Loss Reserving Using Claim-Level Data

James Guszcza, FCAS, MAAA
Jan Lommele, FCAS, MAAA

Abstract

While the actuarial literature devoted to stochastic loss reserving has been developing at an impressive rate, much of this literature has been devoted to the statistical analysis of summarized loss triangles. This restriction limits the benefits that modern statistical techniques can bring to the subject of loss reserving. This paper will sketch one possible framework for estimating future claims payments using claim-level data. The first part of the paper will discuss the use of covariates (or "predictive variables") to improve one's estimates of future payments, especially in cases where the mix of business being analyzed has changed over time. The second part of the paper will describe how the bootstrapping technique can be applied to claim-level data to estimate reserve variability.

Keywords. Reserve Variability; Future Payment Variability; Generalized Linear Model; Over-Dispersed Poisson Model; Bootstrap; Claim-Level Data; Covariates; Predictive Variables; Changing Mix of Business; Chain-Ladder

INTRODUCTION

The recent actuarial literature has enjoyed a growing discussion of statistical methods for performing loss reserve analyses. This discussion has increased the statistical rigor of the subject, and has expanded the set of tools available for estimating reserve variability.

However, much of this recent discussion has been devoted to the statistical analysis of summarized loss triangles. We feel that this limits the potential improvements that predictive modeling can bring to the subject. We will focus on two reasons here. First, summarized loss triangles do not allow the analyst to incorporate predictive variables in his or her reserve analysis. Second, using summarized data limits the accuracy with which an analyst can estimate the variability of his or her loss reserve estimates. It is reasonable to expect that by not "summarizing away" the size-of-loss and loss development information implicit in (un-summarized) claim-level data, potentially better point and variability estimates can result.

Many of the comments in the Discussion of England and Verrall's recent survey paper on stochastic loss reserving [4] expressed this sentiment. Shah's comment is representative:

> The triangulation data that these [Generalized Linear Modeling] techniques have been applied to are just a consequence of history. They come from an era when computing power was expensive. Therefore, I question the value of actually applying such techniques to such limited data. Such sophisticated techniques may be more useful if applied to the underlying claims data, as has been alluded to by several speakers. In view of this, there is a danger that the

> results may be viewed as more scientific than they really are, and may be given more credibility than is truly justified for them.

Tripp's comment also seems to us to be on the mark:

> Why do we throw away information? ... Looking at the life side of our profession, you realise that work like this takes place at policy level detail. If you look within the general insurance part of the actuarial profession, there is a body of thinking that has grown up around premium rating and a body of thinking that has grown up around reserving. Are we getting 'over-siloed'? Could aspects of the methodology and the thinking that has gone into using GLMs for premium rating be brought more into play when it comes to reserving, where, at present, we tend to use aggregated claims data? I wonder whether we are missing out on using information that is available from exposure descriptions and from the circumstances of individual claims.

Motivated by the concerns expressed in these quotes, this paper is an attempt to develop the idea that using un-summarized data will allow one to unleash the full power of modern predictive modeling techniques on the problem of estimating future claim payments. The goals of improving one's reserve point estimates as well as variability estimates will be discussed sequentially in the two parts of this paper.

In Part I we review the well known shortcoming of traditional reserving methods when applied to books of business that have changed over time. A danger of using summarized loss triangles is that they can mask heterogeneous loss development patterns. They also prohibit the use of predictive variables that might be correlated with loss development. We sketch a reserving technique – inspired by the chain-ladder method – that operates on claim-level data. Using simulated data we illustrate how this technique can reflect heterogeneous loss development patterns that the chain ladder misses, resulting in an improved estimate.

We believe that the potential for improved estimates of future loss payments is sufficient motivation to consider the use of claim-level data for reserving. Doing so obviously requires additional effort (not to mention specialist software that goes beyond spreadsheets). But, as Part II of this paper will discuss, it brings a significant side benefit as well.

Namely: once we have claim-level data available for analysis, we can employ the bootstrapping technique (a type of simulation that involves repeatedly sampling with replacement from one's data) to easily compute confidence intervals around our estimates of outstanding losses. Indeed bootstrapping will give us estimates of the entire distribution of our outstanding loss estimator, no matter how complex.

Bootstrapping has been discussed in the recent literature as a promising avenue for estimating reserve variability. But because of the summarized loss triangles that serve as a

starting point for most current discussions of reserving, the resampling step of bootstrapping is typically applied to the residuals of various models fit to loss triangles. The idea pursued here is to resample the underlying data points, and then apply one's chosen reserving technique to each of the resulting pseudo-datasets. This is a flexible and perhaps conceptually simpler method of bootstrapping. Also, because its resampling step occurs prior to the building of any model, the pseudo-datasets that it employs are not in themselves dependent on the correctness of the model being fit to the data.

PART I: SUMMARIZED DATA AND THE PROBLEM OF A CHANGING MIX OF BUSINESS

A common criticism of traditional loss reserving techniques is that they can be slow to incorporate changes in the company's mix of business into their estimates of outstanding losses. This is the point of the actuarial road trip joke involving the salesperson with his foot on the gas, the underwriter with his foot on the brake, and the actuary navigating by looking out the rear window.

Bornhuetter and Ferguson state the problem well in "The Actuary and IBNR" [1]:

> The product mix can be an important factor, not so much because two somewhat dissimilar items are combined, but because they may have different rates of growth. For example, a company may have personal and commercial automobile loss development experience combined over the years although, if it were looked at separately, commercial business would require higher loss development factors. As long as the relative exposure between the two categories remains constant there is no problem; however, picture the situation if personal automobile increased at a 5% annual rate while commercial automobile, although relatively small, is growing at a 25% annual rate.

The obvious thing to do in such a situation would be to analyze commercial and personal auto reserves separately. That is, divide the data into two separate loss triangles and proceed as usual. This is helpful as far as it goes, but the approach has its limits. Bornhuetter and Ferguson continue:

> Of course, the volume of data is an important factor in determining what kinds of breakdowns of the data are feasible. If the data are subdivided so finely that most groups have only a small volume of data, the subdivisions may accomplish nothing useful. Or to quote Mr. Longley- Cook's delightful analogy, "We may liken our statistics to a large crumbly loaf cake, which we

> may cut in slices to obtain easily edible helpings. The method of slicing may be chosen in different ways-across the cake, lengthwise, down the cake, or even in horizontal slices, but only one method of slicing may be used at a time. If we try to slice the cake more than one way at a time, we shall be left with a useless collection of crumbs."

For example, it might be nice to set up separate reserve analyses by both coverage and region. But even adding the single additional dimension of "region" might significantly diminish the credibility of the data and thereby threaten the integrity of one's outstanding loss estimate. The goal of the first part of this paper is to suggest a way beyond this impasse.

Our discussion of changing mixes of business is intended only to motivate the method discussed below. Hopefully the method's usefulness is not restricted to this scenario. For example, it might also be useful when, for example, a company moves into a new region or two companies merge.

ENTER PREDICTIVE MODELING

In modern terms, Longley-Cook's image of the crumbly cake is an illustration of the bias-variance tradeoff in predictive modeling. Stated briefly, a complex model (or multiple models fit on sub-segments of the data) will make predictions that are less biased, but at the same time less certain – i.e., more variable – than a simpler model. The tradeoff is that our model should have sufficient complexity to reflect true statistical regularities in the data (thereby reducing bias), yet not have so much complexity that random patterns in the data overwhelm the model and lead to unreliable results (high variance). This is perhaps a special case of Einstein's dictum, "Everything should be made as simple as possible, but not simpler."

