Premium estimation in the fire insurance through semiparametric bootstrap

W F Adnan¹, U S Pasaribu¹ and H. Husniah²*

¹Actuarial Science, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Jl. Ganesha 10/12 Bandung 40132, Indonesia
²Faculty of Industrial Engineering, Langlangbuana University, Jl. Karapitan 116 Bandung 40261, Indonesia

*h.husniah@unla.ac.id

Abstract. Along with the development of information, science and technology, there is a quite popular developing resampling method, namely bootstrapping. Bootstrap estimates asymptotically against its original value (observation). Thus, the greater bootstrap replication, the resample distribution will be normally distributed. It indicates that the bootstrap estimator gives better results. Based on the goodness-of-fit test by using Kolmogorov-Smirnov test, the severity on fire insurance data follow Weibull 2 parameter distribution. A case study is conducted on reinsurance company’s data for shopping centre’s fire. It is a big data. Since it is a reinsurance company’s data, the data completeness may be inadequate. It causes the severity claim to be processed using semiparametric bootstrap.

1. Introduction

Bootstrap, introduced by Bradley Efron (1979), is a development of the jackknife method coined by Quenouille (1949). Then, it was refined by John Tukey (1958) [1]. The bootstrap method is a resampling technique with n sample size with replacement. At larger sample sizes \( n \geq 20 \), it is easier to implement on a computer. Bootstrap allows researchers to assess the statistical accuracy of complicated procedures and provides an answer where no analytical answer can be obtained [2]. This method is used due to the consideration that the data is not easy to obtain and expensive. So, it is good to use all the available information in resampling. Moreover, along with the development of science and technology, which is not as expensive and slow as in the 18th century. Thus, this method would be easier to use.

Reinsurance company’s data is a big data. Since it is a reinsurance company’s data, the data completeness may be inadequate. This does not guarantee that the data follow the same distribution. Thus, a suitable method to be applied to severity data is semiparametric bootstrapping. It is performed like nonparametric bootstrapping, in some cases noise been added (https://online.stat.psu.edu/stat555/node/119), for example in residual bootstrap. This method was used by Monchuk et al. [3] in analyzing cases of production inefficiency in China’s agriculture by describing the factors with 1000 times bootstrap replication (B).

2. Method

The data is a real data obtained from a reinsurance company in Indonesia. It consists of insured’s information in fire insurance for shopping centre in 2006-2018 given several claim criteria. The method applied to obtain a severity estimation through semiparametric bootstrap, along with an algorithm scheme as shown in Figure 1.
There are several principles in calculating the risk premium\[4\]. To obtain risk premium, the security loading is needed where each insurance company’s security loading can be differ one to another. Reinsurance company’s data consists of data from several insurance companies, so that, the authors prefer estimating the pure premium to risk premium. Resampling with replacement from the method causes the sample from the semiparametric bootstrap to be independent. Thus, the pure premium can be estimated with the formulas [5]:

$$E[S_n] = E[N] E[X],$$

where $N$ denotes the discrete random variable of claim frequency and $X$ denotes the continuous random variable of severity.

**Figure 1.** Scheme of bootstrap semiparametric algorithm in fire insurance data processing.

### 3. Result and Discussion

There are four random variables that influenced the claim criteria for the insured data of fire insurance. There are severity, occurrence date, claim date, and the policy expiration date. The total of fire insurance insured is 179,892. After implementing occurrence date and claim date’s criteria, the amount of data become 2,143. Subsequently, implementing to the policy expiration date’s criteria, the amount of data become 2,110. The last, implementing severity’s criteria, the amount of data become 1,953. So, the amount of insured data on fire insurance $n = 1,953$. So that, the data is quite large, $4\times1,953$. There are two information processed, namely occurrence date and severity. Semiparametric bootstrap is implemented in severity. Data processing was performed by using R software.

