safety

MDPI

Article

Analyzing Large Workers' Compensation Claims Using Generalized Linear Models and Monte Carlo Simulation

Fatemeh Davoudi Kakhki [1,*], Steven A. Freeman [2] and Gretchen A. Mosher [2]

[1] Department of Aviation & Technology, San Jose State University, San Jose, CA 95192, USA
[2] Agricultural & Biosystems Engineering, Iowa State University, Ames, IA 50011, USA; sfreeman@iastate.edu (S.A.F.); gamosher@iastate.edu (G.A.M.)
* Correspondence: fatemeh.davoudi@sjsu.edu

Received: 12 October 2018; Accepted: 21 November 2018; Published: 1 December 2018

Abstract: Insurance practitioners rely on statistical models to predict future claims in order to provide financial protection. Proper predictive statistical modeling is more challenging when analyzing claims with lower frequency, but high costs. The paper investigated the use of predictive generalized linear models (GLMs) to address this challenge. Workers' compensation claims with costs equal to or more than US$100,000 were analyzed in agribusiness industries in the Midwest of the USA from 2008 to 2016. Predictive GLMs were built with gamma, Weibull, and lognormal distributions using the lasso penalization method. Monte Carlo simulation models were developed to check the performance of predictive models in cost estimation. The results show that the GLM with gamma distribution has the highest predictivity power (R^2 = 0.79). Injury characteristics and worker's occupation were predictive of large claims' occurrence and costs. The conclusions of this study are useful in modifying and estimating insurance pricing within high-risk agribusiness industries. The approach of this study can be used as a framework to forecast workers' compensation claims amounts with rare, high-cost events in other industries. This work is useful for insurance practitioners concerned with statistical and predictive modeling in financial risk analysis.

Keywords: predictive generalized linear models; heavy-tailed distributions; Monte Carlo simulation; insurance risk analysis

1. Introduction

According to the National Council on Compensation Insurance Report (2015), among several types of insurance policies, workers' compensation insurance is considered as a unique line of business since it balances the interests of many system stakeholders to protect and retain their jobs, from injured employees and their families to employers, medical providers, insurance companies, regulators, and states. There are three main types of workers' compensation claims: medical only, temporary disability, and permanent disability, among which the greatest costs are imposed by permanent disability [1]. As an economic remediation industry with claims payments as the main cash flow, the insurance industry provides a means of decreasing monetary loss by spreading or pooling the risk over many insurers [2]. The literature in insurance risk management concentrates on the efficiency of insurance companies [3]. Since insurance industry has high financial operating expenses, insurers, investors, and regulators are interested in models used to understand the behavior of expenses [3].

Workers' compensation provides cash and covers medical and indemnity costs for workers who experience injuries or illness during their employment and provides benefits to the survivors of workers killed at work [4]. The total incurred amount of a claim is also called the loss cost, which is defined as the proportion of the premium which covers losses and related expenses [5].

Modeling insurance claims with high accuracy and prediction rate is essential to insurance companies for several reasons: First, the statistical modeling of insurance claims provides useful results in estimating loss cost, which is important in financial planning in the insurance industry [5]. Also, estimating loss cost is significant in the actuarial practice of reserving, where portfolios may be formed that generate cash flows with expected values matching those of the liability cash flow [6]. In addition, the statistical modeling of insurance claims can produce interpretable results about the parameters that affect the workers' health, which influence employers' costs and directly influence premium-setting costs, because the injury record of an employer is used to revise premiums and set new pricing [7].

Generalized linear models (GLMs) with heavy-tailed distributions are widely recognized as an accepted framework for the price modeling of insurance loss costs [5,8–11]. According to Boland [12], Achieng [2], and Packová [13], since insurance data holds large infrequent claim amounts, most heavy-tailed distributions can be used to model claim amounts, including gamma, Weibull, exponential, and lognormal distributions. Frees [14] and Nath and Das [15] stated that applying regression models with generalized distributions is useful in modeling skewed and fat-tailed data. Keatinge [16] stated that the exponential distribution gains better results in analyzing loss data. Ravi and Butar [17] expressed that heavy-tailed distributions are a much better fit for financial data in comparison to the normal distribution, as financial data are usually highly skewed. Nath and Das [15] applied heavy-tailed distributions (Weibull and Burr) to a set of motor insurance claim data due to the highly skewed nature of the claims. Tang [18] studied the tail behavior of a series of Pareto-type claims. Frees et al. [19] assessed the actuarial applications of statistical modeling to study the accident frequency, loss type, and severity by incorporating characteristics such as age, gender, and driving history in automobile insurance claims. Meyers [20] used historical loss claims data to predict future claim severity in general insurance using gamma and lognormal distributions, because loss data contain infrequent but large values, which makes it different from normally distributed data.

