STA521 HW1

[Jae Hyun Lee and jl914]

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?

```
outlook_auto <- summary(Auto)
any(is.na(Auto))
```

```
## [1] FALSE
```

answer : no missing data in this data set

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 1
```

answer: I can find that only name is qualitative predictor which has 304 levels and the other predictors are numeric quantitative.

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.

```
library(knitr)
range_table <- c(colnames(Auto)[1],range(Auto[,1]))
for(i in 2:(ncol(Auto)-1)){
  range_table <- rbind(range_table, c(colnames(Auto)[i],range(Auto[,i])))
}
colnames(range_table) <- c("variable name","min","max")
row.names(range_table) <- 1:nrow(range_table)
kable(range_table)
```

variable name	min	max
mpg	9	46.6
cylinders	3	8
displacement	68	455
horsepower	46	230
weight	1613	5140
acceleration	8	24.8
year	70	82
origin	1	3

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

```
summary_table <- c(colnames(Auto)[1],
                   round(mean(Auto[,1]),digits = 3),round(sd(Auto[,1]),digits = 3))
for(i in 2:(ncol(Auto)-1)){
  summary_table <- rbind(summary_table,
                         c(colnames(Auto)[i],
                           round(mean(Auto[,i]),digits = 3),
                           round(sd(Auto[,i]),digits = 3)))
}
colnames(summary_table) <- c("variable name","mean","standard deviation")
row.names(summary_table) <- 1:nrow(summary_table)
kable(summary_table)
```

variable name	mean	standard deviation
mpg	23.446	7.805
cylinders	5.472	1.706
displacement	194.412	104.644
horsepower	104.469	38.491
weight	2977.584	849.403
acceleration	15.541	2.759
year	75.98	3.684
origin	1.577	0.806

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*

```
library(GGally)
```

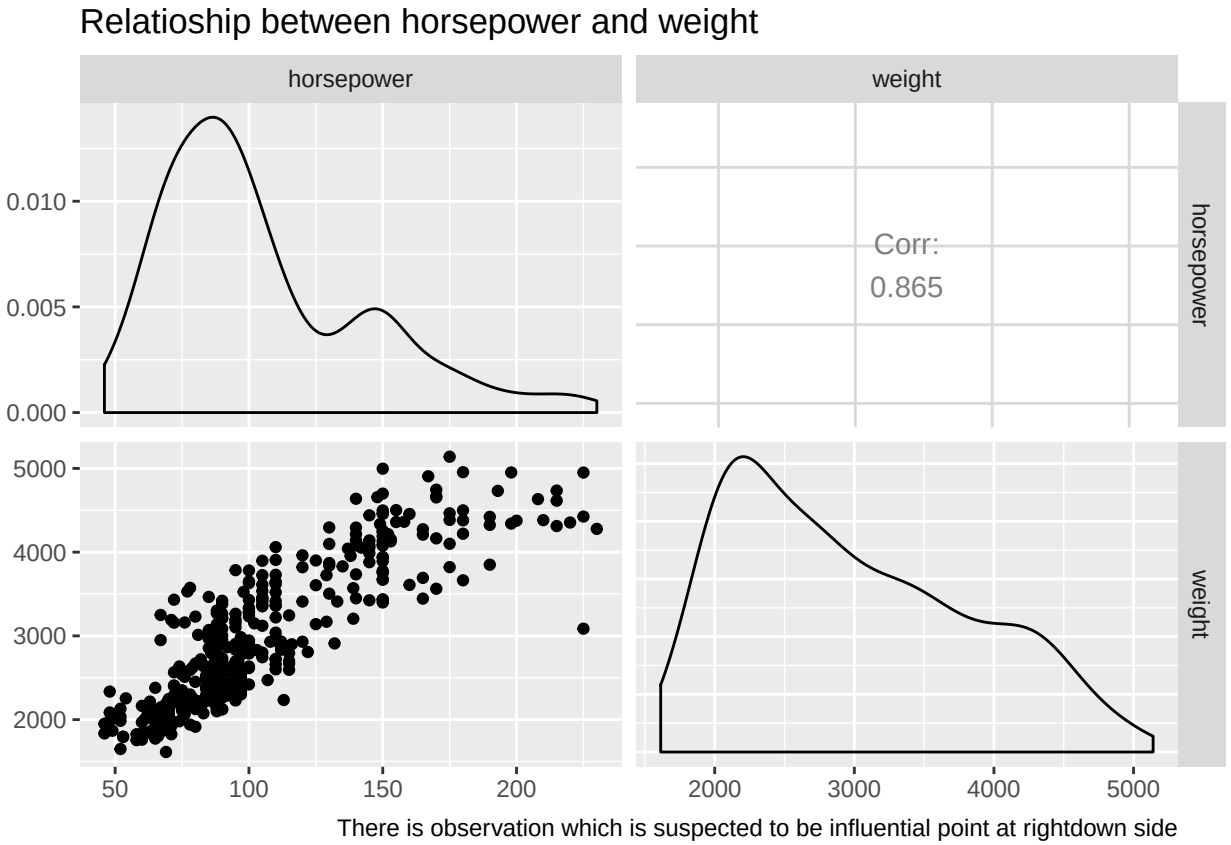
```
## Warning: package 'GGally' was built under R version 3.4.4
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

