Extreme-Value Analysis Problem Sets using R

Eric Gilleland

May 12, 2023

Block Maxima

1. Simulate a sample of size 1000 from a Gamma distribution (e.g., `?rgamma`) with both shape and rate parameters of 1 and another with both parameters of 1/2 (save them as `g1` and `g2`). Fit the GEV distribution to each sample and check the plot diagnostics. Are the assumptions for fitting teh GEV reasonable?

2. Now obtain a sample of size 100 of maxima from samples of size 1000 of the above two distributions and fit the GEV distribution to each (save these as `gmax1` and `gmax2`). Now how do the assumptions look?

3. Use `ci` with argument `type = "parameter"` to obtain normal approximation CI's for the GEV parameters in each of the above fits. Is zero in the interval for the shape parameters?

4. Use `ci` to find the estimated 25- and 100-year return levels (we'll assume each data point represents a year) along with 95% CI's. Hint: see `?fevd` for help.

5. Generate a sample of size 100 by taking maxima of random samples of size 1000 from a GEV($\mu = 0, \sigma = 1, \xi = 1/4$), and fit the GEV distribution to this sample. Check the diagnostics. Are the assumptions reasonable? Given that the GEV distribution is max stable, do the results make sense?

6. Simulate samples from a GEV($\mu = 2, \sigma = 3/2, \xi = -1/2$) of size 10. Fit the GEV distribution using the default MLE method, L-moments (use argument `method = "Lmoments"`), and GMLE (`method = "GMLE"`). Compare the fit diagnostics and estimates. Use `ci` with argument `type = "parameter"` for each fit to compare the resulting uncertainty estimates. How do they compare with each other and the "true" parameters? Use `ci` to obtain the 100-year return level estimate with 95% CI's for each fit. How do they compare (Hint: you can use `rlevd` to obtain the "true" 100-year return level)?

7. Draw a sample of size 100 of maxima from normally distributed samples of size 1000 and fit the GEV to the sample. Given that the Gumbel distribution (i.e., $\xi = 0$) is in the domain of attraction of the normal distribution, is your estimate for $\xi = 0$? Why not? Is zero within the 95% CI's for ξ?

8. Draw samples from the three types of GEV distributions and look at their histograms. What do you notice about the histograms and the tail behavior?

Counting Extremes

1. Load the dataset `FCwx` (i.e., `data( "FCwx" )`, and use `FCwx` to learn about the data).

2. Apply the following code to obtain counts of the number of days that the daily maximum temperature exceeds 95^o F.
```
tempGT95 <- c(aggregate(FCwx$MxT, by = list(FCwx$Year),
    function(x) sum(x > 95, na.rm = TRUE))$x)
yr <- unique(FCwx$Year)
```

3. Plot the counts. Does the frequency appear to be constant over time, or does it appear to be changing? If so, how is it changing?

4. Fit the Poisson distribution to these count data using MLE and test whether or not the mean equals the variance (Hint: use `fpois`).

5. Using the `yr` data from the code above, use `glm` with argument `family = poisson()` to fit a non-homogenous Poisson distribution (a Poisson regression) to the count data using year as a covariate. Use `summary` to test for the inclusion of year as a covariate. Is it significant? Does that jive with your answer to question 3 above? Do the diagnostic plots imply that the assumptions for using this model appear reasonable?

6. Instead of year, tyr fitting a model with an indicator for before 1950 v. after 1950. Is the AIC lower? Do the model assumptions appear reasonable? Any other models to try?

Point Process

1. Load the `Denversp` data set, and plot precipitation against hour. What do you notice?

2. Plot precipitation against `Day`. Plot it against `Year`.

3. Use `mrlplot` to make a mean-residual life plot of the Denver precipitation data. At about what precipitation value does the plot become linear within the uncertainty? See `?mrlplot` for more on what this plot means.

4. Use `threshrange.plot` to fit the GPD to the precipiation data over a range of thresholds and plot the (transformed) scale and shape parameters against the threshold values. How low of a threshold appears reasonable?

5. Use `atdf` to make auto tail dependence plots of the precipitation data using the probability threshold of 0.8 (note that this is a probability threshold and not a direct precipitation threshold). Use `extremalindex` with a threshold of 0.395 mm to estimate the extremal index and, if necessary, obtain an estimated run length should runs declustering be useful.

6. Fit a Poisson point process to the precipitation data using a threshold of 0.395 mm and check the fit diagnostics. Note that the data are hourly (24 hours per day) over only the month of July, which has 31 days, so there are $31 * 24 = 744$ days per year, so you should use `time.units = "744/year"` though it does not seem to matter in this case.

