**STATS 3AT3**

# **Bootstrapping in Insurance**

Team: Liability Lions (Super Wang, Shijun Wu, Oliver Tung, Adamo Dragicevic)

For: Professor, Anas Abdallah

December 5th, 2024

**Introduction:**

Our project investigates the role of bootstrapping in advancing insurance analytics across three key areas. First, we assess insurance reserve estimation by comparing the bootstrapping method with the Bornhuetter-Ferguson (B-F) method. This comparison focuses on regions with differing levels of historical claims data density, offering insights into the reliability and consistency of each method under varying data conditions.

Second, we explore the impact of specific variables on premium prediction by leveraging bootstrapping. This involves identifying key drivers of claims behavior, such as demographic factors, and evaluating how certain variable segments disproportionately influence claim magnitudes. This analysis provides a deeper understanding of the variables most critical to premium prediction.

Lastly, we enhance predictive modeling accuracy by integrating bootstrapping with established models, such as Generalized Linear Models (GLM). Bootstrapping’s iterative resampling capabilities are used to refine model parameters and address variability and uncertainty in the data, leading to more robust and reliable forecasts.

Overall, the project demonstrates the versatility of bootstrapping in tackling challenges in reserve estimation, variable importance analysis, and predictive modeling within the insurance industry.

**Key Concepts**

**Reserving triangle**

Reserving is a strategy that actuaries use to ensure that their company has enough money set aside to make payments on claims while maintaining financial stability. Actuaries use historical data and statistical models to make predictions based on variable frequency, severity and claims frequency. Reserves also consider IBNR (incurred but no reported) claims.

When reserving, reserve triangles filled with historical data are typically used to analyze the development of claims over time. The triangles help in projecting the ultimate costs for fully developed claims allowing actuaries to make informed decisions.

**Bornhuetter-Ferguson and Chain-Ladder Models**

Two methods that are commonly used to fill in the projected values of the reserve’s triangle are the Bornhuetter-Ferguson method and the chain ladder method. Each method is effective given different scenarios.

The BF method uses reported losses and an expected loss ratio to estimate ultimate costs and is ideal when we have limited historical data to work with. It’s useful when we have incomplete or immature data, as well as lines of business that have limited data.

Comparatively, the chain ladder method relies on solid historical claims development patterns to predict future claims. Since the chain ladder method is purely data driven, it is well-suited for when patterns are assumed to continue.

**Bootstrapping**

Bootstrapping is a method that insurers can use to help estimate reserves. The process takes existing data and resamples said data which can be used to create a large number of claims triangles. The large number of pseudo triangles can then be paired with the chain ladder method, resulting in a projection of the ultimate costs for each independent pseudo triangle. The ability to create pseudo triangles allows actuaries to test their reserves under different scenarios. This simulated range of possible outcomes is useful for understanding the risk and volatility in reserves. Additionally, bootstrapping can be paired with a variety of models such as GLMs which we discuss in a later section.

**Stress testing**

Stress testing is a critical tool in risk management and decision-making, designed to evaluate the resilience of systems or models under extreme or adverse scenarios. In the context of insurance and financial modeling, stress testing assesses how predictive models, such as those used for reserve estimation or pricing, perform under volatile conditions or unusual data distributions. This process is particularly relevant when dealing with limited or uncertain data, where bootstrapping methods can amplify the understanding of uncertainty and variability.

**GLM’s and Random Forest**

Generalized Linear Models (GLMs) and Random Forests are two predictive modeling techniques that are widely used in insurance. GLMs are flexible extensions of linear regression models that allow for non-linear relationships and non-normal error distributions, making them well-suited for modeling claims data and estimating insurance reserves. On the other hand, Random Forests, a powerful ensemble machine learning method, provides a non-parametric approach to predictive modeling by combining multiple decision trees. This results in a robust model capable of capturing complex patterns and interactions in the data.

**Methodology and Examples**

**Understanding Bootstrapping Paired with the BF Method**

First, we wanted to examine the differences in results when applying the BF and bootstrapping methods to datasets of different sizes. The BF method is particularly valued in the actuarial field for its ability to handle limited or sparse data, which is crucial in situations where data availability is a challenge—think of low-population-density areas. To test this quality, we created two claims’ triangles from real insurance company data:

* The second claims triangle used a sample of fewer than 1,000 data points, simulating conditions in low-population-density areas where data tends to be sparse.
* The first claims triangle was based on a larger sample size, with more than 1,000 data points, reflecting high-population-density regions where more claims data is available.