An analogy with ratemaking might be helpful. Consider a simple rating plan with the following rating factors:

- Age {<26, 26-50, >50 years}
- Credit {bad, average, good}
- Claim in past 3 years {yes, no}

This rating plan has 3·3·2=18 cells. The most naïve – and over-parameterized – way to proceed would be to simply estimate the loss ratio relativity of each of these cells and base one's rating factors on these parameters. Note that this is equivalent to fitting a regression model with 17 indicator variables. But as Longley-Cook warns, the data in each of these cells

is unlikely to have sufficient credibility to produce stable results. Therefore the variance around the resulting rating factor estimates will be large.

For this reason, the modern approach to ratemaking is to employ Generalized Linear Models [GLMs]. Rather than estimate 3·3·2-1=17 parameters, a GLM model in this scenario would estimate 2+2+1=5 parameters. Extending Longley-Cook's analogy, we now get to have our cake and make multivariate estimates with it too. Rather than estimate each of the 17 rating factors each with its own "crumb" of data, we use the loaf to estimate a more modest 5 parameters.

There are three major advantages of deriving one's rating factors from the parameters of a multivariate model, rather than estimating them directly from small "crumbs" of data:

- The resulting rating factors will have less variability (less parameter risk).
- A larger number of rating factors can be used without running into Longley-Cook's "crumbly cake" problem.
- Factors such as Age and Credit can be treated as continuous predictive variables, rather than being arbitrarily divided into discrete bins.

Returning to loss reserving, it is good and accepted practice to perform separate reserve analyses by line of business and by such important subdivisions as Workers Comp Medical vs. Indemnity claims. As we have discussed, this can only be taken so far. But what if (a) claim development patterns vary by a multitude of factors such as Report Lag, Credit Score, Prior Claim, Policy Age... and (b) the mix of business measured by these factors has changed over time? As Bornhuetter and Ferguson point out, it is essential to reflect this shifting mix of business in one's analysis. But as Longley-Cook points out, dividing the data by many of these dimensions will quickly lead to serious credibility problems.

In the light of the ratemaking analogy above, it is perhaps natural to suggest that the way forward is to somehow incorporate a multivariate predictive model into one's reserve analysis. We will sketch one such model below. This model is offered very much in the spirit of taking a first step. We expect that it could be improved or replaced with a better one. Nevertheless, we hope that sketching a sample multivariate loss reserving model that admits covariates will spark further thoughts on the subject.

THE LEVEL OF DATA NEEDED

Multivariate loss reserving requires that one analyze disaggregated data, at the policy or claim level, rather than summarized loss development triangles. The reason for this is clear: predictive variables such as Age, Credit, and Prior Claim pertain to the policy that made the claim. To incorporate policy-level variables such as these, policy-level data must be used in the analysis. There is no way to "attach" such covariates to summarized data. Similarly, if we wish to incorporate variables such as Report Lag or Injury Type into the analysis, claim-level data must be used. Traditional loss triangles do not allow one to use this potentially useful predictive information.

To summarize what has been said so far:

- The traditional approach of separating one's data and performing separate analyses on the resulting loss triangles is an incomplete answer to the problem of a shifting mix of business.
- A plausible approach to this problem is to incorporate covariates into one's reserving technique – that is, build a multivariate reserving model.
- Doing so requires that we use data at the policy or claim (or indeed claimant) level.

For the remainder of this paper phrases such as "reserving using claim-level data" will serve as shorthand for "reserving using policy- or claim- or claimant-level data".

MODEL DESIGN

In this section we propose a claim-level generalization of the simple chain ladder reserving method. As stated above, this is merely one of many possible starting points. For all of its faults, the chain ladder has the virtues of being simple and familiar. Generalizing the chain ladder therefore gives us an intuitive way of illustrating the benefits of using claim-level data to estimate future claim payments.

As discussed above, we assume we have data at the policy, claim, or claimant level. Of course, the finer the level of summarization of one's data, the broader the array of predictive variables one can include in one's model. Deciding on the level of data is a practical decision that does not substantially affect the discussion below. Let us therefore assume that our data is at the claim level.

We therefore assume that we have a database with one record per claim, and multiple variables on each record. These variables can be categorized into three types:

- Predictive variables (Credit Score, Injury Type, Policy Age...)
- Target variables (Loss at 24 months, Loss at 36 months...)
- Informational variables (Accident Year, Zip Code, Agent Number...)

The "informational" variables can sometimes be used to derive further predictive variables (e.g., by using zip code to match such demographic variables as Population Density onto the records). Other times, they are used simply for analytic purposes (e.g., displaying total losses by accident year).

Let us establish some notation. We attempt to be consistent with the notation of England and Verrall. Let C_j denote cumulative losses evaluated as of j months. For example C_{24} denotes the losses (associated with a particular claim) evaluated as of 24 months. $\{C_j\}$ will serve as the target variables in our model design.

Let $\{X_1, X_2, \ldots, X_N\}$ represent the predictive variables. Each value of each predictive variable X_i will appear on each claim-level record. We also assume that the values of each of the predictive variables are measured either at policy inception, or at the claim report date (whichever is appropriate).

Let U_k denote the total ultimate losses for accident year k, summed across all policies: $U_k=\sum C_\alpha$. Let R_k denote the outstanding losses (ultimate losses minus losses paid to date) for accident year k. Let U and R denote the sums of U_k and R_k respectively across all accident years. The goal of loss reserving is to calculate an estimate r of R as well as an estimate of variability of, or confidence interval around, R. R is often referred to as a "reserve estimate", but to distinguish it from the quantity that is actually booked in the financial statements, it is probably better to call it the "total outstanding losses" or "total future payments" (see [2]). In the remainder of this paper, the three terms will be used synonymously.

In predictive modeling it is typically the case that we are presented with a single target variable Y (such as pure premium or claim frequency or size of loss) and multiple predictive variables $\{X_1, X_2, \ldots, X_N\}$. We might fit a GLM model of the form:

$$g(\mu) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N \equiv \boldsymbol{\beta} \cdot \boldsymbol{X}$$

Where μ denotes E[Y], the expected value of the target variable Y; and g(·) is the link function.

Here, the situation is not so simple. For one thing, we are presented with multiple target variables $\{C_1, C_2, \ldots, C_J\}$ rather than a single target Y. In addition, this (single) target variable is typically the quantity we are ultimately interested in predicting. Here, we are interested in predicting either losses at ultimate or losses as of a certain development period, such as 10 or 20 years. Let us assume that for practical purposes, C_J represents losses at ultimate. That is, $C_J \approx C_\infty$. (That is, let us assume that no tail factors are needed for our analysis.) Then C_J is what we are ultimately interested in predicting; and $\{C_1, C_2, \ldots, C_{J-1}\}$ are intermediate quantities used as stepping stones to estimate C_J.

The reason for this complexity is that C_J is missing on most of the claim-level records in our dataset. Using it as "the" target value analogous to Y in the GLM example above would require us to throw away data points for which Y is unknown. Let us frame our discussion in terms of an example. Suppose we have claim-level records for accident years 1990, 1991,..., 1999. On the 1990 records, we have losses evaluated as of 12, 24, ..., 120 months. On the 1991 records, we have losses evaluated as of 12, 24, ..., 108 months; while losses as of 120 months are unknown ("missing"). On the 1999 records, we have only losses evaluated as of 12 months; $\{C_{24}, C_{36}, \ldots, C_{120}\}$ are all missing.

