HW1

[Rui Wang]

Due Monday September 11, 2017

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.

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
```

```
any(is.na(Auto))
```

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

Answer to Q1

The summary of dataset Auto is created above. Based on the result, there is no missing data in any variable.

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

```
split(names(Auto),sapply(Auto,function(x) paste(class(x),collapse="")))
```

```
## $factor
## [1] "name"
```

```
##
## $numeric
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
```

Answer to Q2

The results above show that predictor "name" is a factor, which is qualitative. Predictors "mpg","cylinders","displacement","hors
and "origin" are numeric, which should be quanlitative. However, the predictor "origin" should be categorical
based on my observation.

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)
Auto_quan=subset(Auto[,1:8])
min_max=apply(Auto_quan,2,range)
kable(t(as.data.frame(min_max)),col.names=c("min","max"),format='latex')
```

	min	max
mpg	9	46.6
cylinders	3	8.0
displacement	68	455.0
horsepower	46	230.0
weight	1613	5140.0
acceleration	8	24.8
year	70	82.0
origin	1	3.0

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

```
mean=apply(Auto_quan,2,mean)
sd=apply(Auto_quan,2,sd)
mean_sd=cbind(mean,sd)
kable(mean_sd,format='latex')
```

	mean	sd
mpg	23.445918	7.8050075
cylinders	5.471939	1.7057832
displacement	194.411990	104.6440039
horsepower	104.469388	38.4911599
weight	2977.584184	849.4025600
acceleration	15.541327	2.7588641
year	75.979592	3.6837365
origin	1.576531	0.8055182

5. Now remove the 10th through 85th observations (try this with `filter` from the `dplyr` package). What is the range, mean, and standard deviation of each predictor in the subset of the data that remains? *Again, present the output as a nicely formated table*

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Auto_filter=filter(Auto,! row(Auto)[,1] %in% c(10:85))

### calculate range
Auto_quan_filter=subset(Auto_filter[,1:8])
min_max_filter=apply(Auto_quan_filter,2,range)
kable(t(as.data.frame(min_max_filter)),col.names=c("min_filter","max_filter"),format='latex')
```

	min_filter	max_filter
mpg	11.0	46.6
cylinders	3.0	8.0
displacement	68.0	455.0
horsepower	46.0	230.0
weight	1649.0	4997.0
acceleration	8.5	24.8
year	70.0	82.0
origin	1.0	3.0

