STA521 HW1

[Wei Zhang wz94]

Due Wednesday September 12, 2018

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-F17 organization site by the deadline (the version that is submitted will be graded)

Exploratory Data Analysis

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

```r
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
```

```r
#check whether any variables have missing data
sum(is.na(Auto$mpg))
```

```
## [1] 0
```

```r
sum(is.na(Auto$displacement))
```

```
## [1] 0
```

```r
sum(is.na(Auto$horsepower))
```

```
## [1] 0
```

```r
sum(is.na(Auto$weight))
```

```
## [1] 0
```

```r
sum(is.na(Auto$acceleration))
```

```
## [1] 0
```

```r
sum(is.na(Auto$year))
```

```
## [1] 0
```

```r
sum(is.na(Auto$origin))
```

```
## [1] 0
```

```r
sum(is.na(Auto$name))
```

```
## [1] 0
```

```r
#we can find that all get results zero, which means none of those variables have missing variables.
```

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

```r
#From results in question 1, we can see that the name variable
#is qualitative, other variables are 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.

```r
library(knitr)
df<-data.frame(matrix(ncol = 3, nrow = 0))
df<-rbind(df,data.frame(t(c("mpg",range(Auto$mpg)))))
df<-rbind(df,data.frame(t(c("cylinders",range(Auto$cylinders)))))
df<-rbind(df,data.frame(t(c("displacement",range(Auto$displacement)))))
df<-rbind(df,data.frame(t(c("horsepower",range(Auto$horsepower)))))
df<-rbind(df,data.frame(t(c("weight",range(Auto$weight)))))
df<-rbind(df,data.frame(t(c("acceleration",range(Auto$acceleration)))))
df<-rbind(df,data.frame(t(c("year",range(Auto$year)))))
df<-rbind(df,data.frame(t(c("origin",range(Auto$origin)))))
varna<-c("Variable name","min", "max")
colnames(df)<-varna
kable(df)
```

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*

