ACTL3162 Task 1

Ray Xu z5476219

Complete Data

Question 1.

To estimate the model parameters for each distribution, we can derive their log-likelihood functions by employing maximum likelihood estimation techniques. The log-likelihood functions for the observations $(N_1, \dots, N_T)$ can be given by:

$$l(\theta) = \sum_{t=1}^{T} \log p_{N_t}^{(t)}(\theta).$$

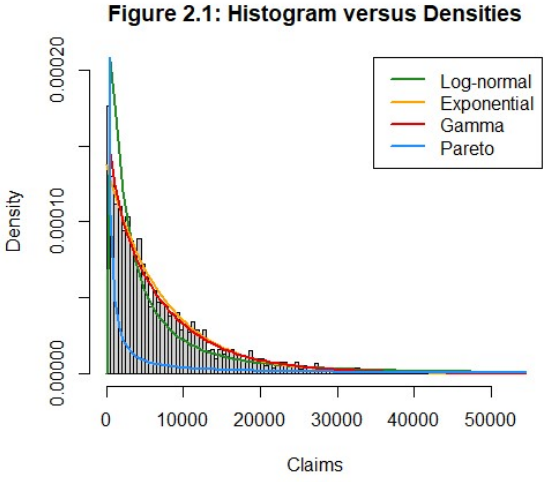
We can differentiate the log-likelihood function with respect to the desired parameters and set these derivatives to 0. By setting the derivative to 0, we can rearrange the equation and solve to obtain an estimate for the parameter(s). The second derivative should be less than 0. This can be done in R using the "fitdistrplus" package's "fitdist" function excluding the pareto distribution which was calculated analytically *(Appendix Complete.1).* We find the following estimates for the different distribution parameters.

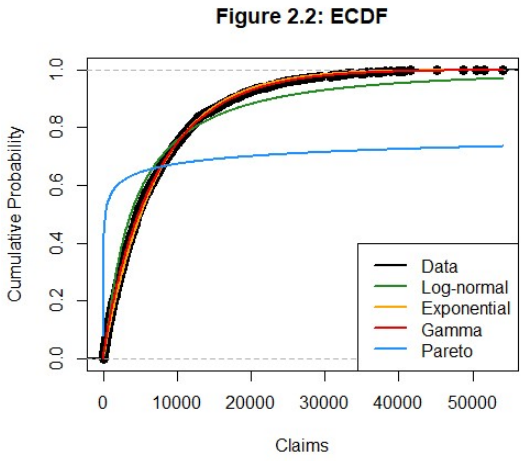
Distribution	Estimated Parameters
Log-normal	$\mu = 8.22, \sigma^2 = 2.0164$
Exponential	$\lambda = 0.000137$
Gamma	$\alpha = 0.87678, \beta = 0.00012$
Pareto	$\alpha = 0.12284, \lambda = 1.08626$

Table 1.1: Estimated parameters for complete data

Question 2.

To evaluate the quality of fit, we can examine the following plots to determine how well the distributions with estimated parameters agree with the actual data. From *Figure 2.1*, the "Loss" data has a flatter distribution compared to the log-normal and pareto fits. We find that the pareto distribution is an unsuitable fit as it is unable to account for most of the data around the mean, further lacking the flatter positive tail required to model insurance claims. This is a similar case for the log-normal distribution where its density falls below the actual claim amounts around the centre, underestimating the actual claim amounts near the mean, however overestimating for extreme claim amounts, both near 0 and for larger values. Both the exponential and gamma distribution better capture the shape and peaks of the data, particularly around the mean while also accounting for the heavy tails in insurance claims.