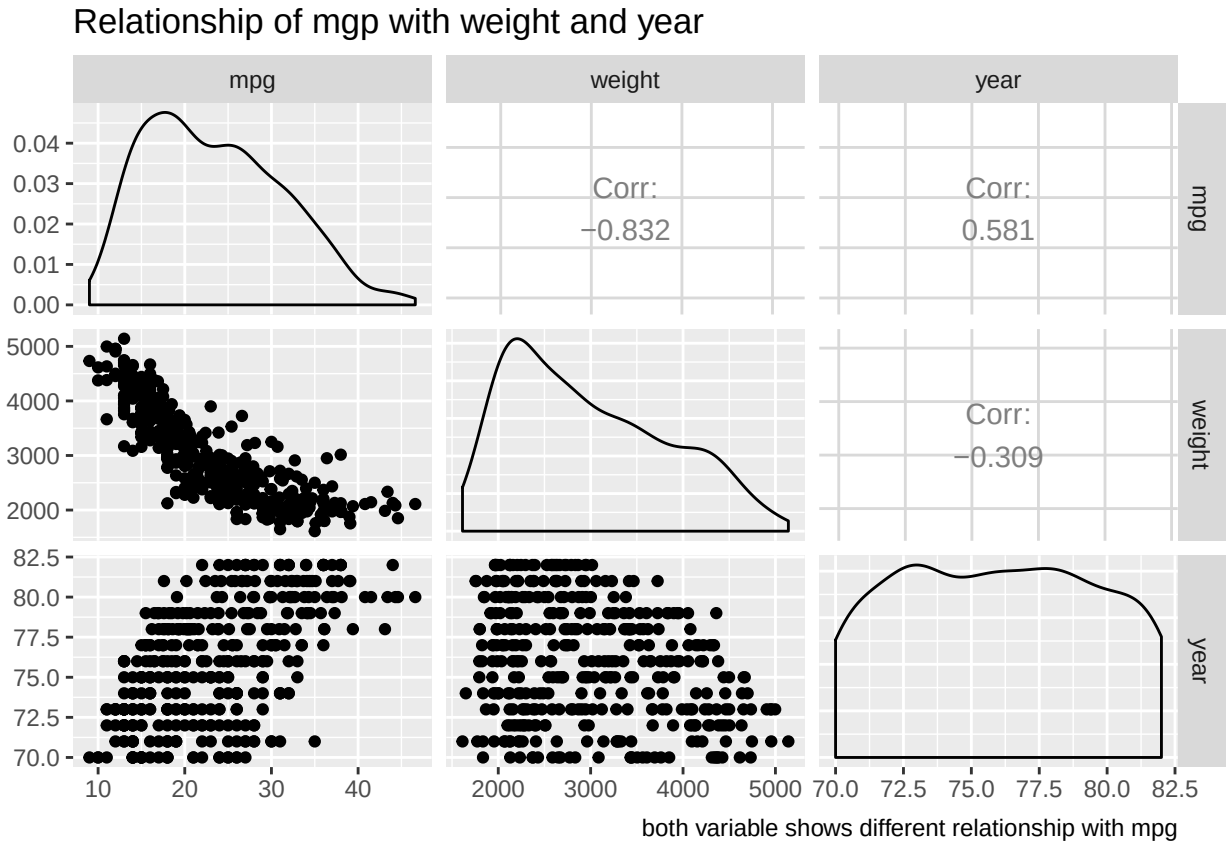
```
all_gp <- ggpairs(Auto, columns = c(1:8))
all_gp <- all_gp +
  labs(title = "Relatioship between all quantative predictors",
       caption = "There is strong positive relationship between horsepower and weight")

hlgt_gp <- ggpairs(Auto, columns = c(4:5))
hlgt_gp <- hlgt_gp +
  labs(title = "Relatioship between horsepower and weight",
       caption = "There is observation which is suspected to be influential point at rightdown side")
hlgt_gp
```



