STA521 HW1

Di Deng dd224

Due Wednesday September 4, 2019

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

There are no missing data.

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

```
str(Auto)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2
```

Mpg, displacement, horsepower, weight and acceleration are quantitative, while cylinders, year, origin and name are qualitative.

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
range.table = range(Auto$mpg)
for (i in c(3,4,5,6)){
  range.table = rbind(range.table, range(Auto[,i]))
}
rownames(range.table) = colnames(Auto)[c(1,3,4,5,6)]
colnames(range.table) = c("min", "max")
kable(range.table)
```

	min	max
mpg	9	46.6
displacement	68	455.0
horsepower	46	230.0
weight	1613	5140.0
acceleration	8	24.8

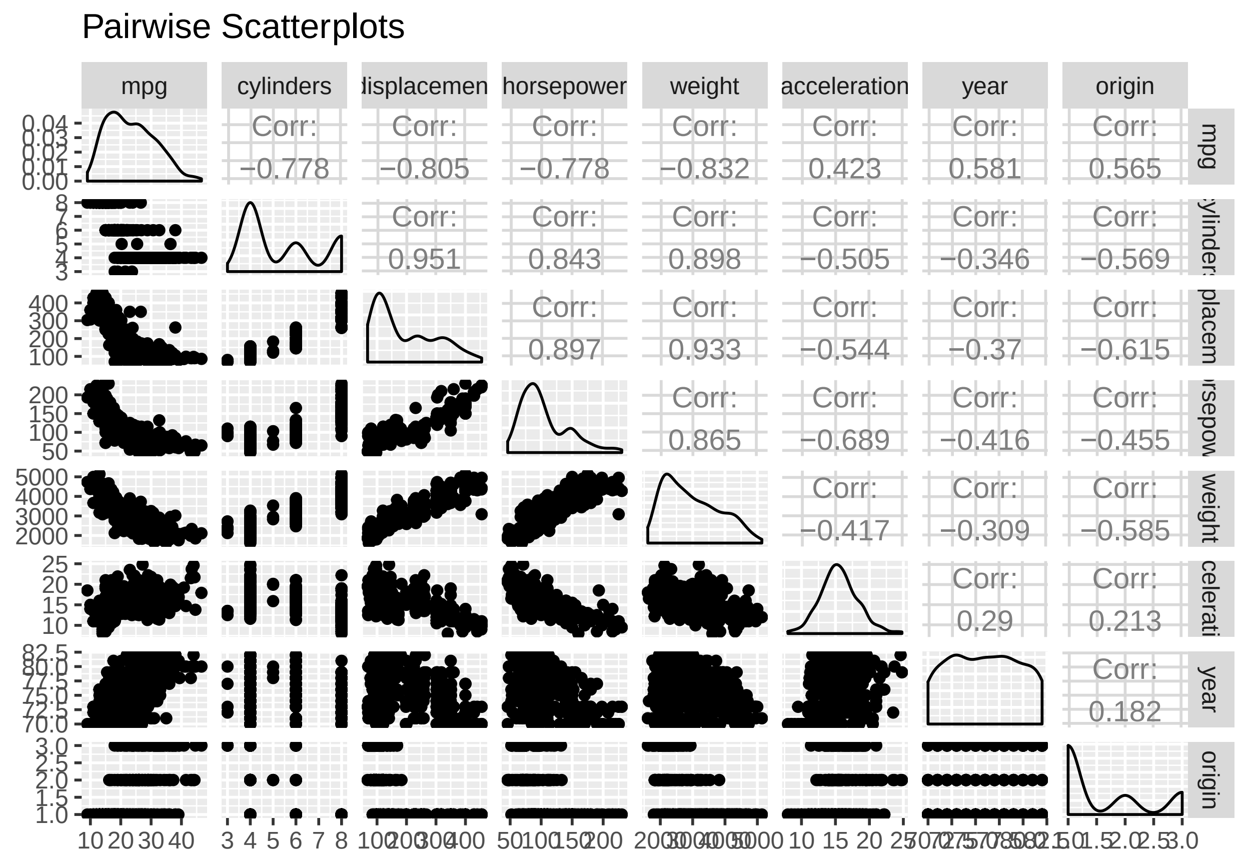
4. What is the mean and standard deviation of each quantitative predictor? *Format nicely in a table as above*

```r
mean = colMeans(Auto[, c(1,3,4,5,6)])
sd = apply(Auto[,c(1,3,4,5,6)], 2, sd)
mean_sd.table = cbind(mean,sd)
kable(mean_sd.table)
```

	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
ggpairs(Auto[,-9], title = "Pairwise Scatterplots", progress = F)
```



Displacement, weight and horsepower are strongly, positively correlated, which makes common sense–heavier cars need more horsepower thus have larger displacement.

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.