Even though there is a great deal of literature on the statistical analysis of loss data with a skewed nature, there has been little research on modeling workers' compensation claims with heavy-tailed distributions or on addressing the effect of both continuous and categorical variables on the cost of claims. This study focused on: (1) finding the proper statistical distribution to explain the behavior of large claims in workers' compensation data, (2) applying generalized linear regression models (GLMs) with proper statistical distributions to detect the important variables that affect the claims' escalation, and (3) applying Monte Carlo simulation for the selected GLMs to estimate the future cost of similar incidents in agribusiness industries.

1.1. Data

The data set used in this study was obtained from a private insurance provider in the Midwest of the United States that specializes in insurance products for agribusiness industries. From 2008 to 2016, more than 35,000 claims were recorded in the data set. Severe claims refer to those with a total cost equal to or in excess of $100,000. Out of all the workers' compensation claims in the eight-year period, 2.82% were classified as severe for both open and closed claims, with a total incurred amount of $278 million. As shown in Table 1, descriptive statistics of the severe claims gives a better understanding of the skewed nature of the data. The high coefficient of skewness suggests that generalized distributions are appropriate for modeling the workers' compensation claims in the data set [13]. The target variable in this study is the summation of expenses, medical costs, and indemnity costs of each claim. The list of variables that were used as inputs were obtained from the data set and are shown in Table 2.

Table 1. Descriptive statistics of claims by year (2008–2016).

Year	Mean	Std Dev	Min	Max	Median	Sample Size	Skewness
2008	$273,965	$215,299	$102,673	$1,105,357	$171,901	80	1.83
2009	$342,128	$940,824	$103,273	$8,151,576	$174,868	74	8.07
2010	$279,556	$319,357	$100,714	$2,615,677	$187,036	90	4.96
2011	$255,055	$180,380	$100,354	$831,617	$191,890	76	1.79
2012	$278,590	$352,159	$100,542	$3,206,900	$209,496	95	6.51
2013	$304,881	$694,877	$100,243	$7,591,850	$170,690	155	8.84
2014	$267,087	$390,138	$100,961	$3,748,887	$173,204	187	6.27
2015	$222,002	$226,601	$100,162	$2,145,148	$152,556	223	5.19
2016	$235,226	$265,838	$101,317	$1,452,000	$146,391	51	3.33
All	$268,622	$451,790	$100,162	$8,151,576	$168,988	1031	11.36

Table 2. Description of predictive variables.

Variable	Description
Agricultural-related Industry	16 levels; grain, agronomy, refined fuel, feed milling, etc.
Gender	Male, female, unidentified
Occupation	104 levels; grain elevator operators, poultry producers, etc.
Injury	7 levels; death, permanent disability, medical only, etc.
Body group	6 levels; lower extremities, trunk, upper extremities, etc.
Cause group	9 levels; burn or heat-scald, etc.
Nature group	3 levels; multiple injuries, occupational diseases, etc.
Body part	49 levels; abdomen, ankle, hip, eye(s), internal organs, etc.
Cause	59 levels; chemicals, dust, lifting, machinery, pushing, etc.
Nature	29 levels; dislocation, amputation, laceration, etc.
Age	min: 17.8 years old; max: 81.7 years old
Tenure	min: 0 years; max: 48 years

1.2. Methods

1.2.1. Generalized Linear Regression Modeling

Although a lot of regression techniques consider the underlying distribution of the response variable as being normal, there are situations where the assumption of normality is not appropriate, such as in insurance claims that are often highly skewed in nature [21]. Using generalized regression methods gives a straightforward way to analyze the effect of many factors on the target variable without the restriction of the normality assumption [22]. According to James [23], applying generalized regression to accommodate nonlinear relationships among variables is an alternative to the least square regression method due to higher prediction accuracy, as well as easier model interpretability due to removing irrelevant variables.

