Risk Management: First Assessment

Lirida Papallazi
lirida.papallazi@studenti.unimi.it

February 2021

1 The Dataset and pre-processing

The dataset chosen to perform this analysis is the US Accidents dataset available on Kaggle. This specific dataset gathers the information of the traffic accidents from 2016 to 2020, with reference to the US territory. Among the variables we find the identifier of the accident, the severity of the same (expressed through a value between 1 and 4), latitude and longitude, City and State, the description of the event, the address.. The pre-processing involved the construction of a new dataset from the starting one, containing the number of accidents per city. The total for each city was subsequently sorted in descending order and given a ranking value.

By computing the logs for both the totals for each city and the ranking we are then able to produce the log-log plot of our data.

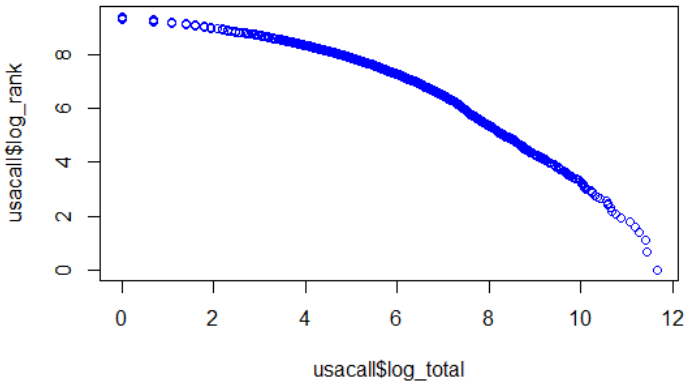

Figure 1: Rank/frequency plot in Log-log scale for the US Car Accidents dataset

Figure 1 seems to confirm that our data follow a regularly varying or heavy tail distribution, where with heavy tail we mean that the tail of the distribution under the analysis is heavier that that of an exponential function.

2 Extreme Value Theory and GDP

Risk management typically focuses on the behavior of the tails, since it is interested in detecting extreme values and therefore, possible huge losses. Extreme Value Theory is the theory which focuses on the the tails of the distribution, taking into account the heavy tails properties. Pareto distributions become of vital importance, and even more its generalisation, namely the Generalized Pareto Distribution, whose df is defined as follows:

$$G_{\xi,\beta} = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - e^{-x/\beta} & \text{if } \xi = 1. \end{cases}$$

Figure 2: Distribution function of a Generalized Pareto Distribution (GDP)

As a matter of fact the GDP can be used to model the tails of distribution and in particular the values exceeding a certain threshold u. It is clear that this method opens up many issues such as:

- the selection of the threshold u, decided by trying a different range of possible thresholds and via QQ-plots
- estimation of the parameters ξ and β, estimated via MLE
- how can we understand if real data follow this model, via mean excess function.

3 Mean excess function

The mean excess function determines both the model validation and the choice of the threshold, and it represents the conditional mean of the exceedance size over the threshold, given that the exceedance occurred. We opted, based on the mef plot, for 4 thresholds at: 6.000, 8.000, 11.000 and 18.500.

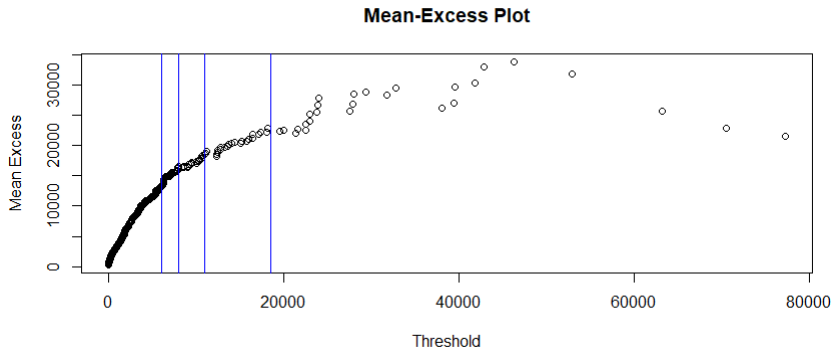

Figure 3: Mean excess function with 4 possible thresholds at 6.000, 8.000, 11.000 and 18.500