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.

```
mpg_gp <- ggpairs(Auto, columns = c(1,5,7))
mpg_gp <- mpg_gp +
  labs(title = "Relationship of mgp with weight and year",
       caption = "both variable shows different relationship with mpg")
mpg_gp
```



answer: weight and year variables show different relation with mpg each other. If I include variables related with enjine size or power, there would be collinearity problem. Thus I choose variable which has the strongest negative linear relationship with mpg. On the other hand, year has shown positive linear relationship with mpg. In consequence it can improve model by providing other information that weight cannot give.

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.

```
lm_mgp <- lm(mpg~horsepower, data = Auto)
cor(Auto$mpg,Auto$horsepower)
```

```
## [1] -0.7784268
```

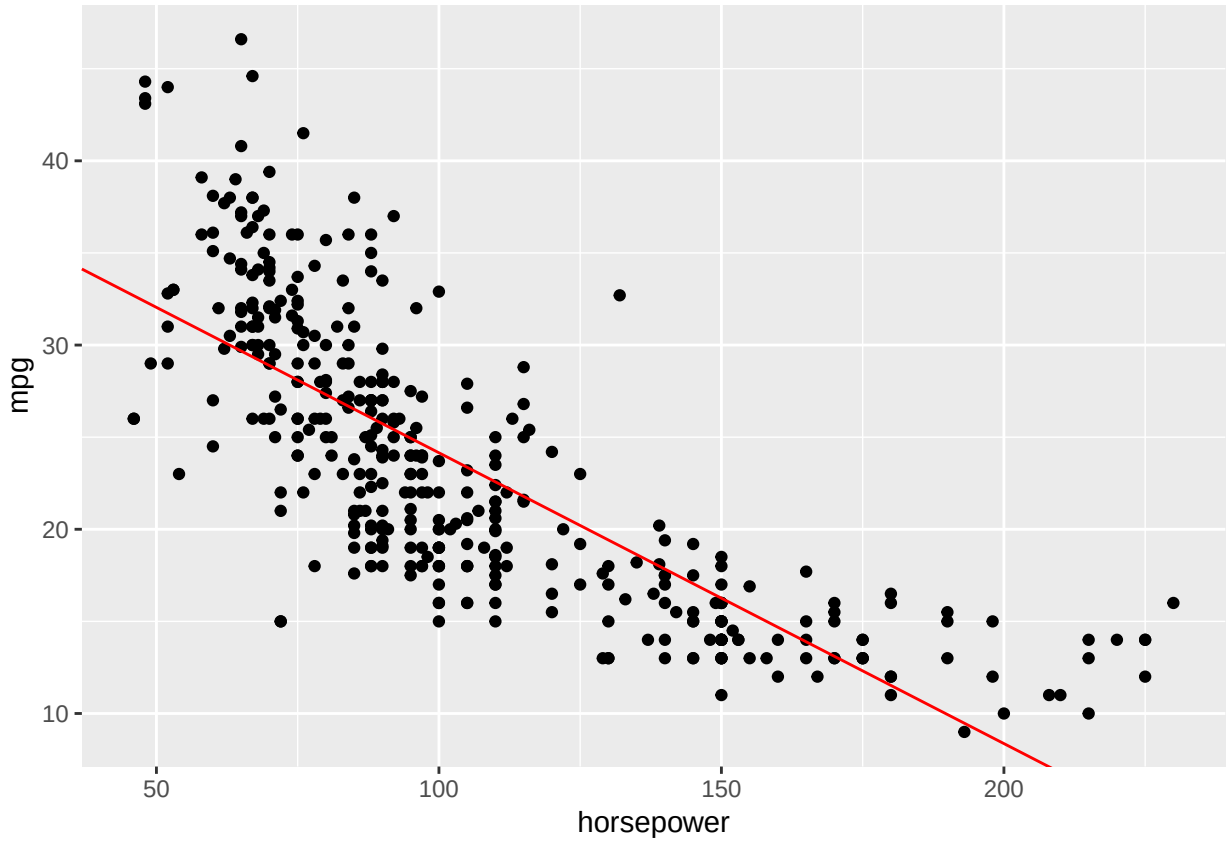
```r
summary(lm_mgp)
```

```
##
## 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
```