1.2.2. Penalization Methods and Variable Selection

The need for using penalization methods (or fitting procedures) in regression modeling is justified by the willingness to accept some bias to reduce variance and avoid overfitting [21]. Overfitting means that the model works well on the observed data, but performs poorly on a new data set. Penalization methods can deal with this issue through subset selection and shrinkage.

According to Tibshirani [24], generalized linear models have distinct advantages in terms of inference and usefulness in real-world problems in comparison to nonlinear models [8]. In order to gain better results in terms of prediction accuracy and model interpretability, penalization methods are important. Regarding prediction accuracy, by constraining or shrinking the estimated coefficients, the overfitting is significantly reduced. This highly improves the prediction of the response variable by applying the model in a new data set. Regarding the model interpretability, by using the shrinkage method, selecting a subgroup of all the input variables leads to omitting the irrelevant and less

relevant variables (to the response variable), and therefore, the unnecessary complexity of the model is decreased.

Based on the work of Crotty and Barker [21], a summary of various penalization methods and their relationships are depicted in Table 3. The shrinkage (also known as regularization) is preferred, since it has the effect of decreasing the variance by shrinking irrelevant estimated coefficients towards zero [23].

Table 3. Penalization methods versus selection and shrinkage.

Method	Selection	Shrinkage
Maximum Likelihood	no	no
Ridge	no	yes
Forward Selection	yes	no
Lasso	yes	yes
Elastic Net	yes	yes

Tibshirani [24] introduced the least absolute shrinkage selection operator (lasso). It is a popular penalization technique as it allows simultaneous estimation and variable selection [25]. Therefore, lasso is used as the fitting procedure applied in building regression models in this study. By using the lasso method for fitting the generalized regression models, a less complex final equation is gained for each model, which includes only a subset of relevant variables as the main predictors of the target variable, or the loss cost in this study.

1.2.3. Quantitative Measure of Performance for Model Selection

Model selection is a process of seeking the model in a set of candidate models that gives the best balance between model fit and complexity [26]. The comparison criterion should be based on knowledge and history of the data as well as personal preference. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are the most common model selection methods; AIC finds the most predictive model, while BIC finds the true model as the final choice [27].

Other measures of model performance are R^2 and the root mean square error (RMSE). Values of R^2 range from 0 to 1, where 1 is a perfect fit and 0 means there is no gain by using the model over using fixed background response rates. It estimates the proportion of the variation in the response around the mean that can be attributed to terms in the model rather than to random error. The RMSE is defined as the standard deviation of the response variable.

When it comes to comparing models, the one with the highest R^2 and the lowest RMSE is preferred. The statistical details of all the model selection criteria are shown in Table 4 (where k is the number of estimated parameters in the model and n is the number of observations in the data set). The model comparison criteria in this study are adopted from [28], and the analyses were done using JMP Pro statistical software (JMP®, Version <*13.2*>. SAS Institute Inc., Cary, NC, 1989-2007).

Table 4. Model comparison criteria.

Criterion	Formula
AIC *	$-2\log(\textit{likelihood}) + 2k$
BIC *	$-2\log(\textit{likelihood}) + k\ln(n)$
RMSE *	$\sqrt{\sum_{i=1}^{n}\frac{(yi-\hat{y}i)^2}{n}}$
R^2	$1-\frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\overline{y}i)^2}$

* AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; RMSE: Root Mean Square Error; k: number of estimated parameters in the model; n: number of observations in the data set

1.2.4. Stochastic Monte Carlo Modeling for Severity Simulation and Risk Analysis

Monte Carlo (MC) simulation is a standard tool in analyzing business and financial problems [29]. The MC method is a computerized probability simulation technique to assess the effect of risk and uncertainty in diverse forecasting models of financial costs and decision-making problems [30]. MC simulation uses a repeated random sampling method with the help of statistical analysis to achieve a probabilistic approximation for a developed model or an arithmetic equation [31]. One distinctive feature of MC simulation studies is the saving of parameter estimates from the analysis of real data to be used as population parameters for data generation in a Monte Carlo simulation study [32]. Unlike the conventional forecasting models, which can estimate fixed values, in stochastic MC simulation processes, a range of estimated values are used as input, and thus the output is also a range of values, which will provide a more realistic picture of the simulation model [30]. As the process is repeated for 5000 iterations, many output values are gained that can be utilized for the description of the likelihood of numerous results in MC modeling [33]. MC simulation studies involve three steps: (i) generating simulated data, (ii) performing some statistical procedures, and (iii) recording the results. After some (necessarily) finite repetitions of these steps, commonly 5000 or 10,000 iterations, a summary statistic is usually calculated and is compared to results from empirical data [34]. Using MC simulation contributes to knowing the characteristics of the simulated data set in advance, which helps to better analyze future cases of similar data [35].