```r
df1<-data.frame(matrix(ncol = 3, nrow = 0))
df1<-rbind(df1,data.frame(t(c("mpg",mean(Auto$mpg),sd(Auto$mpg)))))
df1<-rbind(df1,data.frame(t(c("cylinders",mean(Auto$cylinders),sd(Auto$cylinders)))))
df1<-rbind(df1,data.frame(t(c("displacement",mean(Auto$displacement),sd(Auto$displacement)))))
df1<-rbind(df1,data.frame(t(c("horsepower",mean(Auto$horsepower),sd(Auto$horsepower)))))
df1<-rbind(df1,data.frame(t(c("weight",mean(Auto$weight),sd(Auto$weight)))))
df1<-rbind(df1,data.frame(t(c("acceleration",mean(Auto$acceleration),sd(Auto$acceleration)))))
df1<-rbind(df1,data.frame(t(c("year",mean(Auto$year),sd(Auto$year)))))
df1<-rbind(df1,data.frame(t(c("origin",mean(Auto$origin),sd(Auto$origin)))))
varna1<-c("Variable name","mean", "std")
colnames(df1)<-varna1
kable(df1)
```

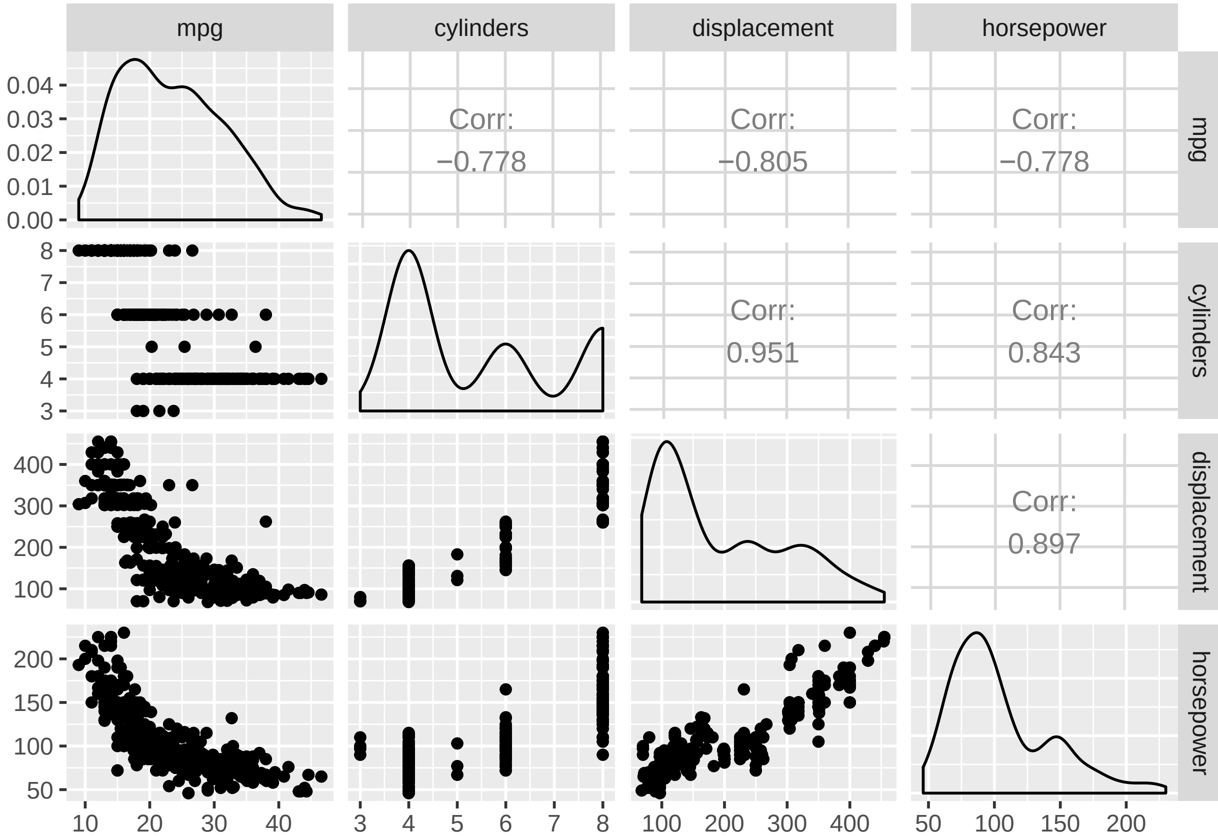
Variable name	mean	std
mpg	23.4459183673469	7.8050074865718
cylinders	5.4719387755102	1.70578324745278
displacement	194.411989795918	104.644003908905
horsepower	104.469387755102	38.4911599328285
weight	2977.58418367347	849.402560042949
acceleration	15.5413265306122	2.75886411918808
year	75.9795918367347	3.68373654357783
origin	1.5765306122449	0.805518183418306

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*

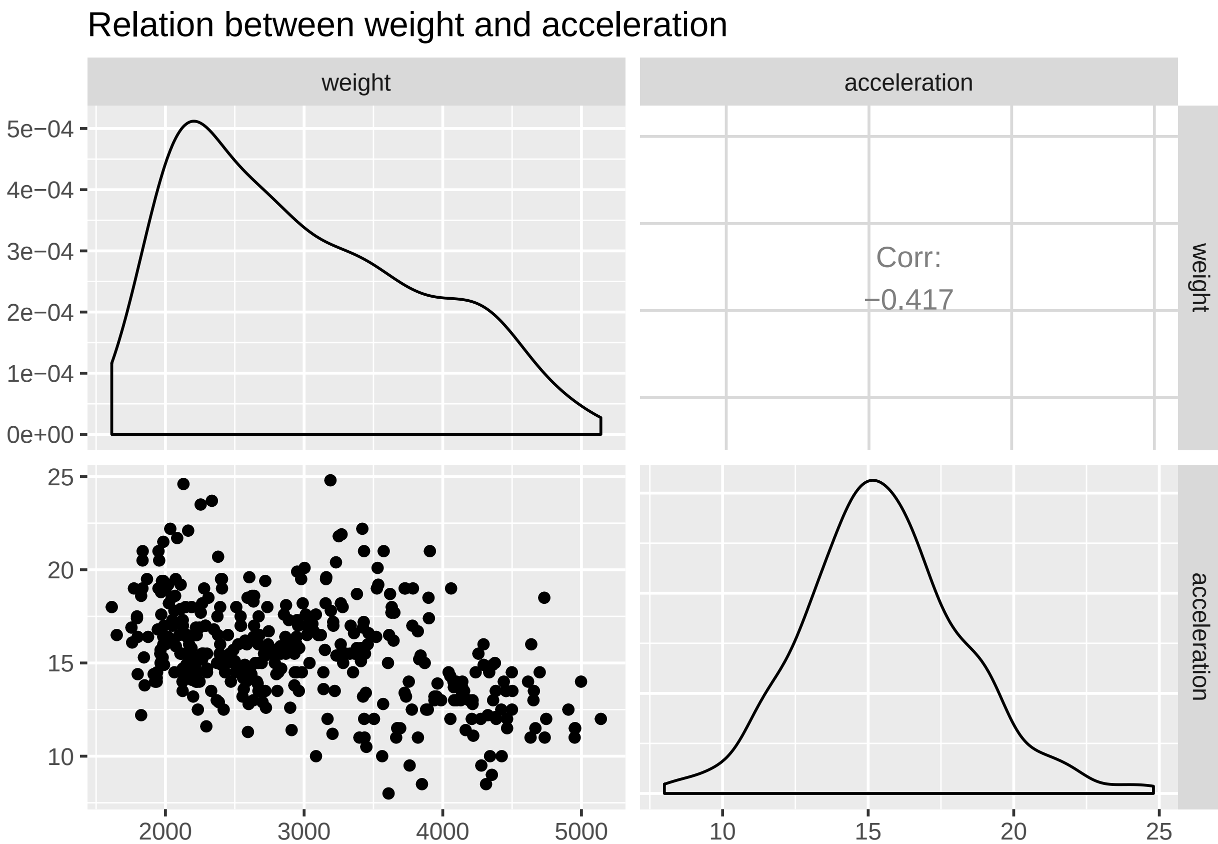
```r
library(GGally)
```

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

