STA521 HW1

[Fan Zhu | netid: fz63]

Due August 28th, 2020

This exercise involves the Auto data set from ISLR. Load the data and answer the following questions adding your code in the code chunks. Please submit a pdf version to Sakai. For full credit, you should push your final Rmd file to your github repo on the STA521-F19 organization site by the deadline (the version that is submitted on Sakai will be graded)

Exploratory Data Analysis

1. Create a summary of the data. How many variables have missing data?

```
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower        weight    
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613  
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804  
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978  
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140  
##                                                                                
##   acceleration        year           origin                      name    
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5  
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5  
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5  
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4  
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4  
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4  
##                                                  (Other)           :365  
```

```
colSums(is.na(Auto))
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##            0            0            0            0            0            0 
##         year       origin         name 
##            0            0            0 
```

```
#None of the variables has missing data.
```

2. Which of the predictors are quantitative, and which are qualitative?

```
help(Auto)
```

Quantitative: mpg, displacement, horsepower, weight, and acceleration

Qualitative: origin, cylinders, year and name

3. What is the range of each quantitative predictor? You can answer this using the `range()` function. Create a table with variable name, min, max with one row per variable. `kable` from the package `knitr`

can display tables nicely.

```r
ranges_Auto <- rbind(range(Auto$mpg),
                     range(Auto$displacement),
                     range(Auto$horsepower),
                     range(Auto$weight),
                     range(Auto$acceleration))
rownames(ranges_Auto) <- c("mpg","displacement",
                           "horsepower", "weight", "acceleration")
knitr::kable(ranges_Auto, "pipe", caption = "Table1: Ranges of the Quantitative Variables",
             col.names = c("Min","Max"))
```

Table 1: Table1: Ranges of the Quantitative Variables

	Min	Max
mpg	9	46.6
displacement	68	455.0
horsepower	46	230.0
weight	1613	5140.0
acceleration	8	24.8

4. What is the mean and standard deviation of each quantitative predictor? Format nicely in a table as above

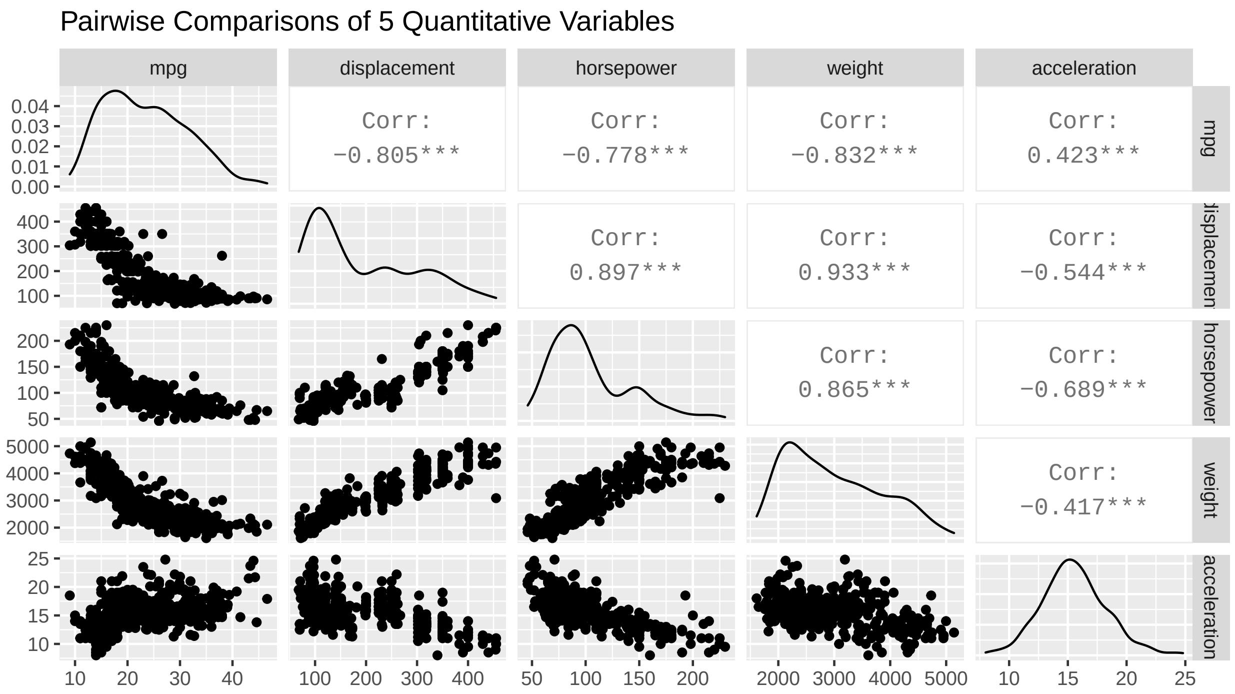
```r
centers_Auto <- rbind(c(mean(Auto$mpg),sd(Auto$mpg)),
                      c(mean(Auto$displacement),sd(Auto$displacement)),
                      c(mean(Auto$horsepower),sd(Auto$horsepower)),
                      c(mean(Auto$weight),sd(Auto$weight)),
                      c(mean(Auto$acceleration),sd(Auto$acceleration)))
rownames(centers_Auto) <- c("mpg", "displacement",
                            "horsepower", "weight", "acceleration")
knitr::kable(centers_Auto, "pipe", caption = "Table2: Central Tendencies for the Quantitative Variables
             col.names = c("Mean","SD"))
```