1.2.5. Development of the MC Simulation Model

Using simulation methodology to determine the distribution of the total costs of claims is very beneficial in the field of insurance risk management [36]. Nath and Das [15], Asmussen [37], and Peters et al. [38] have applied MC simulation in insurance analyses. According to Hahn [39], insurance companies act as the institutional investors in the financial system of a country and risk dispersion is an important segment of their business. Therefore, the ability to make a precise simulation of the financial severity is crucial in reducing the risk of illiquidity in insurance companies' operations management. The MC method has been used in the field of risk reduction within the framework of a developed model in non-life insurance to calculate the values of capital required to ensure solvency in some case studies of insurance companies [36].

2. Results

2.1. Summary of Predictive Modeling Analysis

Fitting the gamma, Weibull, and lognormal distribution to the total cost of claims as the dependent variable showed that the gamma distribution fit the total claims amount with the shape of 1.81 and scale of 148,405 (α = 1.81; σ = 148,405), the Weibull distribution fit the total claim cost with the scale of 282,183 and shape of 1.10 (α = 282,183; β = 1.10), and the lognormal distribution fit the total claim costs with the scale of 12.20 and shape of 0.62 (μ = 12.20; σ = 0.62). In the next step, GLMs with gamma, Weibull, and lognormal distributions were built. The details of the effect tests from all models are summarized in Tables 5–7. All the regression models, no matter the underlying distribution, suggested that the injury characteristics (cause, nature, and body part) are the key factors in predicting the financial severity of a claim's loss. Also, occupation turns out to be a statistically important variable in estimating the loss cost of a claim. Only the Weibull regression model shows the agribusiness industry as a statistically significant factor in the prediction of a future claim's cost.

The results for comparing the performance of the models are depicted in Table 8. Considering all the decision criteria (R^2, RMSE, BIC, and AIC), both the gamma and lognormal regression models show a good fit to the data set. The gamma regression model does a better job of explaining the variability in the data, with a higher R^2. The lognormal model shows lower values for RMSE, BIC, and AIC.

Table 5. Effect test results for the generalized linear model (GLM) with gamma distribution using the lasso penalization method.

Predictor	DF	Wald χ^2	Prob > χ^2 *
Injury	6	1315.03	<0.0001
Cause	50	629.23	<0.0001
Occupation	102	383.51	<0.0001
Body Part	42	165.15	<0.0001
Nature	22	18.92	0.0003
Cause Group	-	-	-
Agricultural-related Industry	-	-	-

* χ^2: chi-square value. DF: degree of freedom for each variable.

Table 6. Effect test results for the GLM with Weibull distribution using the lasso penalization method.

Predictor	DF	Wald χ^2	Prob > χ^2 *
Injury	6	121.12	<0.0001
Cause	50	60.55	<0.0001
Occupation	100	71.51	<0.0001
Body Part	42	61.72	<0.0001
Nature	22	16.51	0.0009
Cause Group	2	13.97	0.0029
Agricultural-related Industry	17	7.12	0.0284

* χ^2: chi-square value.

Table 7. Effect test results for the GLM with lognormal distribution using the lasso penalization method.

Predictor	DF	Wald χ^2	Prob > χ^2 *
Injury	6	55.61	<0.0001
Cause	50	174.28	<0.0001
Occupation	100	67.66	<0.0001
Body Part	42	61.21	<0.0001
Nature	22	11.17	0.0108
Cause Group	-	-	-
Agricultural-related Industry	-	-	-

* χ^2: chi-square value.

Table 8. Quantitative measures of model performance by GLM using the lasso penalization method.

