basic loss example

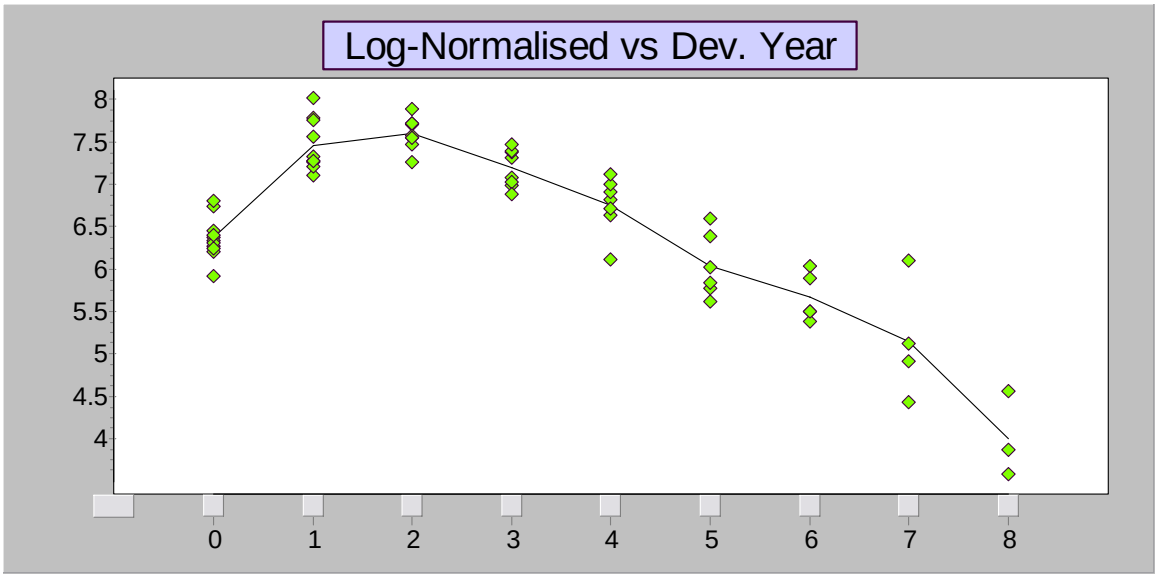
Let's look at some real long-tail data that has been inflation-adjusted and then normalized for a measure of exposure. This is the CTP data that was analyzed in Barnett and Zehnwith (2000).



We see a clear runoff pattern against development year. In this instance the trends in the accident-year and calendar-year directions sufficiently small that we can ignore them for illustrative purposes.

Note that in the figure above, the points have a tendency to "clump" just below the mean and be more spread out above the mean—the normalized data is skewed to the right.

On the log-scale this skewness disappears, and the variance is pretty stable across years. The skewness is removed so the values appear much more symmetric about center and the spread looks fairly constant.



Consider a single development—say DY 3:

AY	Normalized	Log
1	1,489	7.306
2	1,606	7.381
3	1,087	6.991
4	1,628	7.395
5	1,178	7.072
6	1,118	7.019
7	1,761	7.474
8	972	6.879

The maximum likelihood estimates of μ and σ are 7.190 and 0.210, respectively.

(NB: the MLE of σ is s_n, the standard deviation with the n denominator, not the more common s_{n-1}).

Assuming a random sample from a process with constant mean, we would predict the mean for next value as 7.190. However, without some indication of its accuracy, this is not very helpful.

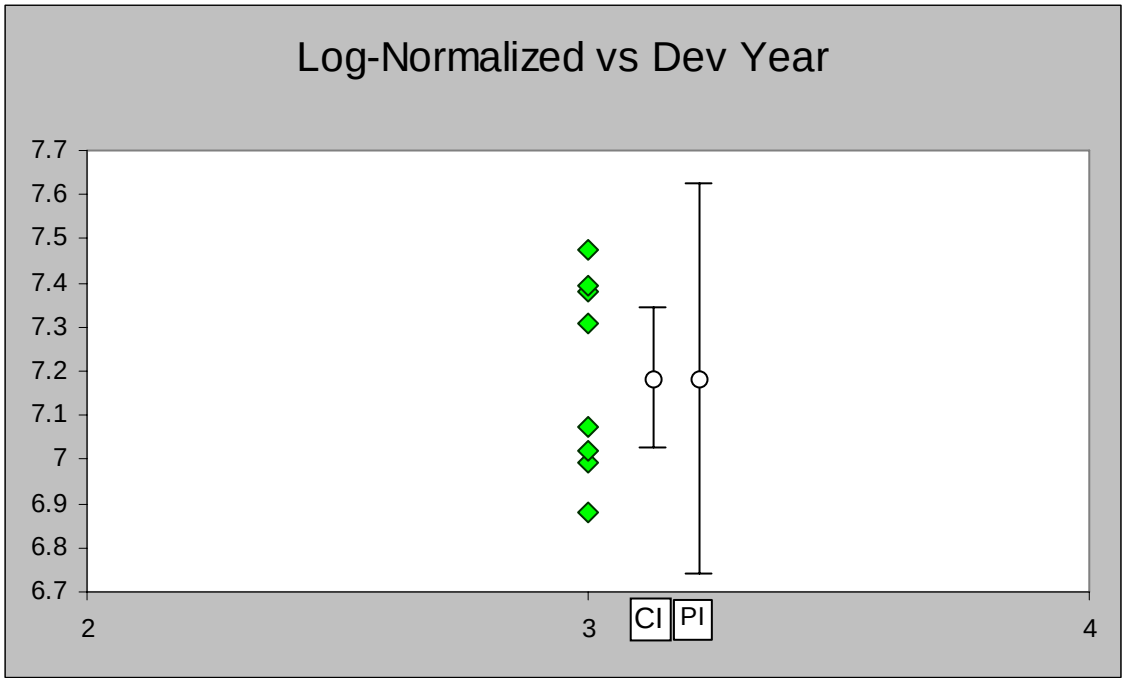
A 95% confidence interval for μ is: (7.034, 7.345).

Prediction:

Recall that the *predictive* variance is $\mathrm{Var}(\hat{\mu}) + \mathrm{Var}(\varepsilon_{n+1})$

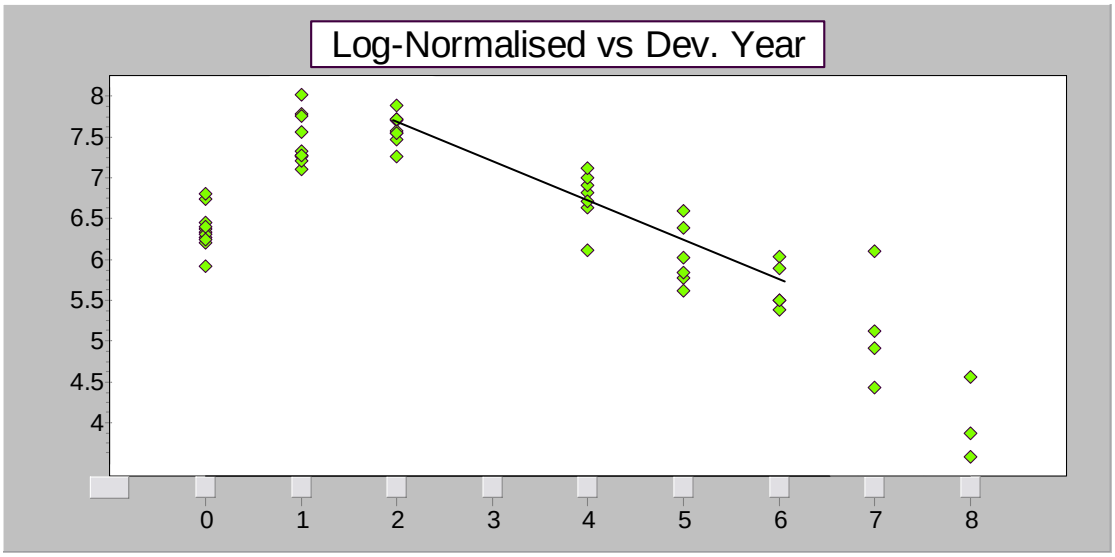
(= parameter uncertainty + process var).

A 95% *prediction interval for* Y_{n+1} is (6.75,7.63). See the figure below.



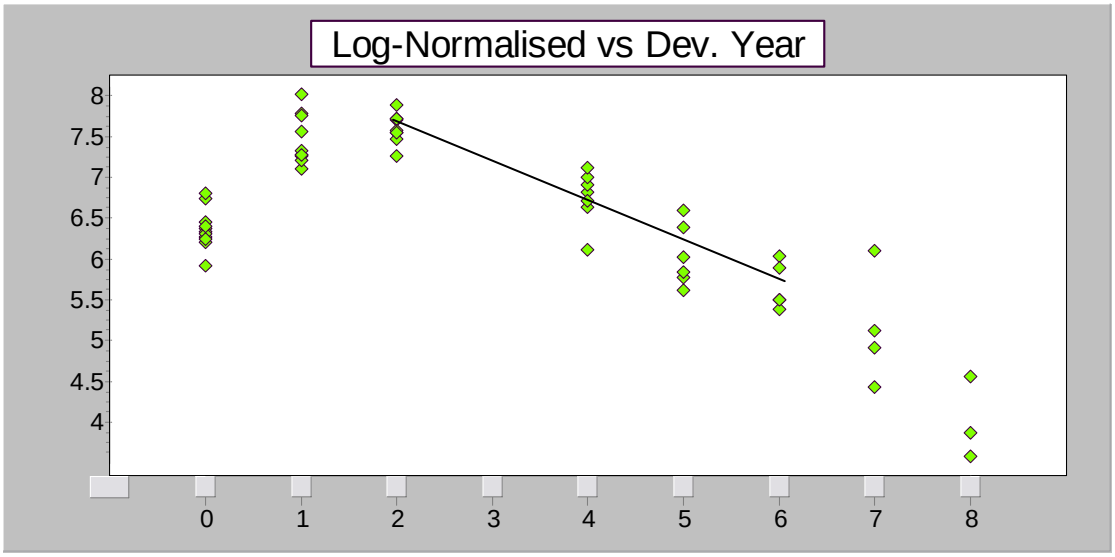
The intervals here are fairly wide. More data reduces parameter uncertainty (e.g., more than 20 tosses of the coin in the earlier trial would make the intervals smaller). In some cases you can go back and get more loss data and eventually you'll have another year of data. However, process variability doesn't reduce with more data—it's an aspect of the process. We can measure it more accurately, but the thing we're measuring is not changing.

