Modeling liability insurance claim severity using distributions from the gamma family [*]

Stanley Sayianka, Lameck Nyambane Kenyaga, Winfred Kinya Bundi, Amos Njagi Kaberia, Khalif Ali Hussein, Emmanuel Watia Wambua *Egerton University*

Liability insurance is a risk-transfer mechanism. It protects the insured from injuries to people and damage to third-party property. This type of insurance is essential in safeguarding from actions arising out of unintended negligence, often leading to infrequent but significant losses. The present study experimentally investigated the claim severities for liability insurance using historical claims data from a liability insurance company in Kenya. The data utilized ranged from the year 2000 to 2020. Five distributions from the Gamma family used to fit the data are Exponential distribution, Gamma distribution, Weibull distribution, Pareto Type II distribution, and Burr XII distribution. The data is split into a training and testing set of 80% and 20%, respectively. The training set was used to fit the five models using Maximum Likelihood Estimation. Graphical methods such as QQ-plots, PP plots, as well as goodness-of-fit tests were used to evaluate model performance. The Bayesian Information Criteria and Akaike's Information Criteria were used for model selection, while a two-sided Kolmogorov-Smirnov test was used to assess the final models. The results indicated that the Burr and Gamma distribution provided the best fit to the data with the Gamma distribution providing a good fit mostly within the center of the data, and the Burr XII distribution providing a good fit to the tails of the data. Based on the findings from this study, we recommend that analysts working with liability claim severity data should analyze data in a two-step method, which involves separating the center and extreme tails of the data then proceed to fit a light tailed distribution to the center of the data and a heavy-tailed distribution to the tails of the data. In addition, future research should concentrate on using other families of distributions such as the Erlang family in modeling claim severity data.

Keywords: Claim Severity, Goodness-of-Fit Tests, Maximum Likelihood Estimation (MLE), Loss Distributions, Tail Weight

Introduction

Liability insurance is a risk-transfer mechanism. It exists to protect the insured from events such as injuries to people, and damage of third-party property. It differs from other forms of insurance in that in the event of the occurrence of the insured risks, the insurance company compensates the affected third parties rather than the insured. Forms of liability insurance include: customer injury lawsuit, property damage lawsuit, indemnity insurance, employer's liability, director's liability, professional indemnity insurance, product liability as well as operations and commercial liability.

Originally, individual companies that faced a common peril formed a group and created a self-help fund out of which to pay compensation should any member incur loss (in other words, a mutual insurance agreement). The modern system relies on dedicated carriers, usually for-profit, to offer protection against specified perils in consideration of a premium.

Liability insurance is one of the fastest growing insurance sectors with a global market size value of more than 25 billion dollars and a projected global market size of 433 billion dollars by

[*] Replication files are available on the author's Github account (http://github.com/stanleyrazor). **Current version**: March 02, 2023; **Corresponding author**: stanleysayankaSSS@gmail.com.

2031. This type of insurance is important in an economy in providing protection from actions, which arise out of negligence, that often give rise to infrequent but large losses. The nature of claim severity coupled with the wide range of cover arising out of liability insurance sets it apart from other forms of general insurance, such as auto-insurance.

Liability insurance actuaries are often interested in accurately modeling claims severity, as well as the extreme events such as the possibility of larger than normal losses in an attempt to depict the uncertain behavior of future claims payments. This uncertainty necessitates the use of probability distributions to model the occurrence of claims, the timing of the settlement and the severity of the claims. In this study, we focus on modeling the severity of liability claims using gamma family of distributions.

Briefly our objectives for this paper are:

1. To fit the five distributions from the Gamma family to the training set of insurance data through maximum likelihood estimation and to assess the goodness of fit through graphical methods and information criteria.

2. To assess the accuracy of the fitted models on the testing set using statistical tests.

Data and methodology

Data

The data used for this study was fetched from a portfolio of liability insurance policies, from a major liability insurance company in Kenya. The data on claim severity ranges from 2000 to 2020, and the following variables are supplied: Claim ID, Claim Date, and Claim Amount.

