Lund Institute of Technology
Mathematical Statistics

Centre for Mathematical Sciences

Statistical Modeling of Extreme Values,
FMSN55/MASM15

Computer Assignment 2

This assignment is a compulsory part of the course. At the end of the session each group's results will be reviewed and graded as *pass* or *fail*. Note that in order to be able to finish the assignment in the specified time above, you have to read through the whole assignment in addition to the necessary parts in the book and manuals described below **before** you attend the computer session.

Threshold Models

Before you start this computer assignment you should read Chapters 4 and 5 in Coles.

1 Getting started

Log in at one of the PCs in the computer room MH:230 or MH:231 using your *STIL*-account. Click on the icon "MClogin" on the desktop and login again with the same user name and password. This will attach the hard drive `"L:"` where your working directory will be saved. Note that you need to do this **before** you start up the software package `R`. Choose the latest version of `R` from the `Start` menu. If you have problems either logging in or starting `R`, ask for help.
Some tips and hints when typing in `R` code:

- `R` is case sensitive (so `LM()` and `lm()` are not equivalent).
- `R` is tolerant to the use of spaces, so `x <- 1` and `x<-1` are equivalent; though, the former being considered to be more readable.
- You can use the arrow keys to speed things up. The 'up' arrow gives you the previous command that you typed.

- The usual prompt sign for `R` is `>`. If you get a `+` prompt sign instead, it means that `R` is awaiting the completion of the previous command that you typed in. This can happen because you have forgotten to close parentheses, for instance. Just type in the remainder of the command.

2 Daily Rainfall Data

We will use the data on the daily rainfall series at a location in south-west England recorded over the period 1914-1962. The dataset was introduced first in Example 1.6, page 9 of the book and the extreme value analysis of data has been discussed in Section 4.4.1, pages 84-86. To repeat this analysis, start `R` and write `library(in2extRemes)` to attach the related libraries and get access to the dataset and `R` functions. To start the graphical user interface run the command `in2extRemes()` in `R` window.
The `R` source object `rain.R` contains daily rainfall accumulation which will be analyzed in this part of the assignment. As usual, the dataset can be loaded from the following locations

- `/usr/common/extremvarde/R/datasets/` (from Linux computers in the labs)
- `P:\Rdata` (from Windows computers in the labs)
- `http://www.maths.lth.se/matstat/kurser/fms155mas231/datasetsR.html`.

To read the dataset into `R` choose `File`→`Read Data` in `in2extRemes` window. Note that the dataset has been saved as `R` source so you should choose `R source` under `File Type` in the window which pops up. You do not need to change any other options but do not forget to assign a name of your choice to the dataset under `Save As(in R)`.

2.1 Statistical analysis

It is important to understand how you can access different results of your analysis in `extRemes` package. In the following we will assume that you have saved the dataset with the name `rain` in `R`. Otherwise you just need to change the following commands accordingly.
After reading the dataset you can run the command `names(rain)` to see the names of different parts in the list which has been created. To access each part of the list you need to add $ and the corresponding name to the object. For instance, `rain$data` contains the data.

Analyze the data by modeling exceedances of observations over a threshold according to the following. Assume all confidence intervals have confidence level 95%.

1. Plot the rainfall data against observation number (`Plot`→`Scatter Plot`).

2. Plot the mean residual life plot for the dataset (`Plot`→`Mean Residual Life Plot`). What is an appropriate threshold for this dataset according to this approach?
   Answer:

3. Plot the model based approach plot for threshold choice for the dataset (`Plot`→`Fit POT model to a range of thresholds`). Note that for each threshold a GPD has to be fitted to to the exceedances over that threshold. In order to avoid numerical difficulties choose about 20 thresholds in the range of $[0, 50]$. What is an appropriate threshold for this dataset according to this approach?
   Answer:

4. Find maximum likelihood estimates of the parameters for generalized Pareto distribution (`Analyze`→`Extreme Value Distributions`). Do not forget to choose Generalized Pareto (GP) as `Model Type` otherwise the default distribution (GEV) will be fitted to the dataset. Run the command `names(rain)` again. Note that a new component with the name `models` has been added to the object. Writing `names(rain$models)` will show you the models which you have fitted so far. If this is your first analysis you will see only `fit1`. Write `names(rain$models$fit1)`

to see which parts are included as the result of the fit. For instance `rain$models$fit1$results` will print the parameter estimates, negative log likelihood, Hessian matrix and so on.
Answer:

5. Check that you can get the same results by running the command `fevd(x = value, data = rain$data, threshold = 30, type = "GP")` in `R` (change the threshold if you have chosen another value than 30) .
Answer:

6. Find the covariance matrix of ML-estimates (`Analyze`→`Fit Summary`). You can also find the covariance matrix by using either `summary(rain$models$fit1)` or `parcov.fevd(rain$models$fit1)`.
Answer:

7. Note that the covariance matrix in `in2extRemes` is calculated based on the inverse of observed Fisher information. To verify this, calculate the inverse of the Hessian matrix by `solve(rain$models$fit1$results$hessian)`. Does the result agree with your calculations in the previous part?
Answer:

8. Confidence intervals for the parameters in GPD based on asymptotic Normal distribution of MLEs (`Analyze`→`Parameter Confidence Intervals`).
   Answer:

9. Check that your results for confidence intervals agree with what you get from this command: `ci(rain$models$fit1,type="parameter")`.
   Answer:

10. Confidence interval for the shape parameter in GPD based on profile likelihood (`Analyze`→`Parameter Confidence Intervals`). Change the "Number of points at which to calculate profile likelihood" to 100.[1]
    Answer:

[1] Note that according to the `in2extRemes` tutorial (page 35) `in2extRemes` only allows for changing the search limits for return levels (for which this approach is arguably more necessary) and not individual parameter estimates. The software will attempt to find an appropriate search range, but will not always succeed; leading to intervals that are clearly incorrect (e.g., the estimated parameter does not fall inside the limits). In such a case, the search range can be changed in the command-line code of `extRemes`, but not from the `in2extRemes` window.

11. Compare your result in the previous part with this command:

    ```
    ci(rain$models$fit1,type="parameter",method="proflik",
    which.par=2,xrange= c(-0.5,1),nint=100).
    ```

    Answer:

12. Maximum likelihood estimates of 10- and 100-year return levels and their variances. Note that for matrix multiplication in `R` you have to use "`%*%`" . See pages 81-82 in the book for the formulas.

    Answer:

13. Confidence intervals for 10- and 100-year return levels based on delta method (`Analyze`→`Parameter Confidence Intervals`).

    Answer:

14. Confidence intervals for 10- and 100-year return levels based on profile likelihood method (`Analyze`→`Parameter Confidence Intervals`). Change the "Number of points at which to calculate profile likelihood" to 100. Note that you can specify "Profile Search Range" for return level. Choose the range $[50, 200]$ in both cases.

    Answer:

15. Diagnostic plots for the GPD fit with discussion of different plots (`Plot`$\rightarrow$`Fit Diagnostics`).
Answer:

16. Maximum likelihood estimate of σ under the hypothesis $H_0 : \gamma = 0$. Note that this is the special case in GPD. Write the density function for $\gamma = 0$ and find the ML-estimate of σ. Note that you can access the whole dataset by `rain$models$fit1$x` so for example all observations which are greater than 30 can be found by

`rain$models$fit1$x[rain$models$fit1$x-30>0]`.

Note that you have to subtract the threshold from the data to obtain the exceedances.
Answer:

17. Check that you can get the same results by running the command `fevd(x = value, data = rain$data, threshold = 30, type = "Exponential")` in `R` (change the threshold if you have chosen another value).
Answer:

18. Confidence intervals for σ under the hypothesis $H_0 : \gamma = 0$.
Answer:

19. Likelihood ratio test for hypothesis $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$.
Answer:

20. QQ-plot for the case $H_0 : \gamma = 0$. Compare this with the corresponding plot for the GPD fit. You can construct a QQ-plot in `R` by using functions `sort` and `qexp`; see the functions help pages for details.
Answer:

3 Wooster Temperature Series

Daily minimum temperature (degrees below 0 F.) are given in the `R` source file `wooster.R` (available at the usual locations; see page 1). Start with loading the dataset to `in2extRemes` first. Note that the file is saved as `R` source so follow the instructions in Example 1, page 3, of `in2extRemes` tutorial to load the data. Below we will extract the temperature data corresponding to winter sessions in the dataset.