As we can see in the figure below, nearby developments are related: if DY 3 was all missing, you could take a fair guess at where it was.



So in this case we *do* have more data!

To take full advantage of this, we need a model to relate the development years. Even just fitting line through DY2-4 has a reasonably large effect on the width of the *confidence interval* (the grey bars shift inward, to the black bars).



However, it only changes the *prediction interval* by ~2%—so calculated VaR hardly changes.

Note that so far this prediction interval is on the *log* scale. To take a *prediction* interval back to the normalized-dollar scale, we just back-transform the endpoints of the prediction interval. To produce

a *confidence interval* for the *mean on the normalized-dollar* scale is harder. We *can't* just backtransform the limits on the confidence interval—that's going to give an interval for the *median*, not the mean. However, we can scale the interval for the median to produce an interval for the mean.

Further, to convert the interval for these scaled dollars to original dollars, we need to re-scale the interval for the inflation and exposures.

There are some companies around for whom (for some lines of business) the process variance is very large—some have a coefficient of variation near 0.6 (so the standard deviation is > 60% of the mean). That's just a feature of the data. You may not be able to control it, but you sure need to *know* it.

Why take logs?

Taking logs tends to stabilize variance. Multiplicative effects (percentage changes, including economic effects such as inflation) become additive. Exponential growth or decay becomes linear. Skewness is usually eliminated. Distributions tend to look near normal, making least squares reasonable. Using logs is a familiar way of dealing with many of these issues—indeed, it's standard in many parts of finance.

Note that for these benefits to work, we have to take logs of *incremental* figures (such as incremental paid), rather than cumulative paid or incurred losses. For example, inflation in the past period affects payments now, but not past payments—so cumulatives (which are also present in incurred figures) will contain a mix of payments across past rates.

SECTION 3: DIAGNOSTIC DISPLAYS FOR CHAIN LADDER MODELS

In this section we consider two models that reproduce chain ladder forecasts, the regression model (Mack model, Mack, 1993) and the quasi-Poisson GLM (Hachemeister and Stanard, 1975, and apparently independently by Renshaw and Verrall, 1994).

Many common regression diagnostics for model adequacy relate to analysis of residuals, particularly residual plots. In many cases these work very well for examining many aspects of model adequacy. When it comes to assessing predictive ability, the focus should, where possible, shift to examining the ability to predict data not used in the estimation. In a regression context, a subset of the data is held aside and predicted from the remainder. Generally the subset is selected at random from the original data. However, in our case, we cannot completely ignore the time-series structure and the fact that we're predicting outside the range of the data. Our prediction is always of future

calendar time. Consequently the subsets that can be held aside and assessed for predictive ability are those at the most recent time periods.

This is common in analysis of time series. For example, models are sometimes selected so as to minimize one-step-ahead prediction errors. See, for example, Chatfield (2000).

Out of sample predictive testing

The critical question for a model being used for prediction is whether the estimated model can predict outside the sample used in the estimation. Since the triangle is a time series, where a new diagonal is observed at each calendar period, prediction (unlike predictions for a model without a time dimension) is of calendar periods after the observed data. To do out-of-sample tests of predictions, it is therefore important to retain a subset of the most recent calendar periods of observations for post-sample predictive testing. We refer to this post-sample-predictive testing as *model validation* (note that some other authors use the term to mean various other things, often related to checking the usefulness or appropriateness of a model).

Imagine we have data up to time t. We use only data up to time $t–k$ to estimate the model and predict the next k periods (in our case, calendar periods), so that we can compare the ability of the model to predict actual observations not used in the estimation. We can, for example, compute the prediction errors (or validation residuals), the difference between observed and predicted in the validation period. If these prediction errors are divided by the predictive standard error, the resulting standardized validation residuals can be plotted against time (calendar period most importantly, and also accident and development period), and against predicted values, (as well as against any other likely predictor), in similar fashion to ordinary residual plots. Indeed, the within-sample residuals and "post-sample" predictive errors (validation residuals) can be combined into a single display.

One step ahead prediction errors are related to validation residuals, but at each calendar time step only the next calendar period is predicted; then the next period of data is brought in and another period is predicted.

In the case of ratio models such as the chain-ladder, prediction is only possible within the range of accident and development years used in estimation, so out of sample prediction cannot be done for all observations left out of the estimation. The use of one step ahead prediction errors maximizes the number of out-of-sample cases that have predictions. Further, when reserving, the liability for the next calendar period is generally a large portion of the total liability, and the liability estimated will typically be updated once it is observed; this makes one-step-ahead prediction errors a particularly useful criterion for model evaluation when dealing with ratio models like the chain

ladder.

For a discussion of the use of out of sample prediction errors and in particular one-step-ahead prediction errors in time series, see Chatfield (2000), chapter 6.

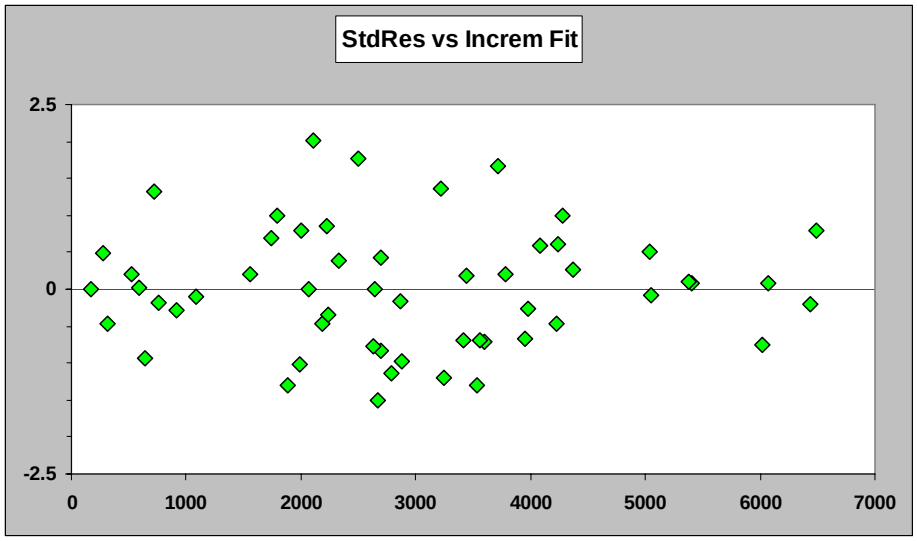
For many models, the patterns in residual plots when compared with the patterns in validation residuals or one step ahead prediction errors appear quite similar. In this circumstance, ordinary residual plots will generally be sufficient for identifying model inadequacy.

Critically, in the case of the Poisson and quasi-Poisson GLM that reproduce the chain ladder, the prediction errors and the residuals *do* show different patterns.

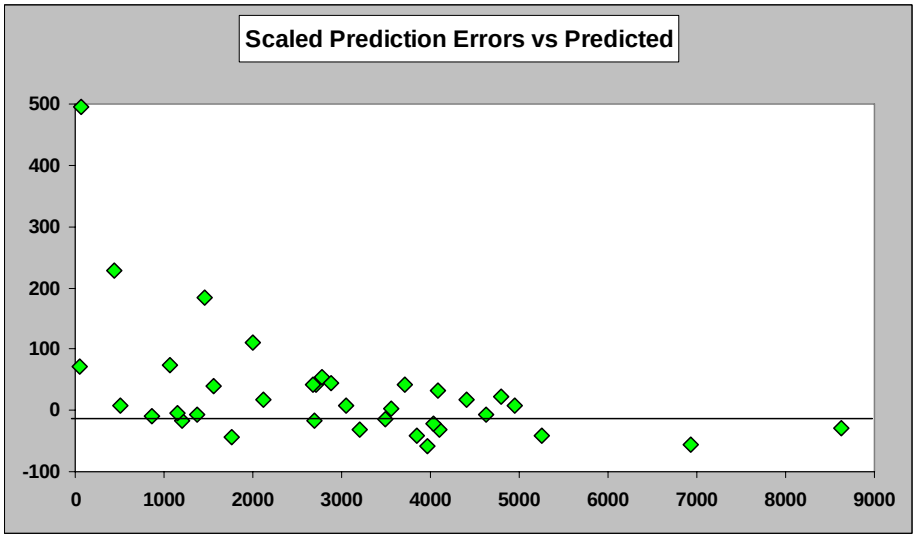
Illustration:

This data was used in Mack (1994). The data are incurred losses for automatic facultative business in general liability, taken from the Reinsurance Association of America's Historical Loss Development Study.

