Modelling the tail of Insurance data

Khaya Mpehle

In this report, we statistically analyse a sample of insurance data. We are interested particularly in the tail behaviour of the claims, with tail behaviour referring to *large* insurance claims. Clearly the tail behaviour is of interest to the risk manager of an insurance portfolio. If it turns out that the chances of large claims are high, this will have implications for the risk manager's business model. A key question is whether the data are heavy-tailed, with heavy-tailed distributions having a higher-liklihood of large claims occuring than 'light-tailed' distributions. Using various graphical methods and statistical methods, we investigate the question of heavy-tailed data and subsequently implement appropriate statistical fitting procedures. It is worth highlighting that, in what follows, all statistical procedures are implemented by the author from scratch using Python.

1 The data

The data are a collection of $n = 2167$ Danish insurance claims related to fire incidences, the claims taken at irregular times from 3 January 1980 until 31 December 1990. These claims were made at irregular times. The units of the data are Millions of Danish Kroner, and are given in 1985 prices.

2 Explatory analysis

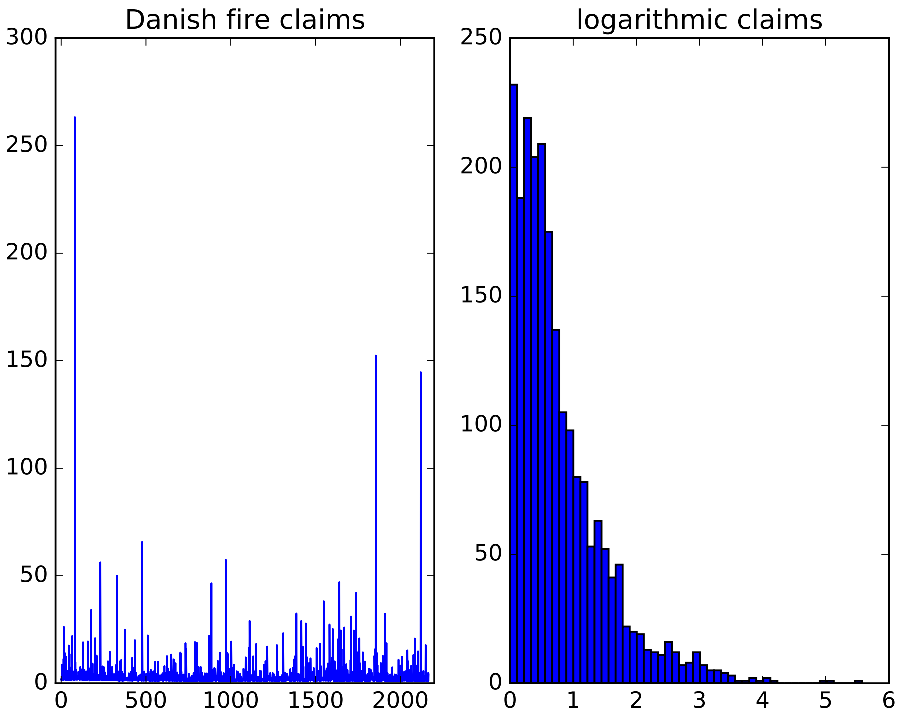

Figure 1: (Left) The Danish fire insurance claims. (Right) a histogram of the log transformed claims. Notice the presence of outliers in both the hisotgram and the sequence plot

To start off, we investigate whether the Insurance data should could be treated as heavy-tailed using several exploratory techniques. Here, a Heavy-tailed distribution is a probability distribution whose cumulative distribution (df), $F(x) = \mathbb{P}(X \leq x)$, decays slower than an exponential function. For example, the df

$$F(x) = 1 - x^{-1},\, x \geq 1$$

is a Heavy tailed distribution. In figure 1 we have a plot of the sequence of insurance claims. Both plots in the figure immediately suggest that the data could be heavy tailed, with multiple, large claims visible in both the sequence plot and several outliers in the right tail of the histogram. As a second test, we show that an exponential distribution fit does not perform well. The exponential distribution of rate parameter λ has df

$$F(x) = 1 - e^{-\lambda x},\, x > 0,$$

and is therefore a light tailed distribution. In the left panel of figure 2, we see that the claims are poorly fitted by the exponential distribution, as the QQ-plot of the insurance claim quantiles against the quantiles of a standard exponential do not lie on a straight line.