The claim severity is first transformed using log-transformation, which takes care of the extreme skewness of the data. The data is then split into a training set and a testing set, whereby the training set comprises 80% of the total dataset, while the testing set comprises 20% of the data. The splitting was done after randomization of the claim severity data in order to remove any bias. The training set is used to fit the five models, after which the two best models are then evaluated on the testing set to determine the best fitting model in terms of chosen metrics.

Distributions

The gamma family of distributions is a large family of distributions with several distributions ranging from the simple one parameter exponential distribution to more complex distributions such as the Burr XII distribution. The Gamma family of distribution is characterized by the gamma function shown below:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

This family of distributions has high positive skewness, hence suitable for modeling strictly positive random variables. The distributions are also categorized by tail weight. The concept of tail weight is used to compare distributions based on the probability assigned to large values, so that: distributions which assign higher probabilities to larger values are said to be heavier-tailed.

In the gamma family of distributions, light-tailed distributions include: the exponential distribution, the chi-squared distribution and the gamma distribution, while heavy-tailed distributions include: the Weibull distribution, the Pareto distribution and the Burr XII distribution.

Exponential distribution

The exponential distribution is one of the elementary models for claim severity since it is a simple distribution with one parameter. The distribution has the following properties:

The probability density function is given by:

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

where the parameter λ is the rate parameter.

For the exponential distribution, the maximum likelihood estimation is derived as follows:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

$$\log L(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

Gamma distribution

The gamma distribution is a central distribution in loss modeling. The gamma distribution is a two-parameter distribution with the shape(α) and scale(β) parameters and is an extension of the exponential distribution since it is the sum of independently distributed exponential random variables. This makes it more suitable for loss modeling since it has two parameters, hence more adaptive. The distribution has a density function of the form:

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

For $x \geq 0, \beta > 0, \alpha > 0$, the maximum likelihood estimation for the parameters α and β is shown below:

$$L(\alpha,\beta) = \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}$$

$$\log L(\alpha,\beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + n(\alpha - 1) \log x_i - \beta \sum_{i=1}^{n} x_i$$

Differentiating with resepct to α gives:

$$\digamma(\alpha) = \log \beta + \sum_{i=1}^{n} x_i$$

Where $\digamma$, is the Di-gamma function.

The above function results in a non-linear equation in α which cannot be solved in closed form, necessitating the use numerical methods to find the parameter estimates.

Differentiating with respect to β gives:

$$\hat{\beta} = \frac{\hat{\alpha}}{\bar{x}}$$

Weibull distribution

The Weibull distribution is a heavy tailed distribution with a wide variety of applications ranging from survival analysis to loss modelling. The Weibull distribution is a transformed distribution, obtained by raising the exponential distribution to a power. The distribution has the following density function:

$$f(x;\gamma,\beta) = \frac{\gamma(x/\beta)^{\gamma}e^{-(x/\beta)^{\gamma}}}{x}$$

Where: γ is the shape parameter and β is the scale parameter. The maximum likelihood estimation for the parameters of the Weibull distribution is shown below:

$$L(\gamma,\beta) = \prod_{i=1}^{n} \frac{\gamma(x_i/\beta)^{\gamma}e^{-(x_i/\beta)^{\gamma}}}{x_i}$$

$$\log L(\gamma,\beta) = n\log\gamma - \gamma n\log\beta + (\gamma-1)\sum_{i=1}^{n}\log x_i - \sum_{i=1}^{n}(x_i/\beta)^{\gamma}$$

The estimates for the parameters are then estimated from the above non-linear equation using Newton-Raphson's iterative algorithm, given possible starting points.