We then compared the reserve estimates generated by the BF method on its own and in combination with bootstrapping. To illustrate this, we plotted box plots showing the range of reserve estimates for both dataset sizes. The results for the large sample size are displayed on the left, and those for the small sample size are on the right. **Pictures(1,2).**

A graph of a comparison of bootstrapping reserves

Description automatically generated***A graph with a bar chart and a diagram

Description automatically generated with medium confidence***

Picture 2

Picture 1

One of the most notable takeaways from the large sample size is how the BF method produces relatively stable reserve estimates, as seen in the tight whiskers of the box plots. This stability is a hallmark of the BF method and demonstrates its reliability when data is abundant. When we applied bootstrapping to this large dataset, the range of reserve estimates widened, which is shown by the box plot becoming taller and broader. Bootstrapping, as you may know, involves resampling the data multiple times to generate a distribution of possible outcomes, providing a better sense of the variability in the estimates.

Looking at the smaller dataset on the right-hand side of the graph, we observed a far greater discrepancy between the bootstrapped and non-bootstrapped reserve estimates. This discrepancy highlights an important limitation of the BF method when working with sparse data. While the BF method maintains its stable output, it doesn't fully capture the variability inherent in limited data. Bootstrapping, however, compensates for this limitation by incorporating the uncertainty into the reserve range, thereby improving the actuary’s judgment. Essentially, this highlights how the BF method might underestimate reserves when data is sparse, whereas the addition of bootstrapping provides a more realistic, risk-aware estimate.

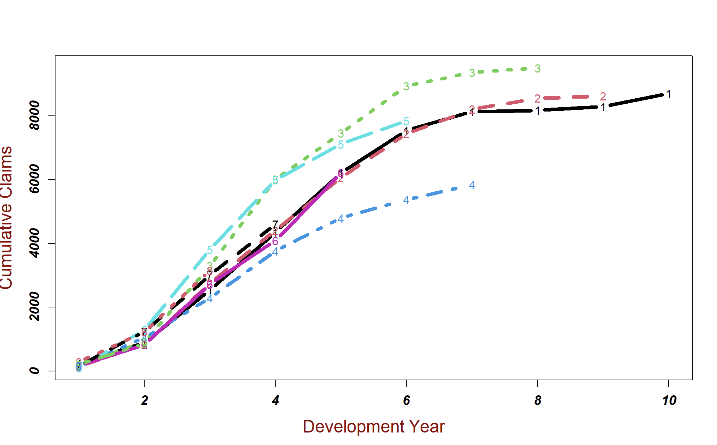
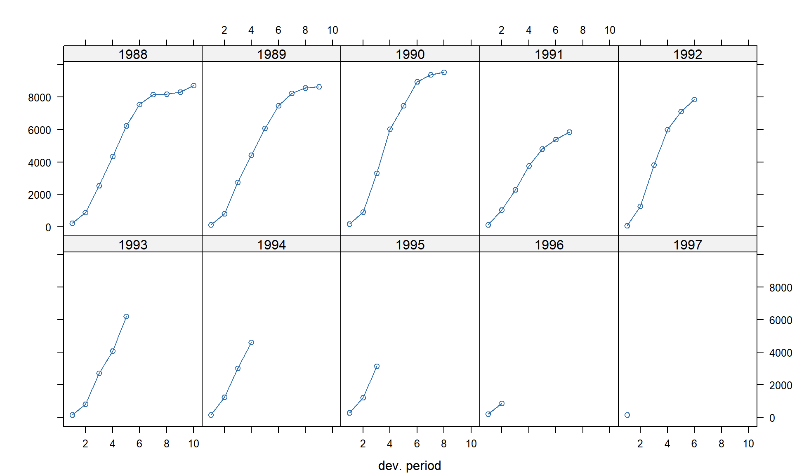
**Interpretations of Chain-Ladder Method Combined with Bootstrapping**

Next, we extended our analysis by exploring the application of bootstrapping with the chain-ladder method. For this part, we utilized a dataset provided by the CAS website, which also happens to be a standard sample dataset for the chain-ladder method in R-Studio. From this, we constructed a claims triangle. **Picture(3)*.***

