HW2

STAT-340 Applied Regression Method

Yerim Oh

Details

Due Date

This assignment is due at 11:59 PM on the 19 of October.

Grading

The problems will be graded for correctness. An emphasis will be placed on full explanations of your thought process. You should write complete sentences and make sure the steps can be followed in the proofs.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Submit in Gradescope

The problems in this assignment require writing mathematical formulas. It can be done directly in RMarkdown using LaTeX coding. For instance, the code:

```
\begin{flalign}
\frac{d}{dx} f(x) &= \frac{d}{dx} (10 x^2 + e^{3x} - log(x)) \notag \\
&= (20 x + 3e^{3x} - 1/x). \notag
\end{flalign}
```

results in the following output:

$$\begin{aligned}\frac{d}{dx} f(x) &= \frac{d}{dx}(10x^2 + e^{3x} - log(x)) \\ &= (20x + 3e^{3x} - 1/x).\end{aligned}$$

You may write your answers up using LaTeX. However, you may write your answers to all problems by hand and include a picture in your Markdown document. The code to include a figure is as follows (from `\begin{figure}` to `\end{figure}`):

```
\begin{figure}[!ht]
\centering
\includegraphics[width=8cm]{NamePicture.png}
\end{figure}
```

I prefer you spend time understanding the problems rather than formatting them. Pick the easier way for you, but make sure your answers are neatly presented and legible.

Problem 1: Interpretation of the leverage

Outliers and their influence As discussed in class, the Least Squares estimate for β under linear regression is as follows:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Consequently, the predicted mean response is $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$, where:

$$H = X(X'X)^{-1}X' = \begin{bmatrix} h_{11} & h_{12}\dots & h_{1n} \\ h_{21} & h_{22}\dots & h_{2n} \\ \vdots & \vdots & \vdots \\ h_{n1} & h_{n2}\dots & h_{nn} \end{bmatrix}_{n\times n}$$

The leverage for observation i is the value on the diagonal of the H matrix, i.e., h_{ii}.

(a) Using the information above, write $\hat{y}_i$ as a linear combination of y_j for $j = 1, 2, ..., n$.

$$\hat{y} = Hy$$

Then,

$$\hat{y_1} = h_{11}y_1 + \cdots + h_{1n}y_n \hat{y_2} = h_{21}y_1 + \cdots + h_{2n}y_n \vdots \hat{y_n} = h_{n1}y_1 + \cdots + h_{nn}y_n$$

Therefore,

$$\begin{aligned} \hat{y_i} &= h_{i1}y_1 + \cdots + h_{in}y_n \\ &= \sum_{j=1}^{n} h_{ij}y_j \end{aligned}$$

- when h_{i1} is large $\rightarrow$ y_1 has a large impact on the $\hat{y_i}$

(b) Suppose that $n = p + 1$ and X is invertible. Prove that, under those circumstances, $H = I_{p+1}$, where I_{p+1} is the $(p + 1)$ dimensional indentity matrix. You may use the following results:

(1) $(AB)^{-1} = B^{-1}A^{-1}$

(2) $(AB)' = (B'A')$

(3) $(A')^{-1} = (A^{-1})'$

$$\begin{aligned} H &= X_{n\times(p+1)}({X'}_{(p+1)\times n}X_{n\times(p+1)})^{-1}{X'}_{(p+1)\times n} \\ &= X_{n\times(p+1)}(X_{n\times(p+1)})^{-1}({X'}_{(p+1)\times n})^{-1}{X'}_{(p+1)\times n} \text{ according to result(1)} \\ &= I(X')^{-1}X' \text{ since } XX^{-1} = I \text{ and } X \text{ is a squared matrix} \\ &= I(X^{-1})'X' \text{ according to result(3)} \\ &= I(XX^{-1})' \text{ according to result(2)} \\ &= I \cdot I \text{ since } XX^{-1} = I \\ &= I_{p+1} \end{aligned}$$

(c) Using parts (a) and (b), describe the influence of y_i on $\hat{y}_i$ when $n = p+1$ and X is invertible.