Pareto distribution

The two parameter Pareto distribution also known as the Lomax or the Pareto Type II is a mixture distribution obtained when an exponential distribution is mixed with a gamma distribution. The Pareto distribution first emerged as a distribution for modeling the wealth distribution, due to its heavy-tailed nature. The Pareto distribution is also used in loss modeling due to its heavy-tailed nature and its adaptability. The distribution has the following probability density function:

$$f(x;\alpha,\lambda) = \frac{\alpha\lambda^{\alpha}}{(x+\lambda)^{\alpha+1}}$$

Where: α is the shape parameter, and λ is the scale parameter, for $x > 0, \alpha > 0$. The maximum likelihood estimation for the parameters is given below:

$$L(\alpha,\lambda) = \prod_{i=1}^{n} \frac{\alpha\lambda^{\alpha}}{(x+\lambda)^{\alpha+1}}$$

$$\log L(\alpha,\lambda) = n\log\alpha + n\alpha\log\lambda - (\alpha+1)\sum_{i=1}^{n}\log(\lambda+x_i)$$

We re-write it as:

$$\log L(\alpha,\lambda) = n\log\alpha + n\alpha\log\lambda - (\alpha+1)S(\lambda)$$

Taking the derivatives we obtain:

$$\hat{\alpha} = \frac{n}{S(\lambda) - n\log\lambda}$$

Solving for $\hat{\lambda}$ requires a numerical approach as it lacks a closed form solution.

Burr XII distribution

The Burr XII distribution also known as the Burr Type XII or the Singh–Maddala distribution is a common distribution used in loss modeling both for insurance and re-insurance events. This distribution is heavy-tailed and can be obtained by raising the two parameter Pareto distribution to a positive power, hence it can be regarded as a transformed Pareto distribution.

The probability density function is given as:

$$f(x;\alpha,\gamma,\theta) = \frac{\alpha\gamma(x/\theta)^{\gamma}}{x(1+(x/\theta)^{\gamma})^{\alpha+1}}$$

The maximum likelihood estimation for the parameters of the Burr XII distribution is shown below:

$$L(\alpha,\gamma,\theta) = \prod_{i=1}^{n} \frac{\alpha\gamma(x_i/\theta)^{\gamma}}{x_i(1+(x_i/\theta)^{\gamma})^{\alpha+1}}$$

$$\log L(\alpha,\gamma,\theta) = n\log\alpha + n\log\beta + \sum_{i=1}^{n}\log\int_0^{\infty} x_i^{\alpha-1}(1+x_i)^{-\beta-1}\mu(x)dx$$

The MLE estimates for α , β and γ are then obtained by taking partial derivatives and equating to 0, but since this is a non-linear likelihood function, then a closed form solution for the Maximum Likelihood Estimates cannot be obtained and hence, the Newton-Raphson's iteration technique is used to find the solution.

Model fitting process

The process used when fitting the models on the claim severity data is detailed below:

1. Split the dataset randomly into a training and testing dataset. The training dataset contains 80% of the original data, and the testing set 20%.

2. Select a model from the model family chosen.

3. Fit the model to the training dataset, using maximum likelihood estimation.

4. Specify criteria for comparing the models fitted: The criteria to choose the model, will be based on graphical methods such as using QQ-plots, PP-plots, as well as information criteria such as the Bayesian Information Criteria, and the Akaike's Information criteria.

In testing the model using the goodness of fit tests, a two-sided Kolmogorov-Smirnov test is used. Given the sample claims severity data from the training set of the form: $X_1, X_2, ..., X_n$ assumed to follow a particular population distribution F, for any particular chosen distribution F_o , the hypotheses statements are:

$$H_o : F = F_o$$

$$H_a : F \neq F_o$$

The test statistic used under the null hypothesis is denoted D and is given by:

$$D = \sqrt{n}|F_n(x) - F_o(x)|$$

For this given test, if the null hypothesis is true, then the test statistic D tends to be small, while if the null hypothesis is not true, then the test statistic takes on large values. The information criteria are given by:

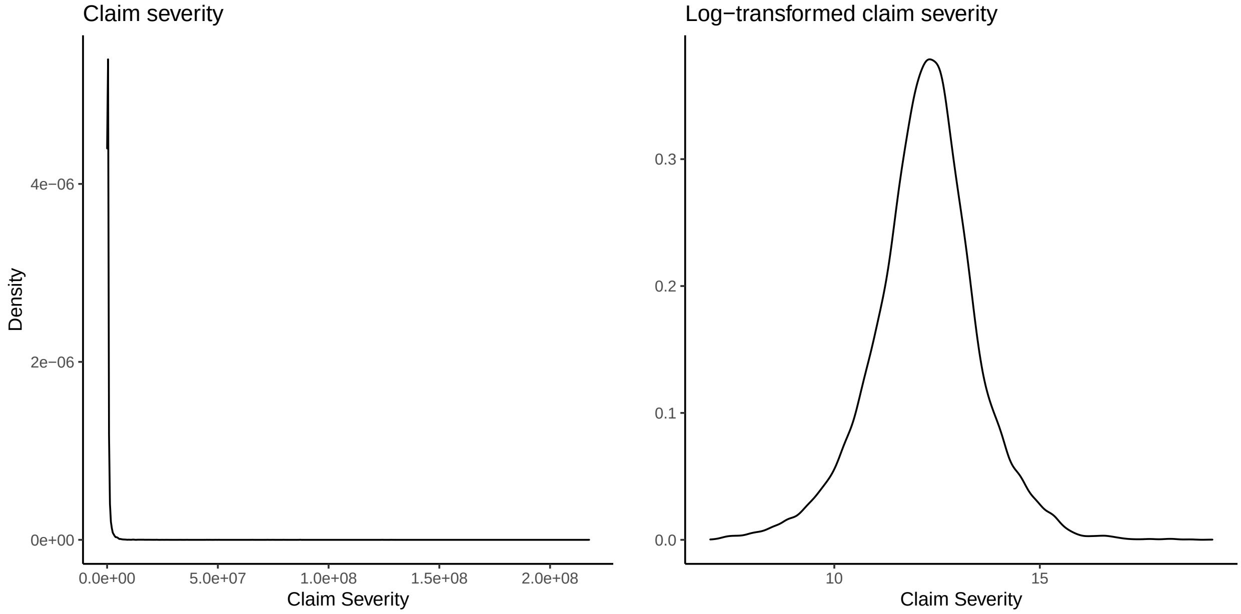
$$AIC = -2\log Likelihood + 2k$$
$$BIC = -2\log Likelihood + k * \log n$$

Where k is the number of parameters in the fitted model, and n is the number of observations.

The Information criterion tend to take smaller values for better models, hence models with lower AIC and BIC scores are preffered.

Results

Claim severity



The claim severity density plot is highly skewed, and with fat tails with the distribution ranging from Kshs. 1000 to Kshs. 220 million, however after applying a log-transformation, the claim severity distribution becomes approximately symmetric, and within a good range of 6 to 20, which is important for ensuring numerical stability in Maximum Likelihood Fitting of the models.