#### 3.1. Overall Severity Information

Based on Figure 2, the largest severity of shopping centre’s fire is occurred in the Indonesian capital city, Jakarta. So, it is very reasonable for insurer to pay the severity that much. Figure 2 also showed that the data has outliers. Detection outliers can be obtained by using formulas [6],

Lower Fence (LF) = $Q1 - (\text{multiplier} \times \text{IQR})$ and Upper Fence (UF) = $Q3 + (\text{multiplier} \times \text{IQR})$  \(2\)

Where IQR = $Q3 - Q1$ and the multiplier’s factor 1.5 applied for normal outliers, meanwhile the multiplier’s factor 3 applied for extreme outliers.

LF and UF with multiplier’s factor 1.5 called as inner fence, meanwhile LF and UF with multiplier’s factor 3 called as outer fence. The observations between an inner fence and nearby outer fence known as “outside” potential outliers, and anything beyond outer fences as “far out” problematic outliers or probable outliers (http://cc.oregon.edu/cnews/spring2000/outliers.html). In this case study, by using (2), there are 25 possible outliers and 254 probable outliers, in total 279 outliers. However, these are not indicating that the data is bad and should be removed / treated. Because it is the severity which is crucial information.
Figure 2. Boxplot and stem-leaf plot for severity in fire insurance for shopping centre in 2006-2018 where stem unit is $10^8$ and leaf unit is $10^7$.

Based on Figure 3, there are 215 insured with no claim during the observation. Therefore, non-zero claim after sorting the data can be found in $\frac{216}{1953} \approx 11\%$, at the 11th percentile where the severity IDR 1,384,240. As the total data $n = 1,953$, then the 99th percentile consists of 1,933 data. It is clear that the frequency claims of the last 1% of the data are 20. This indicates that severity with the large values are exist. Figure 3 also shows that there is accumulate data in the tail, and the data tends to positively skewed, proved by $\gamma_3 = 5.1$. The distribution has many data around small values, shown by 10% of the data is zero. This indicates a very large kurtosis, indicated by a kurtosis value of more than 3, $\gamma_4 = 31.8$.

**Table 1.** The results of goodness-of-fit test on the overall data using the Kolmogorov-Smirnov test and the Maximum Likelihood Estimator (MLE) method.

| Distribution       | Weibull 2P | Pareto (Type II, Lomax) | Log-Logistik |
|--------------------|------------|--------------------------|--------------|
| Statistics         | 0.13739    | 0.11009                  | 0.12421      |
| P-Value            | 0.3714     | 0.0275                   | 0.1092       |
| AIC                | 17,744     | 17,688                   | 19,532       |
| BIC                | 17,752     | 17,696                   | 19,360       |
| Parameter estimation: |
| a. Shape           | 0.53873    | $1.0042 \times 10^7$    | 0.94816      |
| b. Scale           | $1.0046 \times 10^8$ | 0.61465 | $3.6202 \times 10^7$ |

Weibull distribution is accepted where the significance level $\alpha = 0.1$. Based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, the smallest value is in
the Weibull distribution. Then the Weibull distribution fits the overall severity distribution better than the other distributions tested. So that the assumption for the overall data uses the Weibull \( (\hat{\theta}, \hat{\xi}) \) distribution with the scale and shape parameter estimators 1.0046 \( \times 10^8 \) and 0.53873, respectively.

\[
\bar{x} \pm 1.839 \times 10^0, \hat{\mu} \pm 1.7669 \times 10^0, n = 1,953
\]

Figure 4. Fitting distribution using Weibull\((\hat{\theta}, \hat{\xi})\) of severity and its parameter estimation. The value on the brackets are the theoretical estimates.

Due to amount of data \( n = 1,953 \), so the minimum number of classes on the histogram is 12. Based on Figure 4, the results of the statistical and theoretical estimates differ considerably. This assures the author that the data is not fully follow the distribution. Since it is a reinsurance company’s data, completeness may be inadequate. This does not guarantee that the data follow the same distribution. Thus, a suitable method to be applied to severity data is semiparametric bootstrapping.

