Homework 1: Linear Smoothers
Due April 19th
PART I

1. The Strontium data set was collected to test several hypotheses about the catastrophic events that occurred approximately 65 million years ago. The data contains Age in million of years and the ratios described here. There is a division between two geological time periods, the Cretaceous (from 66.4 to 144 million years ago) and the Tertiary (spanning from about 1.6 to 66.4 million years ago). Earth scientist believe that the boundary between these periods is distinguished by tremendous changes in climate that accompanied a mass extension of over half of the species inhabiting the planet at the time. Recently, the compositions of Strontium (Sr) isotopes in sea water has been used to evaluate several hypotheses about the cause of these extreme events. The dependent variable of the data-set is related to the isotopic make up of Sr measured for the shells of marine organisms. The Cretaceous-Tertiary boundary is referred to as KTB. The data show a peak at this time and this is used as evidence that a meteor collided with earth.

   The data presented in the figure represents standardized ratio of strontium-87 isotopes (^{87}Sr) to strontium-86 isotopes (^{86}Sr) contained in the shells of foraminifera fossils taken form cores collected by deep sea drilling. For each sample its time in history is computed and the standardized ratio is computed:

$$^{87}\delta\text{Sr} = \left(\frac{^{87}\text{Sr}/^{86}\text{Sr sample}}{^{87}\text{Sr}/^{86}\text{Sr sea water}} - 1 \right) \times 10^5.$$

   Earth scientist expect that $^{87}\delta$Sr is a smooth-varying function of time and that deviations from smoothness are mostly measurement error.

   Consider the model

$$y = f(x) + \epsilon, x \in I = [a, b] \subset \mathbb{R}.$$

   and assume that the ϵ's are IID $N(0, \sigma^2)$.

   (a) Consider the least squares estimates obtained when we assume that $f \in \mathcal{G}$, the space of polynomials of order 4, 6, 8 and 12. Discuss the estimates obtained under these assumptions. (Hint: talk about degrees of freedom, over-fitting and under-fitting)

(b) Given a sequence $a = t_0 < t_1 < \ldots < t_m < t_{m+1} = b$, construct $m + 1$ (disjoint) intervals

$$I_l = [t_{l-1}, t_l), 1 \leq l \leq m \text{ and } I_{m+1} = [t_m, t_{m+1}],$$

whose union is $I = [a, b]$. Define the piecewise polynomials of order k

$$g(x) = \begin{cases} g_1(x) = \theta_{1,1} + \theta_{1,2}x + \ldots + \theta_{1,k}x^{k-1}, & x \in I_1 \\ \vdots & \vdots \\ g_{m+1}(x) = \theta_{m+1,1} + \theta_{m+1,2}x + \ldots + \theta_{m+1,k}x^{k-1}, & x \in I_{k+1}. \end{cases}$$

Is the space containing these function a linear space? If so write out a basis matrix and explain how you would estimate the parameters $\boldsymbol{\theta} = (\theta_{1,1}, \ldots, \theta_{m+1,k})'$ if we were to assume that f is in this space.

(c) In the previous exercise are there any conditions you need for the $\{X_i\}$'s and I_m's for the solution to be unique?

(d) Try fitting the following piecewise polynomials (you choose the break-points)

i. k = 1, 2-pieces,

ii. k = 2, 2-pieces,

iii. k = 1, 4-pieces, and

iv. k = 2, 4-pieces.

Comment on the fits and compare them to the ones obtained in exercise 2 in Part I.

(e) Construct a test to see if the two quadratic pieces in 4b are the same. Explain why this test doesn't really make sense in practice.

2. Find an estimate of the regression function f defined for the Strontium data set using natural smoothing splines. Use cross-validation to choose a penalty coefficient λ. Make a plot of CV(λ) vs. λ and compare the "smooths".

3. Use loess to estimate the regression function f defined for the Strontium data set.

(a) Use cross-validation to choose the smoothing parameter (what R calls `span`).

(b) Assume the errors are normal and construct point-wise confidence intervals make a plot.

4. Repeat exercise problem 2 for the CD4 cell count data. Discuss why the regression model used in problem 2 isn't as useful for this case. How does this affect the interpretation of the results.

5. Derive the following formulas for a linear smoother estimate $\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$:

   (a)

$$\begin{aligned} \text{MSE}(\lambda) &= n^{-1} \sum_{i=1}^{n} \text{var}\{\hat{f}_\lambda(x_i)\} + \text{ave}(\mathbf{v}_\lambda^2) \\ &= n^{-1}\text{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda')\sigma^2 + n^{-1}\mathbf{v}_\lambda' \mathbf{v}_\lambda \\ \text{PSE}(\lambda) &= \{1 + n^{-1} tr(\mathbf{S}_\lambda \mathbf{S}_\lambda')\}\sigma^2 + n^{-1}\mathbf{v}_\lambda' \mathbf{v}_\lambda. \end{aligned}$$

   (b)

$$\text{E}\{\text{ASR}(\lambda)\} = \left\{1 - n^{-1}\text{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda \mathbf{S}_\lambda')\right\} \sigma^2 + n^{-1}\mathbf{v}_\lambda' \mathbf{v}_\lambda$$

   (c)

$$\text{CV}(\lambda) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}} \right\}^2$$

   (d)

$$\text{E}[\text{CV}(\lambda)] \approx \text{PSE}(\lambda) + 2\text{ave}[\text{diag}(\mathbf{S}_\lambda)\mathbf{v}^2].$$

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