Table 2: Table2: Central Tendencies for the Quantitative Variables

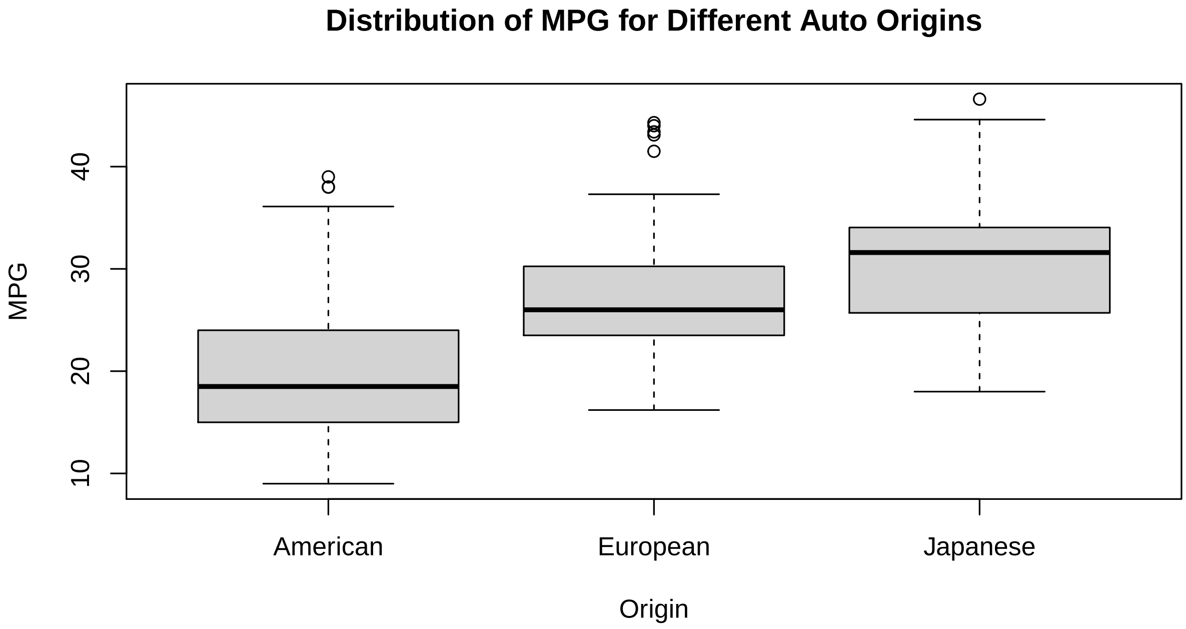
	Mean	SD
mpg	23.44592	7.805008
displacement	194.41199	104.644004
horsepower	104.46939	38.491160
weight	2977.58418	849.402560
acceleration	15.54133	2.758864

5. Investigate the predictors graphically, using scatterplot matrices (`ggpairs`) and other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings. Try adding a caption to your figure