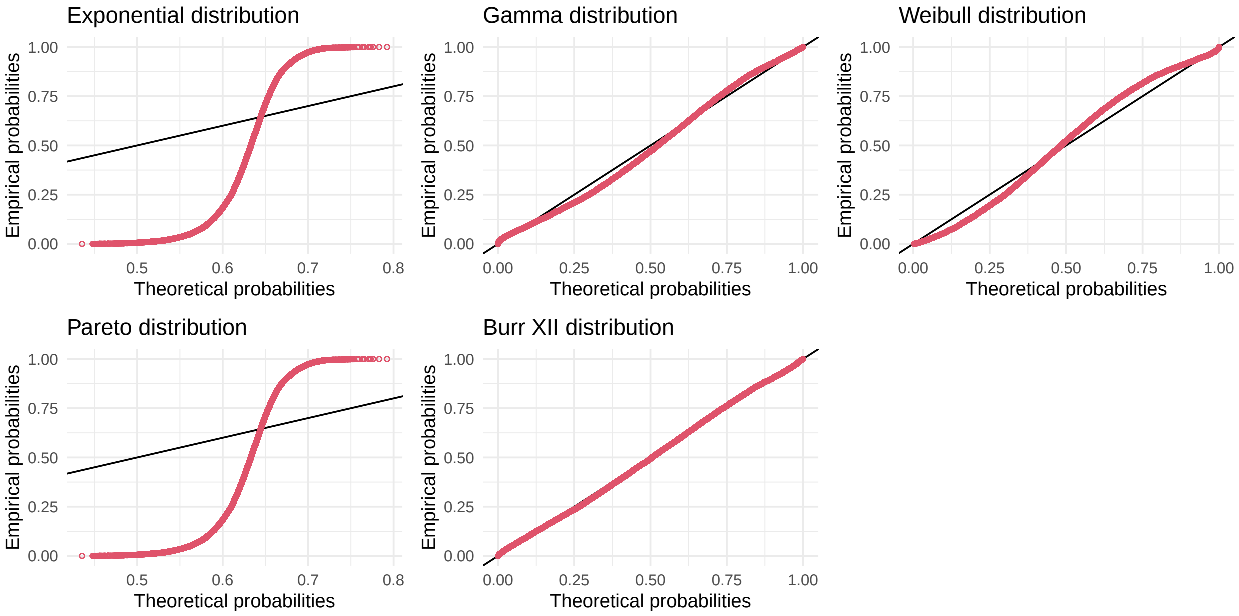
Fitted model statistics

This section covers the parameter estimation summaries for the five models by means of maximum likelihood estimation.

For the Pareto, and Burr XII distribution, reasonable starting values were provided, and the Newton-Raphson's iterative scheme applied the starting values as its initial parameters in maximizing the log-likelihood.

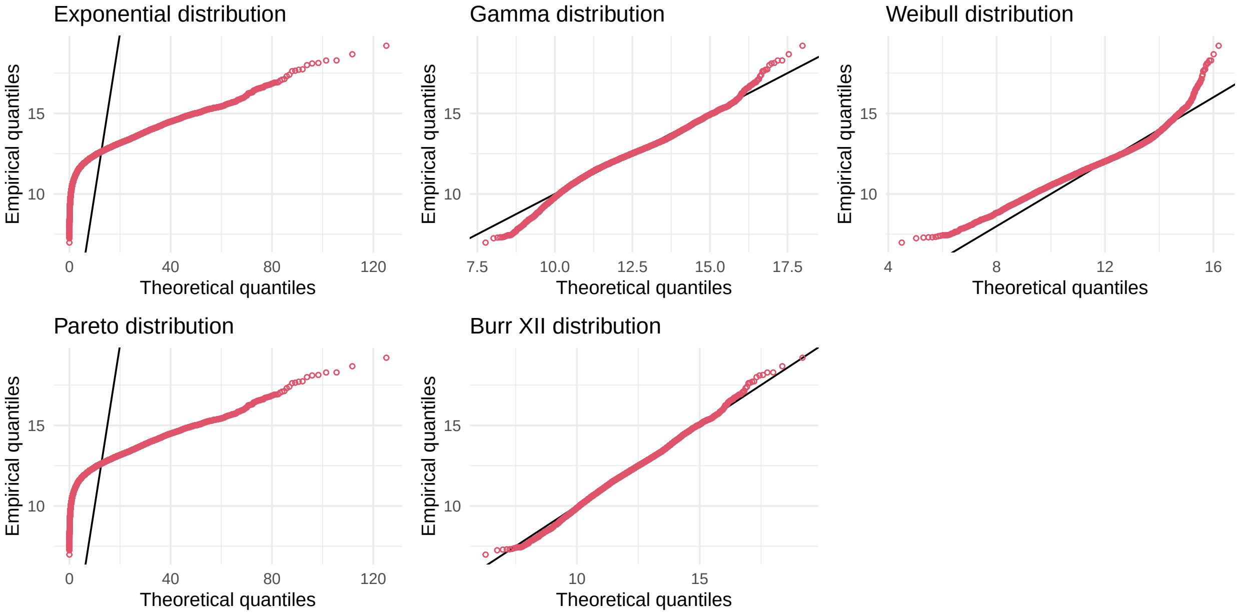
Graphical comparisons of fit

Probability-Probability plot



The PP-plot is a useful tool in comparing the cumulative density function for the log-transformed claims, and the fitted distributions, especially in the regions of high probability density in data. It is evident that the Burr XII distribution provided a good fit, while the Gamma distribution provided a near-good fit to the data, as compared to the rest which gave poor fits.