```r
ggpairs(Auto,columns=1:4)
```




```r
ggpairs(Auto,columns=5:6,title="Relation between weight and acceleration" )
```




```r
#we can see that the horsepower predictors has positive correaltion with
#cylinders and displacement.
#Also from the second scatter plot we can conclude that there is negative
#correlation between weight and acceleration.
```

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.

```r
#Yes, from the first graph I drew, I found mpg might be negatively
#correlated with cylinders, horsepower and displacement. This means
#Those three variables may have the ability to predict mpg.

model=lm(mpg~cylinders+displacement+horsepower,data=Auto)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7144  -3.1391  -0.3149   2.3481  16.5726
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.305268   1.324633  29.673  < 2e-16 ***
## cylinders    -0.719431   0.434180  -1.657 0.098331 .
```

```
## displacement -0.029120   0.008623  -3.377 0.000807 ***
## horsepower   -0.059935   0.013498  -4.440 1.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.523 on 388 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6641
## F-statistic: 258.7 on 3 and 388 DF,  p-value: < 2.2e-16
```

```
#we can find that the pval for displacement and horsepower are realy
#small which means they are significant. Additionally, the fval is
#very high which means those predictors are useful in predicting mpg.
```

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
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
```

```
#(a)Yes, there is negative relation between mpg and horsepower.
# and the relation is significant as the pval is smaller than 5%.
#(b)It is very strong as the pval is smaller than 5% which means
#it is significant.
#(c)The relation is negative as the coefficient for horsepower
#is around -0.15.
#(d)The coefficient of -0.15 means that if horsepower incresae 1,
```

```r
#then the mpg will decrease by aroud -0.15.
#(e)
predict(model1,data.frame(horsepower=c(98)),interval='confidence')
```