7. Use the relation $\log \lambda = -1/\xi \cdot \log\{1 + \xi(u - \mu)/\sigma\}$ to estimate the Poisson rate parameter.

8. Create an object that indicates day or night using `daynight <- (Denversp$Hour <= 5) | (Denversp$Hour >17)`, attach it to the `Denversp` data frame, and fit the Poisson to the precipitation data again but allowing the location parameter to vary according to day or night. Hint: use argument `location.fun = ~ dayornight`. Use a different name for the fitted object than you used for the previous fit. Does the model have a lower AIC/BIC? Use `lr.test` to perform a likelihood-ratio test on whether the inclusion of the diurnal cycle is significant or not.

And now for something difficult...

Here, we will use the R library `astsa` by Stoffer and Poison (use `citation( "astsa")` to see the full reference). In particular, we will look at the fish recruitment and Soutern Oscillation Index data. We will look at both the extreme high recruitment and extreme low recruitment. First, it is helpful to follow some of the code from the book by Shumway and Stoffer (2017):

Shumway, R. H. and D. S. Stoffer, 2017. *Time Series Analysis and its Applications: With R Examples.* Springer International Publishing, Switzerland, 562 pp. (Fourth Edition).

Plot the two time series.

```
 par( mfrow = c(2,1) )
plot( soi, ylab = "", xlab = "", main = "Southern Oscillation Index" )
plot( rec, ylab = "", xlab = "", main = "Recruitment (new fish)" )
```

Plot the ACF and cross-correlation plots.

```
par( mfrow = c(3,1) )
acf( soi, lag.max = 48, main = "SOI", xaxt = "n" )
```

```
axis( 1, at = 0:4, labels = 0:4 * 12 )
acf( rec, lag.max = 48, main = "Recruitment", xaxt = "n" )
axis( 1, at = 0:4, labels = 0:4 * 12 )
ccf( soi, rec, lag.max = 48, main = "SOI v.  Recruitment",
    ylab = "CCF", xaxt = "n" )
axis( 1, at = (-4):4, labels = (-4):4 * 12 )
```

Plot the lagged scatter plots with trend lines and cross lag plots with trend lines.

```
lag1.plot( soi, 12 )
lag2.plot( soi, rec, 8 )
```

Add a lag-6 SOI column and line things up properly. Then fit a linear regression between fish recruitment and the lag-6 SOI.

```
fish <- ts.intersect( rec, soiL6 = lag( soi, -6 ), dframe = TRUE )
summary( fit1 <- lm( rec   soiL6, data = fish, na.action = NULL ) )
```

Add a dummy variable that simply states whether SOI is less than zero or not, and do another regression.

```
dummy <- ifelse( soi < 0, 0, 1 )
fish <- ts.intersect( rec, soiL6 = lag( soi, -6 ), dL6 = lag( dummy, -6),
dframe = TRUE )
fit <- lm( rec   soiL6 * dL6, data = fish, naaction = NULL )
summary( fit )
plot( rec   soiL6, data = fish )
lines( lowess( fish$soiL6, fish$rec ), col = 4, lwd=2 )
points( fish$soiL6, fitted( fit ), pch = "+", col=2 )
plot( resid( fit ) )
acf( resid( fit ) )
hist( fish$rec)
```

Now that we've thoroughly looked at the main part of the data, time to tackle the extremes. Add a variable to the `fish` data frame that is the negative of recruitment. Doing so allows us to use exactly the same mechanics we used for threshold exceedances for threshold deficits. We just need to keep in mind that the functions will not re-negate anything for us later. The scale and shape parameters will not need to change, but locations and thresholds will need to be negated, as well as return levels.

I suggest trying to fit the EVD's to the negative values first. You might even use `blockmaxxer` to find annual minima (i.e., maxima of the negated values) and fit the GEV to these data first. Note, however, how many annual minima there are. In trying to fit EVD's to these data, understand that it will not go smoothly and may not have a good solution. When fitting to the non-negated data, keep in mind that the MLE does not exist when $\xi < -1$. Some things to try when trying to get a good fit include, but

are not limited to: change the optimization routine with `optim.args`, change the initial values to the optimization routine (see `?fevd`), try using L-moments (with GPD or GEV only), try using GMLE and Bayesian estimation. Also try incorporating the SOI lagged values.

The point of this exercise is to learn all of the techniques at your disposal to try to arrive at a fit. Many times one or more of the methods will work, but sometimes it is impossible.