Quantile-Quantile plot



The QQ-plot is useful in capturing the fit of the models with regards to the skewness of the log-transformed claims, as well as the location and scale of the data. It is evident that the Burr XII distribution provided the best fit, since it captured the tail of the data at hand quite well, as well as the center of the data, as compared to the rest of the distributions. The Gamma distribution came close in modelling accuracy, although it only captures the center of the distribution, and does not fit the tails of the data distribution adequately.

Statistical Goodness-of-Fit tests

The goodness of fit statistics for the fit of the five distributions is shown below:

	Exponential	Gamma	Weibull	Pareto	Burr
Kolmogorov-Smirnov statistic	0.5165	0.0498	0.0693	0.5165	0.0172
Cramer-von Mises statistic	1136.4411	11.0478	28.3433	1136.1905	1.2092
Anderson-Darling statistic	5254.5446	64.4542	Inf	5253.5440	7.7555
Akaike's Information Criterion	98665.2304	46876.9629	48159.9486	98667.5702	46109.6867
Bayesian Information Criterion	98672.7833	46892.0686	48175.0543	98682.6759	46132.3453

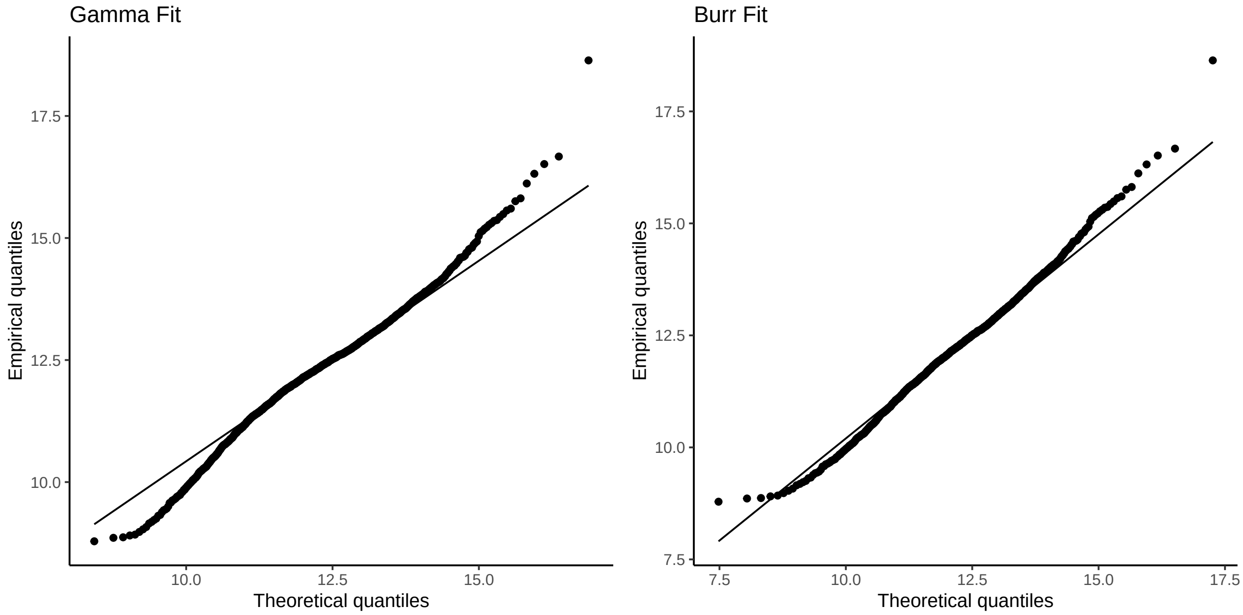
The best fitting distribution based on the three tests: Kolmogorov-Smirnov, Cramer-von-Mises, and Anderson Darling tests is the Burr XII distribution, as evidenced by the low test statistics. The next best fitting distribution is the Gamma distribution, and finally the Weibull distribution. The Exponential and Pareto distribution did not provide a good fit to the data as indicated by their very large test statistics, and hence, we proceeded with only the top two distributions: Burr and Gamma distribution.

