Natural Hazards and Extreme Statistics Training
R Exercises

Hugo Winter

Introduction

The aim of these exercises is to aide understanding of the theory introduced in the lectures by giving delegates a chance for some hands-on experience with a statistical computing program. The most common statistical programming language is R. For more information see https://www.r-project.org/.

I do not assume any previous experience with R and the aim of the course is not to learn the syntax of R so that you can do your own analyses. As such, most of the code to execute the commands is given in this document and can be copy and pasted into the R terminal. It is more important to broadly appreciate the different steps to an extreme value analysis (EVA) to better understand the results in reports that you may see in the future.

1 Annual maxima

Walkthrough

The first set of exercises will concern the concept of modelling maxima. We are going to analyse the daily rainfall data (in mm) from south-west England used as an example in the lectures. This data The data set comes from the ismev package and can be loaded in with the following commands:

```
install.packages("ismev")  # Only needs to be run once
library(ismev)
data(rain)
```

A variable called rain should now have been loaded into the workspace (type ls() into the terminal to check if this is the case). We shall plot the data to get a feel of what the data look like.

```
plot(rain)
```

Firstly we are going to model the annual maxima of the rainfall data. Before we can fit a distribution, the annual maxima need to be extracted:

```
years <- rep(1:48, rep(c(365,365,366,365), times = 12))[-17532]
rain.ann.max <- unlist(lapply(X = split(rain,years), FUN = max))
```

You do not need to know what the code does, just know that we now have a variable `rain.ann.max` that contains annual maxima. Maxima can be modelled using the Generalised Extreme Value (GEV) distribution with parameters μ, σ and ξ which refer to the location, scale and shape parameters respectively. It is possible to fit a GEV distribution with likelihood using the `fevd` function from the `extRemes` package. To load in the package:

```
install.packages("extRemes")  # Only needs to be run once
library(extRemes)
```

Then to fit the GEV you need to type:

```
gev.fit <- fevd(x = rain.ann.max, type = "GEV", time.units = "years")
```

The output from the fit is contained within the variable `gev.fit`. To obtain diagnostic plots to assess the fit then type:

```
plot(gev.fit)
```

From the output we can see that the QQ-plot and return level plot suggest that distribution provides a good fit to the data. We shall now look directly at the estimates of the location, scale and shape parameters (μ, σ, ξ) by typing

```
gev.fit$results$par
```

The estimates are given as $\mu = 40.78$, $\sigma = 9.73$ and $\xi = 0.11$ (to two decimal places). The location and scale parameters of the GEV (very) loosely define the mean and spread of the distribution fitted to the annual maxima. The most important parameter to interpret is the shape parameter ξ. Here, $\xi > 0$ which suggests that the distribution has a heavy upper tail. This suggests that trying to fit a light-tailed distribution might lead to underestimation of extremal quantities such as the return level.

It is important to assess the uncertainty associated with these parameter estimates. This uncertainty arises from the fact that the data we observed is just one possible data sample. One common way to quantify uncertainty is to estimate 95% confidence intervals. These are the bounds within which 95 times out of 100 the parameter estimate should lie in. To obtain estimates for these intervals we run:

```
ci.fevd(x = gev.fit, alpha = 0.05, type = "parameter")
```

One important fact to notice is that the 95% confidence intervals for the shape parameter contain 0 and as such we cannot be certain about the shape of the upper tail from the observed data. Further study would be required to conclude whether a heavy tailed model is required. Most uncertainty in estimates of parameters and return levels is generated by an uncertain shape parameter. A desire to reduce the uncertainty in ξ motivates the regional frequency analysis approach detailed later.

Finally, we need to estimate return levels outside the scope of our data using the fitted GEV model. To do this we shall use the following command:

```
gev.rl <- return.level(x = gev.fit, return.period = c(10,100,1000,10000),
                       do.ci = TRUE, alpha = 0.05)
```

By typing `gev.rl` into terminal we can see the 10, 100, 1000 and 10000 year return levels with 95% confidence intervals. The estimate of the 10,000-year return level is given as $z_{10000} = 194$mm (21, 366). The 95% confidence interval is very wide in this situation due to the sparsity of the data used to fit the model and the uncertain shape parameter. At this stage, if assessing the report, it is important to decide whether the results are physically realistic and if not, question the study. In this situation better approaches could be used to obtain less uncertain estimates for the 10,000-year return level.

Exercises

The `ismev` package also contains the data set `portpirie` which contains annual maximum sea levels (in m) at Port Pirie in south Australia over the period 1923-1987. These data are already annual maxima so you don't need to extract them. After running `data(portpirie)` the data can be accessed by typing `portpirie$SeaLevel` into the terminal.

1.1 Plot the sea level data.

1.2 Fit the GEV distribution to the data. What are the parameter estimates telling you about the shape of the distribution?