When $\boldsymbol{n = p + 1}$ and $\boldsymbol{X}$ is invertible, we have proved in part (b) that $\boldsymbol{H = I_{p+1} = I_n}$.

Since $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$,

$$\hat{\boldsymbol{y}} = \boldsymbol{I_n y}$$

$$\begin{bmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_n} \end{bmatrix} = \begin{bmatrix} 1_1 & 0 & \cdots & 0 \\ 0 & 1_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 + 0 + \cdots + 0 \\ 0 + y_2 + \cdots + 0 \\ \vdots \\ 0 + \cdots + 0 + y_n \end{bmatrix}$$

$$= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\boldsymbol{y_i}$ is the only factor that has an influence on $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$.

Problem 2: Weighted Least Squares

Non-constant variance The linear regression model assumes that the variance of the error term is constant for all observations (i.e., $\mathbf{Var}(\varepsilon_i) = \boldsymbol{\sigma^2}$) and that the errors are uncorrelated (i.e., $\mathbf{Cov}(\varepsilon_i, \varepsilon_j) = \mathbf{0}$). Equivalently, these assumptions may be written as:

$$\mathbf{Cov}[\varepsilon] = \begin{bmatrix} \boldsymbol{\sigma^2} & \mathbf{0} \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma^2} \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} \ldots & \boldsymbol{\sigma^2} \end{bmatrix}_{n \times n} = \boldsymbol{\sigma^2 I_n},$$

where $\boldsymbol{I_n}$ is the $\boldsymbol{n \times n}$ identity matrix. Suppose that the constant variance assumption is unmet so that:

$$\mathbf{Cov}[\varepsilon] = \boldsymbol{\sigma^2 W} = \boldsymbol{\sigma^2} \begin{bmatrix} \boldsymbol{w_{11}} & \mathbf{0} \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{w_{22}} \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} \ldots & \boldsymbol{w_{nn}} \end{bmatrix}_{n \times n},$$

where $\boldsymbol{W}$ is a matrix with known diagonal entries equal to $\boldsymbol{w_{ii}}$. Stated equivalently, we can say that $\mathbf{Var}(\varepsilon_i) = \boldsymbol{w_{ii}\sigma^2}$ (i.e., non-constant variance or heteroscedasticity $\boldsymbol{: w_{11} \neq w_{22}}$) and that the errors are uncorrelated, i.e., $\mathbf{Cov}(\varepsilon_i, \varepsilon_j) = \mathbf{0}$.

One approach to address the heteroscedasticity when $\boldsymbol{W}$ is known is the Weighted Least Squares (WLS). The WLS method consists of transforming both $\boldsymbol{Y}$ and $\boldsymbol{X}$ such that $\boldsymbol{Y^* = W^{1/2} Y}$ and $\boldsymbol{X^* = W^{1/2} X}$ and fitting the following model:

$$\boldsymbol{Y^* = X^* \beta^* + \varepsilon^*}.$$

It can be shown that the WLS eliminates the heteroscedasticity, i.e., $\mathbf{Cov}[\boldsymbol{\varepsilon^*}] = \boldsymbol{\sigma^2 I_n}$.

(a) Write down the entries in the matrices Y^* and X^* (function of the w_{ii}'s for $i = 1, ..., n$.)

$\boldsymbol{W}$ is a diagonal matrix with known entries $\boldsymbol{w_{ii}}$, then $\boldsymbol{W^{1/2}}$ is the matrix formed by taking the square root of each diagonal entry of $\boldsymbol{W}$.

$$\boldsymbol{W^{1/2}} = \begin{bmatrix} \sqrt{w_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{w_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{w_{nn}} \end{bmatrix}$$

Entries of Y^*:

$$\begin{aligned} \boldsymbol{Y^*} &= \boldsymbol{W^{1/2}Y} \\ &= \begin{bmatrix} \sqrt{w_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{w_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{w_{nn}} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sqrt{w_{11}}y_1 + 0 + \cdots + 0 \\ 0 + \sqrt{w_{22}}y_2 + \cdots + 0 \\ \vdots \\ 0 + \cdots + 0 + \sqrt{w_{nn}}y_n \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{w_{11}}y_1 \\ \sqrt{w_{22}}y_2 \\ \vdots \\ \sqrt{w_{nn}}y_n \end{bmatrix} \end{aligned}$$

Therefore, entries of $\boldsymbol{Y^*}$ are $\sqrt{w_{ii}}y_i$ for $i = 1, \cdots, n$.