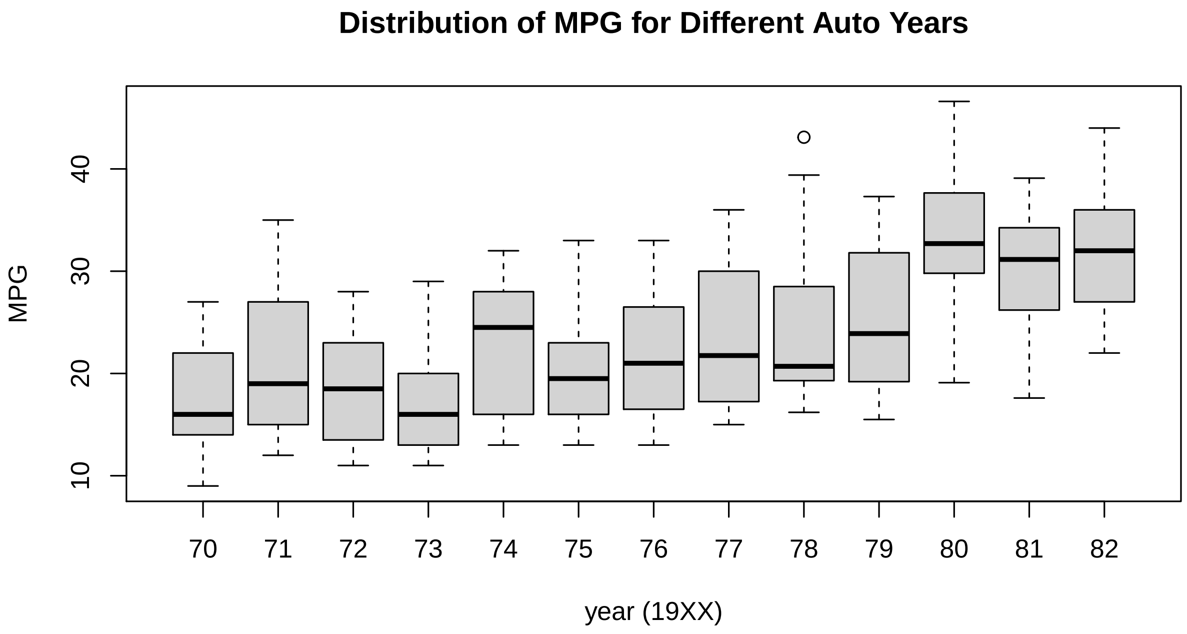
```r
#create pairwise comparisons scatterplot mattrices using the "ggpairs" function
library(ggplot2)
library(GGally)
ggpairs(Auto[,c(1,3,4,5,6)],title = "Pairwise Comparisons of 5 Quantitative Variables", progress = FALS
```



```
#create a boxplot of MPG for different Auto Origins
boxplot(mpg~origin,
data=Auto,
main="Distribution of MPG for Different Auto Origins",
xlab="Origin",
ylab="MPG",
names=c("American","European","Japanese")
)
```



```r
#create a boxplot of MPG for different Auto Years
boxplot(mpg~year,
data=Auto,
main="Distribution of MPG for Different Auto Years",
xlab="year (19XX)",
ylab="MPG"
)
```



In the pairewise scatterplot matrices I compared each pair of the quantitative variables. From the plot we can observe whether there exists a positive or negative linear relationship between each pair of variables. We can see that most of the scatterplots exhibited some forms of patterns. The right upper matrices showed the correlations between each pair of quantitative variables. All of the correlations here are statistically significant (significantly different from zero). And the correlation between the car's weight and displacement is the strongest. In the of MPG for different Auto Origins, we can observe that the Japanese cars have the highest average mpg among the cars from different origins and the American cars have the lowest average mpg. The interquartile range of the Japanese cars is higher than that of the American and European cars. In the boxplot of MPG for different Auto Years, we can see that the cars from 1980 tend to have the highest average mpg, while the cars from 1970 tend to have the lowest average mpg, which might imply that the older the car the lower mpg. The cars from 1974 have relatively unusual high mpg compared to the cars from 1973~1975.

6. Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables using regression. Do your plots suggest that any of the other variables might be useful in predicting mpg using linear regression? Justify your answer.

Yes. The scatterplot matrices contain scatterplots between each pair of variables, from which we can observe whether there exists a positive or negative linear relationship between these variables or not. The plots also contain correlations, which we can use to find which independent variable is correlated with the response/dependent variable. Besides, if the correlations between two independent variables are too high, then there is a multicollinearity issue. Since the scatterplot matrices in Q5 shows that some of the relationships might be not linear, we would need to log transform the variables.