1.3 Does the GEV provide a good fit to the observed data?

1.4 Can you obtain 95% confidence intervals for the parameters?

1.5 What are the 95% confidence intervals for the 50-year and 10,000-year return levels? Set the argument `alpha = 0.3` in the function `return.level`. What has this done to the intervals? Why?

2 Threshold exceedances

Walkthrough

When modelling maxima, much data is discarded by just using the annual maxima; this is an inefficient way of estimating parameters. There may be other large values that are not the maximum but could still provide important information about the distribution of large values. To overcome this problem we now turn to the problem of modelling extremes using threshold exceedance methods. The Generalized Pareto distribution (GPD) is the main distribution for modelling exceedances above a high threshold (Davison and Smith, 1990).

One of the main issues when trying to fit the GPD is what level to set the threshold. Two diagnostics for choosing the threshold were given in the lecture, which we apply here for the daily rainfall data contained in `rain`:

```
mrl.plot(rain)   # Mean residual life plot
gpd.fitrange(rain, umin = 20, umax = 50)   # Parameter stability plot
```

It is important to note that the diagnostics do not suggest a particular threshold to use, rather a set of potential candidates. In the mean residual life plot we are looking for the lowest level above which a straight line can be drawn that doesn't bisect either of the confidence intervals. In the parameter stability plot we are looking for the lowest threshold above which a horizontal line crosses through all the confidence bars.

As in Coles (2001) we could choose the threshold $u = 30$mm:

```
u <- 30
```

Now we can fit the GPD using the command:

```
gpd.fit <- fevd(x = rain, threshold = u, type = "GP", time.units = "days")
```

As with the GEV fit, diagnostic plots to assess the fit can be obtained using `plot(gpd.fit)`. The diagnostics suggest that the GPD is providing an adequate fit to the data. Estimates and confidence intervals for each of the parameters can be obtained in a similar way as above:

```
gpd.fit$results$par
ci.fevd(x = gpd.fit, alpha = 0.05, type = "parameter")
gpd.rl <- return.level(x = gpd.fit, return.period = c(10,100,1000,10000),
                       do.ci = TRUE, alpha = 0.05)
```

The parameter estimates are given as $\sigma_u = 7.44$ and $\xi = 0.18$ (to two decimal places). The use of additional data has led to a different estimate for the shape parameter, but what has happened to the uncertainty estimates? These have reduced for both parameters, although zero is still contained within the 95% confidence intervals for the shape parameter. However, it is looking more likely that the shape parameter is positive and there is indeed a heavy tail. However, uncertainty bounds are still very wide for the return levels which suggests that regional frequency analysis could still be useful.

Up until this point we have fitted the GPD to all the exceedances and not applied any declustering. To apply declustering to the data we need to run the following commands:

```
decl.data <- decluster(x = rain, threshold = u, method = "runs", r = 1)
gpd.fit.decl <- fevd(x = decl.data, threshold = u, type = "GP",
                     time.units = "days")
```

Above we have set the run length to 1, i.e. after 1 non-exceedance a cluster is terminated. This will provide slightly different results (the difference often depends on amount of clustering in the data set). In most situations some declustering will be required prior to fitting a threshold model.

Exercises

The choice of threshold is observed to be a very uncertain exercise and subjective. Lets see what happens when the threshold is set at the wrong level.

2.1 Set the threshold at `u <- 50`. To see how many data points are used in the analysis enter `sum(rain>u)`. What happens to the estimates for the parameters and return levels?

2.2 Now set the threshold at `u <- 3`. Plot the model fit diagnostics, are there any issues? What proportion of the data are being used in the model fit?

You may have noticed when obtaining the confidence intervals above that the lower bound is often negative and thus we are effectively saying that we can have negative rainfall! This occurs as the standard approach to defining confidence intervals is to use something called the delta method. Without going into too much detail, this approach estimates the variability in the parameter or return level from the data and adds and subtracts this from the estimate. As such the confidence interval will be symmetric and often will include values that are impossible.

One approach to remedy this is to use bootstrapping. This approach resamples the observed data to try and take account for variability in the underlying sample. The parameters and return levels can then be estimated for each resampled data set and as such we avoid impossible values.

2.3 What return levels do you get from typing in the following? (NB: this can be slow)

```
gpd.rl.bs <- return.level(x = gpd.fit, return.period = 10000, do.ci = TRUE,
                          alpha = 0.05, method = "boot")
```

2.4 How does `gpd.rl.bs` differ from `gpd.rl`?

Back in the lectures, diagnostics were shown using a light-tailed distribution. This was the exponential distribution and can be fitted using the function `fevd` by supplying the argument `type = "Exponential"`. The exponential distribution only has a rate (or scale) parameter λ and no shape parameter (as only light-tailed behaviour is permitted).