***A screenshot of a computer screen

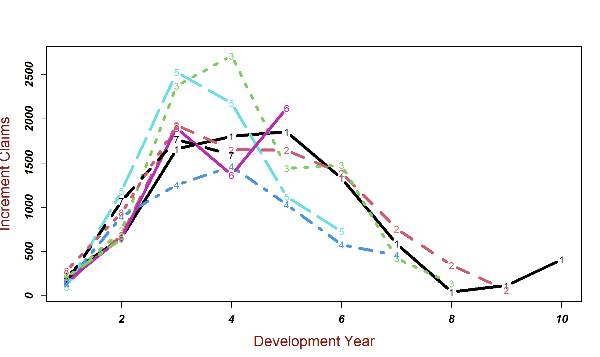
Description automatically generated***

Picture 3

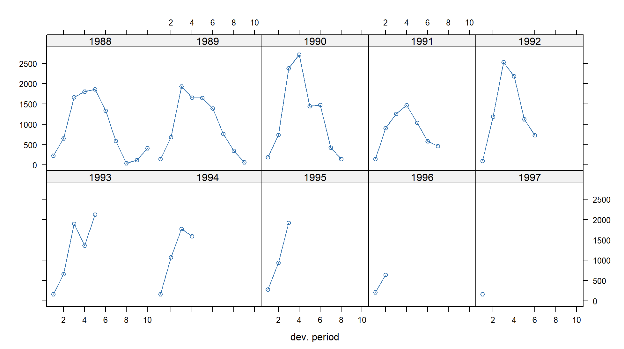
******To better grasp the data’s behavior, we plotted the cumulative claims across accident years. This visualization revealed that most claims take six years to settle. Interestingly, the growth rate of claims accelerates significantly in the first three years, before eventually leveling off as claims close out. This pattern is essential for actuaries because it provides insights into the lifecycle of claims and the time periods where reserves need the most attention. **Pictures (4,5).**

Picture 5

Picture 4

******We also plotted the incremental increases in claims—essentially the differences between cumulative claims from one development year to the next. This graph showed that across accident years, the highest incremental increases typically occur between the third and fourth development years. This consistency in claim development trends confirms the validity of using the chain-ladder method, paired with bootstrapping, to estimate reserves in this dataset. **Pictures(6,7).**

Picture 6

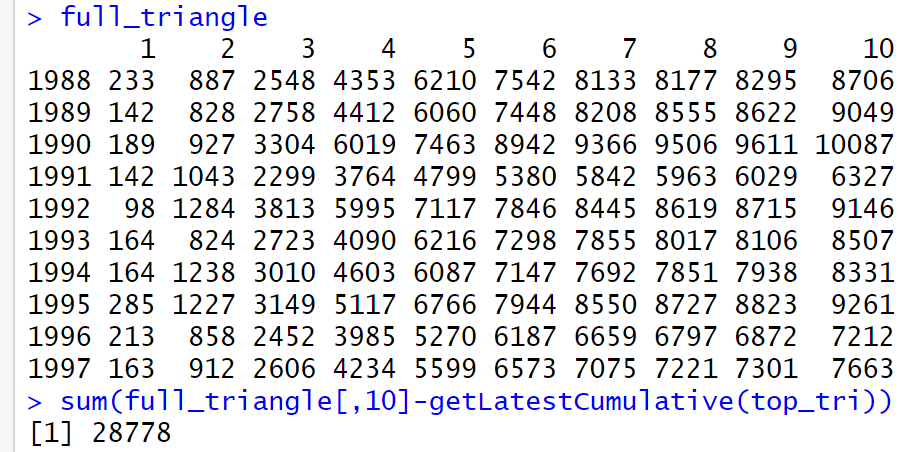
******

Picture 7

The value of these graphs lies in their ability to highlight predictable trends in claim development. These trends reassure us that development factors remain stable, providing a solid foundation for applying our chosen methods. Since the data is well structured, Combining the chain-ladder method with bootstrapping allowed us to incorporate a more robust reserve estimate.

Our findings underscore the importance of pairing actuarial methods like the BF method or chain-ladder with bootstrapping. The BF method shines in scenarios with limited data but benefits significantly from bootstrapping when data variability is high. Meanwhile, the chain-ladder method, with its ability to capture predictable claims trends, is further enhanced by bootstrapping, which provides a more nuanced picture of reserve ranges. Ultimately, these combinations allow actuaries to make more informed decisions, balancing stability with a realistic assessment of risk and uncertainty.