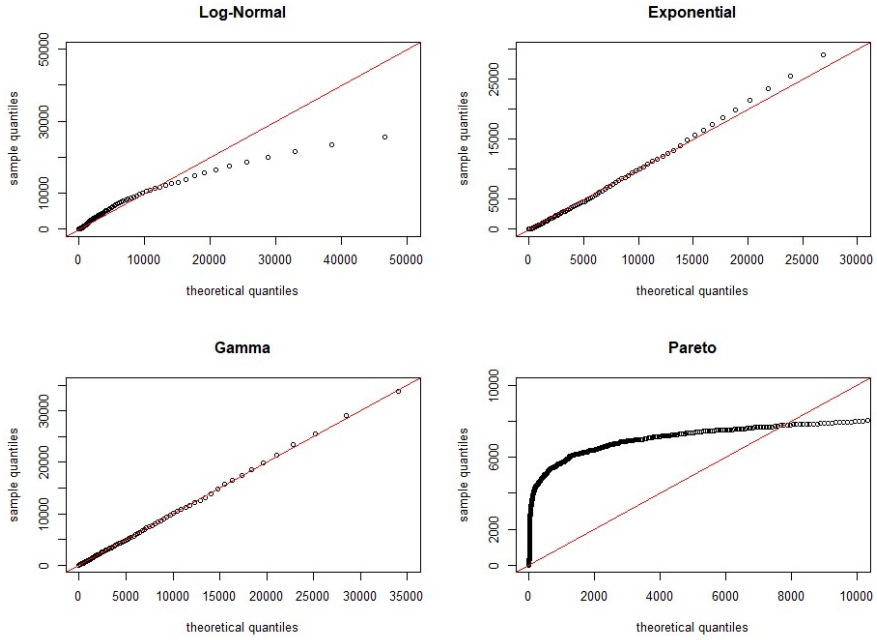


From the *Figure 2.2,* the fitted pareto distribution does not align with the empirical distribution, unable to account for the longer and heavier tails of the data's distribution. While the log-normal distribution from the ECDF plot indicates that for lower claims the data is well-accounted for, the right positive tail is not captured. Both the exponential and gamma distributions provide similar results, capturing the peaks and tails of the data while the gamma lies closer to the empirical distribution around the central region of the data, where the cumulative probability rises from 0 to 0.7 for claims from 0 to 10,000.

From *Figure 2.3* we find that the log-normal distribution can moderately model claims under 10,000, the data points quickly fall below the trend-line towards the top-right, as the sample quantile is less than the theoretical quantile, indicating that the right tail of the data is lighter than the fitted distribution. For the exponential plot, we find that the data is well-fitted, where the actual data deviates above the trend line towards the top-right, indicating that the data's slightly heavier tail in comparison to the fitted distribution. On other hand, the gamma distribution is able to capture the flatter distribution of claims around the mean in addition to the matching the tails of the actual data. The pareto distribution is

unable to account for the higher frequency of lower claim amounts, in addition, with data points falling below the trend-line, possesses heavier tails than the data.

Figure 2.3: QQ Plots



Conducting the Kolmogorov-Smirnov (KS) test, we assume that the "Loss" data is a sample from an unknown continuous distribution G and the corresponding empirical distribution. Setting $H_0: G = G_0$ versus $H_1: G \neq G_0$, we can calculate the KS test statistic and determine if the null hypothesis is rejected at any given significance level α with $D_n > \frac{K^{-1}(1-\alpha)}{\sqrt{n}}$ where $D_n = sup\left|\widehat{G_n}(y) - G_0(y)\right|$ and $K(y) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp\{-2j^2y^2\}$, the Kolmogorov distribution K. Similarly, for the Anderson-Darling (AD) test, the test statistic $A_n = n\int \frac{(\widehat{G_n}(y)-G_0(y))^2}{G_0(y)(1-G_0(y))} dG_0(y)$ can be used to determine the goodness of fit of the fitted distribution. Using R's "goftest", "actuar" and "stats" packages, we can determine the test statistic and corresponding p-value for the test.

Distribution	KS test statistic	AD test statistic	KS test p-value	AD test p-value
Log-normal	0.0802	28.5	1.32e-11	3e-07
Exponential	0.0316	4.87	0.0366	0.00333
Gamma	0.0125	0.23	0.911	0.98
Pareto	0.447	640	0	3e-07

Table 1.2: Hypothesis test results

Assuming a significance level of 5%, from *Table 2.1*, we find that the Pareto distribution as expected does not model the data adequately, with the highest test statistics and small p-values. Similarly, the log-normal distribution's test statistics also produce small enough p-values to reject the null hypothesis. The exponential distribution under the KS test does not reject the null hypothesis contrary to the AD test. This could be attributed to the difference between the emphasis of the tests. The AD test places greater emphasis on good-fit in the tails than the middle of the distribution, which the KS test does not take into account. The gamma distribution fit produces the lowest KS and AD test statistics, and the null hypothesis is not rejected, indicating the model's ability to provide an adequate fit for both the middle and tails of the data.

Question 3.

In addition to both graphical analysis and hypothesis tests, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) of the all the models were calculated by deriving their log-likelihood function (*Appendix Complete.1)*

From graphical analysis in examining the histogram, ECDF plot and QQ plots, we can determine that the pareto distribution is not an appropriate fit as it is unable to account for the flatter nature of the data, in addition to having a

heavier tail, where it overestimates the probability of higher extreme claim values. Similarly, the log-normal distribution contains heavier tails than the data, however, better fits the centrality of the data.

Both the exponential and gamma distributions better fit the central data in addition to the tails, however the QQ plot identifies exponential's lighter tails than the data, indicating a more appropriate fit from the gamma distribution. From the hypothesis tests, we can similarly see that both the pareto and log-normal distributions produce test statistics that reject the null hypothesis under 5% significance level. The exponential distribution only rejects the null under the AD test, while the null hypothesis is not rejected under both tests for the gamma distribution. This is rather important as the AD test places emphasis on good-fit in the tails, which the KS test does not account for, which in the insurance industry is important to insurers to consider large claim values. The gamma distribution has the lowest AIC and BIC, a difference of 22 and 16 to the exponential distribution, indicating a better fit for the data, despite a greater amount of estimated parameters. While both the exponential and gamma distribution capture the flatter data around the mean well, the model that best fits the provided data is the gamma distribution due to its ability to fit for the tails which the exponential distribution lacks. While the gamma distribution is the best fit for this dataset, limitations exist in the potential overfitting to the dataset using MLE estimation. Due to the small size of the dataset and potential for large tails in insurance claims, the estimate can be biased where other distributions pose a greater fit for different datasets. Furthermore, to apply MLE, data is assumed to be independently distributed, which may not be the case in reality where claims are likely to be correlated.

Distribution	AIC	BIC
Log-normal	39951	39962
Exponential	39552	39558
Gamma	39530	39542
Pareto	45261	45272

Table 1.3: AIC and BIC values

Censored Data

Question 1.

To calculate the MLE estimates for the exponential and pareto distributions for right-censored data, we can derive their likelihood and hence log-likelihood function then differentiate with respect to the estimated parameter to obtain required estimates. *(Appendix Censored.1)*

$$L = \prod_{i=1}^{k} g(y_i)\,[\,\bar{G}(M)]^{n-k}$$

As a result, we obtain the following estimated parameters for the exponential and pareto distributions.

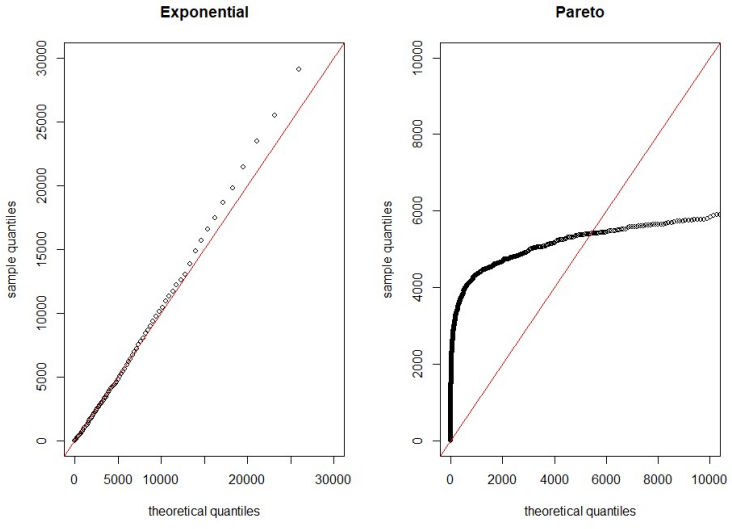
Distribution	Estimated Parameters
Exponential	$\lambda = 0.0001419$
Pareto	$\alpha = 0.093282, \lambda = 1.08626$

Table 1.4: Estimated parameters for censored data

Question 2.

From *Figure 2.4*, we can see that the exponential distribution models the data well for majority of claim sizes around the centre of the data, however the data points deviate above the trend line towards the top-right, indicating that the data has a heavier tail than that of the exponential distribution. Comparing this to the plots of the complete data, *Figure 2.3*, we see that the actual data deviates further from the fitted distribution, as the right-censoring causes the fitted distribution to account less for the heavier right tail. For the pareto plot, the data further deviates from the trend line at lower claim amounts and exhibits flattening quicker as the distribution is fitted to right-censored data, there is less emphasis on the right tail, leading to greater values being unaccounted for and more emphasis on fitting high frequency smaller claim amounts.

Figure 2.4: QQ Plots for Censored data



Appendix

Complete.1

Log-likelihood functions used to estimate parameters for each distribution:

Log-normal:

$$\ell = -\frac{n}{2}\log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log(x_i) - \mu)^2 - \sum_{i=1}^{n}\log(x_i).$$

Exponential:

$$\ell = nlog(\lambda) - \lambda\sum_{i=1}^{n} x_i.$$

Gamma:

$$\ell = n\alpha\log(\beta) - \mathrm{n}\log\big(\Gamma(\alpha)\big) + (\alpha - 1)\sum_{i=1}^{n}\log(x_i) - \beta\sum_{i=1}^{n} x_i.$$

Pareto:

$$\ell = nlog(\alpha) + n\alpha\log(\lambda) - (\alpha + 1)\sum_{i=1}^{n}\log(x_i).$$

$$\hat{\lambda} = \min\{y_1, \dots y_n\}.$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n}(\log(x_i) - log(\hat{\lambda}))}.$$

Censored.1

Assuming k claims below M and $n-k$ claims above M, we can obtain the following likelihood function for the exponential distribution:

$$L = \prod_{i=1}^{k} g(y_i)\,[\,\bar{G}(M)]^{n-k}$$
$$= \prod_{i=1}^{k} \lambda e^{-\lambda y_i}\big(e^{-\lambda M}\big)^{n-k}.$$

Then we the log-likelihood function can be obtained:

$$\ell = klog(\lambda) - \lambda\sum_{i=1}^{k} y_i - (n-k)\lambda.$$

Differentiating with respect to λ we get:

$$\frac{d\ell}{d\lambda} = \frac{k}{\lambda} - \sum_{i=1}^{k} y_i - M(n-k).$$

Setting the derivative to 0, we find the MLE estimate for λ:

$$\hat{\lambda} = \frac{k}{\sum_{i=1}^{k} y_i + M(n-k)}.$$

Repeating this for the pareto distribution, we can find the likelihood function:

$$L = \prod_{i=1}^{k} \frac{\alpha\lambda^{\alpha}}{y_i^{\alpha+1}} \left(\frac{\lambda}{M}\right)^{\alpha(n-k)}.$$

The log-likelihood function can be obtained:

$$\ell = klog(\alpha) + k\alpha \log(\lambda) - (\alpha + 1) \sum_{i=1}^{k} y_i + \alpha(n-k) \log\left(\frac{\lambda}{M}\right).$$

Since we are maximizing the likelihood, a greater λ will always result in a higher likelihood, then we can set $\hat{\lambda} = \min\{y_1, \dots y_n\}$ as $x \geq \lambda$.

Differentiating the log-likelihood function with respect to α we get:

$$\frac{d\ell}{d\alpha} = \frac{k}{\alpha} + klog(\lambda) - \sum_{i=1}^{k} y_i + (n-k) \log\left(\frac{\lambda}{M}\right).$$

Setting the derivative to 0, we find that the MLE estimate for α:

$$\hat{\alpha} = \frac{k}{\sum_{i=1}^{k} \log(y_i) - klog(\lambda) - (n-k)\mathrm{l} \log\left(\frac{\hat{\lambda}}{M}\right)}.$$