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

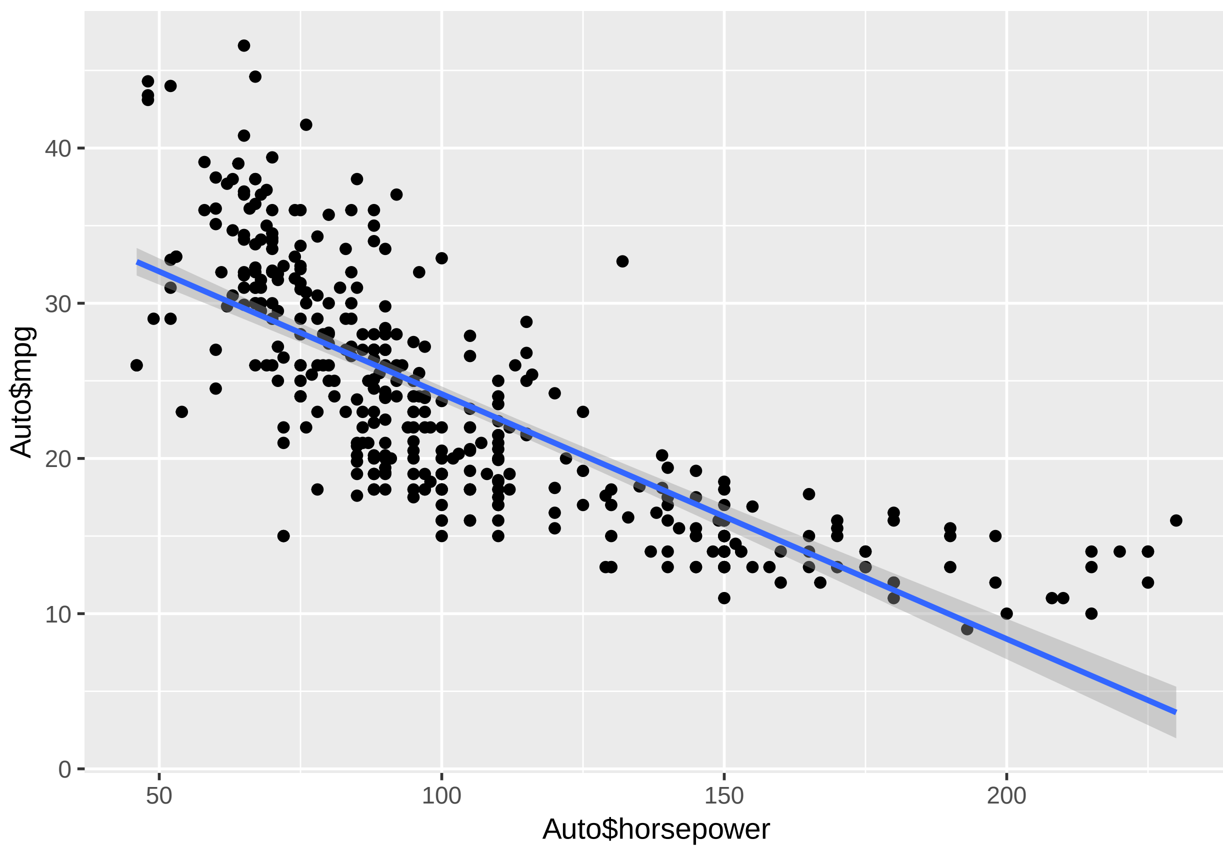
```r
predict(model1,data.frame(horsepower=c(98)),interval='prediction')
```

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

```r
#We can see the predicted mpg is around 24.46. The 95% confiednce and
#prediction interval are shown above. The 95% confidence interval means
#based on dist of the fitting model there are 95% chance the mpg will
#live in between 23.9 and 24.9. The 95% prediction interval means if
#we consider the dist of the prediction(which is differnet form the fitting),
#there is 95% chance that mpg will live in between 14.8 and 34.1 if the
#horsepower is 98.
```

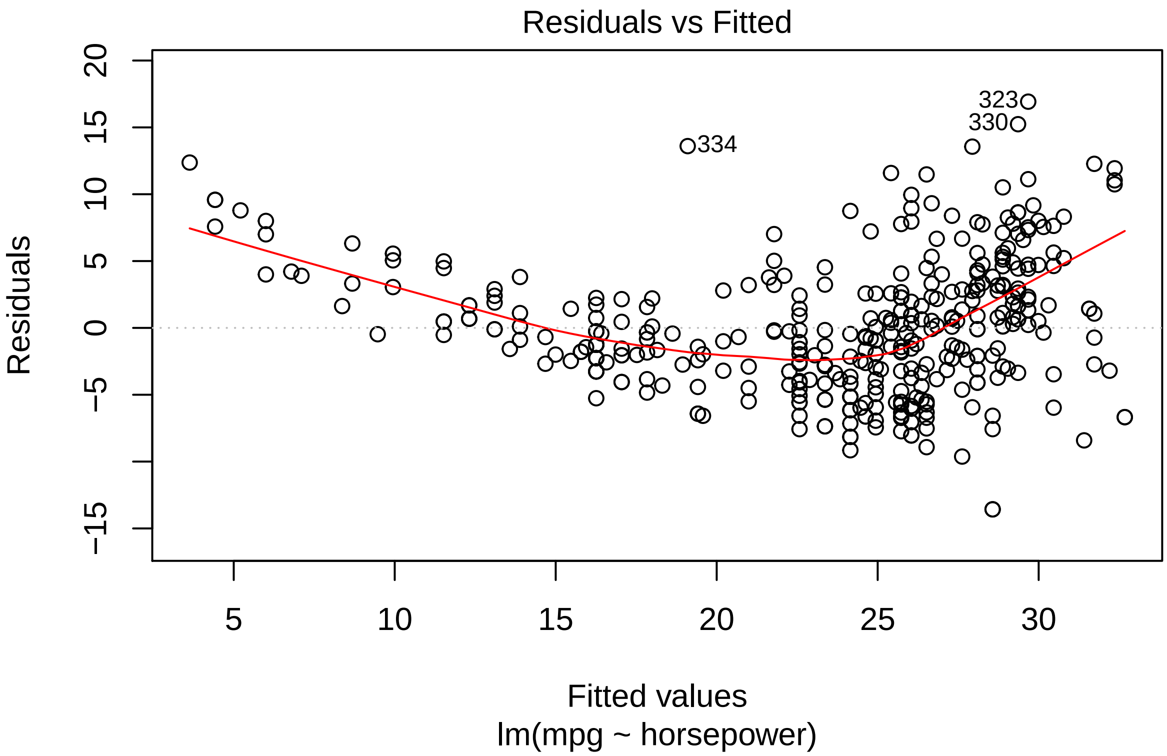
8. Plot the response and the predictor using `ggplot`. Add to the plot a line showing the least squares regression line.