Simple Linear Regression

7. Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:
   (a) Is there a relationship between the predictor and the response?
   (b) How strong is the relationship between the predictor and the response?
   (c) Is the relationship between the predictor and the response positive or negative?
   (d) Provide a brief interpretation of the parameters that would suitable for discussing with a car dealer, who has little statistical background.
   (e) What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals? (see `help(predict)`) Provide interpretations of these for the car dealer.

```r
#run a regression with mpg as the response and horsepower as the predictor.
model1 <- lm(mpg ~ horsepower, data = Auto)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```r
#compute the 95% confidence interval given horsepower = 98
conf_int <-  predict(model1, data.frame(horsepower = 98), interval = "confidence")
conf_int
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```r
#compute the 95% prediction interval given horsepower = 98
pred_int <-  predict(model1, data.frame(horsepower = 98), interval = "prediction")
pred_int
```

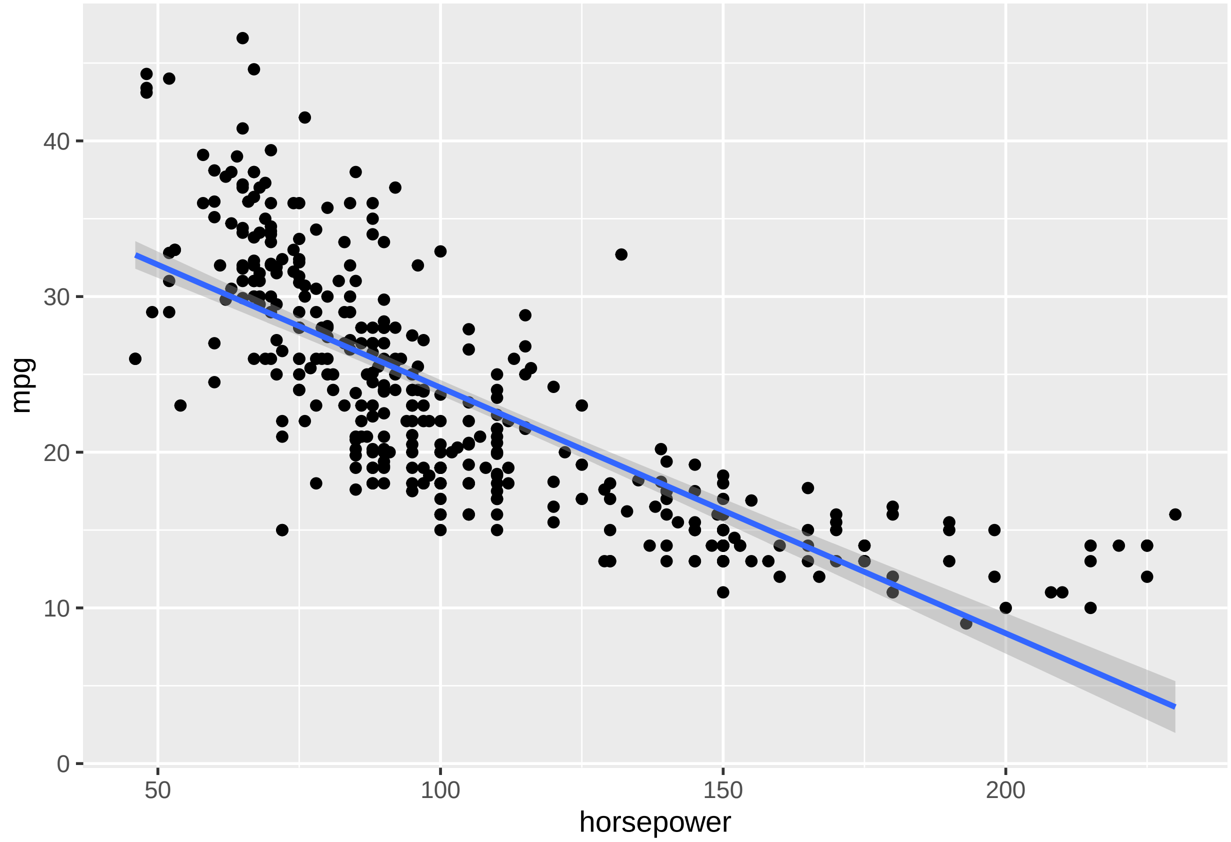
```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

7(a) Yes. There is a relationship between the predictor and the response. The P value for the beta in the regression is smaller than 0.05. Thus, we can reject the null hypothesis that the beta is zero, which means that there is a significant nonzero relationship between teh predictor and the response. 