Criteria	Gamma	Weibull	Lognormal
R^2	0.79	0.46	0.53
RMSE	163,002	245,624	145,974
BIC	27,386	27,410	26,809
AIC	27,145	27,145	27,079
−LL	13,519	13,514	13,345

Further analysis of the regression models detected the most important predictive variables. The type of injury predicts up to 89% of the severity loss, followed by the occupation class code, which contributes up to 16%. Other key factors with high predictive importance are the cause of injury (12%), the injured body part (6%), and the nature of the injury (2%). Both gamma and lognormal GLMs suggest that permanent total disability and permanent partial disability contribute highly to the escalation of a claim's cost. Also, injuries characterized by amputation, respiratory disorders, vision loss, and contusion that are caused by welding operations, explosion, flareback, collapsing materials, and flying or falling objects have a significant effect on increasing the total loss. Such injuries occurred more often in fingers, hands, wrists, neck, and trunk body parts.

2.2. Summary of the Developed MC Model Analysis

According to the model performance criteria shown in Table 8, the generalized regression model with gamma distribution was the best predictor of the financial severity of potential future claims. These claims are large, but rare. Therefore, simulating the predictive model with a bigger size than the original size contributes to checking the model credibility in estimating the severity of future incidents. To ensure that the generalized regression model with the gamma distribution is the most predictive and reliable in estimating the severity of future cases, its performance in the simulation should also be better than the other two distributions (lognormal and Weibull) regarding prediction accuracy and model interpretability. Thus, the equations developed for the generalized regression models with gamma, lognormal, and Weibull distributions were used to generate stochastic MC simulation input and output values. To ensure that all the combinations were randomly selected, 5000 iterations were performed for each model. The number of iterations in a MC simulation is not fixed and varies based on the nature of the application in numerical or categorical data sets [32,35,37].

The comparison between the descriptive statistics for GLMs from empirical data and those gained from the three simulation models are summarized in Table 9. All three models show a mean value which is slightly smaller than the mean value of the empirical data. However, the simulated values for standard deviation, standard error of the mean, and upper and lower 95% mean parameters are all smaller than the same parameters values from the empirical data. Considering the numerical differences between the parameter values estimated from the simulation models and those from the empirical data, the simulated generalized regression with gamma distribution has the smallest values compared to the simulated generalized regressions with lognormal and Weibull distributions.

Table 9. Descriptive statistics for GLM simulation models (values shown are in US$).

Descriptive Statistics	Empirical Data	Gamma	Weibull	Lognormal
Mean	268,622	257,505	257,947	249,064
Standard Deviation	451,790	364,631	264,264	256,901
Standard Error Mean	14,070	5157	3737	3633
Upper 95% Mean	296,232	267,615	265,273	256,187
Lower 95% Mean	241,012	247,396	250,620	241,942
N(Sample Size)	1031	5000	5000	5000

According to Das and Halder [40], the performance of the simulation models is evaluated by the relative bias and root mean square error values, which are defined as:

$$Bias = \left(\frac{\theta estimated - \theta true}{\theta true}\right)$$
$$RMSE = \sqrt{(\theta estimated - \theta true)^2}$$

where $\theta estimated$ and $\theta true$ represent the values of the descriptive statistics parameters from the simulation data and the empirical data, respectively. The results for the bias values are depicted in Table 10, and the RMSE values are shown in Table 11. Both simulated models with gamma and Weibull distributions have the same bias in the mean parameter of -4%, while the other simulated parameters' values are closer to the empirical parameters in the gamma simulation model. The lognormal simulated model does not show a satisfactory performance and has the highest RMSE for both the mean and standard deviation values compared to the other two models. The RMSE values for the gamma simulation data are also the smallest among all the models. Thus, it is reasonable to conclude that the MC simulated model with the gamma-distributed response variable has the best performance criteria for estimating the financial severity of the claims.

Table 10. Comparison of bias between empirical data GLMs and simulation data GLMs.

Descriptive Statistics	Gamma	Weibull	Lognormal
Mean	−4.14%	−3.97%	−7.28%
Standard Deviation	−19.29%	−41.51%	−43.14%
Standard Error Mean	−63.35%	−73.44%	−74.18%
Upper 95% Mean	−9.66%	−10.45%	−13.52%
Lower 95% Mean	2.65%	3.99%	0.39%
N (Sample Size)	5000	5000	5000

Table 11. Comparison of root mean square error (RMSE) between empirical data GLMs and simulation data GLMs.