(a) I can find that they have relationship with strong evidence of t-test's p-value
(b) Although, I cannot find the strength of relationship through regression summary, I can find it using correlation function
(c) regression result show negative relationship
(d) There is tendency that the cars with small horsepower usually have higher mpg. That is they are more efficient car.

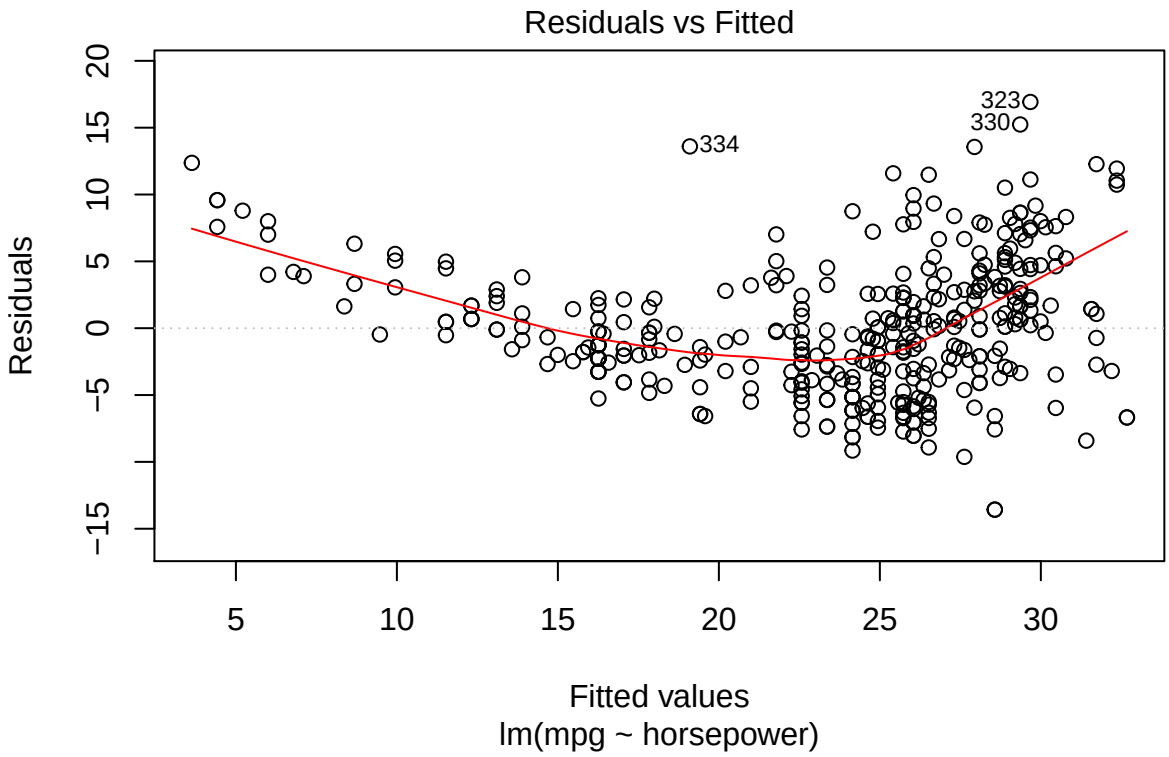
8. Plot the response and the predictor using `ggplot`. Add to the plot a line showing the least squares regression line.