The chain-ladder method resulted in the following ultimate claims table. ***Picture(8).***

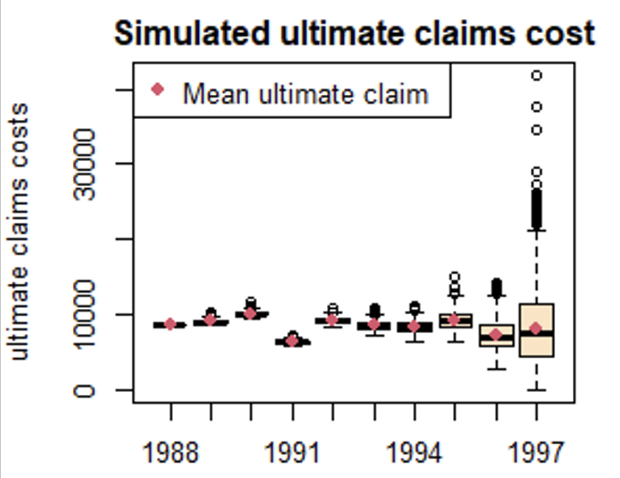
******

Picture 8

The overall estimated IBNR (Incurred but not reported) is $28,778 (in thousand).

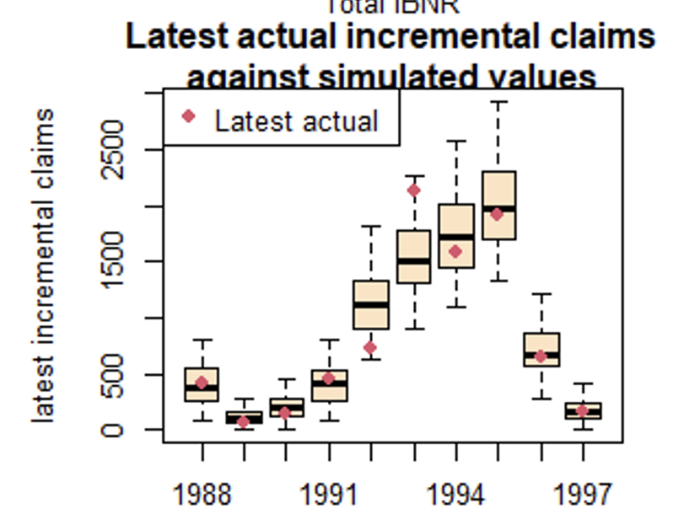
Next, we combined bootstrapping with chain ladder.

A simple chain ladder method is used to calculate the Pearson residuals (representing random fluctuations around the fitted model), and then bootstrapping the residuals to generate multiple incremental claim pseudo-triangles. Each resampling will generate a new dataset using randomly sampled residuals to fit the values. These many, many triangles simulate alternative outcomes for future claims payments.

******This package gives us some charts. The graph on the left shows the final claim cost distribution for each accident year in all guided simulations. Each boxplot shows the change in the final claim cost in the lead simulation for a particular year, with the red dots representing the average final claim for that year. The height of each boxplot indicates the change in the total final claim cost for that accident year. For example, later accident years tend to be more variable due to greater uncertainty. The red dot locates in the center of most boxplots, indicating that the bootstrap distribution is fairly symmetrical for each year. ***Picture(9)***