7(b) The beta is -0.157845, which means that holding everything else constant, every one unit increase in the horsepower will decrease the mpg by 0.157845. And as we have discussed in 7(a), the beta is significant. 7(c) The relationship between the predictor and the response is negative because the beta is negative. 7(d) Holding everything else constant, every one unit increase in the horsepower will decrease the mpg by 0.157845. That is, higher

horsepower leads to more fuel consumption (less miles per gallon). There is a trade-off between performance and the car's costs on fuels. 7(e) The predicted mpg associated with a horsepower of 98 is 24.46708. We can tell the dealer that with a horsepower of 98, the car would run 24.46708 miles per gallon of gasoline. The lower bound of the 95% confidence interval is 23.97308 and the upper bound of the confidence interval is 24.96108. The lower bound of the 95% prediction interval is 14.8094 and the upper bound of the prediction interval is 34.12476.
A more stats-version of the confidence interval is that: if we take the samples repeatedly and computed the 95% confidence interval for each sample, 95% of the intervals would contain the true population mean. In other words, we are 95% confidence that the average mpg of a car, whose horsepower is 98, will be between 23.97308 and 24.96108. we are also 95% confident that a predicted/new car with 98 horsepower will have its mpg between 14.8094 and 34.12476.

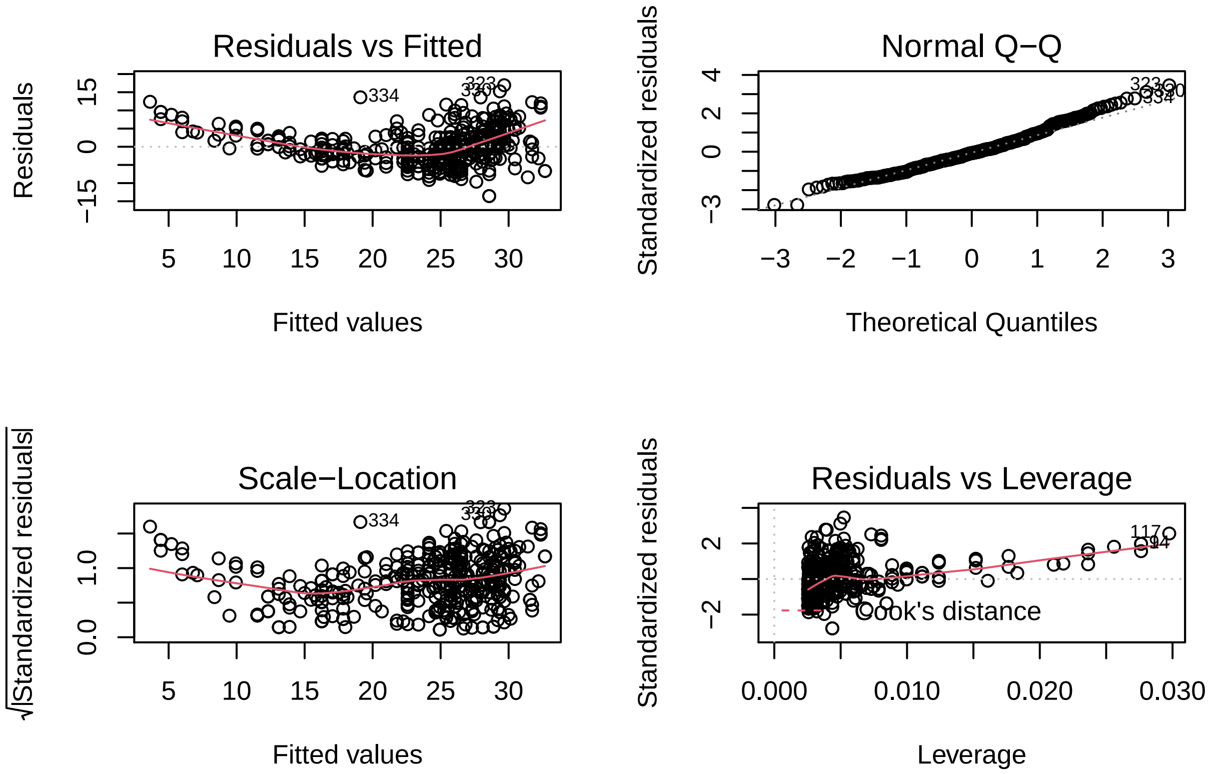
8. Plot the response and the predictor using `ggplot`. Add to the plot a line showing the least squares regression line.