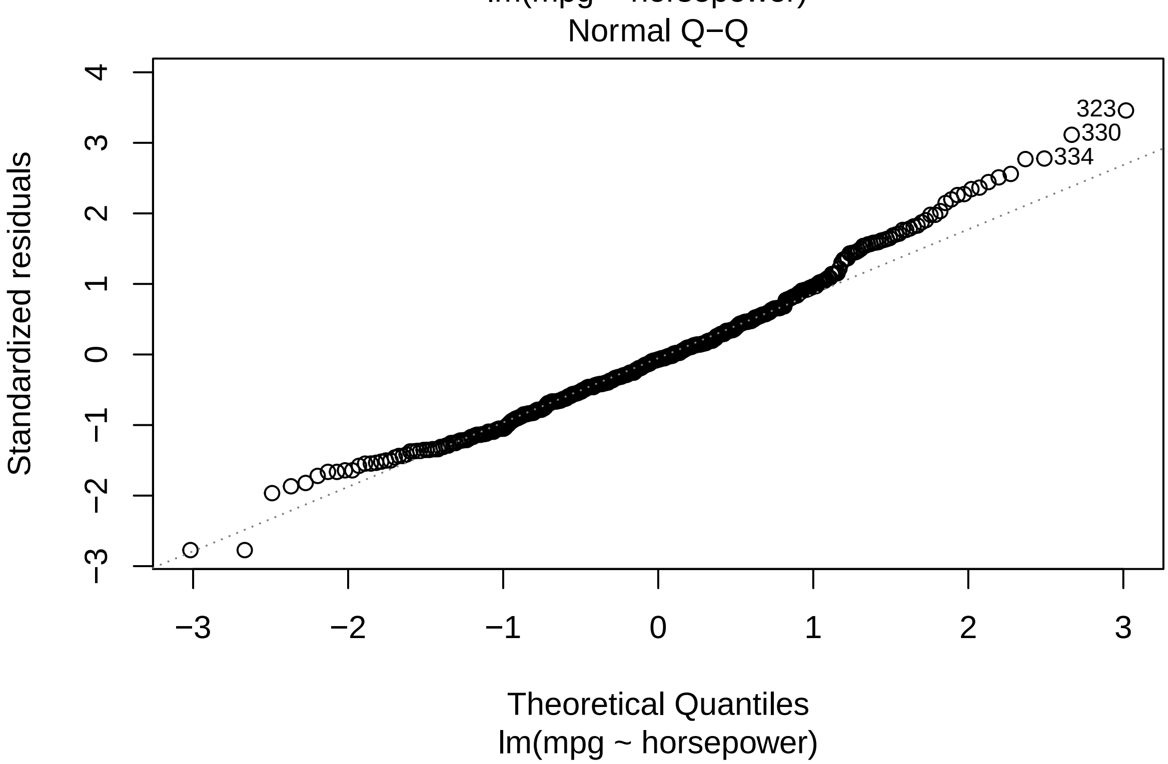
```r
library(ggplot2)
ggplot(Auto,aes(Auto$horsepower,Auto$mpg))+geom_point()+geom_smooth(method='lm')
```