Descriptive Statistics	Gamma	Weibull	Lognormal
Mean	11,117	10,676	19,558
Standard Deviation	87,159	187,526	194,889
Standard Error Mean	8914	10,333	10,437
Upper 95% Mean	28,618	30,959	40,045
Lower 95% Mean	6384	9608	929
N (Sample Size)	5000	5000	5000

3. Discussion

This study described a straightforward approach for modeling inflated claims in workers' compensation data. It relied on GLMs to address the skewed nature of large claims. The GLM model with gamma distribution was selected as the most predictive model. Using the lasso shrinkage technique, all the less-effective factors are assumed to have zero effect and only the variables with higher effect are retained in the final model. The gamma GLM model shows that injuries in the grain-handling sector have the highest cost.

The prediction formula can be used to estimate the cost of an injury. For example, the cost of a medical injury in the shoulder for a worker employed in the grain-milling class within the grain-handling industry is estimated as being $222,155, while the same injury has an estimated cost of $234,631 if it resulted in permanent partial disability. The same injury will have a cost of $3,942,872 if it causes permanent total disability. Such a substantial increase is due to the exponential nature of the gamma GLM prediction formula.

This clarifies the importance of the injury type as the key contributor to claim cost prediction. Occupational class codes with the highest frequency are grain elevator operations, chauffeurs and helpers, hay grain or feed dealers, and farm machinery operations. However, corn product manufacturing and food manufacturing class codes have the highest coefficient, despite being less frequent. The nature of the injury that increases the claim cost includes amputation and respiratory disorders. The injuries with the highest influences on the claim cost are caused by explosion or flareback, vehicle crash, moving parts of machines, cold objects or substances, and collapsing materials. The body part injuries contributing the greatest cost are multiple neck injuries and whole body and skull injuries, compared to fingers, multiple trunk injuries, and shoulders, which are among the body part injuries that have less effect on cost.

Looking at the age factor, the mean age of injured workers in the grain-handling industry is 49.5 years old, with a median of 52 years old. The fact that the population of workers is aging may explain its exponentially raising claim cost of injuries. Referring to the age of the most frequent occupation class codes, the grain elevator operations workforce has a mean age of 47 and median age of 49 years old. Chauffeurs and helpers have the same mean and median age of 48.5 years old. Hay grain or feed dealers have the mean age of 48 and median age of 52 years old, while farm machinery operations have the youngest workforce, with mean age of 46 and median age of 45 years old,

respectively. The corn product manufacturing workforce is the oldest, with mean and median age of 54 years old.

The results are useful for insurance companies in developing future financial plans and managing claim costs and premiums based on the analysis of historical data. From the insurance business analytics perspective, the results of the study will help in changing the e-mod rate for specific high-cost agribusiness sectors. Although the conclusions do not present a generalized character, a similar approach can be taken in addressing the underlying factors that cause loss escalation in other industries.

In addition, safety analysts may find this study useful, as safety professionals have long aimed at adding prediction to safety. The management of occupational safety risks is a significant component of any business [41]. Analysis of the incidents helps occupational risk managers identify which hazards have contributed and led to the most frequent occupational accidents, and thus determine appropriate preventative actions [42]. Analyzing empirical data to extract risk indicators adds predictivity to risk scenarios and helps in efficiently planning and modifying loss approaches in agribusiness industries. Based on the results, further investigations can be done in the highest-risk occupation environments to focus specific safety intervention efforts. Integrating the analysis of empirical data with the knowledge of safety practitioners in safety regulations and the training and education of employees is expected to decrease the rate and alleviate the outcomes of severe injuries.

4. Conclusions

The results of this study showed that having access to adequate incident description and using proper statistical and analytical methods can lead to reliable probabilistic forecasts of the outcome of a future incident being made with high certainty. The study showed that modeling the workers' compensation claims used in this study with the gamma distribution gives the best fit when compared to lognormal and Weibull distributions. The results, however, are limited to the data set that was used for the analyses throughout this study; nevertheless this modeling strategy and approach can be used in different data sets that include enormous amounts of claims to determine which GLM yields better performance in predicting the costs of potential future claims. Future studies should focus on applying other types of regression models to evaluate their performance to check the consistency of the results.