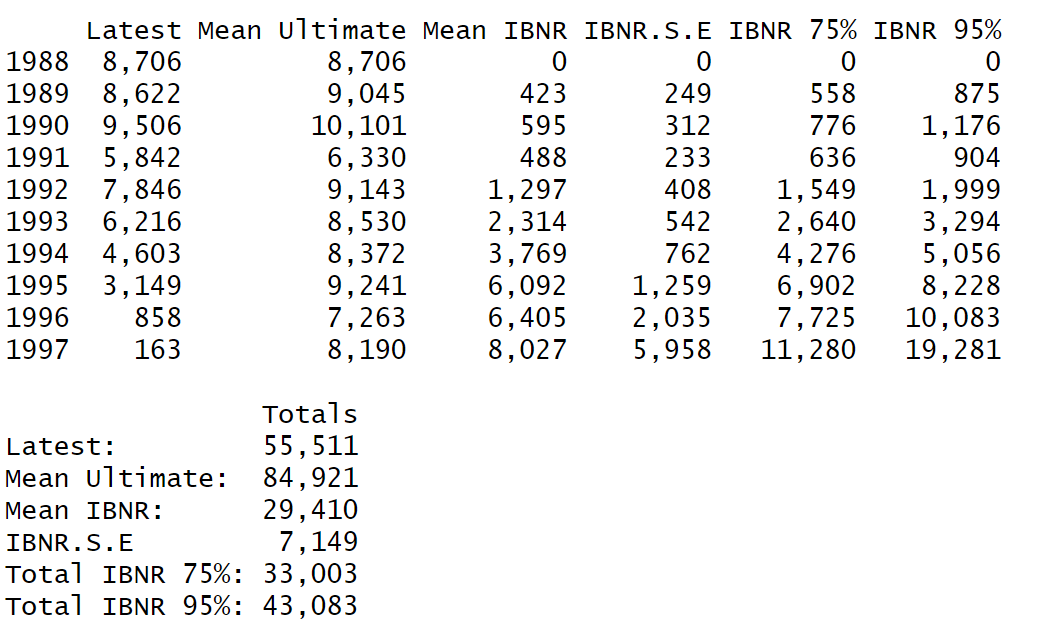
Picture 9

The graph on the right compares the actual incremental claims with the simulated incremental claims distribution for each accident year. Each box plot shows the distribution of simulated incremental claims for the latest development year for each accident year. The boxplot for early origin years is narrower because the claims are more well-established and less uncertain. In years with high growth rates, the red dots deviate from the central range of the boxplot, indicating that claims may be under - or overestimated at this time. ***Picture(10)***

******

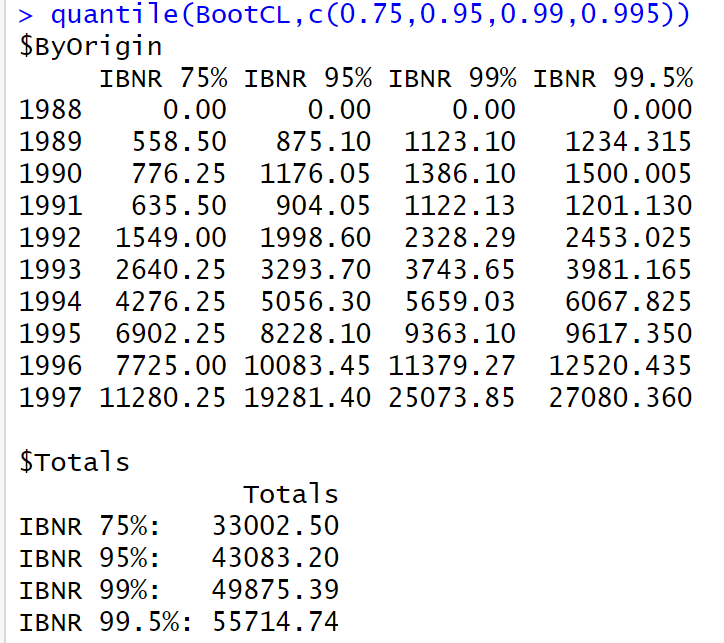
Picture 10

Combining bootstrapping with the simple chain ladder method can provide a distribution of reserves rather than just a single point estimate, and bootstrapping requires fewer stringent assumptions than other models. In addition, this method provides a range of reserves, such as percentiles ***Picture(11).***

