Homework 8

Due Thursday, 11/5/20

1. (Leave one out cross validation) Let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$ be a full rank design matrix, $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$, h_{ii} be the ith diagonal element of $\boldsymbol{H}$ and $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$ be the response. For each $i = 1, \ldots, n$, let $\boldsymbol{X}_{(-i)} \in \mathbb{R}^{(n-1)\times p}$ be the design matrix with the ith row removed and $\boldsymbol{Y}_{(-i)} \in \mathbb{R}^{n-1}$ be the response with the ith element removed. Define

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\boldsymbol{X}_{(-i)}^T\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}^T\boldsymbol{Y}_{(-i)}$$

to be the ordinary least squares estimator that ignores the ith sample.

(a) Show that $\left(\boldsymbol{X}_{(-i)}^T\boldsymbol{X}_{(-i)}\right)^{-1} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} + (1-h_{ii})^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

(b) Show that the leave one out cross validation error,

$$PRESS = \sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(-i)}\right)^2,$$

can be written as

$$PRESS = \sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1-h_{ii}}\right)^2,$$

where $\hat{Y}_i$ is the ith element of $\boldsymbol{HY} \in \mathbb{R}^n$.

2. Consider the file "Gambling.txt", which contains teenage gambling data in Britain. The variables are

- sex: 0=male, 1=female
- status: Socioeconomic status score based on parents' occupation
- income: in pounds per week
- verbal: verbal score in words out of 12 correctly defined
- gamble: expenditure on gambling in pounds per year

We are interested in understanding how sex, status, income and verbal predict gambling expenditures.

(a) Regress gamble onto the other four predictors. Do you see any evidence that the mean model or constant variance assumption is violated? Are the errors normally distributed?

(b) Use Box Cox with $\lambda > 0$ to suggest a transformation of the response, $\tilde{Y}$, so that $\tilde{Y}$ satisfies the usual mean and variance assumptions. Plot $\hat{\tilde{Y}}$ vs. the estimated residuals. Does your new model appear to satisfy the constant variance assumption? Does $\tilde{Y}$ appear to be normally distributed? (Hint: the Box Cox example code can be found in Transformations.Rmd on blackboard. You may need to add a small $\delta > 0$ to gamble to get the function to work, since the function requires $Y > 0$. δ can be chosen to be arbitrarily small, like 10^{-8}.)

(c) Compute the hat matrix and plot a histogram of the leverage scores.

(i) Why should one be concerned if there are any abnormally large leverage scores? Do you see any evidence of large leverage points in these data? lecture 16, p7

(ii) Re-estimate the model from part (b) after removing the points with leverage scores $> \frac{2p}{n}$. Do the parameter estimates or standard errors change substantially?

(d) Compute the Cook's distance for each of the n points. Do any of the points appear to be influential points?

3. Non parametric regression. Assume throughout that $\boldsymbol{Y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and let $x_1, \ldots, x_n \in \mathbb{R}$. Assume that

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} N\left(0, \sigma^2\right).$$

throughout this problem for some unknown, but smooth, function f. Throughout the problem, let $K(x)$ be a kernel function that satisfies

- $K(x) \geq 0$.
- $\arg\max_{x \in \mathbb{R}} K(x) = 0$.
- $\int K(x)dx = 1$.
- For any bandwidth $h > 0$, let $K_h(x) = h^{-1}K(x/h)$.

You might find slides 14-16 of Lecture 18 useful when answering the below questions.

(a) Define $\hat{f}^{(h)}$ to be the local polynomial smoothing estimator for f with degree $d > 0$ using kernel function K_h.Show that

$$\hat{\boldsymbol{Y}} = \begin{pmatrix} \hat{f}^{(h)}(x_1) \\ \vdots \\ \hat{f}^{(h)}(x_n) \end{pmatrix} = \boldsymbol{L}^{(h)}\boldsymbol{Y}$$

for some matrix $\boldsymbol{L}^{(h)} = \left(L_{ij}^{(h)}\right) \in \mathbb{R}^{n \times n}$. What are the rows of $\boldsymbol{L}^{(h)}$ in terms $x_1, \ldots, x_n, h$ and K?

(b) Let $\boldsymbol{W}_i^{(h)} = \text{diag}\{K_h(x_1 - x_i), \ldots, K_h(x_n - x_i)\}$ and

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & x_1 - x_i & \cdots & (x_1 - x_i)^d \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n - x_i & \cdots & (x_n - x_i)^d \end{pmatrix}$$

and assume $\boldsymbol{X}_i^T \boldsymbol{W}_i^{(h)} \boldsymbol{X}_i$ is invertible for each $i = 1, \ldots, n$

(i) Show that $0 < L_{ii}^{(h)} \leq 1$.

(ii) For bandwidth, h, let df_h be the effective number of degrees of freedom defined on the midterm. Compute df_h. Does df_h need to be an integer?

(c) The Google Ngram Viewer (https://books.google.com/ngrams) lets you search for the yearly frequencies of any word or short phrase appearing on different printed sources. Notice that you can set a "smoothing" parameter when generating the plot in the Google Ngram Viewer. This smoothing procedure is a simple moving average.

For this problem, choose your own word or phrase that you consider interesting, and fit a non parametric local linear regression to the word frequency over time. To extract the raw data, type in your choice of words in the Google Ngram Viewer, set the smoothing parameter to be 0, and look at the source file of the webpage, where you can find the frequency data (right click and select "View Page Source").

Plot your estimated function on top of the raw data and report how you chose the bandwidth h, the effective degrees of freedom of your fitted function and if your fit was dependent on the choice of kernel K. Some useful kernels are

- Gaussian: $K(x) \propto \exp\left(-\frac{1}{2}x^2\right)$
- Epanechnikov: $K(x) \propto \begin{cases} 1 - x^2 & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$
- Cosine: $K(x) \propto \begin{cases} \cos\left(\frac{\pi}{2}x\right) & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$,

where $K_h(x) = h^{-1}K(x/h)$.

(d) **PhD problem**: Assume $K(x_1) \leq K(x_2)$ for $x_1 \leq x_2 \leq 0$ and $K(x_1) \geq K(x_2)$ for $0 \leq x_1 \leq x_2$. Show that if $h_2 \leq h_1$ and $\boldsymbol{X}_i^T \boldsymbol{W}_i^{(h_1)} \boldsymbol{X}_i$, $\boldsymbol{X}_i^T \boldsymbol{W}_i^{(h_2)} \boldsymbol{X}_i$ are invertible for each $i = 1, \ldots, n$, then $df_{h_1} \leq df_{h_2}$. Does this agree with your intuition? Explain.