```
ggplot(Auto, aes(x = horsepower, y = mpg)) +
  geom_point() + geom_smooth(method = "lm", formula = y ~ x)
```



9. Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the model regarding assumptions for using a simple linear regression.

```
par(mfrow = c(2, 2))
plot(model1)
```



Interpretations on Q9 plots: The "Residual vs. Fitted" plot exhibits a non-linear trend between the residuals and the fitted values, which implies a violation of the linearity assumption of the model. The normal Q-Q plot shows that most of the data fit the diagonal line except some data on the left and right ends. The ideal case of the Q-Q plot is a perfect match to the diagonal line. Our Q-Q plot shows the data is not very far from a normal distribution but there are still some deviations. The Scale-Location plot shows an U-shaped curve, which implies that the variance of the residuals does not stay constant along the fitted values. Therefore, the assumption of homoscedasticity (equal variance) is violated. The Residuals vs. Leverage plot shows that almost all the data points are within the Cook's distance lines, which indicates that there is no influential case. In conclusion, from the diagnostics plots above, we can find that the residuals may not be normally distributed with a mean of zero and a constant variance.

Theory

10. Show that the regression function $E(Y \mid x) = f(x)$ is the optimal optimal predictor of Y given $X = x$ using squared error loss: that is $f(x)$ minimizes $E[(Y - g(x))^2 \mid X = x]$ over all functions $g(x)$ at all points $X = x$. Hint: there are at least two ways to do this. Differentiation (so think about how to justify) - or - add and subtract the proposed optimal predictor and who that it must minimize the function.

$$E[(Y - g(x))^2 \mid X = x] = E[(Y - f(x) + f(x) - g(x))^2 \mid X = x]$$

$$= E[((Y - f(x)) + (f(x) - g(x)))^2 \mid X = x]$$

$$= E[(Y - f(x))^2 + (f(x) - g(x))^2 + 2(Y - f(x))(f(x) - g(x)) \mid X = x]$$

$$= E[(Y - f(x))^2 \mid X = x] + E[(f(x) - g(x))^2 \mid X = x] + E[2(Y - f(x))(f(x) - g(x)) \mid X = x]$$

$$((f(x) - g(x)) \text{ can be treated as a constant because they are conditional on } X = x)$$

$$= E[(Y - f(x))^2 \mid X = x] + (f(x) - g(x))^2 + 2(f(x) - g(x))E[Y - f(x) \mid X = x]$$

$$= E[(Y - f(x))^2 \mid X = x] + (f(x) - g(x))^2 + 2(f(x) - g(x))(E(Y \mid X = x) - f(x))$$

$$(Because\ E(Y \mid X = x) - f(x))$$

$$= E[(Y - f(x))^2 \mid X = x] + (f(x) - g(x))^2 + 2(f(x) - g(x))(f(x) - f(x))$$

$$= E[(Y - f(x))^2 \mid X = x] + (f(x) - g(x))^2$$

$$\geq E[(Y - f(x))^2 \mid X = x]$$

Based on the abovementioned calculation, we can deduce that $E[(Y-g(x))^2 \mid X = x] = E[(Y-f(x))^2 \mid X = x]$ only if $f(x) = g(x)$ for all x. Thus, we can conclude that $f(x)$ minimizes $E[(Y - g(x))^2 \mid X = x]$ over all functions $g(x)$ at all points $X = x$.