******

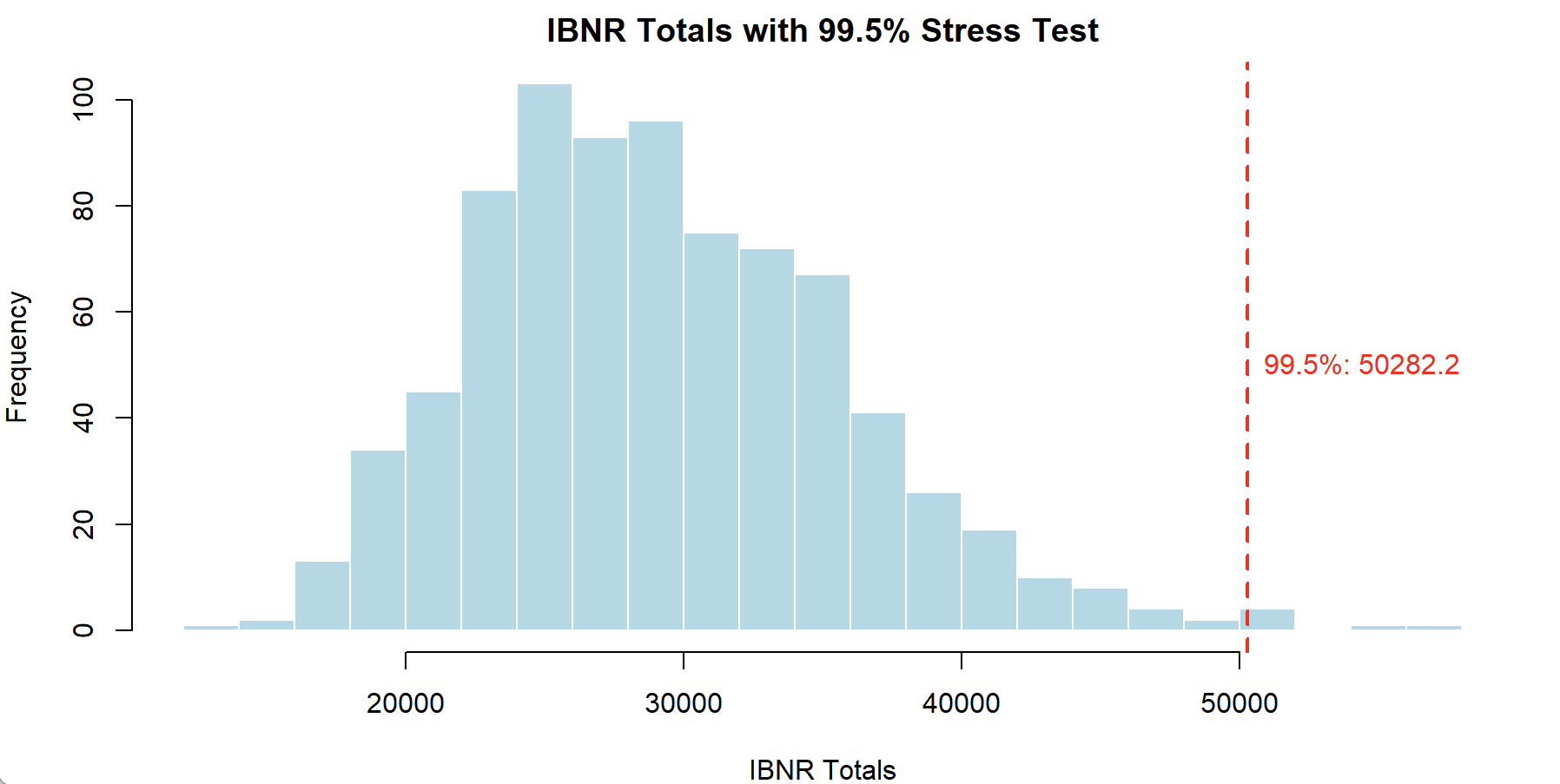
Picture 11

We can do more with these percentiles, like stress testing ***Picture(12)***,



Picture 12

which is the amount of capital that insurance companies hold against extreme conditions. For example, Solvency ll under the EEA (European Economic Area); Insurance company are required to hold enough capital to cover adverse outcomes within 99.5% confidence level for a once-in-200-year event. ***Picture(13)***

******

Picture 13

# **Insurance Claim Prediction: Random Forest vs. GLM**

We aim to compare two different models—Random Forest and Generalized Linear Model (GLM)—for predicting insurance claim amounts. Insurance data often features a wide range of claim amounts, from minor claims to very large payouts, which makes choosing the right analytical method crucial. We specifically selected Random Forest because it incorporates a technique called bootstrapping, which enhances model stability and prediction accuracy. This aligns well with the characteristics of insurance data and the analytical needs of the industry.

## **Data Overview**

The dataset used in this analysis is called “3years\_Claims,” containing 14,174 records. The target variable is `claim\_amount`, which represents the amount claimed by a customer during an insurance claim. The statistical summary of the data is as follows:  
- **Mean**: 1187.76  
- **Standard Deviation**: 1586.39  
- **Minimum Value**: 0.02  
- **Maximum Value**: 90,433.60  
  
These numbers reveal a wide distribution of claim amounts, ranging from very small claims to extreme outliers. This complexity poses significant challenges for predictive models.

## **Methods**

**Random Forest**

Random Forest is a method that combines the predictions of multiple decision trees to arrive at a more reliable and accurate prediction. Think of each tree as an independent analyst, and Random Forest aggregates their opinions to deliver a result. Here’s how it works:

* 1. Data Sampling (Bootstrapping):  
  Random Forest randomly selects subsets of data (with replacement) to train each decision tree. This process allows the model to analyze different parts of the dataset from various perspectives.
* 2. Building Decision Trees:  
   - Each tree creates its own 'rules' to predict the outcome based on the sampled data. For example, it might use customer age, region, or vehicle type to estimate claim amounts.
* 3. Combining Results:  
   - After all the trees make their predictions, Random Forest combines them (e.g., by averaging for regression tasks or majority voting for classification) to produce the final prediction.