If we fit a quasi- (or overdispersed) Poisson GLM and plot standardized residuals against fitted values, the plot appears to show little pattern:

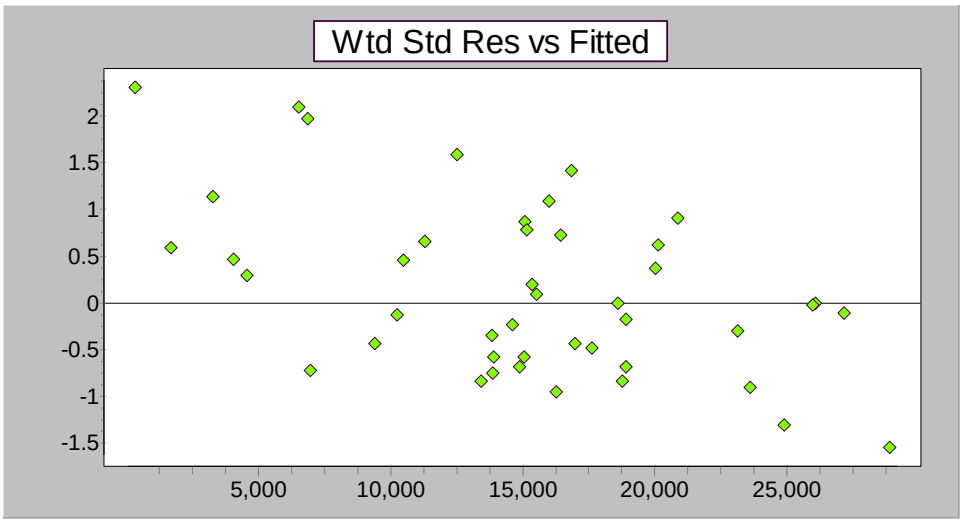


However, if we plot *one step ahead prediction errors* (scaled by dividing the prediction errors by $\hat{\mu}^{1/2}$) against predicted values, we *do* see a distinct pattern of mostly positive prediction errors for small predictions with a downward trend toward more negative prediction errors for large predictions:



Prediction errors above have not been standardized to have unit variance. The underlying quasi-Poisson scale parameter would have a different estimate for each calendar-year prediction; it was felt that the additional noise from separate scaling would not improve the ability of this diagnostic to show model deficiencies. On the other hand, using a common estimate across all the calendar periods would simply alter the scale on the right-hand side without changing the plot at all, and has the disadvantage that for many predictions you'd have to scale them using "future" information. On the whole it seems prudent to avoid the scaling issue for this display, but as a diagnostic tool, such scaling is not a major issue.

This problem of quite different patterns for prediction errors and residuals does not generally occur with the Mack formulation of the chain ladder, where ordinary residuals are sufficient to identify this problem:



As noted in Barnett and Zehnwirth (2000), this downward trend is caused by a simple failure of

the ratio assumption—it is not true that $E(Y|X) = \beta X$, as would be true of any model where the next cumulative is assumed to be (on average) a multiple of the previous one. (For this data, the relationship between a cumulative and the previous cumulative does not go through the origin.)

The above plot is against cumulatives because in the Mack formulation, that's what is being predicted. (Note that the quasi-Poisson GLM residuals vs. cumulative fitted rather than incremental fitted still looks flat.)

Why is the problem obvious in the residuals for the Mack version of the chain ladder model, but not in the plots of GLM residuals vs. fitted (either incremental or cumulative)?

Even though the two models share the same prediction function, the *fitted values* of the two models are quite different.

On the cumulative scale, if X is the most recent cumulative (on the last diagonal) and Y is the next (future) one, both models have the prediction-function $E(Y|X) = \beta X$.

However, *within* the data, while the Mack model uses the same form for the fit—$E(Y|X) = \beta X$, the GLM does not—you can write it as $E(Y) = \beta E(X)$, which seems similar enough that it might be imagined it would not make much difference, but the right-hand side involves "future" values not available to predictions. This allows the fit to "shift" itself to compensate, so you can't see the problem in the fits. However, the out-of-sample prediction function is the same as for the Mack formulation, and so the predictions from the GLM suffers from *exactly* the same problem—once you forecast future values, you're assuming $E(Y|X) = \beta X$ for the future—and this assumption needs to be checked! It cannot be assessed using a *within-data* (residual) analysis of the GLM; it needs to be assessed by checking the actual assumption, whether via use of prediction residuals, or by checking Mack residuals.

Adequate model assessment of quasi-Poisson GLMs therefore *requires* the use of some form of out-of-sample prediction, and because of the structure of the chain ladder, this assessment seems to be best done with one-step-ahead prediction errors. For many other models, such as the Mack model, this would be useful but not as critical, since we can identify the problem even in the residuals.

SECTION 4: THE BOOTSTRAP

The bootstrap is, at heart, a way to obtain an approximate sampling distribution for a statistic (and hence, if required, produce a confidence interval). Where that statistic is a suitable estimator for

a population parameter of interest, the bootstrap enables inferences about that parameter. In the case of simple situations the bootstrap is very simple in form, but more complex situations can also be dealt with. The bootstrap can be modified in order to produce a predictive distribution (and hence, if required, prediction intervals).

It is predictive distributions that are generally of prime interest to insurers (because they pay the outcome of the process, not its mean). The bootstrap has become quite popular in reserving in recent years.

The bootstrap does not require the user to assume a distribution for the data. Instead, sampling distributions are obtained by resampling the data.

However, the bootstrap certainly does not avoid the need for assumptions, nor for checking those assumptions. The bootstrap is far from a cure-all. It suffers from essentially the same problems as finding predictive distributions and sampling distributions of statistics by any other means. These problems are exacerbated by the time-series nature of the forecasting problem—because reserving requires prediction into never-before-observed calendar periods, model inadequacy in the calendar-year direction becomes a critical problem. In particular, the most popular actuarial techniques—those most often used with the bootstrap—don't have any parameters in that direction, and are frequently mis-specified with respect to the behavior against calendar time. The bootstrap does not solve this problem.

Further, commonly used versions of the bootstrap can be sensitive to overparameterization—and overparameterization is a common problem with standard techniques.

A basic bootstrap introduction

The *bootstrap* was devised by Efron (1979), growing out of earlier work on the *jackknife*. He further developed it in a book (Efron, 1982), and various other papers. These days there are numerous books relating to the bootstrap, such as Efron and Tibshirani (1994). A good introduction to the basic bootstrap may be found in Moore et al. (2003); it can be obtained online.

The original form of the bootstrap is where the data itself is resampled, in order to get an approximation to the sampling distribution of some statistic of interest, so an inference can be made about a corresponding population statistic.

For example, in the context of a simple model $E(X_i) = \mu$, $i = 1, 2, \ldots, n$, where the Xs assumed to be independent, the population statistic of interest is the mean, μ, and the sampling statistic of

interest would typically be the sample mean, $\bar{x}$.

Consequently, we estimate the population mean by the sample mean ($\hat{\mu} = \bar{x}$) —but how good is that estimate? If we were to collect many samples, how far would the sample means typically be from the population mean?

While that question could be answered if we could directly take many samples from the population, typically we cannot resample the original population again. If we assume a distribution, we could infer the behavior of the sample mean from the assumed distribution, and then check that the sample could reasonably have come from the assumed distribution.

(Note that rather than needing to assume an entire distribution, if the population variance were assumed known, we could compute the variance of the sample mean, and given a large enough sample, we might consider applying the central limit theorem (CLT) in order to produce an approximate interval for the population parameter, without further assumptions about the distributional form. There are many issues that arise. One such issue is whether or not the sample is large enough—the number of observations per parameter in reserving is often quite small. Indeed, many common techniques have some parameters whose estimates are based on only a single observation! Another issue is that to be able to apply the CLT we assumed a variance—if instead we estimate the variance, then the inference about the mean depends on the distribution again. As the sample sizes become large enough that we may apply Slutsky's theorem, then for example a *t*-statistic is asymptotically normal, even though in small samples the *t*-statistic *only has a t-distribution if the data were normal*. Lastly, and perhaps most importantly when we want a *predictive distribution*, the CLT generally cannot help.)

In the case of bootstrapping, the *sample* is itself resampled, and then from that, inferences about the behavior of samples from the population are made on the basis of those resamples. The empirical distribution of the original sample is taken as the best estimate of the population distribution.

In the simple example above, we repeatedly draw samples of size *n* (with replacement) from the original sample, and compute the distribution of the statistic (the sample mean) of each resample. Not all of the original sample will be present in the resample—on average a little under 2/3 of the original observations will appear, and the rest will be repetitions of values already in the sample. A few observations may appear more than twice.

The standard error, the bias and even the distribution of an estimator about the population value can be approximated using these resamples, by replacing the population distribution, *F* by the

empirical distribution F_n.

For more complex models, this direct resampling approach may not be suitable. For example, in a regression model, there is a difficulty with resampling the responses directly, since they will typically have *different* means.

For regression models, one approach is to keep all the predictors with each observation and sample them together. That is, if $\underline{X}$ is a matrix of predictors (sometimes called a *design matrix*) and $\underline{y}$ is a data-vector, for the multiple regression model $\underline{Y} = \underline{X}\,\beta + \underline{\varepsilon}$, then the rows of the augmented design matrix $[X|y]$ are resampled. (This is particularly useful when the Xs are thought of as random.)

A similar approach can be used when computing multivariate statistics, such as correlations.

Another approach is to resample the *residuals* from the model. The residuals are estimates of the error term, and in many models the errors (or in some cases, scaled errors) at least share a common mean and variance. The bootstrap in this case assumes more than that—they should have a common distribution (in some applications this assumption is violated).

In this case (with the assumption of equal variance), after fitting the model and estimating the parameters, the residuals from the model are computed: $e_i = y_i - \hat{y}_i$, and then the residuals are resampled as if they were the data.

Then a new sample is generated from the resampled residuals by adding them to the fitted values, and the model is fitted to the new bootstrap sample. The procedure is repeated many times.

Forms of this *residual resampling* bootstrap have been used almost exclusively in reserving, even when the other form of the bootstrap could be used.

If the model is correct, appropriately implemented residual resampling works. If it is incorrect, the resampling scheme will be affected by it, some more than others, though in general the size of the difference in predicted variance is small. More sophisticated versions of this kind of resampling scheme, such as the second bootstrap procedure in Pinheiro et al. (2003) can reduce the impact of model misspecification when the prediction is, as is common for regression models, within the range of the data. However, the underlying problem of amplification of unfitted calendar-year effects remains, as we shall see.

For the examples in this paper we use a slightly augmented version of *Sampler 2* given in Pinheiro et al.—the prediction errors are added to the predictions to yield bootstrap-simulated predictive

values, so that we can directly find the proportion of the bootstrap predictive distribution below the actual values in one-step-ahead predictions.

In the case of reserving, the special structure of the problem means that while often we predict inside the range of observed accident years, and usually also within the range of observed development years, we are always projecting *outside* the range of observed calendar years—precisely the direction in which the models corresponding to most standard techniques are inadequate.

