Applications of population sampling to IBNR reserving: Extending and blending the Chain-Ladder with individual reserving methods

Sebastián Calcetero Vanegas, X. Sheldon Lin, Andrei L. Badescu
Department of Statistical Sciences
University of Toronto
Toronto, Ontario

Executive Summary

Accurate claim reserve estimation is crucial for insurers to maintain solvency and comply with regulations. Traditional reserving relies on macro-level models, like the Chain-Ladder method, which estimate reserves using aggregate data, and micro-level models, which analyze individual claims for greater precision. While macro-level models are simple and widely used, they may overlook claim-level patterns. Micro-level models, though more accurate, require complex, data-intensive processes and are not widely adopted in practice due to their complexity.

This research introduces a novel statistical framework that enhances reserving methods while preserving simplicity and interpretability. By treating reported claims as a sample from a broader population, akin to a population sampling problem, the approach improves reserve accuracy and robustness in handling data variability. Using advanced statistical tools like double-robust estimators and debiasing techniques, the method adapts to dynamic conditions more effectively. This report demonstrates the practical implementation of this unified approach through a structured discussion of the problem, notation, and methodology, followed by an extensive numerical study using simulated and real data. Our goal is to equip practitioners with tools to leverage granular datasets for more accurate reserve calculations, with full technical details available in the referenced paper.

The claim reserving problem

Suppose an insurance company is evaluating its total liabilities from claims with accident times between $t = 0$ and $t = \tau$, where τ is the valuation time set by the actuary. In general insurance, there is often a delay between when an accident occurs and when it is reported to the company. Then, at valuation time τ, the insurer knows only the claims that have been reported and the partial amount paid to each one. Therefore the insurer aims to estimate the total amount for unreported claims to calculate the reserve for outstanding claims. Let us describe the claim process as follows:

- $N(\tau)$: Total number of claims with accident times before τ.
- Y_i: Sequence of incurred losses, where $i = 1, \ldots, N(\tau)$.
- T_i: Sequence of accident times for each claim.
- U_i: Sequence of reporting delay times for each claim.

- $\boldsymbol{x}_i$: Sequence of attributes related to the accident, claim type or policyholder.
- $N^R(\tau)$: Number of claims reported by τ, and $N^{IBNR}(\tau)$: Number of unreported claims.

The total liability for accidents occurring before τ, denoted $L(\tau)$, is:

$$L(\tau) = \sum_{i=1}^{N(\tau)} Y_i.$$

The portion of liability known to the company by τ (from reported claims), denoted $L^R(\tau)$, he portion of liability known to the company by τ (from reported claims), denoted $L^R(\tau)$, and the liability for non-reported claims, denoted $L^{IBNR}(\tau)$, are respectively

$$L^R(\tau) = \sum_{i=1}^{N^R(\tau)} Y_i \qquad L^{IBNR}(\tau) = L(\tau) - L^R(\tau) = \sum_{i=1}^{N^{IBNR}(\tau)} Y_i.$$

Stochastic reserving models focus on estimating reserves through individual claims, characterizing the number of claims per individual, claim amounts, and reporting delays. These models apply flexible machine-learning methods to accurately capture data behaviors, providing point estimates $\hat{Y}_i$ for claim amounts, total claims, and non-reported claims $\hat{N}^{IBNR}(\tau)$. Reserves are then estimated by summing over the predicted quantities of nonreported claims as $\sum_{i=1}^{\hat{N}^{IBNR}(\tau)} \hat{Y}_i$.

Models for the reported claims data are fitted using maximum likelihood estimation, involving the components of frequency, severity and delay, respectively. The independence of these components, especially between severity and delay, is often assumed but may introduce bias due to the truncation of non-reported claims. This independence assumption, rarely questioned in reserving literature, contrasts with population sampling, where accounting for sampling design is crucial to avoid bias. The truncated nature of reserving data (i.e., only reported claims observed) means it may not fully represent the population if dependencies between variables exist. Consequently, failing to adjust for this can yield biased reserve estimates, as characteristics of reported claims differ from unreported ones. In response, the model-assisted approach in population sampling emphasizes adjusting for this data bias. The following section explores integrating these approaches into reserving to improve reserve estimation accuracy.

IBNR reserving as a population sampling problem and methods

Reserving can seen as a situation in which we are interested in finding a population total, based only on a sample. Here, we treat all $N(\tau)$ claims as the population, and the $N^R(\tau)$ claims reported by the valuation date as the sample. One can define the membership indicator variable (belongs or not to the sample) for each claim as $\mathbf{1}_i(\tau) = \mathbf{1}_{\{T_i+U_i \leq \tau\}}$, and their inclusion probabilities $\pi_i(\tau)$ as

$$\pi_i(\tau) = P(U_i \leq \tau - T_i | \boldsymbol{x}_i, T_i, Y_i).$$

These probabilities, which depend on valuation time and vary by claim due to differing types and risk profiles, are must be estimated from the data. This component is usually estimated as part of any reserving model itself. This can be achieved by recognizing that the reporting delay time is a time-to-event random variable under a right-truncation scheme. Therefore one can use existing approaches from survival analysis to obtain an estimation of the cumulative distribution function, from which the desired inclusion probabilities can be obtained, e.g, Cox regression models. The important fact to keep in mind is that the estimation of the inclusion probabilities in a proper manner is a key component from the population sampling perspective.