Citation: Professor Mukherjee taught us this computational trick during this office hour. Thank you.

11. (adopted from ELS Ex 2.7) Suppose that we have a sample of N pairs x_i, y_i drwan iid from the distribution characterized as follows

$$x_i \sim h(x), \text{ the design distribution}$$

$$\epsilon_i \sim g(y), \text{ with mean 0 and variance } \sigma^2 \text{ and are independent of the } x_i$$

$$Y_i = f(x_i) + \epsilon$$

(a) What is the conditional expectation of Y given that $X = x_o$? ($E_{Y|X}[Y]$)

$$E[Y \mid X = x_o] = E[f(X) + \epsilon \mid X = x_o]$$

$$= E[f(X)] + E[\epsilon \mid X = x_o]$$

$$= f(x_o) + E[\epsilon \mid X = x_o]$$

$$\text{(Because } \epsilon \text{ and X are supposed to be independent)}$$

$$= f(x_o) + E[\epsilon]$$

$$= f(x_o) + 0 = f(x_o)$$

(b) What is the conditional variance of Y given that $X = x_o$? ($\text{Var}_{Y|X}[Y]$)

$$\text{Var}[Y \mid X = x_o] = \text{Var}[f(X) + \epsilon \mid X = x_o]$$

$$= \text{Var}[f(x_o) + \epsilon \mid X = x_o]$$

$$\text{(Because } Var(f(x_o) = 0)$$

$$= \text{Var}[\epsilon \mid X = x_o]$$

$$\text{(Because } \epsilon \text{ and X are supposed to be independent)}$$

$$= \text{Var}[\epsilon] = \sigma^2$$

(c) show that for any estimator $\hat{f}(x)$ that the conditional (given X) (expected) Mean Squared Error can be decomposed as

$$E_{Y|X}[(Y - \hat{f}(x_o))^2] = \underbrace{\text{Var}_{Y|X}[\hat{f}(x_o)]}_{\textit{Variance of estimator}} + \underbrace{(f(x) - E_{Y|X}[\hat{f}(x_o)])^2}_{\textit{Squared Bias}} + \underbrace{\mathsf{Var}(\epsilon)}_{\textit{Irreducible}}$$

Hint: try the add zero trick of adding and subtracting expected values

$$
\begin{gathered}
E_{Y|X}[(Y-\hat{f}(x_o))^2] \\
= E_{Y|X}[(Y-f(x_o)+f(x_o)-E_{Y|X}[\hat{f}(x_o)]+E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))^2] \\
= E_{Y|X}[(Y-f(x_o))^2]+E_{Y|X}[(f(x_o)-E_{Y|X}[\hat{f}(x_o)])^2]+E_{Y|X}[(E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))^2)^2] \\
+2E_{Y|X}[(Y-f(x_o))(f(x_o)-E_{Y|X}[\hat{f}(x_o)])] \\
+2E_{Y|X}[(f(x_o)-E_{Y|X}[\hat{f}(x_o)])(E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))^2)] \\
+2E_{Y|X}[(Y-f(x_o))(E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))^2)] \\
= E_{Y|X}[\epsilon^2]+(f(x_o)-E_{Y|X}[\hat{f}(x_o)])^2+\mathrm{Var}_{Y|X}[\hat{f}(x_o)] \\
+(f(x_o)-E_{Y|X}[\hat{f}(x_o)])E_{Y|X}[\epsilon]+(f(x_o)-E_{Y|X}[\hat{f}(x_o)])(E_{Y|X}[(E_{Y|X}[\hat{f}(x_o)]-E_{Y|X}[\hat{f}(x_o)]) \\
+E_{Y|X}[(E_{Y|X}[\hat{f}(x_o)]E_{Y|X}[\epsilon]-E_{Y|X}[\hat{f}(x_o)\epsilon] \\
= \mathrm{Var}(\epsilon)+(f(x_o)-E_{Y|X}[\hat{f}(x_o)])^2+\mathrm{Var}_{Y|X}[\hat{f}(x_o)]+0+0+0 \\
= \mathrm{Var}(\epsilon)+(f(x_o)-E_{Y|X}[\hat{f}(x_o)])^2+\mathrm{Var}_{Y|X}[\hat{f}(x_o)]
\end{gathered}
$$

Q.E.D

(d) Explain why even if N goes to infinity the above can never go to zero. e.g. even if we can learn $f(x)$ perfectly that the error in prediction will not vanish.

