Homework 1: Linear Smoothers
Due April 5th
PART I

1. In R create a 10000×10 matrix and subtract from each element of a row the mean of that row using for-loops, `apply` and `sweep`. Do this 25 times and make a histogram of how much time it takes each one to run. Do the same for a 10×10000 matrix.

2. The intensity data set contains Modified Mercalli (MM) intensities (first column) recorded after the 1989 Loma Prieta earthquake in California. The numbers 0 through 9 indicate the MM intensity in different locations. Assume the intensities are observations of random variable defined via a model with 9 latent classes:

$$\text{Prob}\{I_j = i\} = \text{Prob}\{a_{i-1} < z < a_i\} = \pi_i, i = 1, \ldots, 9$$

   for observations $j = 1, \ldots, n$ with

$$\log(-\log(1 - \text{Prob}\{I_j > i\})) = \alpha_i$$

   Find the Maximum Likelihood Estimates (MLE) of the α_is using the function `glm()`. Hint: Look at this paper: McCullagh, P. (1980). Regression models for ordinal data. JRSC B. 42, 109-127.

3. For the CD4 cell count data set compute a smooth using

   (a) a line, a parabola,

   (b) a bin smoother dividing the data into 3, 6, 12, and 24 parts,

   (c) a running mean using the 50,100,500,1000 nearest neighbors,

   (d) a kernel smoothers using span h=1, 2, 3, and 4 years,

   Compare these estimates.

4. Prove that parametric, bin, running mean, and kernel smoothers are all linear smoothers. For all these smoothers, define the matrices $\mathbf{S}$ so that we can write the smooth as as $\mathbf{Sy}$. Make two plots of $S_j(x)$ as functions of j. One for an x in the "middle" and another for an x at the border.

5. Suppose we add a point (x_{j+1}, y_{j+1}) to a neighborhood containing j points. If for these j points, the sample means of X and Y, sample variance of X and sample covariance of X and Y are denoted with $\bar{x}_j$, $\bar{y}_j$, Σ_j^x, and Σ_j^{xy} respectively, show that

$$\begin{aligned}
\bar{x}_{j+1} &= (j\bar{x}_j + x_{j+1})/(j+1) \\
\bar{y}_{j+1} &= (j\bar{y}_j + y_{j+1})/(j+1) \\
(j+1)\Sigma_{j+1}^x &= j\Sigma_j^x + \frac{j+1}{j}(x_{j+1} - \bar{x}_{j+1})^2 \\
(j+1)\Sigma_{j+1}^{xy} &= j\Sigma_j^{xy} + \frac{j+1}{j}(x_{j+1} - \bar{x}_{j+1})(y_{j+1} - \bar{y}_{j+1})
\end{aligned}$$

   (a) Derive the corresponding equations for deleting a point. Show that together these can be used to update the least square slope and intercept and hence the entire smooth for a running line smoother can be obtained in O(n) operations... instead of $O(n^2)$ if we did it the naive way.

   (b) Given the answer to the previous question, why is using a Gaussian kernel smoother really slow compared to something like a box kernel?

6. Prove Theorem 2.1.

Useful R functions: `read.table, scan, apply, sapply, poly, cut, cbind, rep, lm, sort, order`