Entries of X^*:

$$\begin{aligned} \boldsymbol{X^*} = \boldsymbol{W^{1/2}X} &= \begin{bmatrix} \sqrt{w_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{w_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{w_{nn}} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{w_{11}}x_{11} & \sqrt{w_{11}}x_{12} & \cdots & \sqrt{w_{11}}x_{1p} \\ \sqrt{w_{22}}x_{21} & \sqrt{w_{22}}x_{22} & \cdots & \sqrt{w_{22}}x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{w_{nn}}x_{n1} & \sqrt{w_{nn}}x_{n2} & \cdots & \sqrt{w_{nn}}x_{np} \end{bmatrix} \end{aligned}$$

Therefore, entries of $\boldsymbol{X^*}$ are $\sqrt{w_{ii}}x_{ij}$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$.

(b) The WLS estimate for $\hat{\beta}^*$ is obtained by fitting the model $Y^* = X^*\beta^* + \varepsilon^*$ using the LS method, i.e., $\hat{\beta}^* = (X^{*'}X^*)^{-1}X^{*'}Y^*$. Find an expression for $\hat{\beta}^*$ that only depends on Y, X (i.e., the untransformed variables), and W.

$$\begin{aligned} \hat{\beta}^* &= (\boldsymbol{X^{*'}X^*})^{-1}\boldsymbol{X^{*'}Y^*} \\ &= \left[(\boldsymbol{W^{1/2}X})'(\boldsymbol{W^{1/2}X})\right]^{-1}(\boldsymbol{W^{1/2}X})'(\boldsymbol{W^{1/2}Y}) \end{aligned}$$

Simplify:

1. $$\begin{aligned} (\boldsymbol{W^{1/2}X})'(\boldsymbol{W^{1/2}X}) &= \boldsymbol{X'}(\boldsymbol{W^{1/2}})'(\boldsymbol{W^{1/2}})\boldsymbol{X} \\ &= \boldsymbol{X'WX} \end{aligned}$$

   Since $\boldsymbol{W}$ is a diagonal matrix, $(\boldsymbol{W^{1/2}})' = \boldsymbol{W^{1/2}}$. So $(\boldsymbol{W^{1/2}})'(\boldsymbol{W^{1/2}}) = \boldsymbol{W}$

2.

$$(\boldsymbol{W^{1/2}X})'(\boldsymbol{W^{1/2}Y}) = \boldsymbol{X'}(\boldsymbol{W^{1/2}})'(\boldsymbol{W^{1/2}})\boldsymbol{Y}$$
$$= \boldsymbol{X'WY}$$

Therefore,

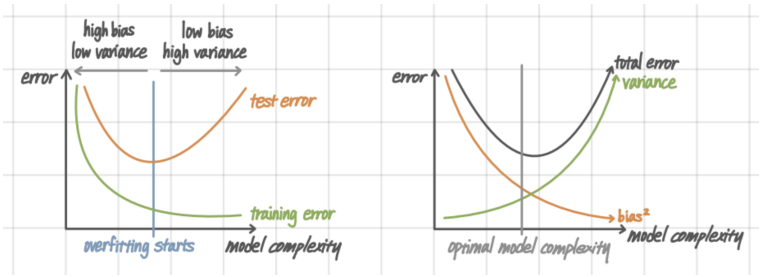
$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'WY}$$

Problem 3: $\boldsymbol{MSE_{te}}$

You fitted a KNN model to a dataset containing $\boldsymbol{n}$ observations and $\boldsymbol{p}$ predictors. You used $\boldsymbol{K = 5}$. Someone is asking you to use $\boldsymbol{K = 20}$ instead. Please describe how you expect this change to affect the bias-variance trade-off. In particular, how do you expect this change to affect:

- simple models: large bias in the prediction
- flexible models: large variance in the prediction

graph:



(a) The bias.

