Graybill VIII: 6th International Conference on Extreme Value Analysis

Workshop 22 June, and Conference 23-26 June 2009

Practice Sets for Workshop on
"An introduction to the analysis of extreme values using `R` and `extRemes`"

1 R Preliminaries

For an introduction to using R code, see the manuals available at `http://www.R-project.org`. The following exercises are designed to introduce you to the way R works, and how to get help when faced with new challenges. You can also use the R search engine from the R project web page given above to try to find answers to specific questions about how to do something in R.

1. Give the matrix `y` created in the lectures (i.e., `y <- cbind( c( 2,1,5), c(3,7,9))`) column names, and write the matrix out to a file (**Hint**: See the help files for `colnames` and `write.table`).

2. Now write `y` out to a file as a comma separated file.

3. Read the files created in 1 and 2 above back into R, and assign them different names (**Hint**: See the help files for `read.table` and `read.csv`).

4. Check the class of these objects read from 3 above.

5. See the help file for the class type found in 4 above. This is a very important class in R. It is a cross between a matrix and a list object (see help files for these types as well). Unlike a list object, it must have the same numbers of rows for each column, and all columns must be a vector (e.g., a list can have wildly different component types, such as a function as one component, and a matrix as another). Unlike a matrix, a data frame can have different types of vectors across columns, such as a character vector in one column and a numeric vector in another.

6. If `y` still has -999.0 as one component (as in the lecture), set it to NA. If it does not have an NA component, add one in somewhere. Using this original `y` object, matrix multiply it by the vector `x <- c( 1, 2, 0)`. That is, find $y^T x$ (**Hint**: See the help page for `%*%`, that is, type `?"%*%"` with the quotes). Now do the same for the `y` objects read in for 3 above. Anything unusual?

Did you get an extra row for the csv file? If so, try writing it out again using `row.names=FALSE`. Matrix multiplication is not always possible with data frames. Convert the data frame object to an object of class "matrix" using the `matrix` function.

7. Interpolate the daily maximum 8-hour ozone values for 18 June 1987 from the `fields` data set `ozone2` to a grid using thin-plate spline interpolation, and make a surface plot of the results (**Hint** Run the examples in the help file for `fields`).

8. What class is the object `fit` from 7 above (i.e., `fit` is the name assigned in the help file for `fields`)? (**Hint** use the function `class`).

9. See the help file for `surface`. Not very helpful, is it? See the help file for `surface.Krig` instead. Some functions, known as "method" functions, have specialized functions for different types of objects. Three very common examples are `predict`, `summary` and `plot`. List out all of the methods currently available for the function `plot` (**Hint**: See the help file for `methods`).

10. Methods are common when fitting a statistical model (e.g., a regression). List out all of the methods for objects of class "Krig." For objects of class "lm" ("lm" is the class associated with the main function in R for fitting linear models (e.g., linear regression), called by the same name, `lm`). **Hint**: be sure to specify the `class` argument here.

11. List out the objects in the current working directory for R (**Hint**: `ls()`).

12. Use `search()` to determine the position of the package `fields`, then use the `pos` argument of the `ls` command to list out the functions contained in the `fields` package.

13. Use the `detach` function to detach the `fields` package from the current R session.

2 Fitting to a Stationary GEV Distribution

1. Simulate a sample of size 500 from the GEV distribution with location 3, scale 1.5 and shape 0.1 using the `extRemes` GUI windows, and assign it the name, `prob2.1`. (**Hint** Under the File menu, select Simulate Data then Generalized Extreme Value Distribution (GEV)).

2. Fit the GEV to the simulated data from 1 above, and check the button to plot the diagnostics.

3. Do the assumptions for fitting the GEV to these simulated data appear reasonable?

4. Plot the profile likelihood for the shape parameter found from the fit in 2 above. What are the 95% CI for this parameter? Is this parameter significantly different from zero at the 5% significance level?

5. Using the **Simulate Data** feature under the **File** menu of the `extRemes` GUI window, simulate two series from the GEV distribution. One with sample size of 10, and one with sample size of 100.

6. Fit the above two simulations to a GEV distribution, and check the **Calculate L-moments** check button. Which estimates are closer to the values of the distribution from which the simulations were realized? Note that all p-values, etc. are only for the ML estimates.

7. Try the above two exercises again for different parameter values.

8. What conclusions would you make regarding sample size and estimation method for the GEV parameters based on these simulation exercises?

9. Read in the `SEPTsp` data set using the `extRemes` GUI windows (i.e., Under File, select Read Data, and browse for the file `SEPTsp.R`, etc.).

10. See the help file for this data set to learn what each field represents.

11. Make a line plot of the maximum temperature over a one month period against year for these data.

12. Make a scatter plot of the standard deviation of maximum temperature against the maximum temperature. Does there appear to be much correlation?