If N goes to infinity, then $\mathrm{Var}_{Y|X}[\hat{f}(x_o)]$ will converge to 0. If we learn $f(x)$ perfectly, then $f(x_o) = E_{Y|X}[\hat{f}(x_o)$. Thus, $(f(x_o) - E_{Y|X}[\hat{f}(x_o)])^2 = 0$. Combining the above-mentioned two conditions, we have

$$
\begin{gathered}
E_{Y|X}[(Y-\hat{f}(x_o))^2] = \mathrm{Var}(\epsilon)+(f(x_o)-E_{Y|X}[\hat{f}(x_o)])^2+\mathrm{Var}_{Y|X}[\hat{f}(x_o)] \\
= \mathrm{Var}(\epsilon)+0+0 = \mathrm{Var}(\epsilon) \text{ ,which is irreducible}
\end{gathered}
$$

Therefore, $E_{Y|X}[(Y - \hat{f}(x_o))^2]$ will never equal zero. Q.E.D

(e) Decompose the unconditional mean squared error

$$E_{Y,X}(f(x_o)-\hat{f}(x_o))^2$$

into a squared bias and a variance component. (See ELS 2.7(c))

$$
\begin{gathered}
E_{Y,X}(f(x_o)-\hat{f}(x_o))^2 = E_{Y,X}[(f(x_o)-E_{Y|X}[\hat{f}(x_o)]+E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))^2] \\
= E_{Y,X}[(f(x_o)-E_{Y|X}[\hat{f}(x_o)])^2]+E_{Y,X}[(E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))^2] \\
+2E_{Y|X}[(f(x_o)-E_{Y|X}[\hat{f}(x_o)])(E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))] \\
= (f(x_o)-E_{Y|X}[\hat{f}(x_o)])^2+E_{Y,X}[(E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))^2] \\
+2E_{Y|X}[(f(x_o)-E_{Y|X}[\hat{f}(x_o)])(E_{Y|X}[\hat{f}(x_o)]-\hat{f}(x_o))] \\
= \mathrm{SquaredBias}_{Y|X}(\hat{f}(x_o))+\mathrm{Var}_{Y|X}(\hat{f}(x_o))+2E_{Y|X}[(f(x_o)-E_{Y|X}[\hat{f}(x_o)])*0] \\
= \mathrm{SquaredBias}_{Y|X}(\hat{f}(x_o))+\mathrm{Var}_{Y|X}(\hat{f}(x_o))+0
\end{gathered}
$$

To conlcude,

$$E_{Y,X}(f(x_o)-\hat{f}(x_o))^2 = \mathrm{SquaredBias}_{Y|X}(\hat{f}(x_o))+\mathrm{Var}_{Y|X}(\hat{f}(x_o))$$

Q.E.D

(f) Establish a relationship between the squared biases and variance in the above Mean squared errors. From Q11-part(c), we obtained

$$E_{Y|X}[(Y - \hat{f}(x_o))^2] = \mathrm{Var}(\epsilon) + (f(x_o) - E_{Y|X}[\hat{f}(x_o)])^2 + \mathrm{Var}_{Y|X}[\hat{f}(x_o)]$$

which can be rewritten as

$$E_{Y|X}[(Y - \hat{f}(x_o))^2] = \mathrm{Var}(\epsilon) + E_{Y|X}[f(x_o) - E_{Y|X}[\hat{f}(x_o)]]^2$$

Apply the law of iterated expectations,

$$E_X E_{Y|X}[(Y - \hat{f}(x_o))^2] = \mathrm{Var}(\epsilon) + E_{Y|X}[(f(x_o) - E_{Y|X}[\hat{f}(x_o)]]^2]$$

Combine the result with Q11-part(e),

$$E_X[((f(x_o) - E_{Y|X}[\hat{f}(x_o)])^2] + E_X[\mathrm{Var}_{Y|X}[\hat{f}(x_o)] = \mathrm{Var}(\hat{f}(x_o)) + ((f(x_o) - E_{Y|X}[\hat{f}(x_o)])^2$$

According to the law of total variance,

$$E_X[\mathrm{Var}_{Y|X}[\hat{f}(x_o)] \leq \mathrm{Var}(\hat{f}(x_o))$$

$$E_X[((f(x_o) - E_{Y|X}[\hat{f}(x_o)])^2] \geq ((f(x_o) - E_{Y|X}[\hat{f}(x_o)])^2$$

In conclusion, the unconditional variance is expected to be greater than the conditional variance and the unconditional squared bias is expected to be smaller than the conditional squared bias.

(Thank you for your time and patience in grading this homework.)