As a number of authors have noted, the chain ladder models the data using a two-way cross-classification scheme (that is, like a two-way main-effects ANOVA model in a log-link). As discussed in Barnett et al. (2005), this is an unsuitable approach in the accident and development direction, but the issues in the calendar direction are even more problematic. Even the more sophisticated approaches to residual resampling can fail on the reserving problem if the model is unsuitable.

Assessing bootstrap prediction intervals

When calculating predictive distributions with the bootstrap, we can in similar fashion make plots of standardized prediction errors against predicted values and against calendar years. Of course, since the prediction *errors* are the same, the only change would be a difference in the amount by which each prediction error is scaled (since we have bootstrap standard errors in place of asymptotic standard errors from an assumed model); the broad pattern will not change, however, so the plot based on asymptotic results are useful prior to performing the bootstrap.

Since we can produce the entire predictive distribution via the bootstrap, we can evaluate the percentiles of the omitted observations from their bootstrapped predictive distributions—if the model is suitable, the data should be reasonably close to "random" percentiles from the predictive distribution. This further information will be of particular interest for the most recent calendar periods (since the ability of the model to predict recent periods gives our best available indication if there is any hope for it in the immediate future—if your model cannot predict last year you cannot have a great deal of confidence in its ability to predict *next* year).

We could look at a visual diagnostic, such as the set of predictive distributions with the position of each value marked on it, though it may be desirable to look at all of them together on a single plot, if the scale can be rendered so that enough detail can be gleaned from each individual component. It may be necessary to "summarize" the distribution somewhat in order to see where the values lie (for example, indicating 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles, rather than showing the entire bootstrap density). In order to more readily compare values it may help to standardize by

subtracting the mean and dividing by the standard deviation, though in many cases, if the means don't vary over too many standard deviations, simply looking at the original predicted values on (whether on the original scale or on a log scale) may be sufficient—sometimes a little judgment is required as to which plot will be most informative.

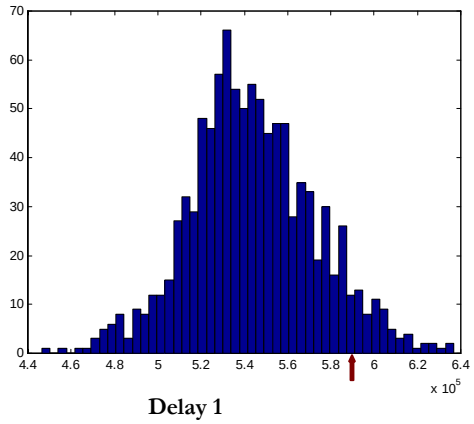
SECTION 5: EXAMPLES

Example 1

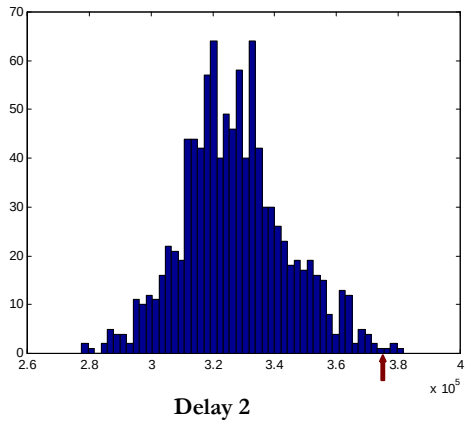
ABC data is workers compensation data for a large company. This data was analyzed in some detail in Barnett and Zehnwirth (2000).

In this example we actually use the bootstrap predictions discussed in the basic bootstrap introduction above, based on the second algorithm from Pinheiro et al. (2003). Below are the predictive distributions for the first two values (after DY0) for the last diagonal, for a quasi-Poisson GLM fitted to the data prior to the final calendar year, which was omitted. The brown arrows mark the *actual* observation that the predictive distribution is attempting to predict.

ABC Predictive distribution for last diagonal - histograms



Delay 1

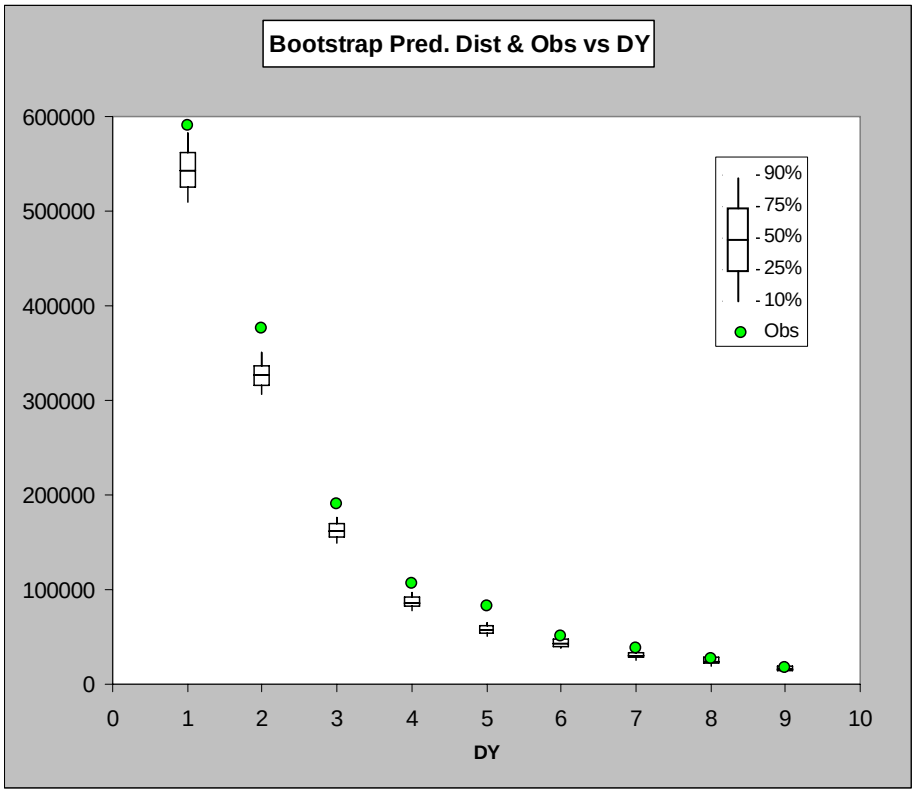


Delay 2

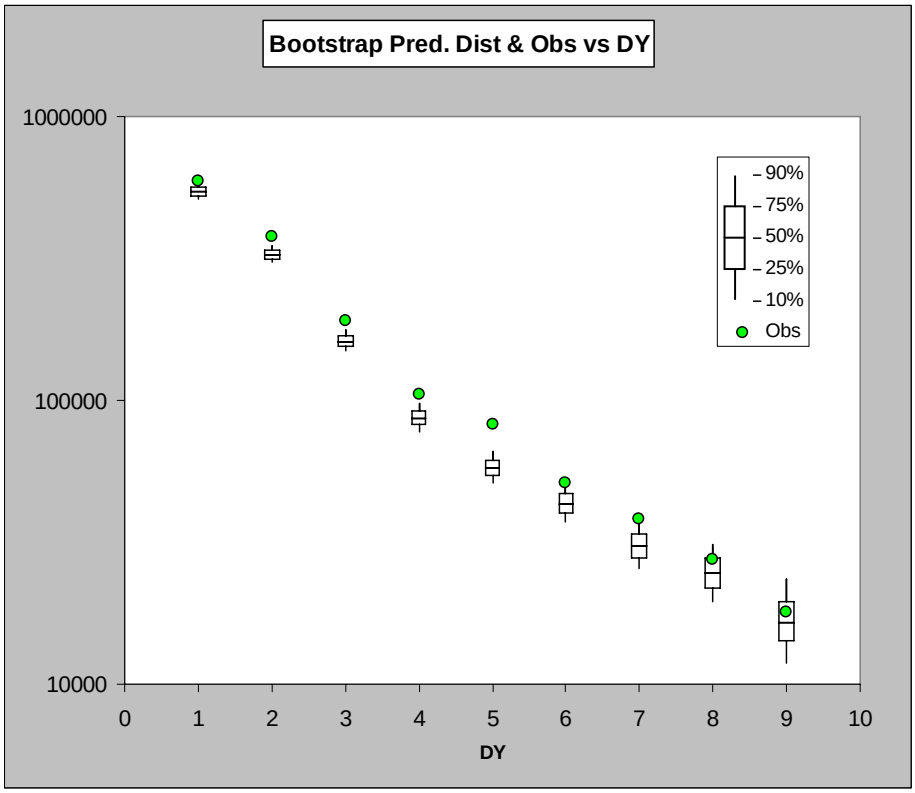
For the two distributions shown, the observed values sit fairly high. For a single observation, this might happens by chance, even with an appropriate model, of course.

The runoff decreases sharply for this triangle, so most of this information in the histograms would be lost if we looked at them on a single plot. Consequently, for a more detailed examination, the bootstrap results are reduced to a five-number summary (in the form of a boxplot) of the percentiles:

ABC Predictive distribution for last diagonal—box and whisker plots



The actual payments for the first seven development periods are all very high, but it's a little hard to see the details in the last few periods. Let's look at them on the log-scale:



Now we can see that in all cases the observations sit above the median of the predictive distribution, and all but the last two are above the upper quartile.