Increasing $\boldsymbol{K}$ from 5 to 20 will make the training data set uses a larger number of neighbors to make predictions. This will cause the model to less likely to be influenced by unusual patterns or outliers in the training data. This leads to smoother decision boundaries because the predictions are averaged over more data. Therefore as $\boldsymbol{K}$ increases, the **bias increases** because the model would potentially miss finer details in the data.

(b) The variance.

As $\boldsymbol{K}$ increases from 5 to 20, the model rely on a larger set of neighbors and is less sensitive to small changes in the data set. It becomes more stable because each prediction is based on a larger number of neighbors. Therefore, the **variance decreases** since larger $\boldsymbol{K}$ reduces the influence of a single data point or outlier in the training data.

(c) The $\mathbf{MSE_{Te}}$. Can you tell if it will decrease, stay the same, or increase?

The graph of the $\boldsymbol{MSE_{te}}$ is parabola, which means that it will be the lowest when the model complexity is optimal (as seen in the graph). So, there are 2 possibilities for the $\boldsymbol{MSE_{te}}$ as $\boldsymbol{K}$ increases from 5 to 20.

1. $\boldsymbol{K = 5}$ is overfitting (low bias and high variance):

   $\rightarrow$ Increasing $\boldsymbol{K}$ to 20 can **reduce the** $\boldsymbol{MSE}$ of the test data by lowering the variance, even though the bias increases slightly

2. $\boldsymbol{K = 5}$ is the optimal model complexity (lowest $\boldsymbol{MSE_{te}}$):

   $\rightarrow$ Increasing $\boldsymbol{K}$ to 20 will **increase the** $\boldsymbol{MSE}$ of the test data by increasing the bias, although the variance decreases slightly.

Problem 4: Estimate with KNN

The table below provides a training data set containing five observations, two predictors (explanatory variables), and one quantitative response variable (knit the document to view the table).

Obs. Index	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{Y}$
1	2	4	18
2	1	2	9
3	-1	0	0
4	4	-1	6
5	3	3	19

Suppose we wish to use this data set to make a prediction for $\boldsymbol{Y}$ when $\boldsymbol{X_1 = 2}$ and $\boldsymbol{X_2 = 1}$ using K nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $(2, 1)$. Show your work.

```
dataset <- data.frame(x1 = c(2, 1, -1, 4, 3),
                      x2 = c(4, 2, 0, -1, 3),
                      y = c(18, 9, 0, 6, 19))
```

Euclidean distance to $\boldsymbol{x_0} = \sqrt{(\boldsymbol{x_1} - \boldsymbol{x_{10}})^2 + (\boldsymbol{x_2} - \boldsymbol{x_{20}})^2}$

```
# Euclidean distance to (2,1)
dataset <- dataset %>%
  mutate(distance = sqrt((x1 - 2)^2 + (x2 - 1)^2))

dataset
```

```
##   x1 x2  y distance
## 1  2  4 18 3.000000
## 2  1  2  9 1.414214
## 3 -1  0  0 3.162278
## 4  4 -1  6 2.828427
## 5  3  3 19 2.236068
```

(b) What is our prediction ($\hat{Y}$) with $K = 1$? Why?

$\boldsymbol{K = 1}$ nearest neighbor: $\boldsymbol{(1, 2)}$

Predicted value at the test point $\boldsymbol{(2, 1) = 9/1 = 9}$

(c) What is our prediction with $K = 3$? Show your work.

$\boldsymbol{K = 3}$ nearest neighbors: $\boldsymbol{(1, 2), (3, 3), (4, -1)}$

Predicted value at the test point $\boldsymbol{(2, 1) = (9 + 6 + 19)/3 = 11.333}$