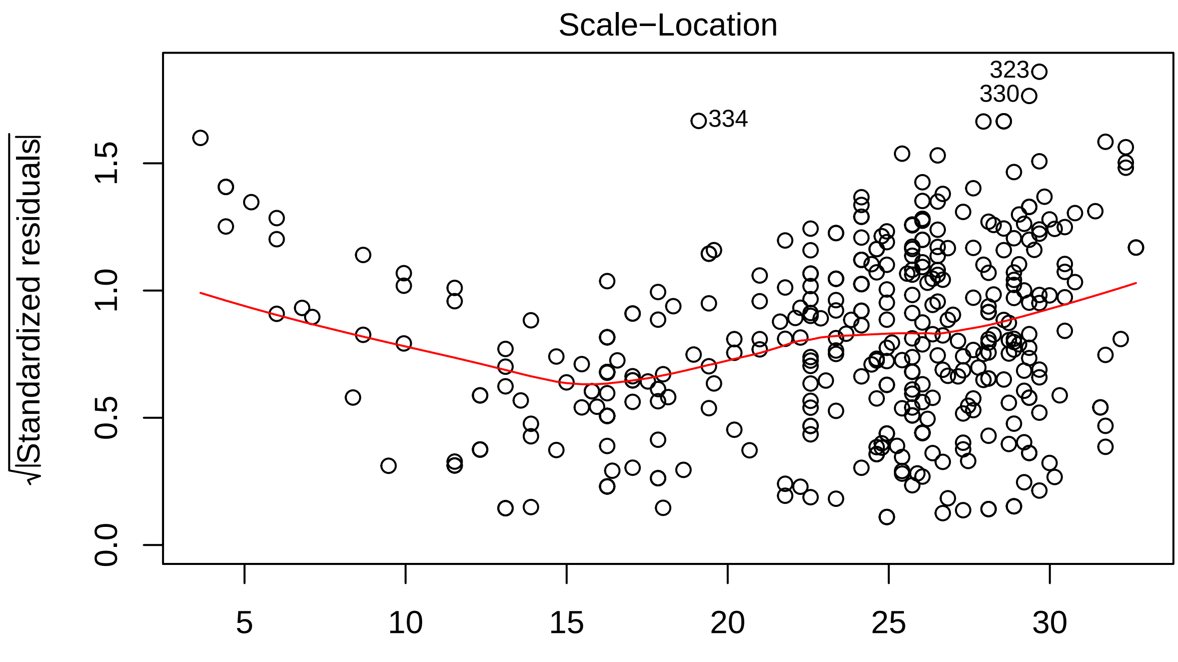


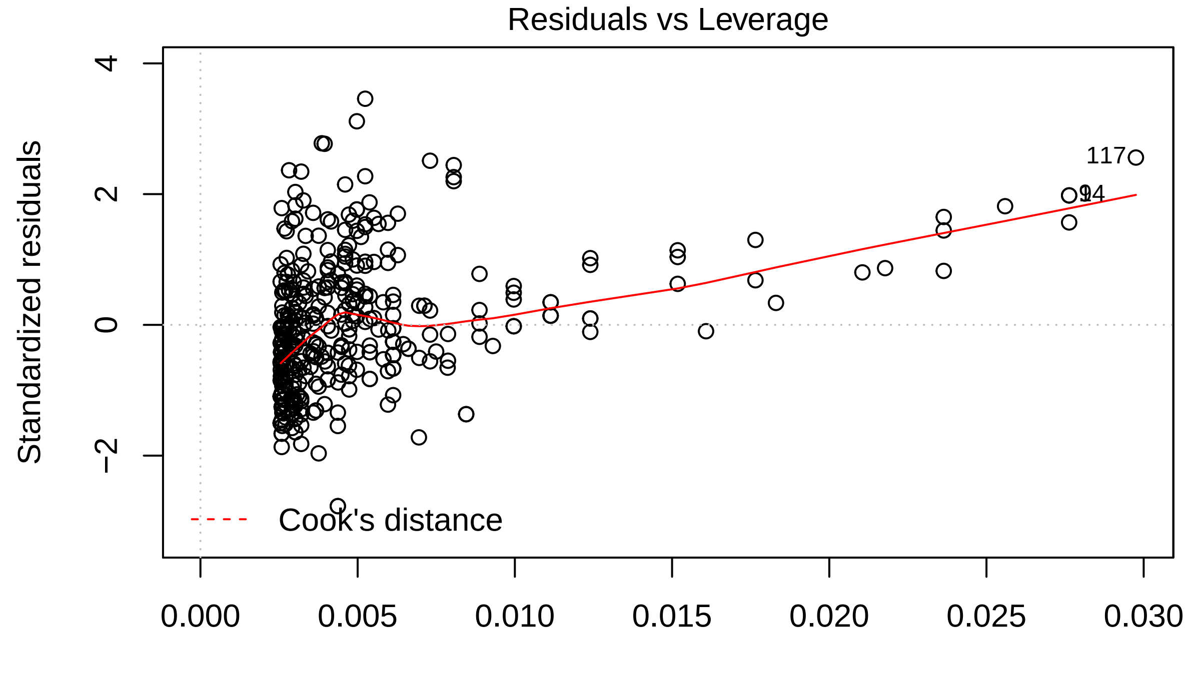
9. Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the model regarding assumptions for using a simple linear regression.

```r
plot(model1)
```










```
#From residual VS Fitted and Scale-Location
#we find that the variance might not be the same
#for the different means. From Normal Q-Q we can see that
#the residual might not be normal distributed.
#For cooks distance we can see most of the data point are fine.
```