I will use cylinders, acceleration, year, origin and two of the horsepower, weight and acceleration–due to high correlations–to fit a linear model for mpg, because from the scatterplots, they all have some associations with mpg. However, further model selection might be needed.

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.

```
mpg.lm1 = lm(data = Auto, mpg~horsepower)
summary(mpg.lm1)
```

```
##
## 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
predict(mpg.lm1, data.frame(horsepower = c(98)))
```

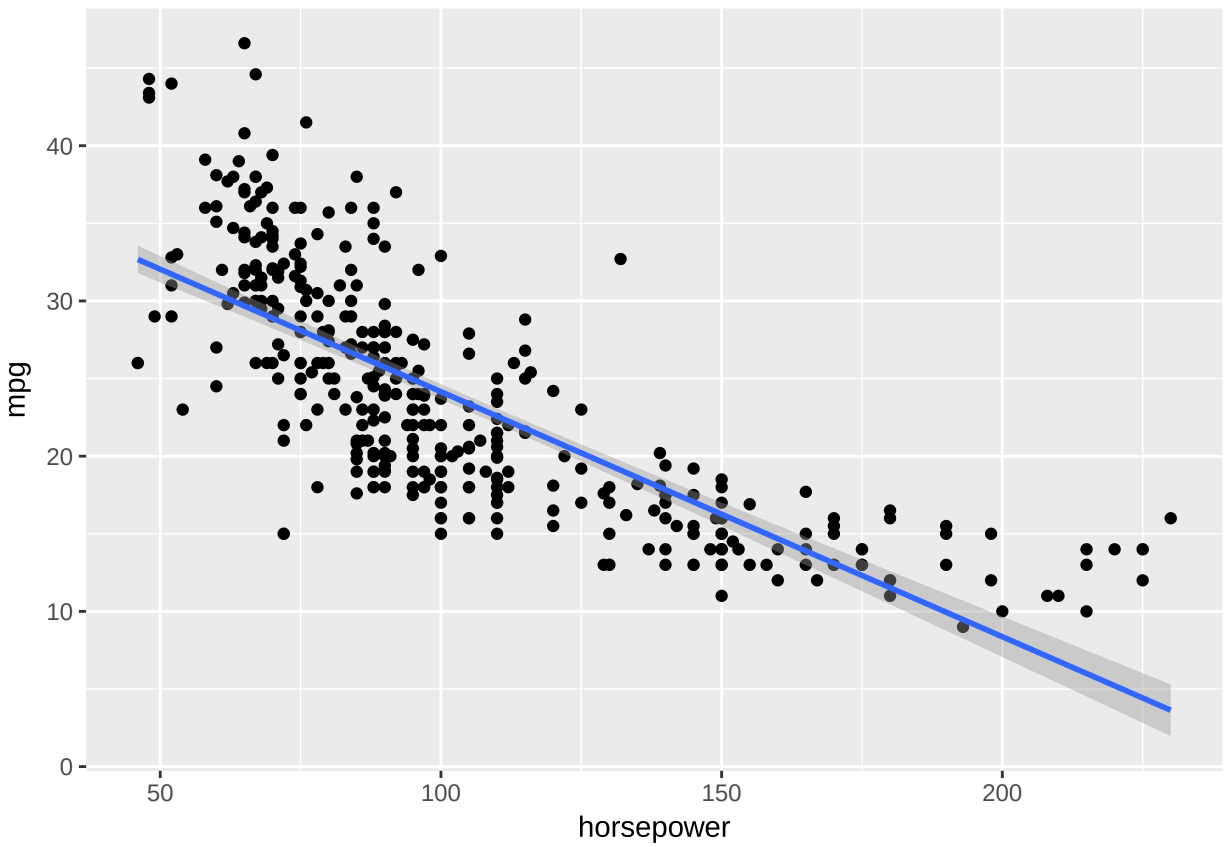
```
##        1
## 24.46708
```

1. The slope is -0.157845 and it is significant according to the test.
2. A-hundred-unit increment of horsepower will lead to 15.7845 decrease in mpg, which is a very significant influence, since the range of horsepower and mpg are (46,230) and (9,46) respectively. In layman's words, higher horsepower will lead to lower mpg.
3. The model suggests that a car with 98 horsepower has a mpg value of 24.46708.

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.