The Akaike's and Bayesian Information criterion prefer the Burr XII distribution as the best fitting distribution, and agree with the analysis generated by the tests above.

Comparison on test data

We proceed with the two best fitting distributions the Burr XII and Gamma distribution and attempt to compare their fit on previously unseen data in order to assess fit. The quantiles of the test claim severity distribution are compared to the quantiles of the two distributions by graphical methods, and then a formal two-sided Kolmogorov-Smirnov test is applied.

Graphical comparison



The chart above gives a good visual representation and indicates that the Gamma and Burr XII distribution accurately capture the dynamics of the test claim severity data.

Comparison using a Kolmogorov-Smironov test

The Kolmogorov-Smirnov test is applied to test whether the quantiles of the test claim severity data came from the same distribution as the quantiles generated by either the Burr XII or Gamma distribution. The test results are shown in the table below;

Distribution	Statistic	P-Value
Gamma	0.0554995	0.0944740
Burr	0.0252270	0.9107503

From the test statistics and the P-values generated, we fail to reject the null-hypothesis and conclude that the quantile distribution of the test claim severity data is not significantly different from the quantile distribution of the fitted Burr XII and Gamma distribution at both the 5 percent significance level.

Conclusions

In this study, we examine the distribution of the log-claims severity from a portfolio of liability insurance policies using data ranging from the year 2000 to 2020. The distributions used to model the data were: Exponential, Gamma, Weibull, Pareto, and Burr XII distribution. The results indicated that the Burr and Gamma distribution provide the best fit to the data with the Gamma distribution providing a good fit mostly within the center of the data, and the Burr XII distribution providing a good fit to the tails of the data. The results are interesting since, the Burr XII is regarded a heavy-tailed distribution, while the Gamma distribution is regarded a light-tailed distribution, yet both give a good fit to liability claim severity data. These results are useful to academic researchers and analysts working within liability insurance sector.

Recommendations

Based on the findings from this study, we recommend the following:

i. Analysts working with liability claim severity data should analyze data in a two-step method, which involves separating the data into the center of the distribution of data, and the extreme tails of the data. The analyst should then proceed to fit a light-tailed distribution to the center of the data distribution, and fit a heavy tailed distribution to the tails of the data.

ii. Future research should concentrate on using other families of distributions such as the Erlang family, and extreme value theory distributions in modelling claim severity data.

References

1. Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? Insurance: Mathematics and Economics, 51(2), 239-248.

2. Karobia, R. J. (2015). Modelling extreme claims using generalized pareto distributions family in an insurance company (Doctoral dissertation, University of Nairobi).

3. Packová, V., & Brebera, D. (2015). Loss distributions in insurance risk management. Recent Advances on Economics and Business Administration, 17-22.

4. Das, J., & Nath, D. C. (2016). Burr XII distribution as an actuarial risk model and the computation of some of its actuarial quantities related to the probability of ruin. Journal of mathematical finance, 6(1), 213-231.

5. Omari, C. O., Nyambura, S. G., & Mwangi, J. M. W. (2018). Modeling the frequency and severity of auto insurance claims using statistical distributions.

6. Ng'elechei, J. K., Chelule, J. C., Orango, H. I., & Anapapa, A. O. (2020). Modeling Frequency and Severity of Insurance Claims in an Insurance Portfolio. American Journal of Applied Mathematics and Statistics, 8(3), 103-111.

7. Okindo, N. F. (2021). Oq-o2-2o21 (Doctoral dissertation, Strathmore University).