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:
We want to want the min of $E[(Y - g(x))^2 \mid X = x]$, the variable is a fucntion $g(x)$, so we differentiate the equation by $g(x)$ and let it to be zero. That is:$\frac{\partial(E[(Y-g(x))^2|X=x)}{\partial(g(x))=0}$. That is $E[-2(Y - g(x)) \mid X = x] = 0$. That is equivalent to $g(x) = E[Y \mid X = x]$. As $\frac{\partial^2(E[(Y-g(x))^2|X=x)}{\partial(g(x))^2} > 0$, which means that $g(x) = E[Y \mid X = x]$ is the minminum point.

11. Irreducible error:
    (a) show that for any estimator $\hat{f}(x)$ that

$$E[(Y - \hat{f}(x))^2 \mid X = x] = \underbrace{(f(x) - \hat{f}(x)))^2}_{Reducible} + \underbrace{\mathsf{Var}(\epsilon)}_{Irreducible}$$

Hint: try the add zero trick of adding and subtracting $E[Y] = f(x)$

Answer:

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

$$= E[(Y - E[Y])^2 \mid X = x] + E(E[Y] - \hat{f}(x))^2 \mid X = x] + 2 * E[(Y - E(Y))(E(Y) - \hat{f}(x)) \mid X = x]$$

$$= E[(Y - E[Y])^2 \mid X = x] + E(E[Y] - \hat{f}(x))^2 \mid X = x] + 2 * E[Y - E[Y]] * E[E(Y) - \hat{f}(x)) \mid X = x]$$

$$= E[(Y - E[Y])^2 \mid X = x] + E(E[Y] - \hat{f}(x))^2 \mid X = x] + 2 * 0 * E[E(Y) - \hat{f}(x)) \mid X = x]$$

$$= E[(Y - E[Y])^2 \mid X = x] + E(E[Y] - \hat{f}(x))^2 \mid X = x] = Var(\epsilon) + (f(x) - \hat{f}(x))^2$$

(b) Show that the prediction error can never be smaller than σ^2:

$$E[(Y - \hat{f}(x))^2 \mid X = x] \geq \mathsf{Var}(\epsilon)$$

Answer:
As we know from part a, that $E[(Y - \hat{f}(x))^2 \mid X = x] = Var(\epsilon) + (f(x) - \hat{f}(x))^2$
Additionally we know that $(f(x) - \hat{f}(x))^2$ is great or equal to zero.
This says that $E[(Y - \hat{f}(x))^2 \mid X = x]$ is great or equal to $Var(\epsilon)$.

e.g. even if we can learn $f(x)$ perfectly that the error in prediction will not vanish.

12. Exercise 9.3 from Weisberg (hint: direct multiplication)

Answer:
We want to show A.37 holds, as we know

$$X_i'X_i = (X'X - x_i'x_i)^{-1}$$

.
We can just show A times A^{-1} equals to 1 where A represent to $X_i'X_i$. That is

$$(X'X - x_i'x_i)$$

times

$$[(X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_{ii}}]$$

is equals to 1.

$$[(X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_{ii}}] * (X'X - x_i'x_i)$$

$$= 1 - (X'X)^{-1}x_ix_i' + \frac{(X'X)^{-1}x_ix_i' - (X'X)^{-1}x_ix_i'(X'X)^{-1}x_ix_i'}{1-h_{ii}}$$

$$= 1 + \frac{-(X'X)^{-1}x_ix_i'(X'X)^{-1}x_ix_i' + (X'X)^{-1}x_ix_i'h_{ii}}{1-h_{ii}}$$

$$= 1 + \frac{-(X'X)^{-1}x_ih_{ii}x_i' + (X'X)^{-1}x_ix_i'h_{ii}}{1-h_{ii}}$$

$$= 1 + \frac{(-(X'X)^{-1}x_ix_i' + (X'X)^{-1}x_ix_i') * h_{ii}}{1-h_{ii}} = 1$$

This says that A.37 holds.