As discussed in Chapter 5 of the book, to obtain an approximately stationary series consider only the temperatures in winter. To obtain winter temperatures in November-February, you can use the following commands in `R`. You can also copy and paste these commands from the file `assignment2.txt` which is stored in the following locations:

- `/usr/common/extremvarde/R/datasets/` (from Linux computers in the labs)
- `P:\Rdata` (from Windows computers in the labs)
- `http://www.maths.lth.se/matstat/kurser/fms155mas231/datasetsR.html`.

You can copy and paste from this file to `R` directly.
Please note the following: before submitting the following commands you need to read the `wooster.R` dataset to `in2extRemes` first. You need also to save this dataset as `wooster` otherwise the following commands will result in an error.

```
winter = c(1:60, 305:365)
wooster.w = wooster
wooster.w$data =   wooster$data[c(winter, winter + 365,
winter + 365 * 2, winter + 365 *3, winter+365*4),]
wooster.w$data[,2]= - wooster.w$data[,2]
wooster.w$data <-
cbind(wooster.w$data,year.number=c(rep(1:5,rep(121,5))))

wooster.w.u.minus10.r.2<-
decluster(wooster.w$data[,"value"],threshold=-10,r=2,
groups=wooster.w$data[,"year.number"])
attributes(wooster.w.u.minus10.r.2) <- NULL
wooster.w.u.minus10.r.2  <-
as.in2extRemesDataObject(wooster.w.u.minus10.r.2)

wooster.w.u.minus20.r.2<-
decluster(wooster.w$data[,"value"],threshold=-20,r=2,
groups=wooster.w$data[,"year.number"])
attributes(wooster.w.u.minus20.r.2) <- NULL
wooster.w.u.minus20.r.2  <-
as.in2extRemesDataObject(wooster.w.u.minus20.r.2)

wooster.w.u.minus10.r.4<-
decluster(wooster.w$data[,"value"],threshold=-10,r=4,
groups=wooster.w$data[,"year.number"])
attributes(wooster.w.u.minus10.r.4) <- NULL
wooster.w.u.minus10.r.4 <-
as.in2extRemesDataObject(wooster.w.u.minus10.r.4)
```

```
wooster.w.u.minus20.r.4<-
decluster(wooster.w$data[,"value"],threshold=-20,r=4,
groups=wooster.w$data[,"year.number"])
attributes(wooster.w.u.minus20.r.4) <- NULL
wooster.w.u.minus20.r.4 <-
as.in2extRemesDataObject(wooster.w.u.minus20.r.4)
```

Note that we have also created a column named `year.number`. As the name suggests this contains values 1-5 which correspond to five years of temperature data. This is used to ensure that the clusters do not cross boundary of each year so we will never have temperatures from two different years in one cluster. The above commands decluster the negated Wooster winter temperature data for all 4 combinations of $u = -10, -20$ and $r = 2, 4$. For instance `wooster.w.u.minus10.r.2` contains the declustered data for $u = -10$ and $r = 2$. These datasets can be used to fit GPD to maximum of each cluster (`Analyze`→`Extreme Value Distributions`).
You can see the plot of the winter data by submitting the following command in `R`: `plot(wooster.w$data[,"value"])`. Use the declustered versions of `wooster.w` created above and analyze the exceedances of the (negated) data over thresholds $u = -10, -20$ and clusters obtained from $r = 2, 4$.

3.1 Statistical analysis

The Wooster dataset contains 605 observations (i.e. 121 observations per year). Note that you will obtain slightly different values than those given in the book due to the different choice of winter data.

1. Repeat the whole analysis in Section 5.3.3 of the book and calculate a table similar to Table 5.1 on page 102. As discussed in the course, in order to calculate return levels we need to take into account the rate at which clusters occur (see formula (5.5), page 103). Note that $\hat{\eta}_u\hat{\theta} = \frac{n_c}{n}$ where $n = 605$ is number of observations and n_c is number of exceedances over u after declustering so for instance the following command in `R`

   ```
   sum(wooster.w.u.minus10.r.2$data[,2]>-10)
   ```

   gives number of clusters for $u = -10$ and $r = 2$. Similarly,

   ```
   sum(wooster.w$data[,2]>-10)
   ```

gives n_u which is the number of exceedances over u in the dataset before declustering.
Answer:

2. Analyze your results and write your conclusions about different models.
Answer: