STAT380: Assignment 1

Vivienne Crowe ID:40071153

25/1/22

1 Part A

1.1

Let x_i denote the i-th row of the matrix$\mathbf{X}$

kNN			
i	$\|\|x_i - x_0\|\|_2$	$\in N_0$ when $k = 1$	$\in N_0$ when $k = 3$
1	1	✓	✓
2	$\sqrt{2}$	✗	✓
3	$\sqrt{5}$	✗	✗
4	$\sqrt{3}$	✗	✓
5	3	✗	✗
6	$\sqrt{5}$	✗	✗

Hence, for $k = 1$ $\hat{y}_0 = y_1 = 0$, and when $k = 3$ $\hat{y}_0 = \frac{1}{1}(0 + 5 - 5) = 0$.

1.2

Parzen Window			
i	$\|\|x_i - x_0\|\|_2$	$\in W_0$ when $D = 2$	$\in W_0$ when $D = 3$
1	1	✓	✓
2	$\sqrt{2}$	✓	✓
3	$\sqrt{5}$	✗	✓
4	$\sqrt{3}$	✓	✓
5	3	✗	✓
6	$\sqrt{5}$	✗	✓

When $D = 2$ the Parzen Window prediction is $\hat{y}_0 = \frac{1}{3}(0 + 5 - 5) = 0$. When $D = 3$ the Parzen Window prediction is $\hat{y}_0 = \frac{1}{6}(0 + 5 + 6 - 5 + 4 + 1) = \frac{11}{6}$

1.3 A.3

First note that the expectation of MSE_{test}, is independent of the number of values in the test set.

$$E\left[\frac{1}{m}\sum_{i=1}^{m}(y_i^* - \hat{\beta}^T x_i^*)^2\right]$$

$$=\frac{1}{m}\sum_{i=1}^{m}E\left[(y_i^* - \hat{\beta}^T x_i^*)^2\right]$$

$$=E\left[(y^* - \hat{\beta}^T x^*)^2\right]$$

Therefore, I can choose $m = n$. Now let $\hat{\beta}^*$ denote the OLS parameters obtained by fitting a linear model to the test set, and note that the following random variables are identically distributed.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{\beta}^{*T} x_i^*)^2$$

Since $\hat{\beta}^*$ is defined as:

$$\arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}(y_i^* - \beta^T x_i^*)^2 := \arg\min_{\beta} f(x^*, y^*, \beta)$$

then the value of f will be larger for any other value of β other than $\hat{\beta}^{*T}$. Therefore,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2\right] \leq E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{\beta}^T x_i^*)^2\right]$$

$$E[MSE_{train}] \leq E[MSE_{test}]$$

1.4 A.4

(a)

$$\min_{\theta}\sum_{i=1}^{n}(f_\theta(x_i) - y_i)^2$$

$$\min_{\theta}\sum_{i=1}^{n}(f_\theta^2(x_i) - 2f_\theta(x_i)y_i + y_i^2)$$

$$\min_{\theta}(nf_\theta^2(x_i) - 2f_\theta(x_i)n\bar{y} + \sum_{i=1}^{n}y_i^2)$$

Note that terms only containing y do not effect which value of θ will minimise the expression so they can be removed, or added.

$$\min_\theta(nf_\theta^2(x_i) - 2nf_\theta(x_i)\bar{y} + \bar{y}^2)$$

$$\min_\theta(f_\theta(x_1) - \bar{y})^2$$

This gives a reduced OLS problem.
Note: $\bar{y} = \sum_{i=1}^n \frac{y_i}{n}$.
(b) The reduced weighted least squares is:

$$\min_\theta \sum_{i=1}^k r_i \left(f_\theta(x_i) - \bar{y}_i\right)^2$$

To show why this is true, divide the summation of the full OLS expression into terms where all the predictor values are the same.

$$\min_\theta \sum_{i=1}^k \sum_{j=1}^{r_i} (f_\theta(x_i) - y_i^{(j)})^2$$

Following the same argument from part (a), we then have

$$\min_\theta \sum_{i=1}^k r_i(f_\theta(x_i) - \bar{y}_i)^2$$

which is a reduced, weight least squares problem.

2 Part B

2.1

Solutions for part (a) and (b) in the .R file.
(c) Yes the results of the regression suggest there is a strong relationship between the predictors and the response. In particular, the variables year, origin, weight and displacement are statistically significant predictors of miles per gallon. The fitted coefficient on the year variable tells us that miles per gallon is improving over time.
(d) There are a few problems.

1. The residuals vs fitted plot suggests non-independence of errors, since smaller fitted values correspond with positive errors.

2. The variance of the residuals is non-constant - it's increasing with the fitted value.

3. There is also one point that (row 14) which has unusally high leverage.

4. The Q-Q shows there's a deviation from the assumption of normal errors.

Note however that observation 14 is within the threshold defined by the Cook's distance threshold that is commonly used to identify outliers.
(e) I first considered all possible pair-wise interactions. Under this model the interaction between displacement and weight was statistically significant. I also looked smaller models, which excluded some predictors, to see how the regression would change. Under these reduced models other interactions, such as between horsepower and acceleration, and displacement and weight became significant.
(f) I found that taking the exponential of the acceleration predictor made it significant. Taking the square root of the horsepower variable increased it's significance. I also tried taking the log of the response variable and found that this increase the F statistics 1.62 times suggesting a better overall model.

2.2

(a) n is 100, p is 2. Let $\epsilon \sim N[0,1]$, then the model is

$$Y = X - 2 \cdot X^2 + \epsilon$$

(b) It looks like a quadratic relationship, values appear more dense around the origin
(c) The LOOVC errors for the models i-iv are: i. 7.288 ii. 0.937 iii. 0.957 iv. 0.954
(d) With the new seed, the errors were the same. This was expected because in the LOOCV procedure all possible subsets of size $n-1$ are considered so the order that they are selected (which is random), does not impacted the minimum error obtained.
(e) The quadratic model has the smallest error, which was expected as this was precisely the underlying model.
(f) The statistical significance for the coeffecients are consistent with the conclusions drawn above, p-values for the intercept, first-, and second-order terms are all very small (on the order of 10^{-8} or less), whereas coefficient estimates for the third-, and fourth-order terms are much larger, around 10^{-1}.