13. Fit the GEV to the maximum temperature field.

14. Make a QQ-Plot for this fit. Do the assumptions for using the GEV appear reasonable for these data?

15. Estimate 95% CI's for the shape parameter. What can you say about the behavior of maximum temperature for Sept-Iles, Québec based on these data?

16. Make a line plot of the year against minimum temperature.

17. Fit the GEV to the minimum temperature data, and check the QQ-Plot (**Hint**: remember to take the negative transformation of the variable using the `extRemes` GUI windows).

18. Estimate a 95% CI for the shape parameter. What can you say about minimum temperature for Sept-Iles, Québec based on these data?

3 Threshold Excess Models

1. Read in the Phoenix minimum temperature data (i.e., from file Tphap.R) using the `extRemes` GUI's or by using `data( Tphap)`. If you use this latter convention, then use `as.extRemesDataObject` to convert the data frame into data recognized by `extRemes` GUI windows.

2. Take the negative transform the variable `MinT` from the `Tphap` data set using the `extRemes` GUI windows.

3. Fit the GPD over a range of thresholds to the (negative) `MinT` variable. Does -73 degrees appear to be a reasonable threshold? (**Hint**: You may need to try different ranges and numbers of thresholds).

4. Fit the GPD using a threshold of -73 degrees to the (negative) `MinT` variable.

5. De-cluster the (negtive) `MinT` series for the threshold of -73 degrees (de-cluster by `Year`), and re-fit the GPD to the de-clustered series. How does the fit compare to the previous one? How do the QQ-Plots for each compare?

6. What conclusions would you make about the appropriateness of a stationary GPD model for these data? Do the underlying assumptions appear to be valid based on the qq-plots?

7. Under the **Plot** menu, select **Return Level Plot** to make a return level graph for the fit to the non-declustered data. From the R prompt, type `X11()` to open a new plot device, and then make a return level graph for the fit to the de-clustered data. Compare the two graphs side-by-side. Do they differ much?

8. Load the data set called `Denversp` from the package `extRemes`.

9. See the help file for this data set to learn what it contains.

10. Make a scatter plot of precipitation against hour. What do you notice?

11. Make a scatter plot of precipitation against day and then year. Any patterns or trends?

12. Use **Fit threshold ranges (GPD)** under the **Plot** menu to choose a threshold for fitting the GPD to these data (**Hint**: use 0.1 and 0.8 as the lower and upper limits). Does 0.395 mm appear to be a reasonable choice for a threshold?

13. Make another scatter plot of precipitation against hour, and add a red horizontal dashed line at 0.395. Do the data appear to be independent over the threshold?

14. Fit a GPD to the Denver precipitation data. Plot the diagnostics for the fit. Is this model reasonable for these data?

15. Estimate a 95% CI for the shape parameter. Is the shape parameter significantly different from zero at the 5% level for either fit? Does this concur with the likelihood ratio test for $\xi = 0$? (**Hint**: Use **Parameter confidence intervals** under the **Analyze** menu).

16. De-cluster the precipitation field from `Denversp` using runs de-clustering with `r=1`.

17. Re-fit the newly de-clustered field to the GPD. Is this fit any different from the previous ones?

18. Estimate the Poisson rate parameter associated to a threshold of 0.395 mm for the de-clustered precipitation data.

19. Find the Poisson rate parameter from the fit in 17 above. Is it nearly the same as the estimate obtained in 18 above? (**Hint**: use the relation $\hat{\lambda} = \left[1 + \frac{\hat{\xi}}{\hat{\sigma}}(u - \hat{\mu})\right]^{-1/\hat{\xi}}$).

20. Make a QQ-Plot for the point process model fit (if you haven't already). Do the assumptions for the model appear to be reasonable?

4 Linear temporal trends

1. Load the `Denmint` data set of the `extRemes` package using `data( Denmint` instead of the GUI windows.

2. Take the annual maximum of the negative of the minimum temperature. (**Hint**: use `DenmintAM <- aggregate( -Denmint$Min, by=list( Denmint$Year), FUN=max, na.rm=TRUE)$x`).

3. Make a line plot of year against negative minimum temperature. Does there appear to be any temporal trend in these data? (**Hint**: use `yr <- unique( Denmint$Year)` and `plot(yr, DenmintAM, type="l")`).

4. Fit a linear regression of year against negative minimum temperature (**Hint**: See the help file for `lm`). Is there a significant linear trend in these data (**Hint**: use the `summary` function on the `lm` fitted object)?

5. Use `cbind` to make a new matrix that has `yr` and `DenmintAM` as its columns, and give the columns names. Use `as.extRemesDataObject` to convert this matrix into one that the `extRemes` GUI windows will recognize. Then, fit the negative minimum temperature to a GEV (without any trend).

6. Look at the QQ-Plot for this fit. Do the model assumptions appear to be reasonable? Did the likelihood-ratio test accept the Gumbel hypothesis?