Below is a summary table of the bootstrap distribution for the final calendar year:

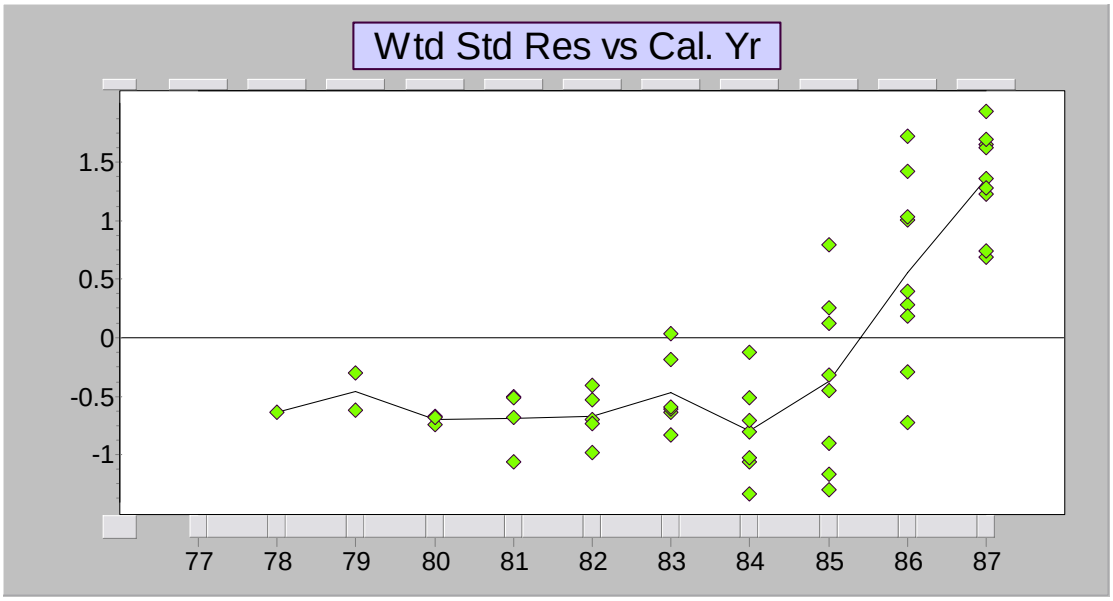
ABC: Bootstrap Predictive distributions for last calendar year

DY	Actual	10%	25%	50%	75%	90%	% ≤obs
0	496200						
1	590400	509620	525430	542150	562070	583270	93.9
2	375400	306580	315890	326600	337060	351080	99.6
3	190400	148750	155240	161520	169110	176290	98.9
4	105600	77760	81850	86220	91330	97340	99.2
5	82400	51050	54270	57740	61590	65730	100
6	51000	37440	40300	43360	46950	51380	89.3
7	38000	25490	27940	30770	33680	37110	92
8	27400	19430	21920	24540	27840	30970	72.9
9	18000	11930	14210	16460	19630	23450	63.9
10	12200						

So what's going on? Why is this predicting so badly?

We would see via one-step-ahead prediction errors that there's a problem with the assumption of no calendar-period trend; alternatively, as we noted earlier, we simply can look at residuals from a

Mack-style model and get a similar impression:

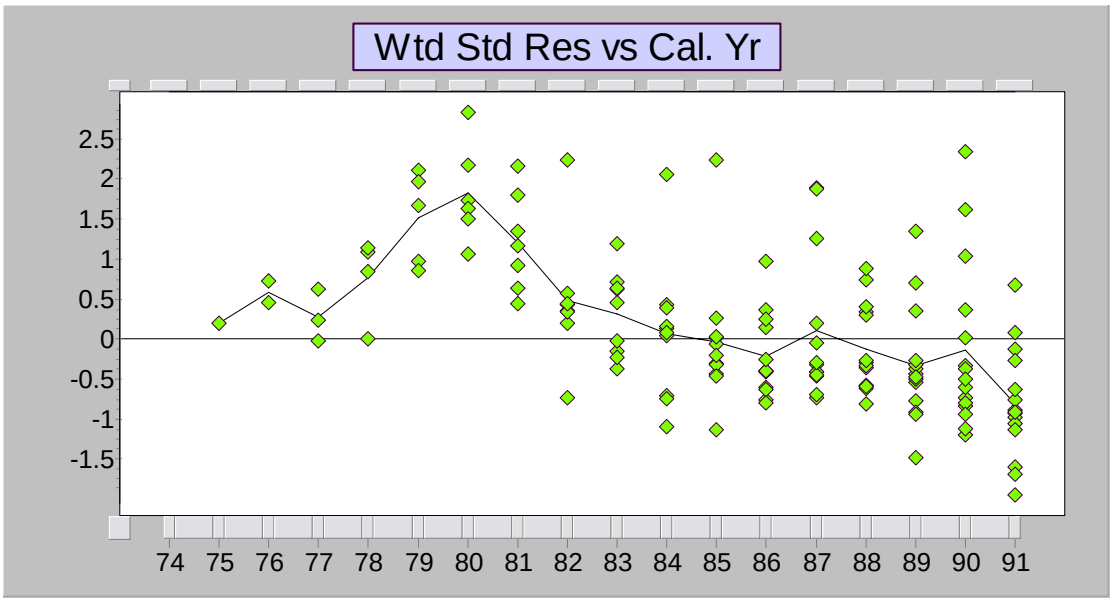


There is a strong trend-change in the calendar-year direction. Consequently, predictions of the last calendar year will be too low. One major difficulty with the common use of the chain ladder in the absence of careful consideration of the remaining calendar-period trend is that there is no opportunity to apply proper judgment of the future trends in this direction. The practitioner lacks a context for seeking all the information relevant to scenarios for future behavior.

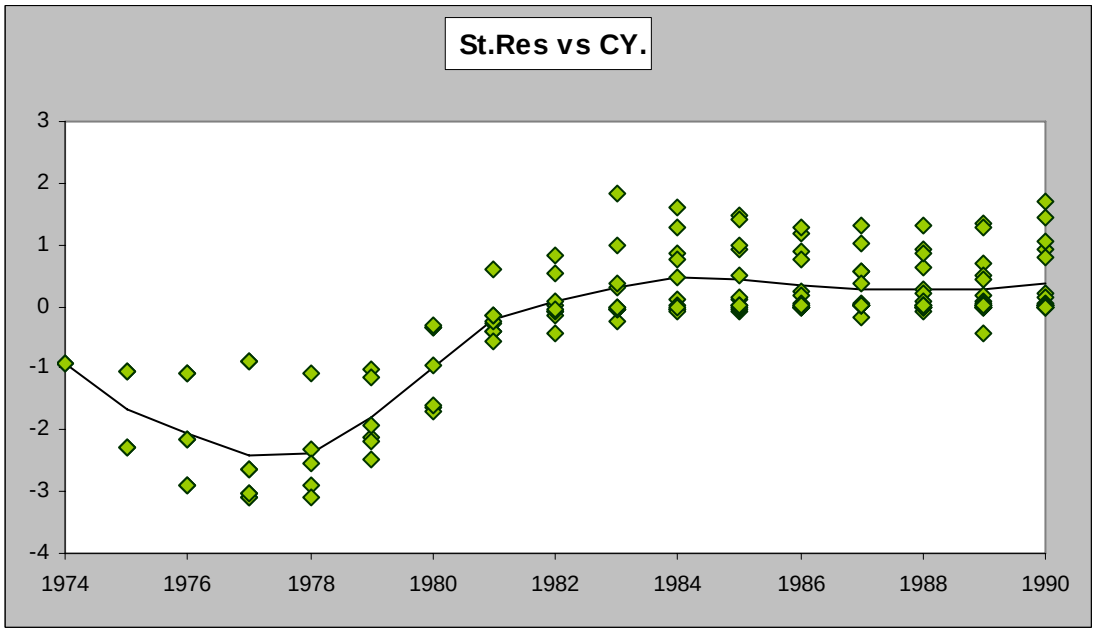
Example 2—LR high

As we have seen, we can look at diagnostics, which would have allowed us to assess *before we try to produce bootstrap prediction intervals* whether we should proceed.

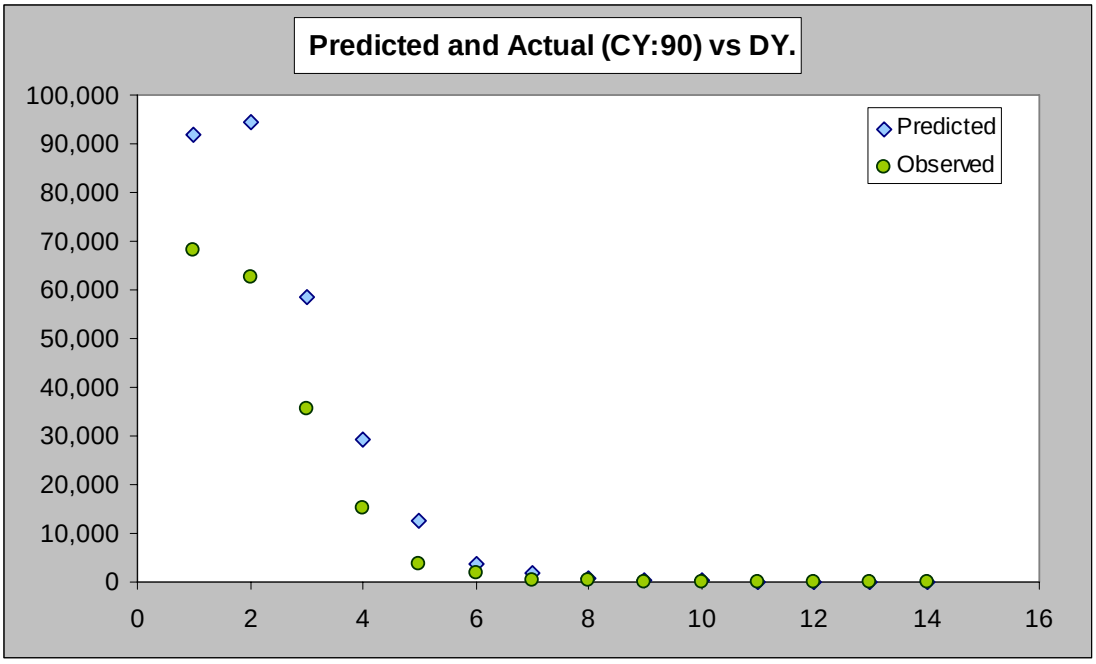
Here are the standardized residuals vs. calendar years from a Mack-style chain ladder fit. As you can see, there's a lot of structure.