Table 1: ξ and β ML estimates for the selected thresholds

	u6000	u8000	u11000	u18500
ξ	6.035809e-01	3.610107e-01	2.955752e-01	1.530367e-01
β	6.370884e+03	1.076801e+04	1.334396e+04	1.920729e+04

4 QQ-plots

We have decided to proceed the analysis considering the thresholds at 6.000 and 8.000 in order to have more data on the tail. To understand whether or not our thresholds are reasonable and if there is one that is better than the other we build the QQ-plots.

Our thresholds seem to be reasonable but probably the one set at 8.000 could be a better choice since it fits our data better.

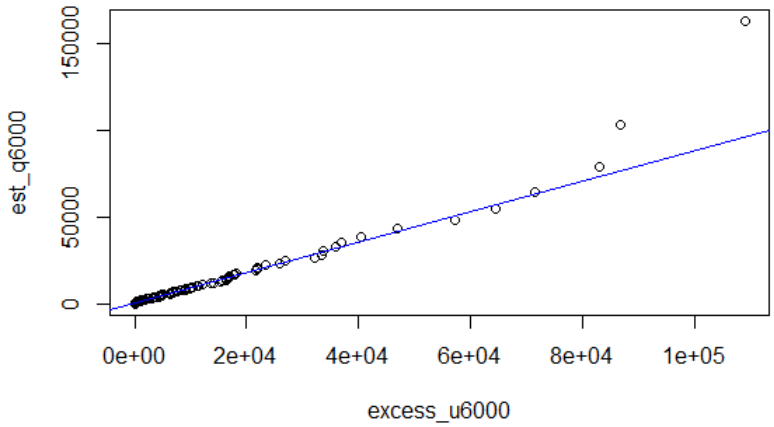

Figure 4: QQ-plot at threshold u=6.000

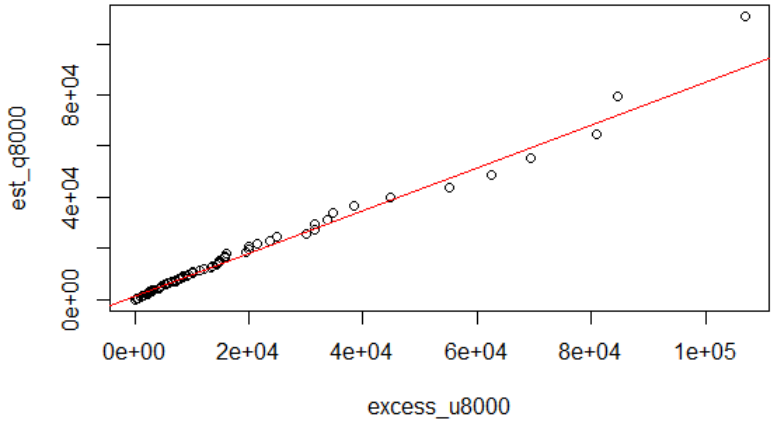

Figure 5: QQ-plot at threshold u=8.000

5 Applicability of the Risk Measures

Lastly, for both our main thresholds, 8.000, we compute the quantiles and the Expected Shortfalls using the riskmeasures function of the QRM package, at a confidence interval of 0.99 and 0.995.

The VaR (0,990) of our distribution considering u8000, tells us that there is a 1% of probability that, in 3 years, there will be more than 2792.148 car accidents in any City of the Database. The VaR (0,995) of our distribution considering u8000, tells us that there is a 0,05% of probability that, in 3 years, there will be more than 9792.105 car accidents in any City of the Database.

6 Conclusion

Based on the analysis conducted, there are strong suggestions that our data follow a heavy tail distribution. In particular, given the values of the ξ estimated - lower than 0.5 but still positive - we could also say that we are facing a heavy tail distribution with finite mean and finite variance.