ENSAI - GS & GR 3A 2023-2024 ACADEMIC YEAR
Gilles Stupfler & José Gregorio Gomez Garcia Email : gilles.stupfler@univ-angers.fr

Extreme Value Theory - Practical session 1: Heavy-tailed models

Supporting data for this practical session are:

- Rainfall data from Orlando, Florida, USA, see `Florida_rainfall.RData`,
- Financial data on the CAC40 index, see `CAC40.csv`,
- The Society of Actuaries (SOA) Group Medical Insurance Large Claims Database from the R package `ReIns`,
- 2010 US Census data ranking US cities by population, see `census_USA_2010.xlsx`.

You *may not* have the time to work on all these real data examples. The idea is to allow you to work on a data set that suits your interests. In each case, it will be important to code the estimators yourself, and to compare your procedures with those of R packages such as `evir`, `evd`, `evt0` and `extRemes`. As a preliminary step, you may therefore want to do the following things:

1. For a fixed sample of data of size n and $1 \leq k < n$, implement the Hill and Weissman estimators. The result should be a pair of vectors of size $n-1$. As one of the samples of data is fairly large, it will be useful to speed the code up as much as you can.
2. Implement diagnostic tools (QQ-plots...) to judge whether the heavy-tailed assumption makes sense.
3. Download a few packages. Good choices in this practical session are `evir` and `evt0`. These will help to see if your results make sense (and you might find that one of the packages does something wrong!)

1. Rainfall data

The variable of interest is `sum_rain_2m_inches`, representing daily total rainfall at a weather station near Orlando, Florida, USA, during the American storm season, which lasts from August to October. Accompanying this variable is the day of recording, from 1st August 1998 to 31st October 2020.

1. Represent the data, first ignoring the time series structure, and then by plotting the data as a time series. Do you notice seasonality, autocorrelation, nonstationarity?

2. Can you find a reasonable statistical model for the whole of the data?

3. Check that there is evidence of a heavy right tail in this data set.

4. Calculate an estimate for the extreme value index of the data. Justify your choice of tuning parameters.

5. Calculate an estimate for the extreme quantiles at level 0.99, 0.995 and 0.999. Do these make sense?

6. Can you provide a confidence interval for the extreme value index and for these extreme quantiles? If so, why? If not, can you suggest a method that would allow you to do so?

2. CAC40 data

Financial data about stock market indices and equity prices are typically not stationary. When using this data, isolate first the `Close` variable, representing daily closing prices. Then, construct the daily log-return variable X_t defined as follows: if S_t is the closing price on day t, the daily log-return on day t is $\log(S_t/S_{t-1})$. This variable X_t will be the variable of interest, rather than the nonstationary price S_t. Accompanying this variable is the date of recording, from 1st March 1990 to 8th July 2021.

1. Represent the data, first ignoring the time series structure, and then by plotting the data as a time series. Do you notice seasonality, autocorrelation, nonstationarity?

2. Can you find a reasonable statistical model for the whole of the data?

3. Check that there is evidence of a heavy right tail in this data set.

4. Calculate an estimate for the extreme value index of the data. Justify your choice of tuning parameters.

5. Calculate an estimate for the extreme quantiles at level 0.995 and $1 - 1/n$, where n denotes the sample size. Do these make sense?

6. Can you provide a confidence interval for the extreme value index and for these extreme quantiles? If so, why? If not, can you suggest a method that would allow you to do so?

3. SOA data

The only variable provided in this example is the amount of money related to a medical claim exceeding \$25,000 in 1991 in the USA. This data set can be loaded by typing `data(soa)` after having loaded the `ReIns` package.

1. Represent the data. Can you find a reasonable statistical model for the whole of the data?

2. Check that there is evidence of a heavy right tail in this data set.

3. Calculate an estimate for the extreme value index of the data. Justify your choice of tuning parameters.

4. Calculate an estimate for the extreme quantiles at level 0.995 and $1 - 1/n$, where n denotes the sample size. Do these make sense? Do you think that the level 0.995 can be considered extreme here?

5. Can you provide a confidence interval for the extreme value index and for these extreme quantiles? If so, why? If not, can you suggest a method that would allow you to do so?

4. US Census data

The variable of interest is the 2010 population size in US cities having more than 50,000 inhabitants. Other variables provided are population projections in subsequent years.

1. Represent the data. Can you find a reasonable statistical model for the whole of the data?

2. Check that there is evidence of a heavy right tail in this data set.

3. Calculate an estimate for the extreme value index of the data. Justify your choice of tuning parameters.

4. Calculate an estimate for the extreme quantiles at level 0.99, 0.995 and 0.999. Do these make sense? Can you provide an interpretation for such quantiles?

5. Can you provide a confidence interval for the extreme value index and for these extreme quantiles? If so, why? If not, can you suggest a method that would allow you to do so?