Exercise 4: Functions

Niels Richard Hansen

December 8, 2016

The *Base R Cheat Sheet* may be helpful.

Problem formulation

The (symmetric) *running mean* of a sequence $y_1, \ldots, y_n$ using m neighbors to either side is the sequence

$$\overline{y}_i = (y_{i-m} + y_{i-m+1} + \ldots + y_{i+m-1} + y_{i+m})/(2m+1).$$

It is undefined for the boundary indices $i = 1, \ldots, m$ and $i = n - m + 1, \ldots, n$. The running mean can be seen as a simple *scatter plot smoother* if the y-sequence is ordered according to an x-variable. That is, if the data points (x_i, y_i) fulfill that $x_1 \leq x_2 \leq \ldots \leq x_n$, then $(x_i, \overline{y}_i)$ is a scatter plot smoother. This exercise is about implementing the running mean scatter plot smoother in R, and it is supposed to be solved using an R script and not R Markdown.

A test data set

You can use the following data set as a test data set.

http://nielsrhansen.github.io/Dong/easysmooth.txt

- Read this data set into R. Make a scatter plot of `y` against `x` using `ggplot` or `qplot`.

Running mean implementation

- Make one implementation of the running mean computation, add the result to the plot. Try different values of m. How did you handle the boundary problem?
- If you didn't do so already, wrap the implementation up into a function, which takes `y` as well as `m` as arguments.
- Did you remember to check if `x` was sorted in increasing order? (See `?sort`).
- Insert a breakpoint inside your function. Evaluate a function call and note the Environment pane.

More advanced problems

You can consider one or more of the following points for additional challenges.

- Implement a function that takes the data set as argument and returns a data set with the running mean added as a column. Use `ggplot` to draw the scatter plot as well as the running mean from the resulting augmented data set.
- Make the return value from your function an object of a class of your choice (you decide the name). Write a plot method for this class.
- Experiment with alternative ways of implementing the running mean. Your first choice might not be the most efficient.