2.5 Fit the exponential distribution to the rainfall data. What difference does this make to the estimates of the return levels? Do the model diagnostics suggest this is a valid model for this example?

We have also observed that it is important to decluster the data to ensure that we fit our extreme value model to independent exceedances.

2.6 What is the effect of using different values of the run length? Try setting the run length to 2 and rerun the code at the end of the walkthrough.

3 Multivariate extreme value analysis

Walkthrough

In the previous sections we looked at how to fit extreme value models to the extremes of a single variable; the examples we investigated were rainfall and sea level. In many situations we wish to estimate the probability of two or more variables occurring at the same instant. To this end, multivariate extreme value approaches are outlined in this section.

Throughout this section we shall be analysing concurrent daily measurements of wave and surge height at a single location off south-west England. This dataset is contained within the `ismev` package that you should have loaded in from the previous exercises. To load in the data type in `data(wavesurge)` which should load in the variable `wavesurge` which has two columns. To plot the data type:

```
plot(wavesurge)
```

From the plot we can observe that there is some dependence between surge and wave height and as such treating them as independent would be an over-simplification and could lead to incorrect estimates of important extreme quantities. One standard measure to assess dependence is the correlation coefficient ρ which can be obtained by typing:

```
cor(wavesurge)
```

The correlation is found to be $\rho = 0.3$, which confirms that the variables are positively related, such that when one variable increases the other is likely to as well. But this measure is driven by central values and therefore does not inform us about the behaviour in the tails.

The extremal dependence measure $\chi(u)$ (Coles et al., 1999) was introduced in the lecture as a measure for investigating the level of dependence at extreme levels. To estimate this measure for the observed sample of wave and surge heights we need to use the `texmex` package. This can be installed in the standard way:

```
install.packages("texmex")  # Only needs to be run once
library(texmex)
```

To estimate the value of $\chi(u)$ for u associated with the 90% quantile run the following lines:

```
ext.q <- 0.9   # Sets the quantile
chi.val <- chi(data = wavesurge, nq = 1, qlim = c(ext.q,ext.q))
chi.val$chi
```

By setting `ext.q` higher and lower we can investigate the extremal dependence behaviour for different extreme quantiles. To look at the dependence associated with the 10,000-year return level, run the following:

```
ext.q <- 1-1/(365*10000)   # Sets the extreme quantile
chi.val <- chi(data = wavesurge, nq = 1, qlim = c(ext.q,ext.q))
```

The code should produce an error message as this level is too high and cannot be used to obtain an estimate of $\chi(u)$ from the observed sample. We need to fit a multivariate extremes model at a lower threshold and use this to extrapolate to such a high level. To fit the conditional extremes approach (Heffernan and Tawn, 2004) for the dataset then type:

```
thresh <- 0.9
mex.fit <- mex(data = wavesurge, which = 1, mqu = thresh, dqu = thresh)
```

Here, the threshold used to fit the model has been fixed at the 90% quantile for each of the variables. If you were completing this analysis it would be necessary to use threshold diagnostics to choose this level. Now we have fitted the model we can obtain diagnostics by typing:

```
plot(mex.fit)
```

The most important plot here is the second one that shows the behaviour of extreme conditional quantiles. These should be seen to capture the extremal behaviour of the data well.

Now, to estimate the value of $\chi(u)$ we can now simulate values conditional on having a wave height greater than the 10,000-year return level. To do this type:

```
ext.q <- 1-1/(365*10000)
num.sim <- 5000   # Choose the number of points to simulate
mex.pred <- predict(mex.fit, pqu = ext.q, nsim = num.sim)
```

The simulated data can be visualised by looking at the final plot returned when you type `plot(mex.pred)` into the terminal. The red dots show the original data set and the grey crosses show the simulated data set above the 10,000-year return level. One interesting by product of the above command are the conditional quantiles that can be found by typing:

```
summary(mex.pred)
```

This shows the quantiles of the distribution of the variables conditional on wave height having exceeded the 10,000-year return level. Finally, $\chi(u)$ can be estimated by typing:

```
mex.fit2 <- mex(data = wavesurge, which = 2, mqu = thresh, dqu = thresh)
surge.rl <- predict(mex.fit2, pqu = ext.q)$data$pth
sum(mex.pred$data$simulated[,2]>surge.rl)/num.sim
```

The estimate of $\chi(u)$ at the 10,000-year return period is found to be 0.012. This value is very close to zero and suggests that it is highly unlikely that the 10,000-year return level will occur concurrently.

References

Coles, S. G. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer Verlag.

Coles, S. G., Heffernan, J. E., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. Extremes, 2(4):339–365.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). Journal of the Royal Statistical Society: Series B, 52(3):393–442.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). Journal of the Royal Statistical Society: Series B, 66(3):497–546.