Of course we have the option of using only the AY 1990 claim records to build a single GLM model; and use this model to predict the ultimate values of the 1991-1999 claims. But in doing so we would throw away the loss development pattern information that traditional reserving methods rely on. This is not a satisfactory option.

Many approaches are possible at this point, but we choose to build – continuing with the same example – 9 successive GLM models, "layered" one on top of the other. Speaking figuratively, we "regress" C_{24} on C_{12}; C_{36} on C_{24}; and so on. Each of these 9 GLM models is analogous to the 9 link ratios in the corresponding chain ladder model that could be run on the summarized 10-by-10 loss triangle. Let us denote these 9 models $M_{24}, M_{36}, \ldots, M_{120}$. The M_{36} model will take as an input either losses evaluated at 24 months (for AY 1990-98); or the predicted value of the M_{24} model (for AY 1999). This is analogous to the way a link ratio is applied in a chain-ladder analysis. Of course in addition to C_{J-1}, the M_J model takes as inputs all of the predictive variables $\{X_1, X_2, \ldots, X_N\}$.

Let us make this abstract discussion more concrete. The motivation for introducing predictive variables is to capture differences in different claims' expected loss development patterns. Given that our basic idea is to "incrementally" model these (potentially heterogeneous) development patterns à la the chain ladder, it makes sense to model each claim's development from period j-1 to period j as a function of several covariates:

$$\frac{C_j}{C_{j-1}} = \varphi(X_1, X_2, ..., X_N)$$

For mathematical convenience, we will further assume that this claim-level "link ratio" is in fact a (pre-specified) monotonic function f of a linear combination of the covariates:

$$\frac{C_j}{C_{j-1}} = f(\alpha + \beta_1 X_1 + \beta_N X_2 + ... + \beta_N X_N)$$

This is of course the familiar linear modeling trick: we reduce the job of estimating the function φ to estimating the parameters $\{\alpha, \beta_1, \beta_2, \ldots, \beta_N\}$. The monotonic function $f(\cdot)$ might, for example, be the natural exponent function $\exp(\cdot)$ or the identity function $\mathrm{id}(\cdot)$. The use of linear models (as opposed to, say, generalized additive models or neural networks) is not essential to the basic idea sketched here. But it is fairly flexible and powerful approach that avoids unnecessary complexity.

The above equation implies that the expected development from period j-1 to j of any given claim is a generalized linear function of the covariates $\{X_1, X_2, \ldots, X_N\}$. We do not need to assume that each claim at period j-1 will have the same expected development to period j. Nor do we need to assume that the mix of these (inhomogeneous) claims will stay the same from one accident year to the next.

Suppose, on the contrary, that we did assume perfect claim homogeneity in the sense that all claims have the same expected development. This is tantamount to assuming no variance in claim-level link ratios; and this in turn implies that no covariate X_i could possibly play a statistically significant role in predicting link ratio. Therefore the above equation reduces to a constant:

$$\frac{C_j}{C_{j-1}} = f(\alpha) = Link\_Ratio$$

Thus the chain ladder's link ratio is equivalent to our generalized linear model form with no covariates.

A few more assumptions will let us use the machinery of Generalized Linear Models to estimate the parameters $\{\alpha, \beta_1, \beta_2, \ldots, \beta_N\}$. Let us assume that the function f is the exponential function. This is equivalent to assuming the log link function from GLM theory. Let us further assume that the variance of C_{j+1} is proportional to its mean. (This assumption is not essential to the general technique we're tying to develop. This familiar assumption is being made for convenience, and could be altered without substantially affecting the discussion to follow.) In other words, we are assuming the over-dispersed Poisson GLM model form:

$$\log\left(E\left[\frac{C_j}{C_{j-1}}\right]\right) = \alpha + \beta_1 X_1 + \beta_N X_2 + \ldots + \beta_N X_N$$

Equivalently,

$$E\left[\frac{C_j}{C_{j-1}}\right] = \exp\{\alpha + \beta_1 X_1 + \beta_N X_2 + \ldots + \beta_N X_N\}$$

Or,

$$\frac{C_j}{C_{j-1}} = \exp\{\alpha + \beta_1 X_1 + \beta_N X_2 + \ldots + \beta_N X_N\} + \delta$$

where δ is an overdispersed Poisson-distributed error term. Given the quantities $\{C_{j-1}, C_j, X_1, X_2, \ldots, X_N\}$, we can estimate the parameters $\{\alpha, \beta_1, \beta_2, \ldots, \beta_N\}$ of model M_j using any standard GLM package. To be explicit, we would make the following specifications:

- Target: (C_j / C_{j-1})
- Covariates: $\{X_1, X_2, \ldots, X_N\}$
- Weight: C_{j-1}
- Distribution: Poisson
- Link: Log

Recasting the above equation as follows will allow us an alternate way of conceptualizing the above model form. Let us multiply both sides of the equation by C_{j-1}:

$$C_j = C_{j-1} \cdot \exp\{\alpha + \beta_1 X_1 + \beta_N X_2 + \ldots + \beta_N X_N\} + \varepsilon$$

which is equivalent to:

$$C_j = \exp\{\log(C_{j-1}) + \alpha + \beta_1 X_1 + \beta_N X_2 + \ldots + \beta_N X_N\} + \varepsilon$$

This is perhaps a more useful conceptualization of our model. The target variable is C_j, there is no weight variable, and $\log(C_j\text{-}1)$ serves as the "offset term". Explicitly:

- Target: C_j
- Offset: $\log(C_{j-1})$
- Covariates: $\{X_1, X_2, \ldots, X_N\}$
- Weight: none
- Distribution: Poisson
- Link: Log

(Note that all standard GLM packages allow one to specify an offset term.) The offset term essentially functions as a regressor whose corresponding "beta" parameter is constrained to be 1. This conceptualization illustrates the chain ladder-esque idea that we are building a model that estimates the expected value of C_j as a "generalized linear link function" $\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N)$ applied to C_{j-1}, the (known or estimated) losses as of j-1.

HOW TO HANDLE IBNR

Note: this short section, and the appendix it refers to, outlines a method for extending the model design to handle IBNR claims. The authors suggest skipping it on the first reading. Indeed, this section can be skipped altogether if the reader takes the attitude that the model outlined can be used for losses on reported claims only; with IBNR claims being estimated in a separate analysis.

This model design also allows us a way of incorporating incurred but not reported (IBNR) losses into our model. For simplicity, let us assume that all claims that are unreported at 12 months are reported by 24 months. Therefore there will be records in our data with C_{12}=0 and C_{24}>0. In the M_{24} model, we add to the database one record for each in-force 1999 policy that had no claim as of 12 months from its effective date. On this record, we would force the offset term $\log(C_{12})$ to be zero. We would also include on all records an indicator variable X_0 as a covariate in M_{24} that takes on the value 1 if C_{12}=0, and 0 otherwise. Finally, we would neutralize all predictive variables that measure claim-level information.

("Neutralize" typically means that we recode missing variables to the median value.) As with all of the other AY 1999 records in the database the values of $\{C_{24}, C_{36}, \ldots, C_{120}\}$ are all missing.

Doing this will "allocate" a portion of the 12→24 IBNR (estimated from the AY 1990-1998 data) to each 1999 in-force policy that has no claim reported as of 12 months. The γ_0 parameter of the X_0 indicator functions in place of the offset term, which was forced to be zero on each of the 1999 zero-claim records. In other words, $\exp(\gamma_0)$ is the average expected 12→24 IBNR for each AY 1999 policy. The expected IBNR for an individual policy is $e^{\gamma_0}\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N) = \exp(\alpha + \gamma_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N)$. The successive models M_{36}, M_{48},... will "develop" this allocated IBNR loss along with the other losses.

An example might clarify this discussion. Suppose that the total IBNR (as of 24 months) from AY 1990-98 was $400,000 and that during this time period, there were 4000 policies without claims as of 12 months. This is an average of $100 per claim-free policy. The value of γ_0 would therefore be $\log(100) \approx 4.6$.

Note that this method of treating IBNR assumes that the covariates $\{X_1, X_2, \ldots, X_N\}$ affect the allocation and development of IBNR in the same way that they affect the development of other losses. We could refine the model by including the interactions $\{X_0 * X_1, X_0 * X_2, \ldots\}$ as further model covariates. These covariates would be non-zero only for the records corresponding to policies with no losses as of 12 months. This idea is more fully explicated in the Appendix.

SIMULATION APPROACH

We will now apply the above model to a (very rudimentary) simulated dataset. The advantage of using simulated data is twofold. First, by construction we know which covariates are truly related to the various claims' differential development over time. Because of this, we can illustrate the operation of the model without the distraction of having to convince ourselves that a set of covariates is reasonably complete or significantly correlated with the claims' differential loss development.

Second, we can simulate our data "to ultimate", and set aside the (otherwise unknown) losses at ultimate as a standard against which we can compare our model's predictions with the predictions of the traditional chain-ladder model.

Of course a major disadvantage of using simulated data is that our sample results will give little indication of the degree to which out proposed model will produce improved predictions on real-world data.

However, it is our hope that the potential of this approach will be intuitive to many readers. The authors' experience in building predictive models for ratemaking and underwriting applications suggests that it is nearly always possible to find traditional and non-traditional predictive variables that are significantly correlated with size-of-loss. Given that larger claims are known to develop more slowly, one expects that that many of these same predictive variables will be correlated with loss development patterns.

SIMULATION ASSUMPTIONS

We illustrate our model with a simulated dataset that is very simple, yet with sufficient structure to illustrate the potential advantage of this model over the traditional chain ladder.

By construction, our claim-level dataset has the following characteristics:

- **Near-homogeneity of data:** the claims in our book of business all have identical expected loss development patterns except for one characteristic: whether the policyholder that made the claim had "good" credit or "bad" credit.
- **Differential development:** The claims of bad credit policies are expected to develop more slowly than the claims of good credit policies.
- **Changing mix of business:** A greater proportion of bad credit policies have been written in recent years.

As Bornhuetter and Ferguson point out, the differential loss development of bad/good credit policies' claims would present no special problem to the traditional methods were it not for the changing mix of business. However, the greater proportion of bad credit policies written in more recent years implies that the overall development patterns will shift from year to year. In particular, the expected development pattern for the most recent accident year will not be adequately represented by an average development pattern derived from the prior accident years' claims in a loss triangle.

The simulation incorporates the idea that a measurable quantity – here, credit – is correlated with loss development. Therefore by including credit in our reserving model, we are reflecting the shifting mix of business in our analysis. Put another way, the shifting proportion of bad credit policies is a "leading indicator" of a slow-down in the book's loss development. Using credit as a covariate in our reserving model allows us to quantify this slow-down, rather than judgmentally adjust for it after a traditional reserving exercise.

We simulate 5000 data points, each representing one claim. By design there are 500 claims for each of the accident years 1990, 1991, ..., 1999. Each of the 5000 records has 10 loss fields C_{12}, C_{24}, ..., C_{120}. We will describe how the values of $\{C_{12}, C_{24}, \ldots, C_{120}\}$ are assigned to each claim.

Finally, two simplifying assumptions are made. First, we assume that there is no IBNR: all claims are reported by 12 months from the beginning of the accident year. (See the discussion above and the Appendix for a discussion estimating IBNR in the current model framework.) Second, we assume that losses are fully developed as of 120 months: for each accident year k, $U_k = \sum C_{120}$.

Next we describe our simulation of the loss fields $\{C_{12}, C_{24}, \ldots, C_{120}\}$. We draw the losses at 12 months (C_{12}) from a lognormal distribution; and then successively apply 9 randomly generated "link" factors to these losses. The means and standard deviations of the distributions used to generate the losses and link factors were selected by judgment.

In more detail, the 5000 values of C_{12} were drawn from a lognormal distribution with parameters $\mu=8$ and $\sigma=1.3$:

$$\log(C_{12}) \sim n(8, 1.3)$$

For good credit claims, the values of $\{C_{24}, \ldots, C_{120}\}$ were determined by the following algorithm:

$$C_{j+1} = C_j * (\text{link}_j^{\text{good}} * \varepsilon_j)$$

The similar algorithm for bad claims is:

$$C_{j+1} = C_j * (\text{link}_j^{\text{bad}} * \varepsilon_j)$$

where

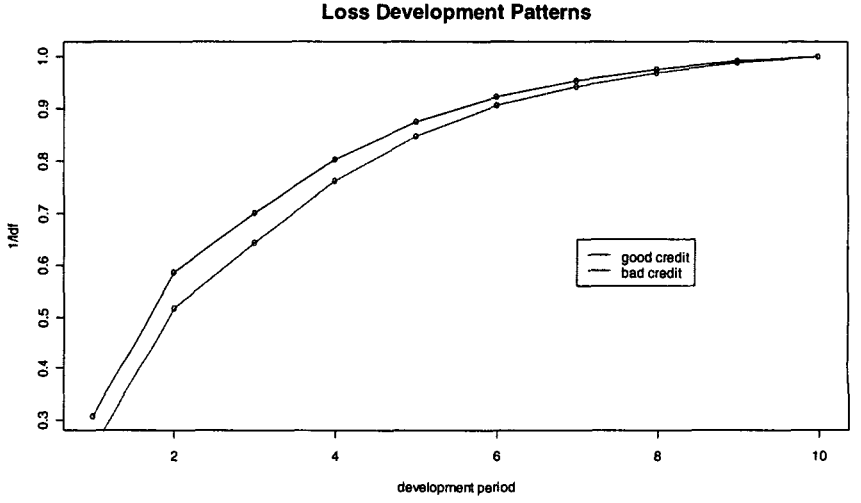
$$\text{link}^{good} = \{1.8, 1.17, 1.13, 1.08, 1.05, 1.03, 1.02, 1.015, 1.008\}$$

and $\text{link}^{bad} = (\text{link}^{good}-1)*1.25 + 1$:

$$\text{link}^{bad} = \{2, 1.2125, 1.1625, 1.1, 1.0625, 1.0375, 1.025, 1.01875, 1.01\}.$$

Finally, ε_j is a normally distributed "shock" term with mean 1 and a standard deviation that is a function of the value of the link ratio.

The development patterns (1/LDF) implied by the above expected link ratios are graphed below. This graph illustrates that by construction, bad credit claims develop more slowly than good credit claims.



In summary, each claim at each time period is assigned its own randomly generated link ratio; but the *expected* link ratios for bad/good credit claims are the ones stated above. (A word about motivation: the number of claims, size-of-loss distribution, and the general magnitude of the link ratios were judgmentally chosen to result in a summarized loss triangle similar to an actual Workers Comp loss triangle studied by one of the authors. The differing link^{bad} and link^{good} development patterns were selected purely judgmentally.)