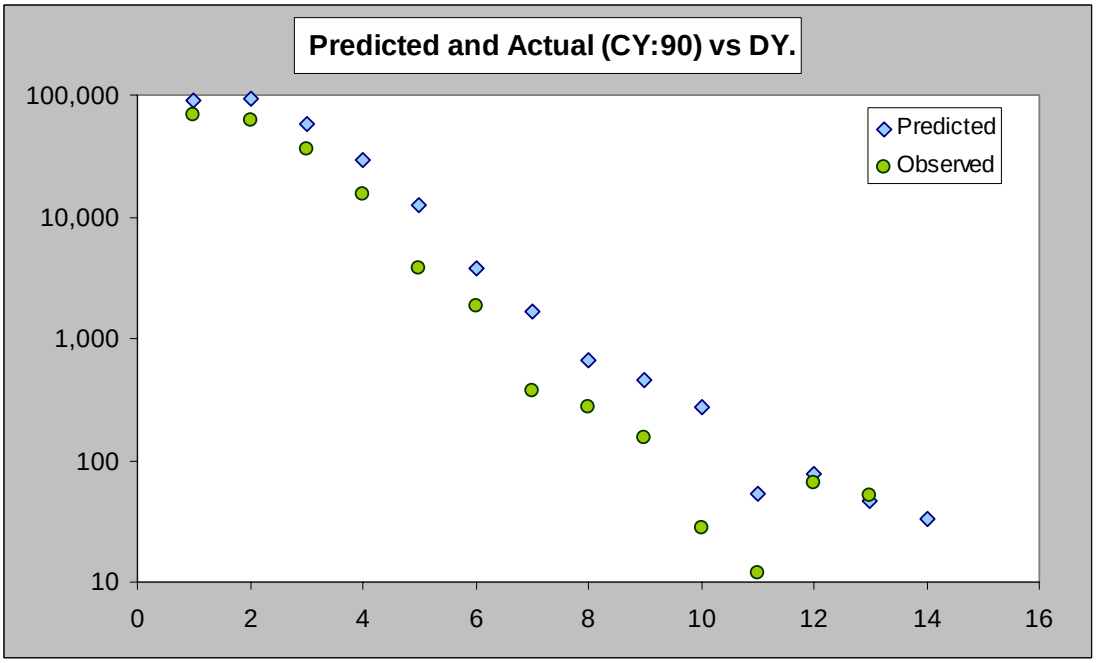
There's also structure in the quasi-Poisson GLM formulation of the chain ladder— residuals show there are strong trend changes in the calendar-year direction:



However, as we described before, this residual plot gives the incorrect impression that the GLM is underpredicting. This impression is incorrect, as we see by looking at the validation (one step ahead predictions) for the last year:

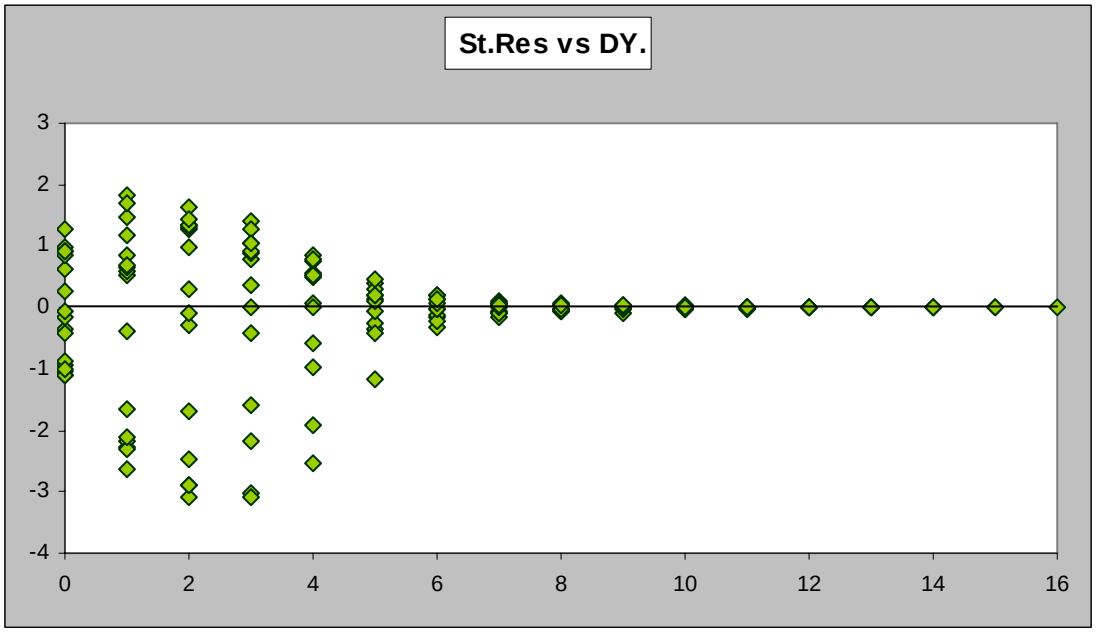


It's a little hard to see detail over on the right, so let's look at the same plot on the log scale:



The Mack-model residual plot gave a good indication of the predictive performance of the chain ladder (bootstrapped or not) for both the Mack model *and* the quasi- (overdispersed) Poisson GLM. It's always a good idea to validate the last calendar year (look at one-step-ahead prediction errors), but a quick approximation of the performance is usually given by examining residuals from a Mack-chain ladder model.

A further problem with the GLM is revealed by the plot of residuals vs. development year. The assumed variance function does not reflect what's in the data—and hence the prediction intervals cannot be correct:



Example 3

The next example has been widely used in the literature relating to the chain ladder. Indeed, Pinheiro et al. (2003) referred to it as a "benchmark for claims reserving models." The data come from Taylor and Ashe (1983).

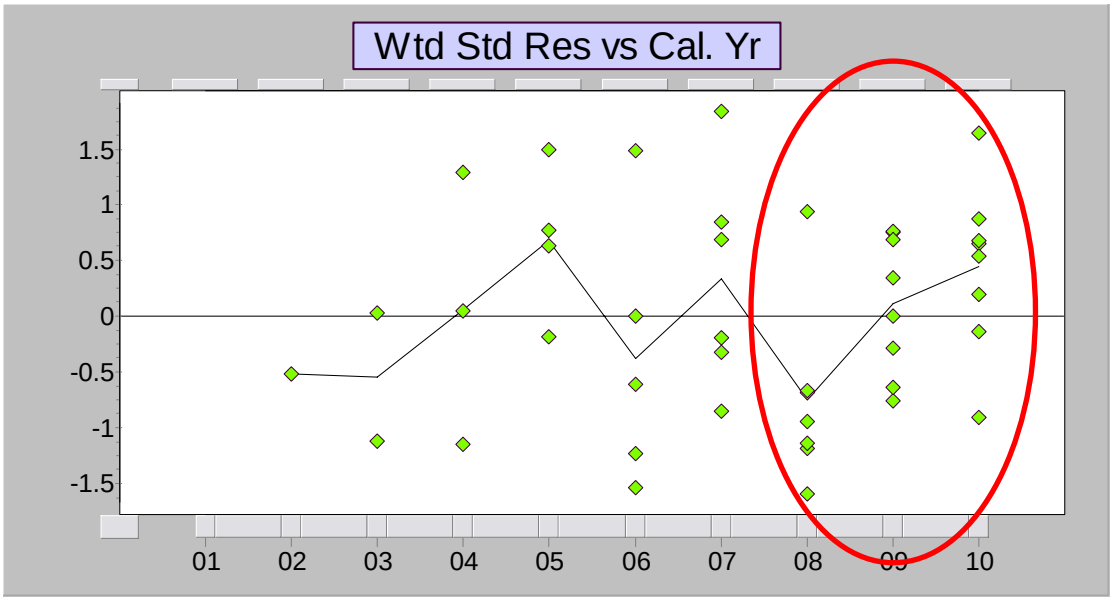
Here are the bootstrap predictive means and s.d.s for the last diagonal (i.e., with that data not used in the estimation) for a quasi-Poisson GLM, and the actual payments for comparison:

DY:	1	2	3	4	5	6	7	8
CL pred	931994	1000686	1115232	482991	325851	443060	231680	309629
mean:	958887	1021227	1114169	490137	328453	452636	242346	327365
stdev:	452285	331706	318026	195225	156289	200080	152170	227644
actual:	986608	1443370	1063269	705960	470639	206286	280405	425046

Firstly, there is an apparent bias in the bootstrap means. The chain ladder predictions sit below the bootstrap means, indicating a bias. Since, the ML for a Poisson is unbiased, if the model is correct, these predictions should be unbiased. This doesn't *necessarily* indicate a bad predictive model, but is there anything going on?

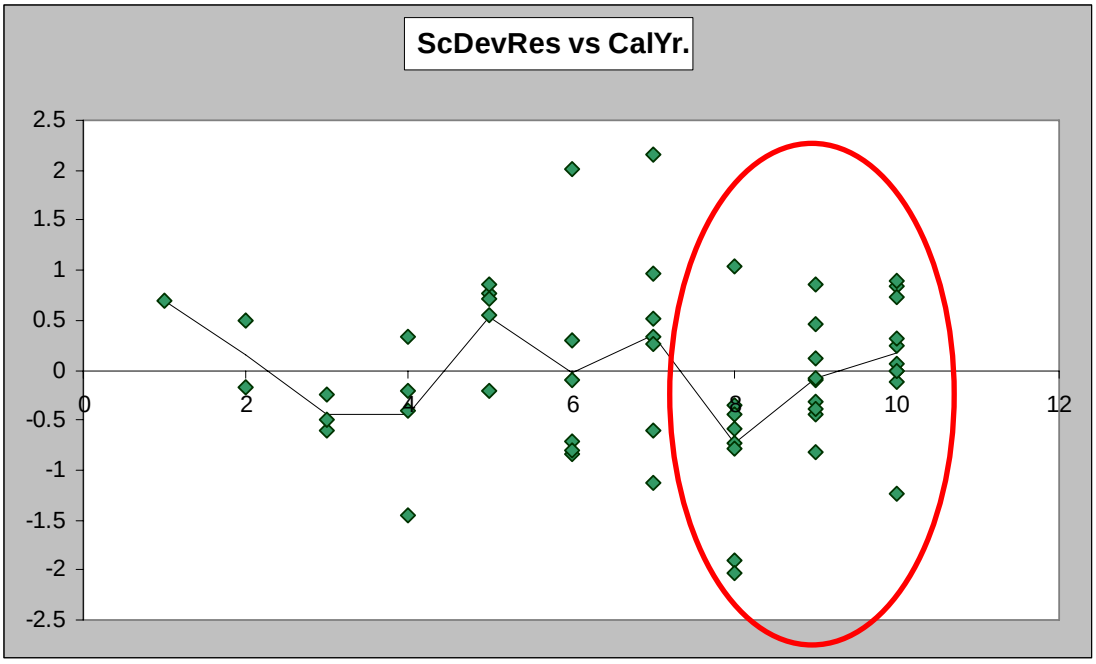
In fact there is, and we can see problems in residual plots.