3.2. Semiparametric Bootstrap

The most important components needed in the use of the bootstrap method are determining bootstrap replication \( (B) \) and sample size \( (n) \). Estimating parameters by using bootstrap can be applied in various fields, such as in psychology [7], biology [8], and health [9]. Bootstrap replication determinations vary depending on needs. Considering the 'thumbs rule', which is \( B \geq 20 \) is already informative [2]. So that the bootstrap replication \( B = \{20, 30, 40, 50\} \) where \( r \) represents the repeated bootstrap. The greater \( B \), the closer bootstrap estimator to the normal distribution and tend to converge [2]. The purpose of doing repeated bootstrap is to show the consistency of the bootstrap estimator. Actually, the parameters that can be considered in the bootstrap are skewness and kurtosis because of their robustness. Whereas mean and variance are less suitable to be put attention because the data have extreme values (zero and very large values).

The repeated bootstrap, denoted by \( r \), shows consistency of the bootstrap replication \( (B) \). From Figure 5, it is obvious \( B = 50 \) has the smallest interval length of the bootstrap estimator of mean. It indicates that it is a good bootstrap estimator. From Figure 6, it is obvious \( B = 40 \) has the smallest interval length of the bootstrap estimator of variance. But it widens at \( B = 50 \). So, the bootstrap estimator for the variance may converges at \( B > 50 \). From Figure 7, it is obvious \( B = 50 \) has the smallest interval length of the bootstrap estimator of skewness. This suggests that large values are often resampled and convinces the authors that the data follows the assumed distribution, the Weibull 2-Parameter. However, the histogram seems still not tend to the normal distribution. From Figure 8, \( B = 30 \) has the smallest interval length of the bootstrap estimator of kurtosis. However, the getting larger \( B \) is has not shown convergent on the bootstrap estimator. The histogram seems still not tend to the normal distribution.
| $B$ | $\hat{\mu}^*$ | $\hat{\sigma}^*$ | $\gamma$ | $\hat{\gamma}$ |
|-----|----------------|-----------------|--------|-------------|
| 20  | $1.814; 1.851 \times 10^8$ | $2.427; 2.536 \times 10^{17}$ | 5.1  | 5.7  |
| 30  | $1.813; 1.851 \times 10^8$ | $2.385; 2.488 \times 10^{17}$ | 5.1  | 5.7  |
| 40  | $1.793; 1.859 \times 10^8$ | $2.436; 2.456 \times 10^{17}$ | 5.1  | 5.7  |
| 50  | $1.831; 1.839 \times 10^8$ | $2.462; 2.515 \times 10^{17}$ | 5.1  | 5.7  |

Figure 5. Bootstrap estimator of mean.

Figure 6. Bootstrap estimator of variance.
**Figure 7.** Bootstrap estimator of skewness.

| $\hat{\gamma}_4$ | $\hat{\gamma}_3 = (5.01; 5.09)$ | $\hat{\gamma}_3 = (5.00; 5.09)$ | $\hat{\gamma}_3 = (5.06; 5.16)$ | $\hat{\gamma}_3 = (5.07; 5.09)$ |
|-------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| $\pi = 1$         | ![Histogram](image1)             | ![Histogram](image2)             | ![Histogram](image3)             | ![Histogram](image4)             |
| $\pi = 2$         | ![Histogram](image5)             | ![Histogram](image6)             | ![Histogram](image7)             | ![Histogram](image8)             |
| $\pi = 3$         | ![Histogram](image9)             | ![Histogram](image10)            | ![Histogram](image11)            | ![Histogram](image12)            |

**Figure 8.** Bootstrap estimator of kurtosis.

| $\gamma_4 = 31.8, \hat{\gamma}_4 = 62.7$ | $n = 1,953$ |
|------------------------------------------|-------------|
| $B = 20$                                 | $\hat{\gamma}_4^* = (29.4; 33.1)$ | $\hat{\gamma}_4^* = (31.7; 31.8)$ | $\hat{\gamma}_4^* = (31.8; 32.9)$ | $\hat{\gamma}_4^* = (30.1; 31.8)$ |
| $\pi = 1$                                | ![Histogram](image13)              | ![Histogram](image14)              | ![Histogram](image15)              | ![Histogram](image16)              |
| $\pi = 2$                                | ![Histogram](image17)              | ![Histogram](image18)              | ![Histogram](image19)              | ![Histogram](image20)              |
| $\pi = 3$                                | ![Histogram](image21)              | ![Histogram](image22)              | ![Histogram](image23)              | ![Histogram](image24)              |
3.3. Claim Frequency Estimation

Suppose $Z$ denotes the random variable fire inter-occurrence date. Inter-arrival time principal is applied to inter-occurrence date. So, $Z^{(j)}, j = 1, 2, 3, ..., k$ is the random variable follows Exponential distribution independently and identically where the mean $\frac{1}{\lambda}$ [10]. The relationship between the Exponential distribution and the Poisson distribution can be obtained. Furthermore, the Poisson distribution is used to claim frequency modelling.