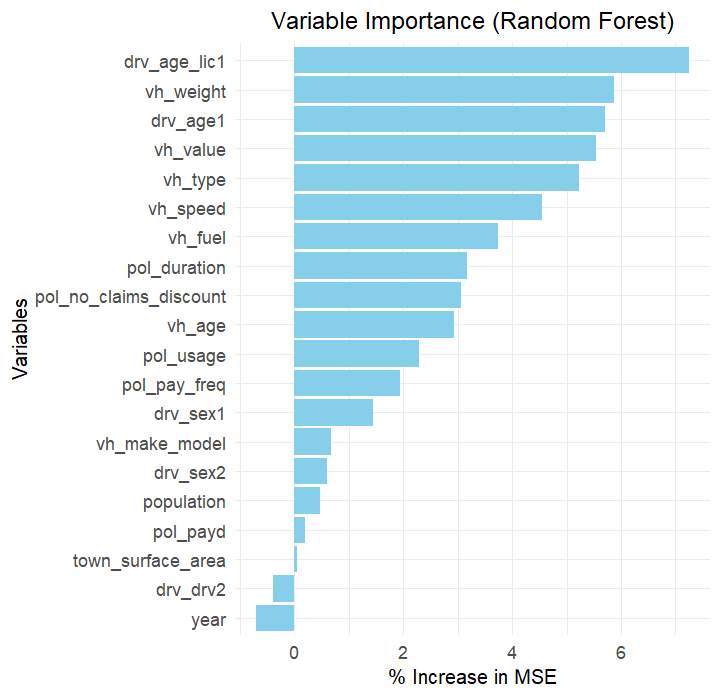
Role of Bootstrapping:  
Bootstrapping is the backbone of Random Forest. It ensures model diversity and stability by training each tree on a different subset of the data. Its benefits include:

* Reducing Overfitting: Bootstrapping minimizes reliance on specific data points, making the model more robust.
* Handling Complex Data: By analyzing multiple subsets, the model captures diverse patterns, even in irregular or complex distributions.
* Improving Stability: The ensemble nature of Random Forest smoothens the impact of outliers or noisy samples, resulting in more consistent predictions.

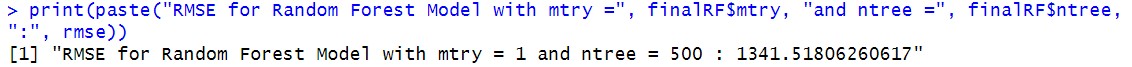
Variable Importance Analysis:  
To further understand which factors have the most significant impact on predicting *`claim\_amount*`, we generated a variable importance chart from the Random Forest model (Picture 14). The chart illustrates the percentage increase in Mean Squared Error (MSE) when each variable is excluded, indicating its importance in the prediction.

**Key Variables. Picture(14)**:

* **drv\_age\_lic1** (driver's age and years licensed): The most influential variable, indicating the importance of driving experience.
* **vh\_weight** (vehicle weight): The second most important factor, showing its critical role in predicting claim amounts.
* **drv\_age1** (driver’s age): Reflects age-related risk behaviors associated with claims.
* **vh\_value** (vehicle value): Demonstrates the relevance of a vehicle's worth in claims.

****

Picture 14

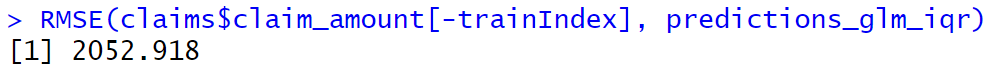
****Results:  
The Random Forest model achieved a Root Mean Squared Error (RMSE) of 1341.52, demonstrating its ability to handle variability and outliers effectively. **Picture(15).**

Picture 15

### **Generalized Linear Model (GLM)**

Principle:  
GLM is a traditional statistical model that aims to establish a mathematical relationship between variables. It can be thought of as creating a formula to predict the target variable. For example:  
*`Claim Amount = (****drv\_age\_lic1*** *× Coefficient\_1) + (****vh\_weight*** *× Coefficient\_2) + … + Constant`*

The model determines the coefficients automatically based on the data to minimize prediction errors.

****Results:  
The GLM model achieved an RMSE of 2052, significantly higher than Random Forest. This suggests that GLM was less effective in capturing the complexity and variability of the data. **Picture(16).**

Picture 16

## **Results Comparison**

Model Performance Summary:

* Random Forest: RMSE of 1341.52, indicating superior performance in handling complex distributions and extreme values.
* GLM: RMSE of 2052, showing limited ability to handle wide-ranging and irregular data distributions.

## **Findings**

This project compared Random Forest and GLM in predicting insurance claim amounts. Random Forest, with its use of bootstrapping and decision tree ensembles, demonstrated exceptional adaptability and accuracy, particularly in dealing with extensive data variability and outliers. In contrast, GLM, while offering a straightforward and interpretable formula, struggled with the complexity of the data.  
  