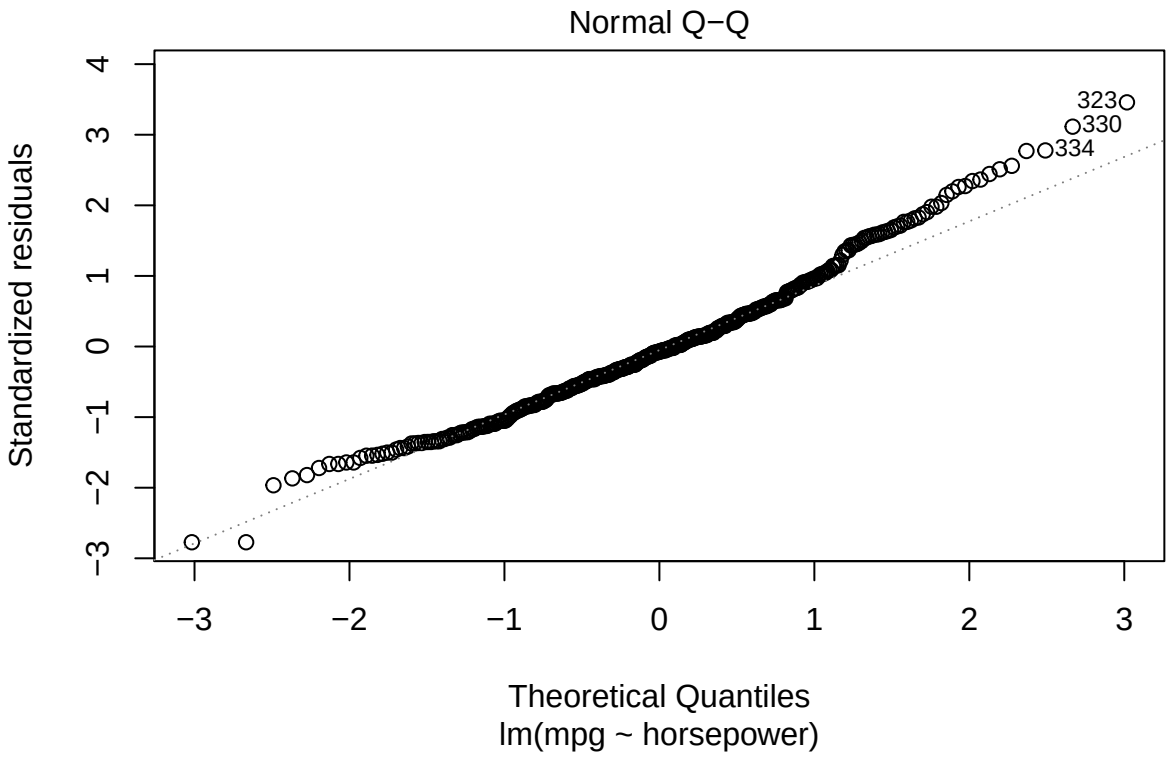
```r
ggplot(data = Auto, mapping = aes(x=horsepower,y=mpg)) +
  geom_point() +
  geom_abline(intercept =lm_mgp$coefficients[1],slope =lm_mgp$coefficients[2],colour="red")
```