$Z \sim \exp(\lambda)$ where $E[Z] = \frac{1}{\lambda}$, so $\lambda = \frac{1}{\mu} = E[N]$. 

Accordingly, the inter-occurrence date as follows,

$$E[Z] = \mu = \frac{1}{\lambda} = \frac{4,680 \text{ day}}{1,953 \text{ occurrence}} = 2.398 \text{ or } \mu = \frac{1}{\lambda} \times 365 \text{ day} = \frac{1}{2.398} \times 365 \text{ day} \approx 152.2393 \text{ or } E[N] \approx 152 \text{ claims per year obtained.} \quad (2)$$

3.4. Premium Estimation

By using (1) and the mean bootstrap estimator closest to the normal distribution as in Figure 5, that is $B = 50$. So, as $E[X] = \hat{\mu}^* = (1.831; 1.839) \times 10^8$ and estimation of claim frequency $E[N] = 152$. By using (1) estimation of the pure premium $E[S_N] = IDR (2.783; 2.795) \times 10^{10}$ can be obtained. Validated with the simulation results of the premium calculations done by the Bandung Institute of Technology’s Statistic Research Group for fire insurance in shopping centre, denotes by $\omega$. Their pure premium is $\omega = IDR 1.109 \times 10^{11}$. 

4. Conclusion

Based on the semiparametric bootstrap simulation, the severity of claims in fire insurance for shopping centre in 2006-2018 follow Weibull 2P distribution. So interval of the pure premium can be obtained. Comparing to the simulation results of the premium calculations done by the Bandung Institute of Technology’s Statistic Research Group, the calculated pure premium has an underestimate value. Therefore, in the future study, bootstrap can be implemented to the claim frequency as well.

Acknowledgements

The authors thank to PT Reasuransi Indonesia Utama who provided crucial information even though it is confidential, to the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia who has funded the work through the scheme of Penelitian Terapan 2020 to HH with contract number 9/E1/Kpt/2020, and to the reviewers for the valuable review for this paper.

References

[1] Teugels J L, and Sundt B 2006 Encyclopedia of Actuarial Science (New York: WILEY)
[2] Efron B and Tibshirani R 1993 An Introduction to the Bootstrap (New York: Chapman and Hall / CRC)
[3] Monchuk D C, Chen Z, and Bonaparte Y 2010 Explaining production inefficiency in China’s agriculture using data envelopment analysis and semi-parametric bootstrapping China Economic Review 21 346–354

[4] Gray R J and Pitts S M 2012 Risk Modelling in General Insurance (Cambridge: Cambridge University Press)

[5] Klugman S A, Panjer H H and Willmot GE 2012 Loss Models: From Data to Decisions (New York: John Wiley & Sons)

[6] Hogg R V, McKean J W, and Craig A T 2012 Introduction to Mathematical Statistics 7th ed. (Boston: Pearson Education)

[7] Wright D B, London K, and Field A P 2011 Using Bootstrap Estimation and the Plug-in Principle for Clinical Psychology Data Journal of Experimental Psychopathology 2 252–270

[8] Calmettes G, Drummond G B, and Vowler S L 2012 Making do with what we have: use your bootstraps: Editorial British Journal of Pharmacology 167 233–237

[9] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman J M, Yan P, and Chowell G 2020 Short-term Forecasts of the COVID-19 Epidemic in Guangdong and Zhejiang, China: February 13–23, 2020 Journal of Clinical Medicine 9 596

[10] Ross S M 1996 Stochastic Processes (New York: John Wiley & Sons, Inc.)