Vishnu Yannam

z5417441

ACTL3162 Report

task 1

**1.0 Executive Summary**

The financial viability of an insurance firm hinges on its ability to manage its resources efficiently, ensuring that the firm's operational expenses, most significantly claim payouts, do not surpass its incoming revenue. Premiums act as the primary source of revenue for insurance firms. Accurate modelling of future claims is paramount for predicting future revenue streams, enabling the firm to maintain its financial health. This report delves into the analysis of an insurance firm's liabilities and offers a methodological approach to simulate future claim sizes. It also finds that the log-normal distribution had the best fit and its parameters were and . This report evaluates the suitability of three different probability distributions (Log-normal, Gamma, and Pareto) as models for describing a set of loss data. The evaluation process consists of the following steps:

1. **Inspecting QQ plots** to visually assess the fit of each distribution to the data.
2. **Analysing AIC and BIC values** to compare the model complexity and goodness-of-fit of each distribution.
3. **Conducting the Kolmogorov-Smirnov (KS) test** to formally assess the goodness-of-fit of each distribution.

**2.0 Methodology for Maximum Likelihood Estimates of Parameters**

For this analysis, Maximum Likelihood Estimation (MLE) was employed as the principal method for this purpose. This section of the report delves into the specifics of the methodology used.

**2.1 Steps Undertaken for MLE Calculation:**

**1. Initialisation:** The dataset of losses was first loaded and cleaned to remove any inconsistencies or outliers that might skew the results. Three distributions were considered: Log-normal, Gamma, and Pareto.

**2. Distribution-specific MLE Computation:**

*Log-normal Distribution*: The mean and standard deviation of the natural logarithm of the data were computed as initial estimates. These were then optimised to maximise the likelihood.

*Gamma Distribution*: The shape and rate parameters were estimated. An iterative method was employed to ensure optimisation.

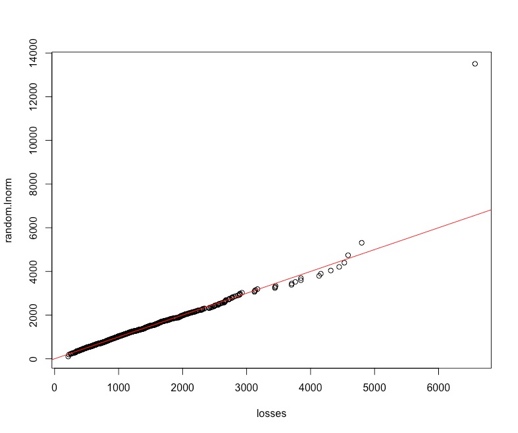
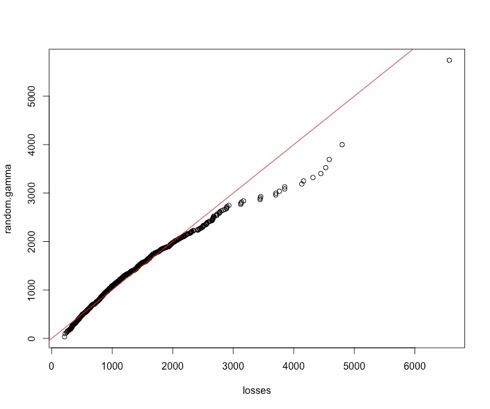
*Pareto Distribution*: For the Pareto, the shape and scale (minimum value) parameters were determined using the dataset. Again, an iterative approach was applied for optimisation.

The parameter estimates are listed in Appendix C of the report.

**3.0 Methodology**

**3.1 QQ Plots Analysis**

**A graph with a number of losses

Description automatically generated with medium confidence******QQ plots provide a graphical comparison of the observed data and the theoretical quantiles of a given distribution. If the distribution is a good fit for the data, the data points will closely follow the theoretical line.

**Figure 1: Quartile-Quartile Plots of Log Normal, Gamma and Pareto Disributions**

*(from left to right)*

**Log-normal distribution:** The QQ plot for the log-normal distribution reveals that the data points closely follow the theoretical red line. This suggests that the log-normal distribution is a good fit for the loss data.

**Gamma distribution:** The QQ plot for the gamma distribution shows that while the majority of data points align with the theoretical line, there are some deviations, especially for higher values. This suggests that the gamma distribution may not be the best fit for the loss data, especially for modelling high-severity losses.

**Pareto distribution:** The QQ plot for the Pareto distribution indicates significant deviations from the theoretical line, suggesting that it may not be a good fit for the loss data at all.

**3.2 AIC and BIC Analysis**

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two model selection criteria that can be used to compare the complexity and goodness-of-fit of different models. Lower AIC and BIC values indicate a better model fit.

The following table shows the AIC and BIC values for the three distributions:

|  |  |  |
| --- | --- | --- |
| Distribution | AIC | BIC |
| Log-Normal | 15424.11 | 15433.92 |
| Gamma | 15492.66 | 15502.47 |
| Pareto | 15604.19 | 15614.00 |

As shown in the table, the log-normal distribution has the lowest AIC and BIC values, indicating that it is the best-fitting model among the three distributions.

**3.3 Goodness-of-Fit Test**

In the process of evaluating the appropriateness of different distributions to model our loss data, it was crucial to select the most appropriate goodness-of-fit test. The Kolmogorov-Smirnov (KS) test was chosen over the Anderson-Darling (AD) test since the KS test is both non-parametric and distribution-free making it universally applicable over various distributions. Furthermore, the results of the KS test are intuitive as they directly measure the maximum difference between the observed and expected cumulative distributions. The KS test is also much less computationally demanding. Despite the AD test's sensitivity to tail deviations, the analysis required a comprehensive review of the entire distribution rather than a narrow focus on the tails. This holistic approach ensured that we did not overlook other potential areas of deviation.

**3.4 Kolmogorov-Smirnov (KS) Test Results**

The Kolmogorov-Smirnov (KS) test is a non-parametric statistical test that can be used to assess the goodness-of-fit of a distribution to a set of data. The test statistic (D) measures the maximum deviation between the cumulative distribution functions (CDFs) of the observed data and the hypothesized distribution.

The following table shows the KS test results for the three distributions:

|  |  |  |
| --- | --- | --- |
| Distribution | D | p-value |
| Log-Normal | 0.0215 | 0.744 |
| Gamma | 0.057792 | 0.002512 |
| Pareto | 0.48034 | < 2.2e-16 |

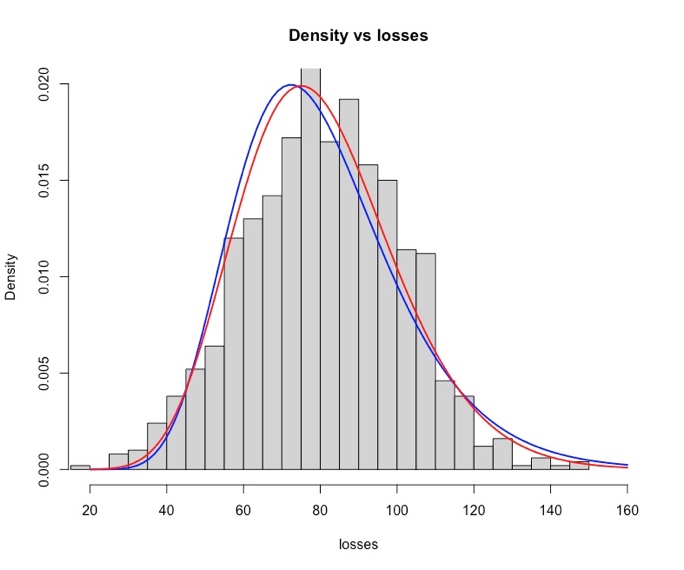
The p-value associated with the K-S test statistic indicates the probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming the hypothesised distribution is true. A low p-value suggests that the hypothesised distribution significantly deviates from the observed data, indicating a poor fit. As shown in the table, the log-normal distribution has the highest p-value, indicating that it is the best-fitting model among the three distributions according to the KS test. The gamma distribution has a slightly lower p-value, suggesting that it may not be as good of a fit as the log-normal distribution. The Pareto distribution has an extremely low p-value, confirming that it is a poor fit for the loss data.

**4.0 Conclusion**

Based on the QQ plots, AIC and BIC values, and the results from the KS test, the log-normal distribution appears to be the most suitable model for the given loss data. Both the gamma and Pareto distributions showed significant deviations in the KS tests, indicating they are less ideal fits for the data.

**Appendix**

**Appendix A: Density vs Losses (Histogram with fitted density lines)**



* Log-normal is blue, Gamma is red, pareto is purple

**Appendix B: Plotted ECDF with Fitted CDFs**

A graph of a graph

Description automatically generated

* Log-normal is blue, Gamma is red, pareto is purple

**Appendix C: Parameter Estimates**

|  |  |  |
| --- | --- | --- |
| Distribution | Parameter 1 | Parameter 2 |
| Log-Normal | Mean Log = 6.909 | SD Log = 0.539 |
| Gamma | Shape = 3.527 | Rate = 0.003 |
| Pareto | Alpha = 1 | X\_min = 212.546 |