13. Verify Equation A.38 in the Appendix of Weisberg

Answer:
we just need to show

$$\hat{\beta}_{(i)} - \hat{\beta} = \frac{(X'X)^{-1}x_i\hat{e}_i}{1-h_{ii}}$$

We know

$$\hat{\beta}_{(i)} - \hat{\beta} = (X'X)^{-1}X'Y - (X_{(i)}'Y_{(i)})^{-1}X_{(i)}'Y_{(i)}$$

$$= (X'X)^{-1}X'Y - [(X'X)^{-1} + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_{ii}}] * X_{(i)}'Y_{(i)}$$

by A.37

$$= (X'X)^{-1}X'Y - (X'X)^{-1}(X'Y - (x_iy_i') - [\frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}}{1-h_{ii}}] * X_{(i)}'Y_{(i)}$$

$$= (X'X)^{-1}x_iy_i' + \frac{(X'X)^{-1}x_ix_i'(X'X)^{-1}x_iy_i' - (X'X)^{-1}x_ix_i'(X'Y)^{-1}X'Y}{1-h_{ii}}$$

$$= (X'X)^{-1}x_iy_i' + \frac{(X'X)^{-1}x_ih_{ii}y_i' - (X'X)^{-1}x_ix_i'(X'Y)^{-1}X'Y}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i y_i' - h_{ii}(X'X)^{-1}x_i y_i' + (X'X)^{-1}x_i h_{ii} y_i' - (X'X)^{-1}x_i x_i'(X'X)^{-1}X'Y}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i y_i' - (X'X)^{-1}x_i x_i'(X'X)^{-1}X'Y}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i(y_i' - x_i'(X'X)^{-1}X'Y)}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i(y_i' - x_i'\hat{\beta})}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i\hat{e}_i}{1-h_{ii}}$$

proved!

14. Exercise 9.4 from Weisberg

Answer:

$$y_i - x_i'\hat{\beta}_{(i)} = y_i - x_i'(\beta - \frac{(X'X)^{-1}x_i\hat{e}_i}{1-h_{ii}})$$

by A.38

$$= y_i - x_i'\beta + \frac{h_{ii}\hat{e}_i}{1-h_{ii}}$$

$$= \frac{\hat{e}_i(1-h_{ii}) + h_{ii}\hat{e}_i}{1-h_{ii}}$$

$$= \frac{\hat{e}_i}{1-h_{ii}}$$

proved!

15. Exercise 9.5 from Weisberg

Answer:
By defintion,

$$D_i = \frac{(\hat{\beta}-\hat{\beta})'(X'X)(\hat{\beta}-\hat{\beta})}{p'\hat{\sigma}^2}$$

by A.37 and A.38 we know:

$$\hat{\beta} - \hat{\beta} = \frac{(X'X)^{-1}x_i\hat{e}_i}{1-h_{ii}}$$

so

$$D_i = \frac{(((X'X)^{-1}x_i\hat{e}_i)/(1-h_{ii}))'(X'X)(((X'X)^{-1}x_i\hat{e}_i)/(1-h_{ii}))}{p'\hat{\sigma}^2}$$

$$= \frac{(\hat{e}_i)' x_i' (X'X)^{-1}(X'X)(X'X)^{-1} x_i \hat{e}_i}{p'\hat{\sigma}^2(1-h_{ii})^2}$$

$$= \frac{(\hat{e}_i)' x_i' (X'X)^{-1} x_i \hat{e}_i}{p'\hat{\sigma}^2(1-h_{ii})^2}$$

$$= \frac{(\hat{e}_i^2) h_{ii}}{p'\hat{\sigma}^2(1-h_{ii})(1-h_{ii})}$$

$$= \frac{h_{ii} r_i^2}{p'(1-h_{ii})}$$

proved!