We proceed to describe the methods from population sampling that can be integrated into reserving. These can be informally classified based on when the technique is applied with respect to the estimation of a model. All of these approaches require the inclusion probabilities as starting point.

- Post-Processing: Modifications are performed after fitting the model to the data. These adjust the resulting point prediction to account for the sampling mechanism.
- In-Processing: Modifications are performed while fitting the model to the data. These involve constraints based on the sampling mechanism while optimizing parameters.
- Pre-Processing: Modifications are performed before fitting the model to the data, e.g., data pre-processing approaches.

For what follows suppose that a reserving model is available (e.g. a frequency-severity model as described in the previous section) that provides estimates of the total number of claims, $\hat{N}(\tau)$, the total number of non-reported claims, $\hat{N}^{IBNR}(\tau)$ and each of the claim amounts, which we will denote as $\hat{Y}_i$. This information could be based on expert opinion, such as is commonly for the so-called case ultimates, or based on a so-called micro-level reserving model.

0.0.1 Post-processing approach: Augmented inverse probability weighting

A well-established estimator when a model is assisting is given by an augmented inverse probability weighting estimator (AIPW) estimator given by

$$\hat{L}(\tau) = \sum_{i=1}^{\hat{N}(\tau)} \hat{Y}_i + \sum_{i=1}^{N^R(\tau)} \frac{Y_i - \hat{Y}_i}{\pi_i(\tau)},$$

where the first term can be thought of as the estimation of the population total based solely on the auxiliary information $\hat{Y}_i$, and the second term can be thought of as an scaling of the "residuals," $Y_i - \hat{Y}_i$, in the sample to match the error in the population. Along those lines, an AIPW estimator of the outstanding claims is therefore

$$\hat{L}^{IBNR}(\tau) = \hat{L}(\tau) - L^R(\tau) = \sum_{i=1}^{\hat{N}^{IBNR}(\tau)} \hat{Y}_i + \sum_{i=1}^{N^R(\tau)} \frac{1-\pi_i(\tau)}{\pi_i(\tau)}(Y_i - \hat{Y}_i),$$

The AIPW estimator acts as a correction mechanism, where the second term addresses errors in the micro-level model. Thus, micro-level reserving models may benefit from the bias correction provided by AIPW estimators, and may provide preferable estimates.

0.0.2 In-processing: Fitting models for the population of non-reported claims

One approach to account for the sampling mechanism in the estimation of a model involves using weighted estimating equations, where the weights are the odds ratio of no inclusion vs inclusion $\frac{1-\pi_i(t)}{\pi_i(t)}$. For example, consider an estimating equation for the severity model based on minimizing the empirical version of a loss function. The following weighted equation is unbiased for the loss function of the non-reported claims population:

$$\sum_{i=1}^{N^R} \frac{1-\pi_i(\tau)}{\pi_i(\tau)} l(Y_i, \theta)$$

In this case, the model parameters θ should be chosen to minimize this weighted loss function rather than the unweighted version and therefore address the sampling bias. This approach is well-documented in the literature on missing data and surveys with non-response.

0.0.3 Pre-Processing: Synthetic data sets of non-reported claims

A broader approach involves estimating the entire population of non-reported claims using data augmentation methods commonly employed when handling missing data. Briefly, each claim Y_i in the sample is expected to appear $\frac{1-\pi_i}{\pi_i}$ times in the population of non-reported claims. Therefore, a synthetic data can be constructed by replicating each reported claim $(\mathbf{x}_i, Y_i)$ in a new dataset a total $\frac{1-\pi_i}{\pi_i}$ times. The distribution function implied on this pseudo-data set would be given by:

$$\hat{F}_Y(y) = \frac{\sum_{i=1}^{N^R(\tau)} \frac{1-\pi_i(\tau)}{\pi_i(\tau)} \mathbf{1}_{\{Y_i \leq y\}}}{\sum_{i=1}^{N^R(\tau)} \frac{1-\pi_i(\tau)}{\pi_i(\tau)}}.$$

which is a self-normalized IPW estimator of the CDF of the non-reported claim amounts, which is known to be a strongly consistent estimator. Therefore, this synthetic data should behave similarly to the unknown non-reported claims and can be treated as the true data, allowing for the estimation of reserves, statistical models and related quantities.