Here is a plot of the residuals vs. calendar year from a Mack-type fit:



Strong calendar-period effects are evident in the last few years. The existence of a calendar-period effect was already noted by Taylor and Ashe in 1983 (who included the late calendar-year effect in some of their models), but it has been ignored by almost every author to consider this data since. If the trend were to continue for next year, the forecasts may be quite wrong. If we didn't examine the residuals, we may not even be aware that this problem is present.

Exactly the same effect appears when fitting a quasi-Poisson two-way cross classification with log-link:



There's a benefit in examining Mack residuals before fitting a quasi-Poisson GLM—the residuals are a little easier to produce, and the plot of residuals vs. fitted has more information about the predictive ability of the model.

Some other considerations

All chain ladder-reproducing models (including both the quasi-Poisson GLM and the Mack model) must assume that the variance of the losses is proportional to the mean (or they will necessarily fail to reproduce the chain ladder). This assumption is found to be rarely tenable in practice—as we saw in example 2—and for an obvious reason. While it can make sense with claim counts—for example, when the counts are higher on average they also tend to be more spread but with lower coefficient of variation. If the counts happen to be Poisson-distributed, the variance will be proportional to the mean (in fact equal to it). Heterogeneity or various forms of dependence in claim probabilities can make the Poisson untenable even for claim numbers. But with claim payments, the amount paid on each claim is itself a random variable, not a constant, and variable claim payments will make the variation increase faster than the mean. Simple variation in claim size (such as a constant percentage change, whether due to inflation effects or change in mix of business or any number of other effects) will make the variance increase as the square of the mean, while claim size effects that vary from policy to policy can make it increase still faster. Dependence in claim size effects across policies can make it increase faster again. Consequently the chain ladder assumption of variance proportional to mean must be viewed with a great deal of caution and carefully checked.

The chain ladder model is overparameterized. It assumes, for example, that there is *no* information in nearby development periods about the level of payments in a given development, yet the development generally follows a fairly smooth trend—indicating that there is information there, and that the trend could be described with few parameters. This overparameterization leads to unstable forecasts.

Finally, in respect of the bootstrap, the sample statistic may in some circumstances be very inefficient as an estimator of the corresponding population quantities. It would be prudent to check that it makes sense to use the estimator you have in mind for distributions that would plausibly describe the data.

SECTION 6: DEALING WITH OBSERVED MODEL INADEQUACY

As we have seen, the chain ladder models considered so far don't predict well with the data triangles we looked at—even though two of them are "standards" for illustrating ratio models.

Calendar-period trend changes

Pattern in the plot of residuals or prediction residuals vs. calendar period indicates calendar-period trend changes.

As we have seen, calendar-period trend changes do show up in real data. Further, because substantive changes don't generally occur frequently (such as every year or two), but more occasionally, changed rates may sometimes be expected to continue for some period (though it depends on the cause of the change—for example, with the triangle ABC, the cause of the calendar-period trend change was a known change in legislation, for which the higher identified rate was not expected to continue; in that case the projections of a ratio method will be too high after the rate drops back. It has sometimes been stated that ratio methods project at an average of past calendar-period rates, but in fact it is not the case that there is a single rate at which future observations are being inflated. Even if it were true, whether the new rate is to continue or discontinue, an average rate would be unsuitable.

Changing inflation can be modeled properly with calendar-year parameters. However, we must beware—the loglinear quasi-Poisson GLM cannot be readily modified in this way. That is, while it is possible to add calendar-year parameters to the GLM (it's no longer chain ladder, of course), the new model is demonstrably unsuitable for inflated payment data. Imagine a triangle with no inflation that otherwise meets the assumptions of the loglinear quasi-Poisson GLM. Now construct a new triangle from the old that has constant nonzero inflation in the last four calendar periods—say

running at 10% per period. Note that $\text{Var}(k.X) = k^2.\text{Var}(X)$. The variance of the inflated observations increases as the square of the factor by which they are inflated. But the quasi-Poisson model requires that the variance increase proportional to the mean, so the model requires $\text{Var}(k.X) = k.\text{Var}(X)$. Since there will be a different factor ("k") for each calendar period in the inflating region, this model *cannot* be consistent with the data.

Trend in the one-step-ahead prediction residuals vs. predicted plot

When there are no changing calendar-period trends, trend in the plot of one-step-ahead prediction residuals vs. predicted values often occurs, and indicates a model inadequacy. In the case of the loglinear quasi-Poisson GLM, this pattern does not appear in the equivalent "within-data" plot—the plot of residuals vs. fitted values.

In the case of the Mack model, if it is present in the plot of prediction residuals, it will generally also be seen in the corresponding residuals vs. fitted plot. In the case of the Mack model, it implies the need for an intercept term (see Barnett and Zehnwirth, 2000, or Murphy, 1994). There does not appear to be a simple modification of the quasi-Poisson GLM that is able to deal with this form of model inadequacy. And ordinary residual plots don't reveal its presence.

Because of the frequent presence of superimposed inflation (claims inflation at different rates to economic inflation), it is necessary to model incremental values. We believe that a log-transform is frequently beneficial both from the point of view of linearizing inflation effects, linearizing trend in the late developments (reducing the number of parameters required), and for stabilizing the variance in terms of the mean.

The Probabilistic Trend Family of models

When models better describe the characteristics of the data, the prediction intervals tend to have the required properties (such as including near to the anticipated proportion of future observations).

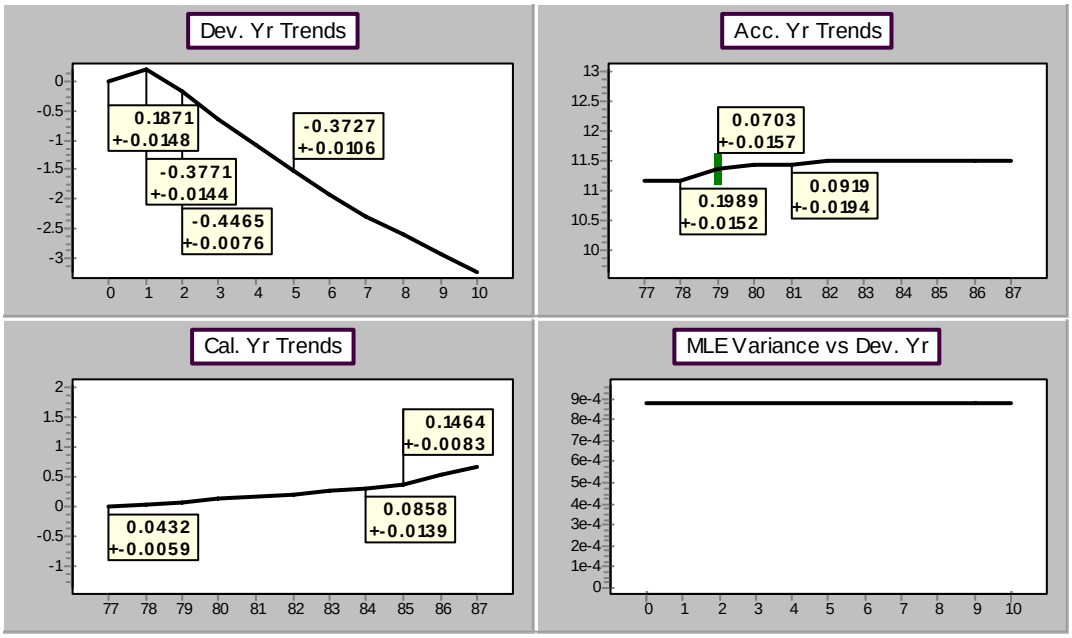
The Probabilistic Trend Family (PTF) models consist of a model for the mean trends in the three directions of the triangle, and a model for the random variability about the trends. It is applied to log-incrementals, adjusted for exposures and economic inflation (where these are available). Many triangles are well described by a few parameters in the early developments (to capture the "run-off"), and where there are trend changes against the accident periods or calendar periods (indeed, in many cases the timing of these may be known in advance), parameters in those directions as well. Often the variability is constant on the log scale, though sometimes it exhibits a variance change against the development periods, requiring some modeling of the variance. The distribution of the inflation-

adjusted and exposure normalized data is assumed to be independent lognormal—which implies normality on the log-scale. This assumption should be checked, but is in practice almost always a good description of the data. The observations are assumed to be independent.