So far, we have discussed the "homogeneity" and "differential development" assumptions. Regarding the "changing mix of business", we randomly apply the "bad" and "good" link ratios in the following proportions across the accident years:

shifting exposure base

year	%bad credit	%good credit
1990	30%	70%
1991	35%	65%
1992	40%	60%
1993	45%	55%
1994	50%	50%
1995	55%	45%
1996	60%	40%
1997	65%	35%
1998	70%	30%
1999	75%	25%

Note that the simulation approach we have laid out allows us to assign values of $\{C_{12}, C_{24}, \ldots, C_{120}\}$ to each claim, *regardless of accident year.* We will apply both our model and the traditional chain ladder to the data elements that would be available in an actual reserving exercise – namely those that form the upper half of the loss triangle. At the same time, we can use the data elements that would be unknown in an actual reserving exercise – the lower half of the triangle – as the "truth" against we can judge the success of both our method and the chain ladder.

The simulated data, summarized to the accident year level, is displayed below:

	Losses in $1000's											
	@12	@24	@36	@48	@60	@72	@84	@96	@108	@120	ultimate	o/s
1990	3,522	6,562	7,766	8,850	9,627	10,144	10,473	10,700	10,875	10,970	10,970	0
1991	3,527	6,623	7,876	9,011	9,817	10,361	10,705	10,942	11,123	11,223	11,223	99
1992	3,681	6,939	8,235	9,428	10,274	10,833	11,194	11,444	11,635	11,739	11,739	295
1993	3,780	7,152	8,539	9,791	10,666	11,262	11,642	11,902	12,100	12,210	12,210	567
1994	2,912	5,563	6,644	7,629	8,329	8,808	9,112	9,321	9,484	9,571	9,571	763
1995	3,724	7,167	8,573	9,850	10,763	11,393	11,796	12,070	12,282	12,397	12,397	1,634
1996	3,213	6,202	7,423	8,540	9,337	9,885	10,232	10,473	10,656	10,757	10,757	2,217
1997	3,335	6,445	7,727	8,887	9,721	10,281	10,643	10,890	11,083	11,187	11,187	3,460
1998	3,596	6,975	8,387	9,662	10,589	11,207	11,604	11,876	12,090	12,204	12,204	5,229
1999	3,327	6,481	7,817	9,018	9,889	10,483	10,860	11,123	11,323	11,432	11,432	8,105
												22,369
implied												
Link ratios	1.964	1.209	1.149	1.094	1.060	1.036	1.022	1.017	1.009	1.000		
LDFs	3.436	1.750	1.448	1.260	1.152	1.087	1.049	1.026	1.009	1.000		

The "unknown" data elements (those that would be known as of 12/31/2000 or after) are shaded, and will not be used to fit models. Note that the "ultimate" column is the same as the

"at 120 months" column, and represents the "true", though unknown ultimate losses (u_k). Similarly, the "o/s" column represents the "true" outstanding losses as of 12/31/1999 (r_k). Thus the "true" value that we wish to estimate is $\varrho=\sum \varrho_k$ =$22.369M.

Note that the link ratios computed from this summarized data are essentially weighted averages of the $link^{bad}$ and $link^{good}$ ratios stated above. This is representative of the way important patterns can be "summarized away" when the data is summarized to the triangle level.

MODEL RESULTS

We applied our sequence of 9 Poisson GLM models to the 5000 simulated data points. The exact steps of this process are sketched below:

Step 1: Regress the 4500 data points with non-missing values of C_{24} (i.e. the claims from AY 1990-98) on credit score, using $\log(C_{12})$ as the offset term. This model is then applied to the 500 claims with unknown values of L_{24} (i.e. the AY 1999 claims) to produce *predicted* values of C_{24}.

Step 2: Regress the 4000 data points with non-missing values of C_{36} (i.e. the claims from AY 1990-97) on credit score, using $\log(C_{24})$ as the offset term. This model is then applied to the 1000 claims with unknown values of L_{36} (i.e. the AY 1998-99 claims) to produce *predicted* values of C_{36}. Note that the AY 1998 values of C_{36} are based on *actual* values of C_{24}; whereas the AY 1999 values of C_{36} are based on *predicted* values of C_{24}.

...

Step 9: Regress the 500 data points with non-missing values of C_{120} (i.e. the claims from AY 1990) on credit score, using $\log(C_{108})$ as the offset term. This model is then applied to the 4500 claims with unknown values of C_{120} (i.e. the AY 1991-99 claims) to produce *predicted* values of C_{120}. Note that the AY 1990 values of C_{120} are based on *actual* values of C_{108}; whereas the AY 1991-99 values of C_{120} are based on *predicted* values of C_{108}.

Step 10: The ultimate loss estimate is the sum of C_{120} across all claims and across all accident years: $u=\sum\sum C_{120}$. The estimate of total outstanding losses r equals u minus the total claims paid as of 12/31/1999.

The way in which the model M_j is applied to the predicted values of model M_{j-1} is analogous to the way the chain ladder's link ratios are multiplied together to produce loss development factors.

The results of these 10 steps, summarized to the accident year level, are displayed below. They can be compared to the display of the "truth" above:

GLM predictions (shaded) Losses in $1000's												
	@12	@24	@36	@48	@60	@72	@84	@96	@108	@120	ultimate	o/s
1990	3,522	6,562	7,766	8,850	9,627	10,144	10,473	10,700	10,875	10,970	10,970	-
1991	3,527	6,623	7,876	9,011	9,817	10,361	10,705	10,942	11,123	11,222	11,222	99
1992	3,681	6,939	8,235	9,428	10,274	10,833	11,194	11,444	11,635	11,738	11,738	294
1993	3,780	7,152	8,539	9,791	10,666	11,262	11,642	11,904	12,106	12,214	12,214	572
1994	2,912	5,563	6,644	7,629	8,329	8,808	9,112	9,323	9,485	9,573	9,573	765
1995	3,724	7,167	8,573	9,850	10,763	11,387	11,786	12,063	12,277	12,392	12,392	1,629
1996	3,213	6,202	7,423	8,540	9,337	9,877	10,222	10,461	10,646	10,745	10,745	2,205
1997	3,335	6,445	7,727	8,896	9,729	10,293	10,654	10,904	11,097	11,202	11,202	3,475
1998	3,596	6,975	8,378	9,665	10,584	11,207	11,605	11,882	12,096	12,212	12,212	5,237
1999	3,327	6,484	7,795	8,999	9,859	10,442	10,816	11,075	11,276	11,384	11,384	8,057
												22,333
implied link	1.954	1.208	1.152	1.093	1.059	1.036	1.023	1.017	1.009	1.000		
LDF	3.422	1.751	1.450	1.258	1.151	1.087	1.049	1.026	1.009	1.000		

Note that the implied LDFs at the bottom of this display were calculated by dividing the predicted ultimate values by the losses for that accident year as of 12/31/99. The implied link ratios were then derived from the implied LDFs.

Finally, the results of a chain ladder exercise are displayed in the following table:

Chain Ladder predictions (shaded) Losses in $1000's												
	@12	@24	@36	@48	@60	@72	@84	@96	@108	@120	ultimate	o/s
1990	3,522	6,562	7,766	8,850	9,627	10,144	10,473	10,700	10,875	10,970	10,970	-
1991	3,527	6,623	7,876	9,011	9,817	10,361	10,705	10,942	11,123	11,223	11,220	97
1992	3,681	6,939	8,235	9,428	10,274	10,833	11,194	11,444	11,635	11,739	11,733	289
1993	3,780	7,152	8,539	9,791	10,666	11,262	11,642	11,902	12,100	12,210	12,200	558
1994	2,912	5,563	6,644	7,629	8,329	8,808	9,112	9,321	9,484	9,571	9,536	728
1995	3,724	7,167	8,573	9,850	10,763	11,393	11,796	12,070	12,282	12,397	12,298	1,535
1996	3,213	6,202	7,423	8,540	9,337	9,885	10,232	10,473	10,656	10,757	10,637	2,097
1997	3,335	6,445	7,727	8,887	9,721	10,281	10,643	10,890	11,083	11,187	11,031	3,304
1998	3,596	6,975	8,387	9,662	10,589	11,207	11,604	11,876	12,090	12,204	11,873	4,898
1999	3,327	6,481	7,817	9,018	9,889	10,483	10,860	11,123	11,323	11,432	10,793	7,466
												20,972
implied link	1.906	1.192	1.146	1.090	1.055	1.033	1.022	1.016	1.009	1.000		
LDF	3.244	1.702	1.428	1.246	1.143	1.083	1.048	1.025	1.009	1.000		

(Note that this calculation can be verified by the reader in a spreadsheet. The spreadsheet-based results will differ from the above o/s loss estimate by $2000 (0.01%). This is due to rounding errors: the above table was generated by a computer program using un-rounded losses in the upper triangle.)

For convenience, the results of both methods – together with the simulated "truth" – are displayed below:

	Losses in $1000's													
	@12	@24	@36	@48	@60	@72	@84	@96	@108	@120		C-L	truth	proposed
1990	3,522	6,562	7,766	8,850	9,627	10,144	10,473	10,700	10,875	10,970	10,970	0	**0**	0
1991	3,527	6,623	7,876	9,011	9,817	10,361	10,705	10,942	11,123		11,220	97	**99**	99
1992	3,681	6,939	8,235	9,428	10,274	10,833	11,194	11,444			11,734	289	**295**	294
1993	3,780	7,152	8,539	9,791	10,666	11,262	11,642				12,200	558	**567**	572
1994	2,912	5,563	6,644	7,629	8,329	8,808					9,537	728	**763**	765
1995	3,724	7,167	8,573	9,850	10,763						12,298	1,535	**1,634**	1,629
1996	3,213	6,202	7,423	8,540							10,637	2,097	**2,217**	2,205
1997	3,335	6,445	7,727								11,031	3,304	**3,460**	3,475
1998	3,596	6,975									11,873	4,898	**5,229**	5,237
1999	3,327										10,792	7,466	**8,105**	8,057
												20,972	**22,369**	22,333
C-L	1.906	1.192	1.146	1.090	1.055	1.033	1.022	1.016	1.009	1.000		**-6.25%**		**-0.16%**
	3.244	1.702	1.428	1.246	1.143	1.083	1.048	1.025	1.009	1.000				
truth	1.964	1.209	1.149	1.094	1.060	1.036	1.022	1.017	1.009	1.000				
	3.436	1.750	1.448	1.260	1.152	1.087	1.049	1.026	1.009	1.000				
proposed	1.954	1.208	1.152	1.093	1.059	1.036	1.023	1.017	1.009	1.000				
	3.422	1.751	1.450	1.258	1.151	1.087	1.049	1.026	1.009	1.000				

Because the chain ladder is slow to pick up the changing mix of business (i.e., increasing proportion of bad credit policies that produce slower-developing claims), its estimates are too low for each accident year. This effect is most pronounced for the later accident years (shaded). In this example, the chain ladder's total outstanding loss estimate is approximately 6% too low.

By comparison, the proposed method's total outstanding loss estimate is almost exactly correct. It goes without saying that this is because our losses were simulated to develop in the multiplicative fashion assumed by the chain ladder; and because by construction only one covariate – credit – has a statistically significant relationship with loss development. Of course real-world data present no such conveniences. The above results are therefore suggestive at best. Still, the point remains that the proposed method is able to reflect changes in the mix of business (assuming that these changes can be measured by covariates capable of being collected and put into a model) that the chain ladder misses.

THE PROPOSED METHOD IS A PROPER GENERALIZATION OF THE CHAIN LADDER

By now it should be clear that the proposed loss reserving framework is intended to function as a GLM/micro-data-based analog of the chain ladder. One can go further and state that it is a true generalization of the chain ladder, in the sense that it produces the same results as the chain ladder when no covariates are present.

We verified this with the simulated data analyzed above. That is, we simply fit the above sequence of 9 GLM models, replacing the credit variable with a constant. The proposed method results in *exactly* the same results as the chain ladder. These results are summarized below.

acc. year	losses @ 12/99	true ultimate	true o/s	our method	chain ladder
1990	10,970	10,970	-	-	-
1991	11,123	11,223	99	97	97
1992	11,444	11,739	295	289	289
1993	11,642	12,210	567	558	558
1994	8,808	9,571	763	728	728
1995	10,763	12,397	1,634	1,535	1,535
1996	8,540	10,757	2,217	2,097	2,097
1997	7,727	11,187	3,460	3,304	3,304
1998	6,975	12,204	5,229	4,898	4,898
1999	3,327	11,432	8,105	7,466	7,466
			22,369	20,972	20,972
				-6.25%	-6.25%

It is generally a bad idea to exclude a statistically significant covariate from the GLM models. Here we see that doing so reproduces the chain ladder's (understated) reserve estimate. This lends a statistical perspective to where the chain ladder goes wrong when applied to a book of business whose development patterns have changed over time.

PART II: THE PROBLEM OF ESTIMATING RESERVE VARIABILITY

From a statistical perspective, R is an *estimator* of outstanding losses. It is a function of the values of the random variables $\{C_{12}, X_1, X_2, \ldots, X_N\}$ for each data point. In other words, it is a complicated function of several random variables. Like any such estimator, it has a probability distribution that is a complicated function of the distributions of the underlying random variables.

As we have demonstrated above, it is fairly straightforward to calculate the expected value of R. This is our outstanding loss *estimate*. It summarizes what the data (and our model) tells us to expect about the amount of future claim payments. But we would also like a measure of how strongly we should believe this estimate. To do this, we need further information – other than the expected value – about the distribution of our estimator of outstanding losses. For example, what are the cutoffs of a 95% confidence interval around the estimate?

This problem – sometimes referred to as the problem of *reserve variability* – has received a lot of attention in the recent loss reserving literature. The recent report of the CAS Working Party on Quantifying Variability in Reserve Estimates [2] puts the matter this way:

> A risk bearing entity wishes to know its financial position on a particular date. In order to do this, among other items it must understand the future payments it will be liable to make for obligations existing at the date of the valuation. For an insurance situation, these future payments are not known with certainty at the time of the valuation.
>
> The fundamental question that the risk bearing entity asks itself is:
> *Given any value (estimate of future payments) and our current state of knowledge, what is the probability that the final payments will be no larger than the given value?*

A full answer to this question would involve the assessment of model risk, and is beyond the scope of this paper. But even a limited answer would go beyond supplying a mere confidence interval or variability estimate. Ideally, we would like an estimate of the entire *probability distribution* of the outstanding loss estimator.

This seems like a lot to ask. After all, both the loss distribution underlying our claims data as well as our estimators of outstanding losses are fairly complex. Surprisingly, modern statistics supplies us with a simulation-based technique – called *bootstrapping* – that allows us to estimate this distribution with fairly little effort.

ENTER THE BOOTSTRAP

The Bootstrap was introduced by Bradley Efron in the late 1970s. Since then, it has become a commonly used technique in any number of problems in applied statistics. The classic text is Efron and Tibshirani [3]. Put briefly, bootstrapping is a simulation-based technique for estimating potentially "difficult" distributional properties – such as the standard deviation or the 90th percentile – of potentially complex estimators. We typically do not know the "true" distribution of such estimators. The basic idea of the Bootstrap is therefore to use the *actual,* empirical distribution (i.e., the data) as a proxy for the true, unknown distribution. Once this conceptual leap is made, many otherwise intractable problems become fairly straightforward exercises in statistical computing.

