Interdisciplinary Workshop: Effects of climate change: coastal systems, policy implications, and the role of statistics

Course March 16-17, and Workshop March 18-20 2009

Practice Sets for Short Course on Statistical Software and Extreme Value Analysis

1 R Preliminaries

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`)?

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.

2 Fitting to a Stationary GEV Distribution

1. Simulate 500 maxima from samples of size 40 from the normal distribution.

2. Simulate 500 maxima from samples of size 40 from the exponential distribution.

3. Fit the GEV to the simulated data from 1 above (**Hint**: use `gev.fit` from package `ismev` or `fgev` from package `evd`).

4. Fit the GEV to the simulated data from 2 above.

5. Plot the QQ-Plot for the fit from 3 above. Do the assumptions for fitting the GEV to these simulated data appear reasonable?

6. Plot the QQ-Plot for the fit from 4 above. Do the assumptions for fitting the GEV to these simulated data appear reasonable?

7. Plot the profile likelihood for the shape parameter found from the fit in 3 above. Use `locator(2)` to estimate a 95% CI for the parameter. Is this parameter significantly different from zero at the 5% significance level?

8. Do the same for the fitted shape parameter in 4 above. Is it significantly different from zero at the 5% level?

9. Simulate samples of size 10, 20, 50, 100 and 500 from the GEV distribution with location parameter 2, scale parameter 1.5 and shape parameter -0.5 (**Hint**: use `gen.gev` from packae `extRemes` or `rgev` from package `evd`).

10. Fit each sample from 9 above to the GEV. Check the QQ-Plots for each, and estimate 95% CI's for the shape parameter in each case. Are the fits reasonable?

11. Load the `SEPTsp` data set from the package `extRemes`. What is the class of this data object? Use `colnames` or `names` to list the available fields.

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

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

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

15. Fit the GEV to the maximum temperature field.

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

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

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

19. Fit the GEV to the minimum temperature data, and check the QQ-Plot (**Hint**: remember to take the negative of the values first).

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

3 Threshold Excess Models

1. Load the data set called `Denversp` from the package `extRemes`.
2. See the help file for this data set to learn what it contains.
3. What is the class for this data set?
4. Make a scatter plot of precipitation against hour. What do you notice?
5. Make a scatter plot of precipitation against day and then year. Any patterns or trends?
6. Use `stats( Denversp)` to see a summary of the data (**Hint**: `stats` is a function in the `fields` pacakge).
7. Use `gpd.fitrange` 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?
8. 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?
9. Use `aggregate` to obtain a vector of maximum daily precipitation (i.e., aggregate by year and day).
10. Make a line plot of these aggregated values against `1:1302`, and add a red dashed horizontal line at 0.395 mm. Do the threshold excesses appear more independent now?
11. Fit a GPD to both the original Denver precipitation data, and to the aggregated data from 10 above. How do the two fits compare? How do the QQ-plots compare?
12. Estimate a 95% CI for the shape parameter from both fits. Is the shape parameter significantly different from zero at the 5% level for either fit?
13. De-cluster the original precipitation field (or the already de-clustered field) from `Denversp` using the `dclust` function from `extRemes` with `r=1`.
14. Re-fit the newly de-clustered field to the GPD. Is this fit any different from the previous ones?
15. Estimate the Poisson rate parameter associated to a threshold of 0.395 mm for the (physically) de-clustered precipitation data.

16. Fit the (physically) de-clustered precipitation data to a point process model.

17. Find the Poisson rate parameter from the fit in 16 above. Is it nearly the same as the estimate obtained in 15 above? **Hint**: use the relation $\hat{\lambda} = \left[1 + \frac{\hat{\xi}}{\hat{\sigma}}(u - \hat{\mu})\right]^{-1/\hat{\xi}}$).

18. Make a QQ-Plot for the point process model fit. Do the assumptions for the model appear to be reasonable?

4 Linear temporal trends

1. Load the `Denmint` data set of the `extRemes` package.

2. Take the annual maximum of the negative of the minimum temperature.

3. Make a line plot of year against negative minimum temperature. Does there appear to be any temporal trend in these data?

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. 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?

7. Estimate a 95% CI for the shape parameter.

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**: the likelihood ratio statistic, D, is given by $D = -2(\ell(M1) - \ell(M0))$, where $\ell(\cdot)$ is the negative log-likelihood for the model with the trend and without the trend, resp. The probability of exceeding this value based on a χ^2_ν distribution, with ν the difference in the number of parameters between $M1$ and $M0$, gives the p-value for the test. Use the `lower.tail=FALSE` argument with `pchisq` to get the p-value). Is the result consistent with the result from the regression fit from 4 above?

13. Simulate 1000 maxima from normal distributions of size 30 with mean increasing at a rate of 25% (i.e., with slope 0.25), and standard deviation of 10.

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. If `Denversp` is not already loaded, re-load it. 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. Next, make two vectors containing the cyclic trends over time, $t = 1, 2, \ldots$ (i.e., create vectors containing $\sin(2\pi t/365.25)$ and $\cos(2\pi t/365.25)$. Now use `glm` with `family=poisson()` and `summary` to fit the model ($\hat{\lambda}(t) = \hat{\lambda}_0 + \hat{\lambda}_1 \sin(2\pi t/365.25) + \hat{\lambda}_2 \cos(2\pi t/365.25)$) and see the results. Is there a significant (at the 5% level) annual cycle in the Poisson rate parameter?

2. Fit the model $\hat{\lambda}(t) = \hat{\lambda}_0 + \hat{\lambda}_1 \sin(2\pi t/24) + \hat{\lambda}_2 \cos(2\pi t/24)$, where $t =$`Denversp$Hour` to the binary vector in 1 above. Is there any significant cyclic trend for this model?

3. Fit the Denver precipitation data to a point process model with no parameter covariates, and threshold of 0.395 mm.

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

5. 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**: use the argument `method="BFGS"`).

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

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?