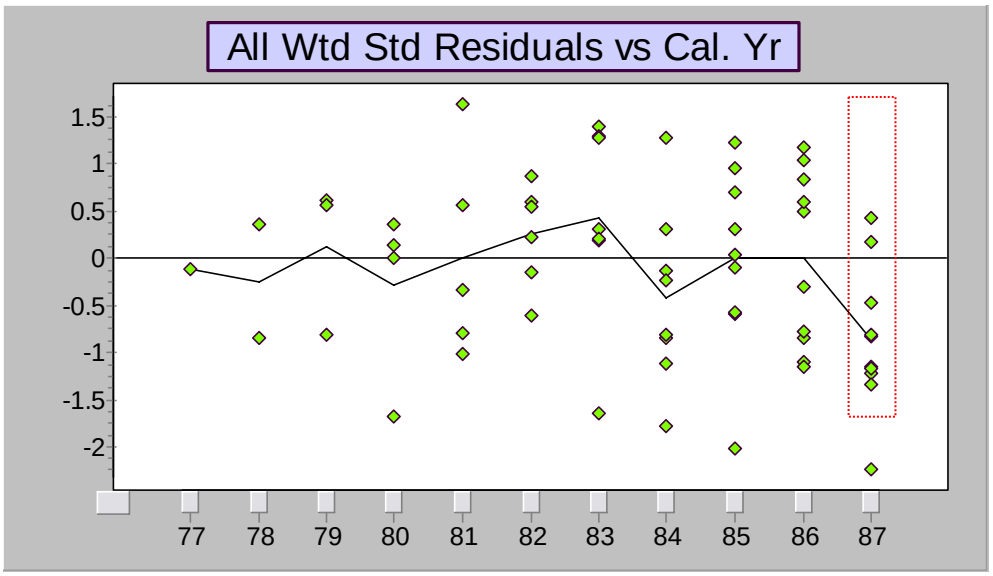
Accident-period parameters represent "levels," while development and calendar-period parameters describe linear trends (in the logs). The Probabilistic Trend Family is described in more detail in Barnett and Zehnwirth (2000).

Because of the simple form of these models, they may be represented pictorially by a decomposition of the model for the mean into trend changes in each direction and the model for the variability about it.

For example, the figure below shows a display of a reasonable model for the ABC data of Example 1. (This model includes treating the observation at 1982 delay five as an outlier and giving it zero weight in the estimation.)

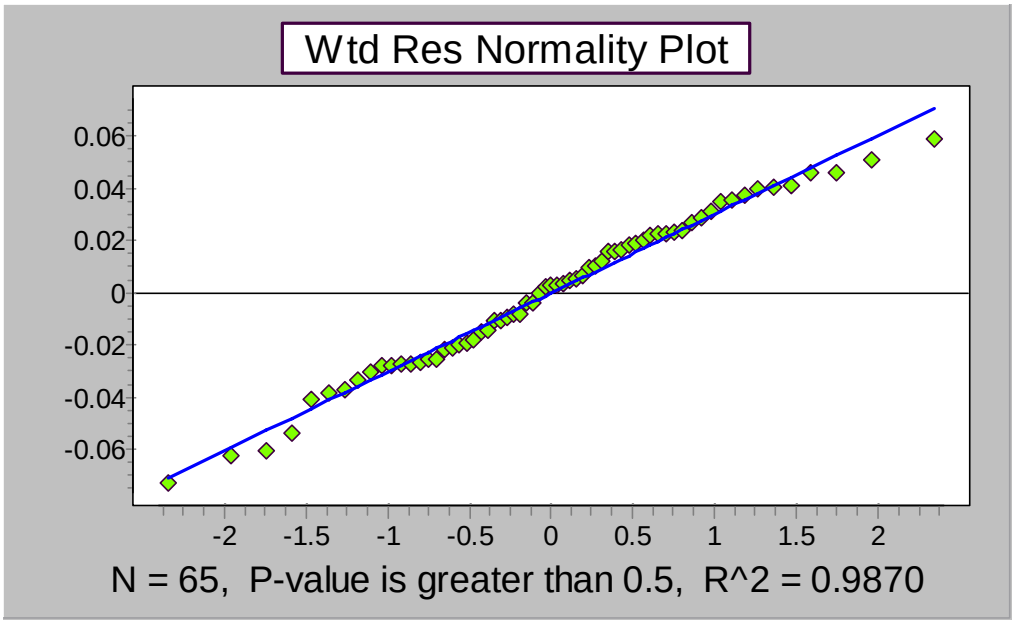


Examination of diagnostic plots indicate that the model is a reasonable description of the data. For example, the next figure shows residuals, with prediction residuals for the final diagonal, against calendar year. The final calendar period is reasonably well predicted by the model, even though those observations are not included in the estimation.



Standardized residuals and one step ahead prediction errors with approximate 90% confidence interval for the predictions.

The final plot clearly shows that the lognormal assumption seems reasonable.



Residuals against expected normal scores for all years

Multivariate PTF

Where related triangles are being analyzed, such as different subsets of a line of business (excess of loss layers, different territories, or different claim types), different lines within a single company,

or related lines across several business units, it is essential to be able to model the related triangles together. It may be that there are related trend changes across triangles and the errors about the models may be correlated. The PTF models may be extended by incorporating correlated error terms and the possibility of related parameters or changes in parameters. When the models are the same (in terms of where trend changes are located), these form generalized least squares (GLS) models. Where parameters are unrelated, these are seemingly unrelated regression (SUR) models.

These models are especially useful for calculation of diversification effects, for example, in risk capital calculations or for reserving.

By providing an adequate description of the ABC data, the identified model from the Probabilistic Trend Family is able to predict the final year (in the sense that the observations are reasonably consistent with the predictive distribution, as indicated by the validation residuals). As long as the model used to predict the future (which will be informed by the model for the past) is valid, confidence intervals and predictive intervals should have close to the right coverage probabilities, making them suitable inputs to the determination of the relevant ranges.

SECTION 7: CONCLUSIONS

Prediction intervals are important components of risk capital calculations, but such intervals rely on the model's predictive assumptions. We frequently find that for commonly used models, those predictive assumptions are violated, and we find that the models often don't predict the most recent data well.

Consequently, when fitting a quasi-Poisson GLM, it's important to check the one-step-ahead prediction errors in order to see how it performs as a predictive model—the residuals against fitted values don't show you the problems. Alternatively, the Mack residuals can be useful approximate diagnostic tools for the *predictive* assumptions of the quasi-Poisson GLM.

Predictive diagnostics should also be looked at before bootstrapping a model and once a bootstrap has been done, you should also validate at least the last year—that is, examine whether the actual values from the last calendar year could plausibly have come from the bootstrap predictive distribution standing a year earlier.

The use of the bootstrap does not remove the need to check assumptions relating to the appropriateness of the model. Indeed, it is clear that there's a critical need to check the assumptions. Use of the bootstrap does not avoid the fact that chain-ladder type models have no simple descriptors of features in the data. We show in several examples that there is much remaining

structure in the residuals.

If it is the predictive behavior that is of interest, prediction errors are appropriate tools to use in standard diagnostics, and they can be analyzed in the same way as residuals are for models where prediction is within the range of the data.

Checking the model when utilizing the bootstrap technique is achieved in much the same way as it is for any other model—via diagnostics—but they should be predictive diagnostics selected with a clear understanding of the problem, the model and the way in which the bootstrap works.

REFERENCES

[1] Ashe, F., "An essay at measuring the variance of estimates of outstanding claim payments," *ASTIN Bulletin* 1986, 16S:99–113.

[2] Barnett, G. and B. Zehnwirth, "Best Estimates for Reserves," *Proceedings of the Casualty Actuarial Society* 2000*,* 87(167):245-321.

[3] England P.D. and R.J. Verrall, "Stochastic Claims Reserving in General Insurance," *British Actuarial Journal* 2002, 8(3):443-518.

[4] Barnett, G., B. Zehnwirth, and E. Dubossarsky, "When can accident years be regarded as development years?" *Proceedings of the Casualty Actuarial Society* 2005, 92(177):239-256, www.casact.org/pubs/proceed/proceed05/05249.pdf

[5] Chatfield, C., *Time-Series Forecasting* (Boca Raton, Florida: Chapman and Hall/CRC Press, 2000).

[6] Efron, B. "Bootstrap methods: another look at the jackknife," *Annals of Statistics* 1979, 7**:**1-26.

[7] Efron, B. *The Jacknife, the Bootstrap and Other Resampling Plans,* vol. 38*,* (Philadelphia, Pa., Society for Industrial and Applied Mathematics, 1982).

[8] Efron, B. and R. Tibshirani, *An Introduction to the Bootstrap* (New York: Chapman and Hall,. 1994).

[9] England, P.D. and R. J. Verrall, "Analytic and bootstrap estimates of prediction errors in claims reserving," *Insurance: Mathematics and Economics* 1999, 25(3):281-293.

[10] Hachemeister, C. and J. Stanard, "IBNR Claims Count Estimation with Static Lag Functions," Presented at the 1975 Spring Meeting of the Casualty Actuarial Society.

[11] Mack, T. "Measuring the variability of chain ladder reserve estimates," *Casualty Actuarial Society Forum*, Spring 1994, vol. 1:101-182.

[12] Mack, T., "Which stochastic model is underlying the chain ladder method?" *Insurance Mathematics and Economics* 1994, 15(2-3):133-138.

[13] Moore, D. S., G.P. McCabe, W.M. Duckworth, and S.L. Sclove, "Bootstrap Methods and Permutation Tests," companion chapter 18 in *The Practice of Business Statistics*, (New York: W. H. Freeman, 2003), http://bcs.whfreeman.com/pbs/cat_140/chap18.pdf.

[14] Murphy, D., "Unbiased Loss Development Factors," *Proceedings of the Casualty Actuarial Society* 1994, 81(154):154–222.

[15] Pinheiro, P.J.R., J. M. Andrade e Silva, and M. L. Centeno, "Bootstrap Methodology in Claim Reserving," *Journal of Risk and Insurance* 2003, 70(4):701-714

[16] Reinsurance Association of America, *Historical Loss Development Study*, (Washington: RAA, 1991).

[17] Renshaw, A.E. and R. J. Verrall, "A stochastic model underlying the chain ladder technique," *Proceedings of the XXV ASTIN Colloquium* 1994*,* Cannes, France.

[18] Stanard, J.N. "A Simulation Test of Prediction Errors of Loss Reserve Estimation Techniques," *Proceedings of the Casualty Actuarial Society* 1985, 72(137):124-148.

[19] Taylor, G.C. and F.R. Ashe, "Second Moments of Estimates of Outstanding Claims," *Journal of Econometrics* 1983, 23(1)37-61.