Homework 9

Due Thursday, 11/14/19

1. 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 Lecture7Code.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?

      (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?

2. Non parametric regression. Assume throughout that $\boldsymbol{Y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times p}$ is full rank. We will assume that

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right).$$

   throughout this problem. Also, define the effective number of degrees of freedom (i.e. effective number of parameters) of an estimator $\hat{f}$ to be

$$df = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{cov}\left\{\hat{f}(\boldsymbol{x}_i), y_i\right\}.$$

(a) Define $\hat{f}^{(h)}$ to be the local polynomial smoothing estimator for f with degree $d \geq 0$ using kernel function K and bandwidth $h \geq 0$. Since weighted least squares only depends on the weights up to a change in scale, we may assume $\int K(x)dx = 1$, and to avoid pathological examples, assume $\boldsymbol{W}_i^{(h)}\boldsymbol{X}_i$ is always full rank for each $i = 1, \ldots, n$ (see slide 13 from the lecture on 10/31). Show that

$$\hat{\boldsymbol{Y}} = \begin{pmatrix} \hat{f}^{(h)}(\boldsymbol{x}_1) \\ \vdots \\ \hat{f}^{(h)}(\boldsymbol{x}_n) \end{pmatrix} = \boldsymbol{L}^{(h)}\boldsymbol{Y}$$

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 $\boldsymbol{X}, h$ and K?

(b) Show that $L_{ii}^{(h)} > 0$ and $\boldsymbol{L}^{(h)}\boldsymbol{1}_n = \boldsymbol{1}_n$. The latter implies the estimator $\hat{f}^{(h)}$ preserves the "location" of the data.

(c) Show that the effective number of degrees of freedom of $\hat{f}^{(h)}$ is $df^{(h)} = \text{Tr}\left\{\boldsymbol{L}^{(h)}\right\}$, and that $df^{(h)}$ is always greater than 0. Does $df^{(h)}$ need to be an integer? Do you expect $df^{(h)}$ to increase or decrease as h decreases? Explain.

(d) 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 simply a 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: $K(x) = \exp\left(-\frac{1}{2}x^2\right)$
- Epanechnikov: $K(x) = \begin{cases} 1 - x^2 & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$
- Cosine: $K(x) = \begin{cases} \cos\left(\frac{\pi}{2}x\right) & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1. \end{cases}$