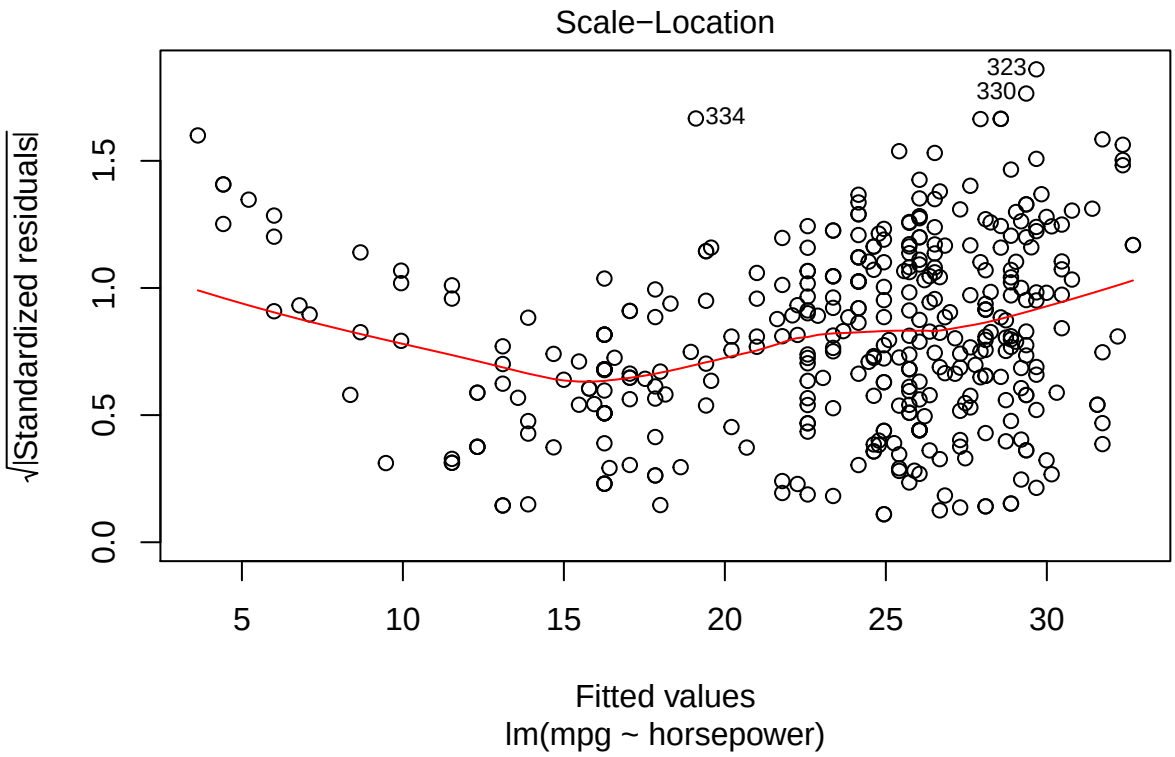


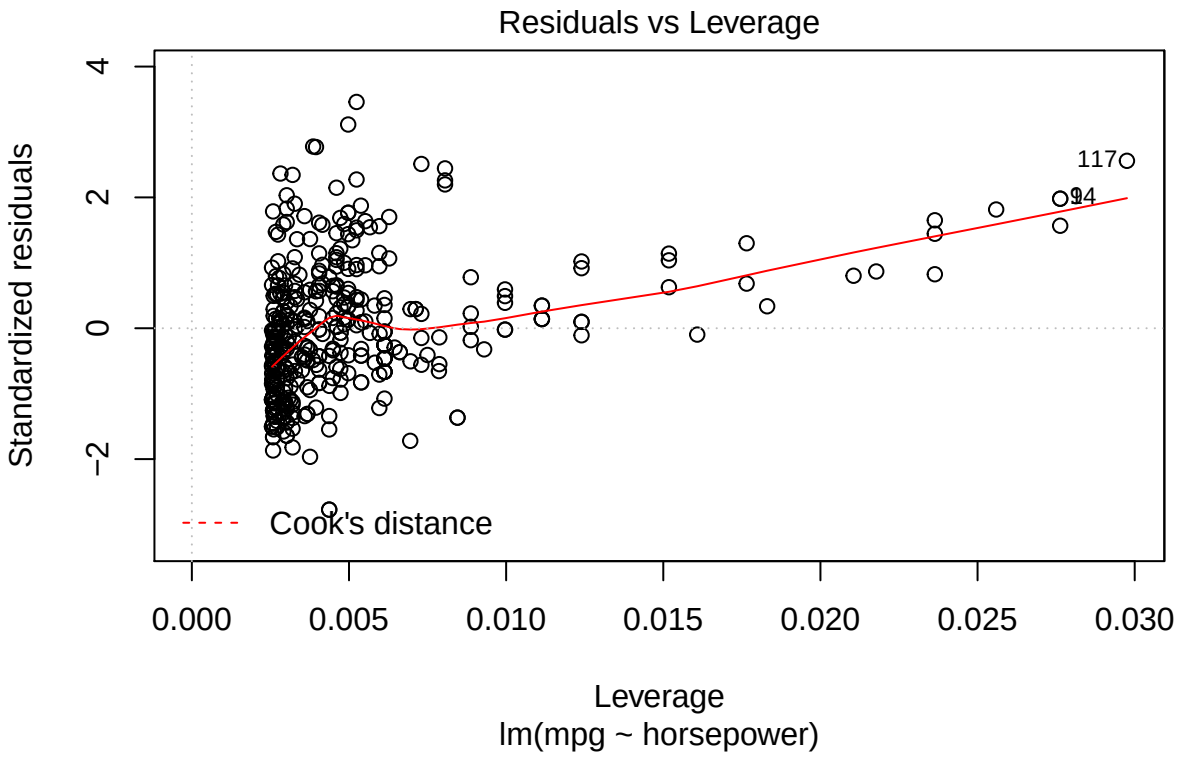
9. Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the model regarding assumptions for using a simple linear regression.

```
plot(lm_mgp)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ horsepower)
323
330
334
−15
−5
0
5
10
15
20
5
10
15
20
25
30


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ horsepower)
323
330
334
−3
−2
−1
0
1
2
3
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
10
15
20
25
30
323
330
334
Fitted values
lm(mpg ~ horsepower)



1. When I see residual vs fitted plot, the line in graph shows that there is unlinear relationship between two variables

2. other graphs show that there isn't severe violation of assumption which are normality, and outliers

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.*

answer: There are two ways to show above results. First,

$$(Y - g(x))^2 = (Y - f(x) + f(x) - g(x))^2 = (Y - f(x))^2 + (f(x) - g(x))^2 + 2(Y - f(x))(f(x) - g(x))$$

Then

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

Last term of above equation is 0 because $f(x) = E(Y \mid X)$ and first term is $var(Y \mid X)$. Furthermore, second term is non-negative because it is quadratic form and it has minimum value when $g(x) = f(x)$. Consequently, $f(x)$ is optimal predictor.

Second,

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