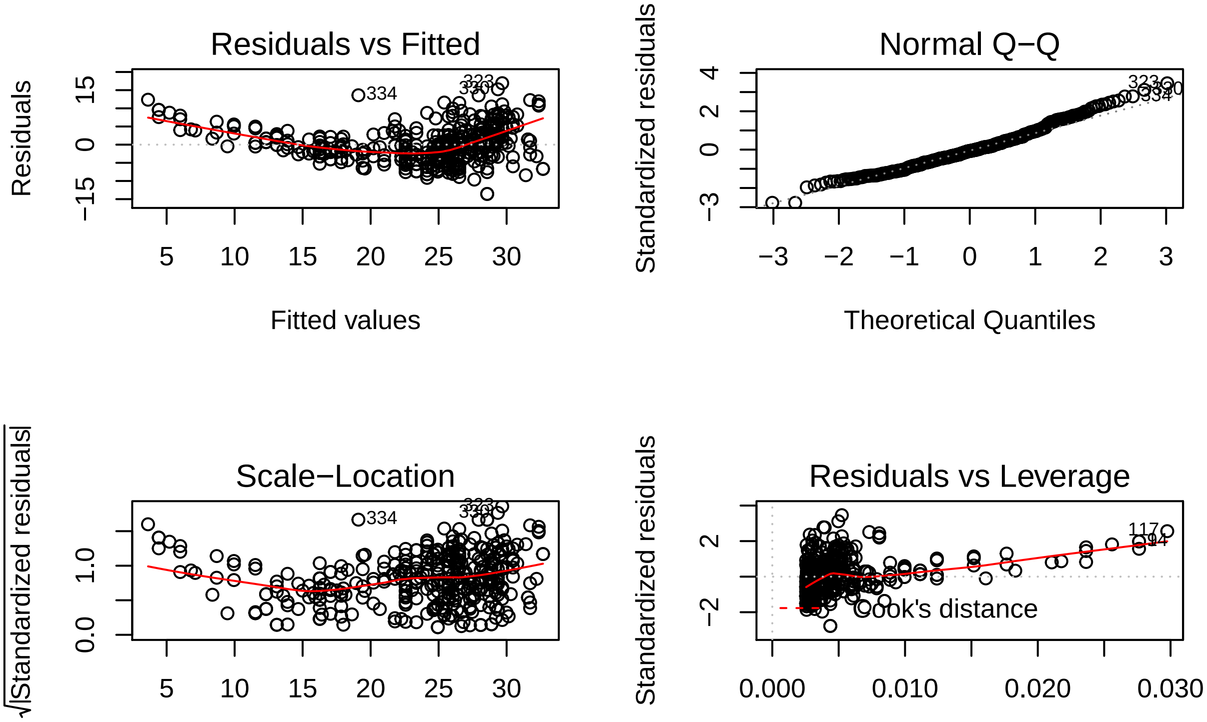
8. Plot the response and the predictor using `ggplot`. Add to the plot a line showing the least squares regression line.

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



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(mpg.lm1)
```



From the Residual-Fitted plot, the homogenuity is violated and the underlyding relationship is not linear–a curve shows up. There are no influential points in the data.

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.

Proof:

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

Since $E[(Y - f(x))(f(x) - g(x)) \mid X] = (f(x) - g(x))(E[Y \mid X] - f(x)) = 0$,

$E[(Y - g(x))^2 \mid X] \geq E[(Y - f(x))^2 \mid X]$, where the equality is satisfied when $g(x) = f(x) = E[Y|X]$.

11. (adopted from ELS Ex 2.6) 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 $X = x_o$? ($E_{Y|X}[Y]$)

$E[Y_i|X = x_0] = E[f(x_0) + \epsilon_i] = f(x_0) + E[\epsilon_i] = f(x_0)$

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

$Var[Y_i|X = x_0] = Var[f(x_0) + \epsilon_i] = Var[\epsilon_i] = \sigma^2$

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

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

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

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

$E_{Y|X}[(Y - f(x))(f(x) - \hat{f}(x_o))] = (f(x) - \hat{f}(x_o))E[Y - f(x)] = 0$

$E_{Y|X}[(Y - f(x))^2] = E[\epsilon^2] = Var[\epsilon]$

$E_{Y|X}[(f(x) - \hat{f}(x_o))^2] = Var_{Y|X}[f(x) - \hat{f}(x_o)] + E_{Y|X}[f(x) - \hat{f}(x_o)]^2$, where $Var_{Y|X}[f(x)] = 0$.

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

(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.

Because ϵ is an intrinsic prediction error that is independent of X—no matter how well we know the true function, it will be there when we try to predict.

(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 2.6(c))

$E_{Y,X}(f(x_o) - \hat{f}(x_o))^2 = Var[f(x_o) - \hat{f}(x_o)] + E[f(x_o) - \hat{f}(x_o)]^2 = Var[\hat{f}(x_0)] + bias^2$

(f) Establish a relationship between the squared biases and variance in the above Mean squared errors.

$MSE = Var + Bias^2$, which is that for any parameter or function to be estimated, we have

$$E[(f - \hat{f})^2] = Var[\hat{f}] + (f - E[\hat{f}])^2$$

Proof:

$Var[\hat{f}] = Var[\hat{f} - f] = E[(f - \hat{f})^2] - E[f - \hat{f}]^2 = MSE - Bias^2$

So, $MSE = Var + Bias^2$