7. Estimate a 95% CI for the shape parameter. Do the results concur with the likelihood-ratio test? (**Hint**: check the **Plot profile likelihoods** button without putting lower and upper limits for xi. Then do it again with -0.5 as the lower limit and 0.5 as the upper limit. Do the resulting CI's change? Does the conclusion concerning the Gumbel hypothesis change?).

8. Make a return level plot for the negative minimum temperature with (delta method) 95% CI's.

9. Interpret this return level plot for a gas/power company wanting to understand the risk of too much demand for gas in Denver in any given year (**Hint**: remember the return levels are for the negative of minimum temperature).

10. Fit the negative minimum temperature data to the GEV with a linear trend in the location parameter for $t = 1, 2, \ldots$.

11. Check the QQ-Plot for this fit. Do the assumptiions for the model fit appear to be reasonable?

12. Perform a likelihood ratio test for $\mu_1 = 0$ in the fit from 10 above (**Hint**: select **Likelihood-ratio test** under the **Analyze** menu. Then, select the `DenmintAM` Data Object, then `gev.fit1` in the first list box (i.e., M0) and `gev.fit2` in the second list box (i.e., M1), and click **OK**). Is the result consistent with the result from the regression fit from 4 above?

13. Simulate 1000 random variables from a GEV with a trend of 0.25 in the location parameter using the `extRemes` GUI windows (i.e., under **File** menu, select **Simulat Data**, etc.).

14. Make a line scatter plot of the resulting sample. Is there a trend?

15. Fit the sample to a GEV without a trend.

16. Check the QQ-Plot. Are the model assumptions reasonable here?

17. Fit the sample to a GEV with a linear trend in the location parameter.

18. Perform a likelihood ratio test for $\mu_1 = 0$ on this fit.

19. Check the QQ-Plot for the fit with a linear trend in the location parameter. Are the model assumptions reasonable?

20. Try fitting the sample to a GEV with a linear trend in the scale parameter (using `siglink=exp`), and check the QQ-Plot. Is this a reasonable model?

5 Cyclic variation

1. Use `data( Denversp)` to load this `extRemes` data set as a regular R data frame. We fit the GPD to these data and de-clustered versions of these data in the threshold excess practice above. Now, let's fit the Poisson rate parameter including an annual cycle using the `glm` function. First, make a binary vector where 1 indicates an excess over 0.395 mm (**Hint**: use `ind <- Denversp$Prec > 0.395`). Next, make two vectors containing the cyclic trends over time (i.e., create vectors containing $\sin(2\pi t/365.25)$ and $\cos(2\pi t/365.25)$, where $t =$ `Denversp$Hour`). Now use `glm` with `family=poisson()` to fit the model $\hat{\lambda}(t) = \hat{\lambda}_0 + \hat{\lambda}_1 \sin(2\pi t/365.25) + \hat{\lambda}_2 \cos(2\pi t/365.25)$ (where $\hat{\lambda}(t)$ is the indicator vector), and use `summary` to see the results. Is there a significant (say at the 5% level) annual cycle in the Poisson rate parameter?

2. Fit the Denver precipitation data to a point process model with no parameter covariates, and threshold of 0.395 mm. (**Hint**: Convert `Denversp` to an `extRemes` Data Object).

3. Check the QQ-Plot. Are the model assumptions reasonable?

4. Fit the Denver precipitation data to a point process model for a threshold of 0.395 mm, and with a cyclic variation in the location parameter as $\hat{\mu}(t) = \hat{\mu}_0 + \hat{\mu}_1 \sin(2\pi t/24)$ for $t =$`Denversp$Hour` (**Hint**: select **Transform Data** under the **File** menu to make a trigonometric transformation of the `Hour` field. For the point process fit).

5. Perform a likelihood ratio test for $\mu_1 = 0$ in the above model. Is the fit significant? Are the model assumptions reasonable?

6. Now do the same but use both the sine and cosine fields. You might need to use `Method = BFGS quasi-Newton`. Are the model assumptions reasonable here?

7. Fit the Denver precipitation data to a point process model with a cyclic trend in the scale parameter (i.e., $\log \sigma(t) = \sigma_0 + \sigma_1 \sin(2\pi t/24) + \sigma_2 \cos(2\pi t/24)$). Is the trend significant? Are the model assumptions reasonable based on the QQ-Plot?

8. Given the results here, and the results from de-clustering previously, which approach would you recommend for these data?

6 More Practice

1. List out the arguments for the function, `optim` (**Hint**: use the `args` function).

2. See the help file for the function, `optim`.

3. Type `date` from the R prompt, and then hit return. What happens?

4. Now, type `date()` and hit return. What happens?

5. See the help file for `extRemes` to see, among other things, a list of the data sets included with the package.

6. Analyze the `Peak` data set. Is a block maxima or threshold excess model more appropriate here? Do there appear to be any tends in the data?

7. Analyze the maximum winter temperature for Sept-Iles. Do any of the other fields included with the data set make sense to try as covariates?