To find minimize above equation, I take derivative regard to $g(x)$

$$\frac{d}{dg(x)}(var(Y \mid X) + f(x)^2 - 2f(x)g(x) + g(x)^2) = -2f(x) + 2g(x).$$

Thus when $g(x) = f(x)$ has minimum.

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

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

answer:

(a) $E(Y \mid X = x_0) = E(f(x_0 + \varepsilon \mid X = x_0) = f(x_0)$

(b) $var(Y \mid X = x_0) = var(f(x_0)) + \sigma^2$

(c) Let $E(\hat{f}(x_0)) = m$, Then

$(Y - \hat{f}(x_0))^2 = (f(x) + \varepsilon - \hat{f}(x_0))^2 = (f(x) + \varepsilon - m + m - \hat{f}(x_0))^2 = (\hat{f}(x_0) - m)^2 + (f(x) - m)^2 + \varepsilon^2 + Crossproduct$

where expectation of crossproduct is 0 Thus expectation of equation is

$$var(\hat{f}(x_0)) + (f(x) - E(\hat{f}(x_0))^2 + \sigma^2$$

$var(\hat{f}(x_0))$ is variance of estimator,
$(f(x) - E(\hat{f}(x_0))^2$ is squared bias,
σ^2 is variance of ε.

(d) If $\hat{f}(x)$ is consistent estimator for $f(x)$ then as $n \to \infty$, bias and variance of estimator will be reduced into 0. But variance of ε is not reduced.

(e) Like (c), let $E(\hat{f}(x_0)$ be m. Then,

$$E[(f(x_0) - \hat{f}(x_0)^2] = E[(f(x_0) - m + m - \hat{f}(x_0)^2]$$

$$= E[(f(x_0) - m)^2] + E[(\hat{f}(x_0) - m)^2] + (f(x_0) - m)E[(\hat{f}(x_0) - m)] = var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2$$

(f) If MSE is fixed, then variance and bias of estimator is trade-off relationship.