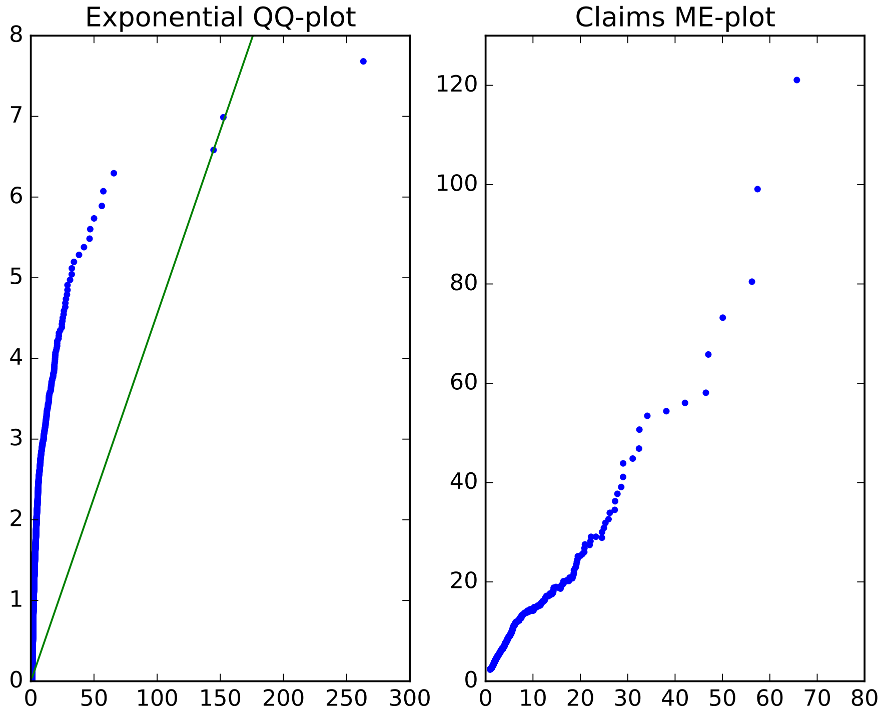

Figure 2: (Left) Exponential QQ-plot of the insurance claims against a standard exponential distribution with a reference straight line. (Right) The mean-excess plot of the insurance claims

As a second test for heavy tailed behaviour, we apply the so called ratio of maximum and sum method. In this method one takes a sub-sample of the data of size k the sample data $X_1,\ X_2, \ldots,\ X_k$

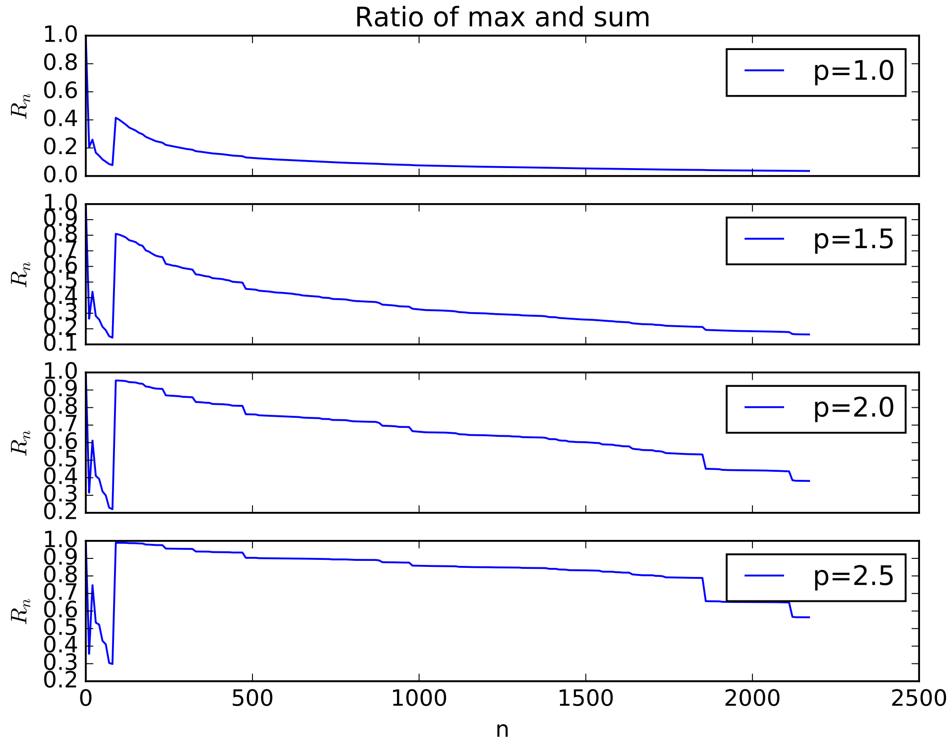

Figure 3: Ratio of the maximum and sum method for moment orders $p = 1$, 1.5, 2, 3. The lack of

and forms the ratio

$$\begin{aligned} R_k(p) &= \frac{M_k(p)}{S_k(p)},\ p > 1 \\ S_k(p) &= |X_1|^p + |X_2|^p + \ldots + |X_k|^p \\ M_k(p) &= \max(|X_1|^p, \ldots, |X_k|^p). \end{aligned}$$

We then plot $R_k(p)$ against various sample sizes k for a given p and see if the ratio decays with increasing sub-sample size. If the ratio does not decay for a given p, this indicates the pth moment of the data does not exist. Moments failing to exist indicates a heavy tailed distribution. For example, the Cauchy distribution does not have a first moment (i.e its mean does not exist) and is a Heavy tailed distribution. Figure 3 shows the ratio of sum and maximum method for a variety of p values. The figure indicates that the second moment does not exist, or at least the moments past $p = 2$ do not exist. At this stage we believe a heavy tailed distrbution would be an appropriate fit to the data, or at least to the right tail of the data where large claims occur. A distribution with heavy tails that has the same qualitative behaviour as the insurance data is the Pareto distribution. A Pareto distribution distribution of shape parameter α has the df

$$F(x) = 1 - x^{-\alpha},\ x \geq 1.$$

It turns out that one can use a plot of the empirical mean excess function to determine if the tail of a given sample follows Pareto or similar distributions. Namely, an approximately linear mean excess plot (ME plot) means suggests an approximately Pareto-like distribution. We have produced an ME plot for the claims data, shown in the right panel of figure 2. One can argue an approximately linear trend in the ME plot. A Pareto distribution is therefore proposed as a parametric model for the insurance data.

3 Estimation of the Pareto shape parameter

At this stage, we suppose a Pareto parametric model for the data, looking to fit a distribution of the form

$$\bar{F}(x) = 1 - F(x) \sim Cx^{-\alpha},\, x \geq u$$

for some constant C, where $\sim$ means ”approximately equal” for large x (quantiles), and u is an exceedance level above which the data are supposed to be Pareto.

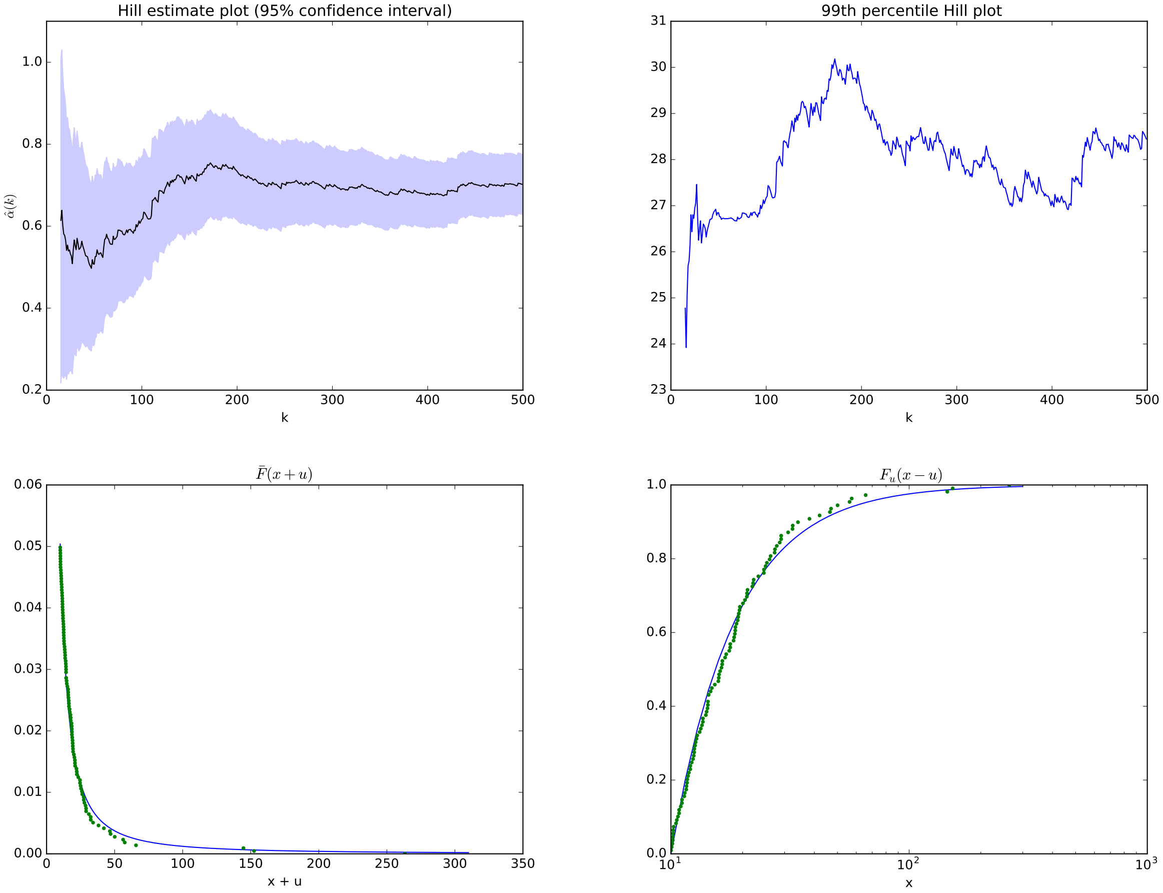

Figure 4: (Left, top) Hill estimate plot for varying sub-sample size. Shaded region is the 95% confidence interval for each estimated shape parameter. (Right, top) 99th percentile $\hat{x}_{0.99}$ as a function of k upper statistics. (Left, bottom) The tail fit $\bar{F}(x)$ plotted for $x \geq u$, or $\bar{F}(x+u)$ plotted for $x \geq 0$. (Right, bottom) The fitted mean excess function $F_u(x-u)$. The bottom row's distribution are fitted for exceedance level $u = 10$, or 109 upper order statistics.

One statistical estimator for the shape parameter α is the Hill estimator. The Hill estimator is calculated using a subsequence of k upper order statistics, i.e the k largest values in the data set, so that different Hill estimator's are obtained using different sample sizes k. This can be thought of as using data above the threshold u, so that the threshold u can be put into correspondence with the k upper order statistics used. For our data, Hill's estimates for varying subsample sizes are shown in figure 4. With a threshold of $u = 10$ set, there are $k = 109$ claims in excess of u. The corresponding Hill estimate of α is $\alpha = 1.451$. With the tail of the distribution fitted to a Pareto, we turn to estimating a high quantile, for example the 99th quantile $\hat{x}_{0.99}$, as a application of the fit. This is a question of significant importance in risk management. Knowing this quantile is like asking the question: what is the minimum size of those large claims that one is likely to get one percent of the time? The top right plot in figure 4 shows a Hill-based estimate of $\hat{x}_{0.99}$ for various exceedance levels. There is a stable, horizontal level of estimates from about $k = 50$ to $k = 110$. At the exceedance level $u = 109$, $\bar{x}_{0.99} = 27.18$. We conclude the introductory investigation here.

4 Conclusion

In this investigation, we have performed an introductory data analysis on a famous the famous Danish insurance data set. A preliminary graphical data analysis suggested that the data were not well suited to a light tailed model fit, such as the exponential distribution, and that the statistical quality of the data is such that moments higher than the second moment of the underlying distribution do not exist. Based on the graphical procedures, a Pareto fit to the tail of the data was suggested. After fitting a power tail distribution to the data, the investigation ended in giving an estimate of the 99th percentile of the data based on the tail fit, with the 99th percentile estimated as 27.18 million Danish Kroner. There are several improvements that can be made. For example, providing confidence intervals on the estimated 99th quantile. Confidence intervals in extreme value theory are important. Usually the confidence intervals are large, giving great uncertainty in the estimates obtained. This is a well known feature of extreme value theory estimates, with some saying that this is why there is ‘no free lunch’ even if one employs the theory. It would also be worth engaging a Generalised Pareto Distribution based investigation.

References

[1] Embrechts, P., Klüppelberg, C., & Mikosch, T. (2013), Modelling extremal events: for insurance and finance.*Springer Science & Business Media*, Vol 33.