An analogy lies at the heart of bootstrapping. Just as our *actual* distribution is one of an infinite number of *possible* draws from the "true" theoretical distribution; we can take a large number of *resamples* of our actual distribution to form an arbitrarily large number of "pseudo-datasets".

Actual distribution : "true" distribution :: resampled datasets : actual distribution

Just as we would know everything we need to know about the "true" distribution if we could draw a large number of samples from it, we can *estimate* much of what we would like to know about the "true" distribution by treating the actual distribution as a proxy, and drawing multiple resamples from it.

We can illustrate this idea by applying it to a very simple problem for which we know the answer in advance. Suppose we draw 500 observations $\boldsymbol{X}=\{X_1,\ldots, X_{5000}\}$ from a normal distribution with μ=5000 and σ=100: $n(5000,100)$. Let m denote the sample average of this data:

$$m=\frac{1}{n}\sum_{i=1}^{5000}X_i$$

m is an estimate of the true value μ, just as we derived an estimate of the "true" outstanding losses in the previous sections. m therefore tells us "what we think" about the true value of μ based on the data. We would also like a measure of "how sure we are". In this simple

example, the obvious thing to do is construct a confidence interval by appealing to the elementary fact that:

$$s.d.(m) = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{5000}} \approx 4.47$$

Let us apply bootstrapping to this problem to see how close we can come to the answer (4.47) that we know in advance.

The following table records some facts about our data:

- # obs: 500
- Mean: 4995.79
- Stdev: 98.78
- 2.5^{th} %^{ile}: 4812.30
- 97.5^{th} %^{ile}: 5195.58

We can resample from this dataset a large number of times to create multiple "pseudo-datasets". "Resampling" means sampling with replacement as many times as there are points in your initial dataset (here, 500). Explicitly: pull a point at random from $\{X_1,\ldots, X_{500}\}$; record it; throw it back in; repeat this until we have our first pseudo-dataset containing 500 observations. Let us denote this pseudo-dataset $\boldsymbol{X}^*_1$.

We now repeat this process as many times as we would like, say 999 additional times. We therefore have 1000 pseudo-datasets $\boldsymbol{X}^*_1,\ldots \boldsymbol{X}^*_{1000}$. We can compute the sample average m on each one of these datasets. Denote these $\{m^*_1,\ldots m^*_{1000}\}$. These 1000 estimates constitute an estimate of the *distribution* of our estimator m. With this distribution $\{m^*_1,\ldots m^*_{1000}\}$ in hand, we can very easily estimate nearly any distributional property of m that we would like. In particular: the sample standard deviation of m based on our 1000 resamples is 4.43:

$$s.d.(m) \approx \frac{1}{999}\sum_{i=1}^{1000}\left(m^*_i - \frac{1}{1000}\sum_{k=1}^{1000} m^*_k\right)^2 \approx 4.43$$

This differs from the true value (4.47) by less than a percentage point.

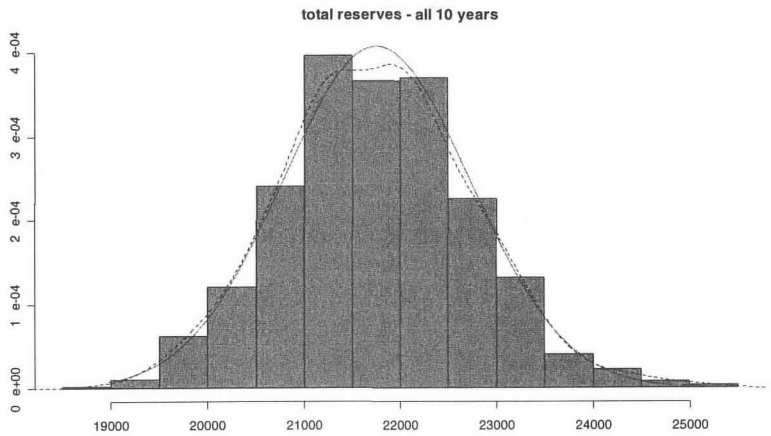
Bootstrapping in this toy example is therefore a complete success. The key point to note is that the unlike our analytic formula for s.d.(m), the bootstrapping technique does not assume any knowledge of the underlying distribution of $\boldsymbol{X}$. All that was required was computing

power. Because of this, it is possible to execute essentially the same process on the loss data analyzed in the previous sections.

BOOTSTRAPPING RESERVE ESTIMATES

Having introduced the concept and run through a simple example, there is little to say in this section, other than to report the results. Let $\boldsymbol{S}$ denote our database of 5000 claims. We resampled $\boldsymbol{S}$ 500 times to get the 500 pseudo-datasets $\boldsymbol{S}^*_1, \ldots \boldsymbol{S}^*_{500}$. We then ran the above 9 GLM models on each of these 500 pseudo-datasets and computed outstanding losses on each pseudo-dataset: $\{R^*_1, \ldots, R^*_{500}\}$. Although it might seem excessive to fit 4500 GLM models to estimate the distribution of outstanding losses, doing so took less than 15 minutes on a standard laptop equipped with the shareware statistical software package R.

The estimated distribution of the outstanding loss estimator R is plotted below:



The bars are simply a histogram of the 500 estimates of outstanding losses. The solid curve is a superimposed normal distribution. The dotted curve is a kernel density estimate of the distribution underlying the histogram. Some basic statistics of this distribution are reported below:

- Mean: $21.751M
- Median: $21.746M
- Stdev: $0.982M

- C.V.: 4.5%

This kernel density estimate in the graph suggests that the distribution of our outstanding loss estimator is normal, to a reasonable degree of approximation. The fact that the mean is nearly equal to the median reinforces this judgment. Therefore a 95% confidence interval around our reserve estimate can be calculated in one of two ways:

- Record the 2.5 and 97.5 percentiles of the bootstrap distribution.
- Calculate 21.751M $\pm$ (1.96)*(0.982M).

Both of these methods produce the same answer, to within the nearest $100K:

($19.8M, $23.7M)

In short (ignoring model risk), we have 95% confidence that the true outstanding loss is within $\pm$ 9% of our estimated value. We remind the reader that this result is based on a rudimentary simulation, and is only intended to be suggestive.

DISCUSSION

Before concluding this paper, we would like to make four points about the bootstrapping technique illustrated above. First, bootstrapping is uncommonly generous to the practitioner in that it gives one an estimate of the entire *distribution* of an arbitrarily complex estimator without asking for *any* knowledge of the distributions underlying the data. Nearly any question we would typically ask about the outstanding loss distribution (standard deviation, skewness, percentiles, probability of ruin...) can be addressed with mere computation.

Second, the bootstrap method illustrated above is not specific to our GLM-based reserving technique. Indeed, if the claim-level data is available, one can also use this technique to bootstrap chain-ladder, Bornhuetter-Ferguson, or any other reserve estimates. To do this, we would summarize each of our pseudo-datasets to the triangle level; and apply our favorite technique to each of the resulting triangles. The 1000 outstanding loss estimates (assuming 1000 pseudo-datasets, as in the above illustration) resulting from each of the 1000 pseudo-triangles will constitute the distribution of our outstanding loss estimate.

Third, bootstrapping has been the subject of some discussion in the recent loss reserving literature. But there is an important difference between these discussions and the technique

illustrated here. To the best of our knowledge, these discussions have been offered in the context of analyses of summarized loss triangles, not claim-level data.