The results highlight the importance of bootstrapping in enhancing model performance. Random Forest’s flexibility and stability make it a powerful tool for the insurance industry, especially when analyzing complex claim patterns. These findings provide valuable insights for insurers looking to refine their predictive models and better understand the factors influencing claims.

**Conclusion**

This report highlights the essential role of bootstrapping in enhancing claims reserving and predictive modeling within the insurance industry. Bootstrapping's ability to generate multiple pseudo datasets allows actuaries to assess variability, quantify uncertainty, and improve the robustness of reserve estimates. When paired with methods like the Bornhuetter-Ferguson (BF) and chain-ladder, bootstrapping addresses the limitations of relying solely on historical data, particularly in scenarios with sparse or volatile datasets.

As a group, we found the implications of bootstrapping to insurance models to be extremely valuable as a seemingly intuitive method adds a large amount of value to the quality of predictive models. As bootstrapping allows for a wide range of flexibility when it comes to enhancing data analysis in several areas, one being stress testing, where actuaries can be flexible and observe ranges of different scenarios it is no surprise that bootstrapping is solidified as a staple in actuarial practice.

Our analysis demonstrated that while traditional models like the BF method and chain-ladder excel in their respective domains, integrating bootstrapping significantly enhances their accuracy and adaptability. Additionally, Random Forest's built-in use of bootstrapping highlights its advantage over Generalized Linear Models (GLMs) in handling complex data distributions and outliers, making it a powerful tool for predictive modeling. These findings underscore bootstrapping’s importance as a versatile and widely used methodology in modern actuarial practices.

Some potential next steps and questions that we have are:

1. Stress Testing: Expand the use of stress testing techniques to evaluate reserve estimates under extreme scenarios, such as changes in claim frequency or severity due to catastrophic events.
2. Enhanced Data Analysis: Incorporate additional variables and explore advanced feature engineering techniques to improve the predictive power of Random Forest and GLMs. This could include deeper analysis of external factors such as economic trends and regulatory changes.
3. Model Comparison in Broader Contexts: Test the performance of bootstrapping with other reserving models, such as Bayesian methods, to explore its effectiveness across different contexts.

Other applications of bootstrapping in the insurance industry:

*1. Reinsurance optimization*

Reinsurance optimization involves structuring treaties in a way that balances cost efficiency with effective risk transfer. By leveraging methods like bootstrapping claims data, insurers can simulate potential outcomes under various reinsurance structures, evaluating critical metrics such as ceded premiums, retained losses, and recoveries. This data-driven approach empowers insurers to tailor reinsurance programs that align with their financial objectives and risk appetite.

*2. Fraud detection*

Fraud detection is essential for safeguarding the integrity of the insurance process. By employing techniques such as bootstrapping historical claims data to identify normal patterns, insurers can effectively flag anomalies that suggest fraudulent activity. Comparing new claims against established patterns allows insurers to uncover irregularities, enhancing their ability to investigate and mitigate fraudulent behavior. Together, these strategies not only protect insurers from financial losses but also improve overall operational efficiency, fostering a more secure and trustworthy insurance environment.

***REFERENCE***

***Data from:***

1. [*https://www.kaggle.com/datasets/mastmustu/insurance-claims-fraud-data?resource=download*](https://www.kaggle.com/datasets/mastmustu/insurance-claims-fraud-data?resource=download)
2. *Maynard, Trevor, Nigel De Silva, Richard Holloway, Markus Gesmann, Sie Lau, and John Harnett. 2006. “An Actuarial Toolkit. Introducing The Toolkit Manifesto.”*[*https://www.actuaries.org.uk/system/files/documents/pdf/actuarial-toolkit.pdf*](https://www.actuaries.org.uk/system/files/documents/pdf/actuarial-toolkit.pdf)
3. *Alessandro Carrato, Fabio Concina, Markus Gesmann, Dan Murphy, Mario Wüthrich and Wayne Zhang*

*https://cran.rstudio.com/web/packages/ChainLadder/vignettes/ChainLadder.html#references*

1. *Glenn Meyers*

[*https://www.casact.org/sites/default/files/2021-02/01-Meyers.PDF*](https://www.casact.org/sites/default/files/2021-02/01-Meyers.PDF)

1. *Paulo Pinheiro, Joao Andrade e Silva, Maria de Lourdes Centeno. Bootstrap Methodology in Claim Reserbing*

*https://www.actuaries.org/astin/colloquia/washington/pinheiro\_silva\_centeno.pdf*