Author Contributions: Conceptualization: all the authors; Methodology: F.D.K., S.A.F.; Software: F.D.K., S.A.F.; Validation: F.D.K., S.A.F.; Formal Analysis: F.D.K.; Resources: S.A.F.; Data Curation: G.A.M.; Writing-Original Draft Preparation and Writing-Review & Editing: all the authors; Visualization: F.D.K.; Supervision, S.A.F.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baldwin, M.L.; McLaren, C.F. *Workers' Compensation: Benefits, Coverage, and Costs (2014 Data)*; National Academy of Social Insurance: Washington, DC, USA, 2016.
2. Achieng, O.M. Actuarial modeling for insurance claim severity in motor comprehensive policy using industrial statistical distributions. In Proceedings of the 2010 International Congress of Actuaries, Cape Town, South Africa, 7–12 March 2010.
3. Shi, P.; Frees, E.W. Long-tail longitudinal modeling of insurance company expenses. *Insur. Math. Econ.* **2010**, *47*, 303–314. [CrossRef]
4. Szymendera, S.D. *Workers' Compensation: Overview and Issues*; (CRS Report R44580); Congressional Research Service: Washington, DC, USA, 2016.
5. Guelman, L. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Syst. Appl.* **2012**, *39*, 3659–3667. [CrossRef]
6. Engsner, H.; Lindholm, M.; Lindskog, F. Insurance valuation: A computable multi-period cost-of-capital approach. *Insur. Math. Econ.* **2017**, *72*, 250–264. [CrossRef]

7. Schwatka, N.V.; Atherly, A.; Dally, M.J.; Fang, H.; Brockbank, C.V.; Tenney, L.; Newman, L.S. Health risk factors as predictors of workers' compensation claim occurrence and cost. *Occup. Environ. Med.* **2017**, *74*, 14–23. [CrossRef] [PubMed]
8. McCullagh, P.; Nelder, J. *Generalized Linear Models. Generalized Linear Models*, 2nd ed.; Chapman and Hall: London, UK, 1989.
9. Anderson, D.; Feldblum, S.; Modlin, C.; Schirmacher, D.; Schirmacher, E.; Thandi, N. *A Practitioner's Guide to Generalized Linear Models*; Syllabus Year; Casualty Actuarial Society (CAS): Arlington County, VA, USA, 2010.
10. Haberman, S.; Renshaw, A. Generalized linear models and actuarial science. *Ournal Royal Stat. Soc.* **1996**, *45*, 407–436. [CrossRef]
11. Xia, M. Bayesian Adjustment for Insurance Misrepresentation in Heavy-Tailed Loss Regression. *Risks* **2018**, *6*, 83. [CrossRef]
12. Boland, P.J. Statistical Methods in General Insurance. 2006. Available online: https://iase-web.org/documents/papers/icots7/5G1_BOLA.pdf (accessed on 25 June 2018).
13. Packová, V. Loss Distributions in Insurance Risk Management. In *Recent Advances on Economics and Business Administration, Proceedings of the International Conference on Economics and Business Administration (EBA 2015), Barcelona, Spain, 7–9 April 2015*; INASE: Barcelona, Spain, 2015; pp. 17–22.
14. Frees, E.W. *Predictive modeling applications in actuarial science. Predictive Modeling Applications in Actuarial Science (Vol. 1)*; Cambridge University Press: Cambridge, UK, 2014; Volume 1.
15. Nath, D.C.; Das, J. Modeling of Insurance Data through Two Heavy Tailed Distributions: Computation of Some of Their Actuarial Quantities through Simulation from Their Equilibrium Distributions and the Use of Their Convolutions. *J. Math. Finance* **2016**, *6*, 378–400. [CrossRef]
16. Keatinge, C.L. Modeling Losses with the Mixed Exponential Distribution. *Proc. Casualty Actuar. Soc.* **1999**, *LXXXVI*, 654–698.
17. Ravi, A.; Butar, F.B. An insight into heavy-tailed distribution. *J. Math. Sci. Math. Educ.* **2010**, *5*, 15.
18. Tang, Q. Heavy Tails of Discounted Aggregate Claims in the Continuous-Time Renewal Model. *J. Appl. Probab.* **2007**, *44*, 285–294. [CrossRef]
19. Frees, E.W.; Shi, P.; Valdez, E.A. Actuarial applications of a hierarchical insurance claims model. *ASTIN Bull. J. IAA* **2009**, *39*, 165–197. [CrossRef]
20. Meyers, G. *On Predictive Modeling for Claim Severity*; Casualty Actuarial Society (CAS): Arlington County, VA, USA, 2017.
21. Crotty, M.; Barker, C. *Penalizing Your Models: An Overview of the Generalized Regression Platform*; SAS Institute: Cary, NC, USA, 2014.
22. Cerchiara, R.R.; Edwards, M.; Gambini, A. Generalized Linear Models in Life Insurance: Decrements and Risk Factor Analysis Under Solvency II. In Proceedings of the 18th International AFIR Colloquium, Rome, Italy, 1–3 October 2008; Available online: http://www.actuaries.org/AFIR/Colloquia/Rome2/Cerchiara_Edwards_Gambini.pdf (accessed on 20 May 2018).
23. James, G.W. *Linear Model Selection and Regularization. An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; pp. 203–264.
24. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.
25. Zou, H. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.* **2012**, *101*, 1418–1429. [CrossRef]
26. Burnham, P.K.; Anderson, D. Model selection and multi-model inference. In *A Practical Information-Theoric Approach*; Springer: Berlin, Germnay, 2003; p. 1229.
27. Burnham, K.P.; Anderson, D. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]
28. *JMP® 11 Fitting Linear Models*; SAS Institute: Cary, NC, USA, 2013.
29. Fish, L.J.; Halcoussis, D.; Phillips, G.M. Statistical Analysis of a Class: Monte Carlo and Multiple Imputation Spreadsheet Methods for Estimation and Extrapolation. *Am. J. Bus. Educ.* **2017**, *10*, 81–96.
30. Armaghani, D.J.; Mahdiyar, A.; Hasanipanah, M.; Faradonbeh, R.S.; Khandelwal, M.; Amnieh, H.B. Risk Assessment and Prediction of Flyrock Distance by Combined Multiple Regression Analysis and Monte Carlo Simulation of Quarry Blasting. *Rock Mech. Rock Eng.* **2016**, *49*, 3631–3641. [CrossRef]
31. Panel, U.E.T. *Guiding Principles for Monte Carlo Analysis*; US EPA: Washington, DC, USA, 1997.
32. Mooney, C.Z. *Monte Carlo Simulation*; Sage Publications: New York, NY, USA, 1997.