The excellent survey paper by England and Verrall [4] is an example. England and Verrall apply a GLM model to a summarized loss triangle, and resample the standardized *residuals* of this model. They resample the distribution of residuals (there will be 55 data points for a 10-by-10 loss triangle) a large number of times. Each time they add the pseudo-dataset of residuals to the original loss triangle to form a pseudo-history to which they can again apply their GLM. Doing so allows them to estimate the prediction error of their estimate.

The difference between England and Verrall's approach and the approach illustrated here is generic, and found in most textbook discussions of bootstrapping. When bootstrapping model predictions, it is possible either to bootstrap *cases* (our approach) or *residuals* (England-Verrall). When dealing with small loss triangles it is not meaningful to bootstrap cases. However bootstrapping cases *is* meaningful when claim-level data is available.

As noted in the final paragraph of the introduction, our approach of resampling cases occurs prior to any reserving model being fit to the data. In other words, the very validity of our pseudo-datasets does not depend on the adequacy of the model being fit. In this sense, the cases-based resampling strategy is less sensitive to the correctness of ones model than the residual-based resampling strategy.

One final comment: bootstrapping is not the last word on the topic of reserve variability. In particular, nothing we have said addresses the problem of *model* risk. Suppose, for example, that we bootstrapped the traditional chain ladder applied to our simulated data. The bootstrapped confidence interval would not reflect the bias due to excluding the credit covariate in our reserving model. What is perhaps the biggest challenge in reserve variability is therefore left untouched by this discussion. Still, by giving us a practical way of estimating the predictive distribution of outstanding losses, bootstrapping potentially allows one to devote more attention to model risk.

CONCLUSION

The traditional summarized loss triangle is in general not a "sufficient statistic" for estimating outstanding losses. There will be times when we can do better by basing reserve and reserve variability estimates on un-summarized claim-level data.

As the first half of our paper illustrates, loss triangles can suppress heterogeneous loss development patterns that could be used to improve our predictions of outstanding losses. At the same time, summarized data does not allow us to use predictive variables that might be correlated with different loss development patterns.

Furthermore, as noted in the second half of our paper, loss triangles potentially summarize away variability information that could be used to make improved estimates of reserve variability. Using claim-level data allows us to bootstrap *cases*, not merely residuals from models applied to loss triangles with small numbers of observations.

In short, the use of claim-level data, together with relevant predictive variables, has the potential to improve actuaries' estimates of outstanding losses. In addition, it makes available a powerful and conceptually simple method for estimating reserve variability.

Acknowledgment

The authors would like to thanks Gerry Kirschner, Keith Curley, Peter Wu, and Frank Zizzamia for helpful conversations.

APPENDIX: ADDING IBNR TO THE MODEL

This appendix outlines a method by which one can enhance the model to predict INBR losses. Alternately, one can simply use the model outlined in the body of this paper to model the development of reported claims (as is done in the simulation example to follow); and build a separate model to estimate IBNR.

The 12→24 model (M_{24}) not modified to reflect IBNR takes the form:

$$C_{24} = \exp\{\log(C_{12}) + \alpha + \beta_1 X_1 + \beta_N X_2 + ... + \beta_N X_N\} + \varepsilon$$

The idea is to introduce a record for each *policy* with no losses as of 12 months (C_{12}=0) from its effective date. (Note that the other records in our database are at the claim level.) We set the offset term $\log(C_{12})$ on these records to be zero. We also include on all records an indicator variable X_0 that takes on the value 1 if C_{12}=0, and 0 otherwise. Finally, on the (claim-free) policy-level records we would neutralize all predictive variables that measure claim-level information. ("Neutralize" might mean that we recode missing values of a variable to the median value of that variable.)

For the 1990-98 policy-level records, we let $\{C_{24}, C_{36}, \ldots, C_{120}\}$ equal the total IBNR evaluated at these various evaluation points. As with all of the other AY 1999 records in the database the values of $\{C_{24}, C_{36}, \ldots, C_{120}\}$ are all missing. We add the indicator variable X_0 in the model. At this point our model takes the form:

$$C_{24} = \exp\{\log(C_{12}) + \alpha + \gamma_0 X_0 + \beta_1 X_1 + \beta_N X_2 + ... + \beta_N X_N\} + \varepsilon$$

Note that in this model form, the offset term only "applies" to the claim-level records with a non-zero value of C_{12}; similarly, the term $\gamma_0 X_0$ "applies" only to the policy-level records with C_{12}=0. The remaining terms apply to both types of records. In other words, each of the β parameters simultaneously models development of losses reported as of 12 months, as well as allocates IBNR losses at 24 months.

If this dual functioning of the β parameters is unsatisfactory, it is possible to let the β parameters *only* model the development of *reported* claims (as in the original model with no IBNR component) by introducing interaction terms. Suppose that $X_1 \ldots X_{N-p}$ are the policy-level covariates (such as policy age and credit score) in the model. (Claim-level variables such

as report lag or injury type do not apply to policy-level records.) We add the interaction terms $X_0 * X_1 \ldots X_0 * X_{N-p}$ into the model:

$$C_{24} = \exp\{\log(C_{12}) + \alpha + \gamma_0 X_0 + \beta_1 X_1 + \ldots + \beta_N X_N + \gamma_1 X_0 * X_1 + \ldots + \gamma_{N-p} X_0 * X_{N-p}\} + \varepsilon$$

If this seems somewhat complex, it is because we have really designed "two models in one". The 12→24 development of a claim C^*_{12} is given by the following equation:

$$C^*_{24} = \exp\{\log(C^*_{12}) + \alpha + \beta_1 X_1 + \ldots + \beta_N X_N\}$$

All of the terms with X_0 drop out because X_0 is assumed to be 0 on (claim-level) records with non-zero C_{12}. In other words, we are back to the model form given at the beginning of this appendix.

On the other hand, the allocated IBNR at 24 months for a policy with no loss at 12 months is given by the following equation:

$$C_{24} = \exp\{\alpha + \gamma_0 + (\beta_1 + \gamma_1) X_1 + \ldots + (\beta_{N-p} + \gamma_{N-p}) X_{N-p} + \kappa\}$$

Here κ denotes the terms $\{\beta_{N-p+1} X_{N-p+1} + \ldots + \beta_N X_N\}$. These terms reduce to a constant κ because the claim-level variables $\{X_{N-p+1} \ldots X_N\}$ were neutralized on the policy-level records. In addition, note that the offset term was forced to be zero on these policy-level records.

It might be helpful to note that $\exp\{\alpha + \gamma_0 + \kappa\}$ is the *average* IBNR allocated to each of the policies that were claim-free as of 12 months. The multiplier $\exp\{(\beta_1 + \gamma_1) X_1 + \ldots + (\beta_{N-p} + \gamma_{N-p}) X_{N-p}\}$ adjusts each policy's allocated IBNR based on the values of the policy-level covariates $X_1 \ldots X_{N-p}$. As with expected claim development, the fact that the allocation of IBNR is "tailored" to the individual policy according to that policy's characteristics allows the model to reflect changes in the mix of business being analyzed.

Models M_{36}, ..., M_{120} can similarly be modified to handle the further emergence and development of IBNR.

REFERENCES

[1] Bornhuetter, R. and R. Ferguson. 1972. The Actuary and IBNR. *PCAS* LIX:181-95.

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

[3] Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap.* San Francisco: Chapman & Hall.

[4] England, P. D. and R. J. Verrall. 2002. Stochastic Claims Reserving in General Insurance. *British Actuarial Journal* 8:443-544.