33. Dunn, W.L.; Shultis, J.K. Monte Carlo Methods for Design and Analysis of Radiation Detectors. *Radiat. Phys. Chem.* **2009**, *78*, 852–858. [CrossRef]
34. Koehler, E.; Brown, E.; Haneuse, S.J.P.A. On the Assessment of Monte Carlo Error in Simulation-Based. Statistical Analyses. *Am. Stat. Assoc.* **2009**, *63*, 155–162. [CrossRef] [PubMed]
35. Mingoti, S.A.; Matos, R.A. Clustering Algorithms for Categorical Data: A Monte Carlo Study. *Int. J. Stat. Appl.* **2012**, *2*, 24–32. [CrossRef]
36. Mucha, V.; Pales, M.; Sakalova, K. Calculation of the Capital Requirement Using the Monte Carlo Simulation for Non-life Insurance. *Èkon. Cas.* **2016**, *64*, 878–893.
37. Asmussen, S. *Conditional Monte Carlo for Sums, with Applications to Insurance and Finance*; Thiele Research Reports; Department of Mathematics, Aarhus University: Aarhus, Denmark, 2017.
38. Peters, G.W.; Targino, R.S.; Wuthrich, M.V. Bayesian Modelling, Monte Carlo Sampling and Capital Allocation of Insurance Risks. *Safety* **2017**, *5*, 53.
39. Hahn, L. Multi-year non-life insurance risk of dependent lines of business in the multivariate additive loss reserving model. *Insur. Math. Econ.* **2017**, *75*, 71–81. [CrossRef]
40. Das, K.P.; Halder, S.C. Understanding extreme stock trading volume by generalized Pareto distribution. *N. C. J. Math. Stat.* **2016**, *2*, 45–60.
41. Kaassis, B.; Badri, A. Development of a Preliminary Model for Evaluating Occupational Health and SafetyRisk Management Maturity in Small and Medium-Sized Enterprises. *Safety* **2018**, *4*, 5. [CrossRef]
42. Comberti, L.; Demichela, M.; Baldissone, G.; Fois, G.; Luzzi, R. Large Occupational Accidents Data Analysis with a Coupled Unsupervised Algorithm: The S.O.M. K-Means Method an Application to the Wood Industry. *Safety* **2018**, *4*, 51. [CrossRef]