```
### calculate mean and standard deviation
mean_filter=apply(Auto_quan_filter,2,mean)
sd_filter=apply(Auto_quan_filter,2,sd)
mean_sd_filter=cbind(mean_filter,sd_filter)
kable(mean_sd_filter,format='latex')
```

	mean_filter	sd_filter
mpg	24.404430	7.867283
cylinders	5.373418	1.654179
displacement	187.240506	99.678367
horsepower	100.721519	35.708853
weight	2935.971519	811.300208
acceleration	15.726899	2.693721
year	77.145570	3.106217
origin	1.601266	0.819910

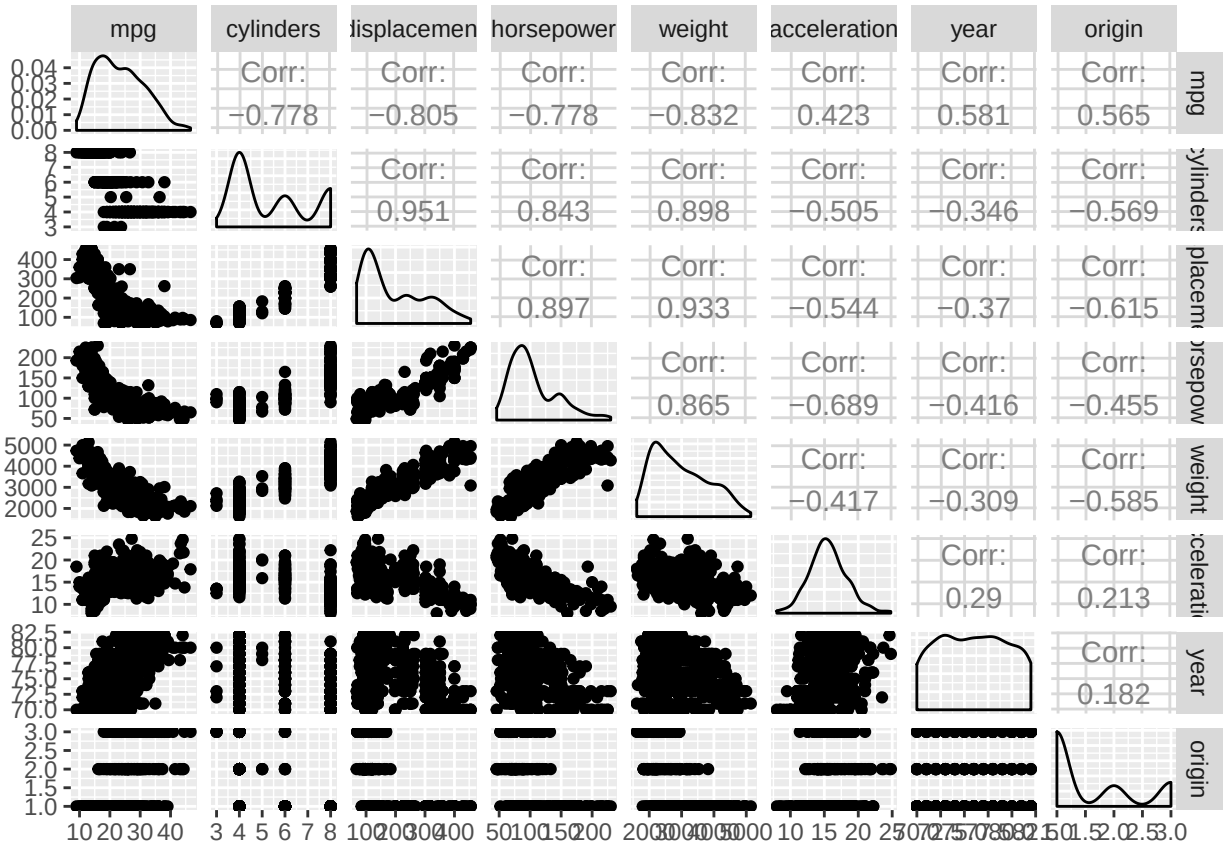
6. Investigate the predictors graphically, using scatterplot matrices (`ggpairs`) and other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings. *Try adding a caption to your figure*

```
library(GGally)
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggpairs(Auto,columns=1:8)
```



Answer to Q6

As we can see from the plot above, the correlation between "displacement" and "cylinders" is 0.951, the correlation between "weight" and "displacement" is 0.933. It means that those two pairs of predictors are highly correlated. Putting them altogether in a model may have the problem of multicollinearity. Some predictors are positively related such as "displacement" and "cylinders","weight" and "displacement".Some predictors are negatively related such as "weight" and "mpg", "displacement" and "mpg". In addtion, some predictors may not have a strong relationship such as "cylinders" and "year".

7. 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.

Answer to Q7

According to the plot in Q6, "displacement","horsepwer" and "weight" might be useful in predicting mpg using linear regression. The correlation between "mpg" and "displacement","horsepwer" and "weight" are -0.805,-0.778 and -0.832 resepctively, which indicate strong relationship.

Simple Linear Regression

8. Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` nction to print the results. Comment on the output.

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

```
##
## 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
```

Answer to Question 8

(a) Is there a relationship between the predictor and the response?

Answer: Yes, there is. The correlation between the predictor and the response is -0.15784.

(b) How strong is the relationship between the predictor and the response?

Answer: The significance of p-value shows that the relationship between the predicrtor and the response is strong.

(c) Is the relationship between the predictor and the response positive or negative?

Answer: The relationship is negative.

(d) Provide a brief interpretation of the parameters that would suitable for discussing with a car dealer, who has little statistical background.

Answer: The correation between "horsepower" and "mpg" is negative, which means that a car with larger horsepower spend more gallon per mile than a car with small horsepower. t-value together with p-value indicate that this relationship is very strong.

(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 cardealer.

Answer:The predicted mpg associated with a horsepower of 98 is 24.46708. The prediction interval is (14.8094, 34.12476), which means that the predicted value of mpg with a horsepower of 98 lie in this interval in 95%. The confidence interval is (23.97308, 24.96108), which means that you would expect 95% of this interval to include the true value of the mean of predicted mpg with a horsepower of 98.

```
new=data.frame(horsepower=98)
### calculate prediction interval
pred_plim <- predict(lm_reg, new, interval = "prediction")
kable(pred_plim,format="latex",caption = "prediction interval")
```

Table 1: prediction interval

fit	lwr	upr
24.46708	14.8094	34.12476

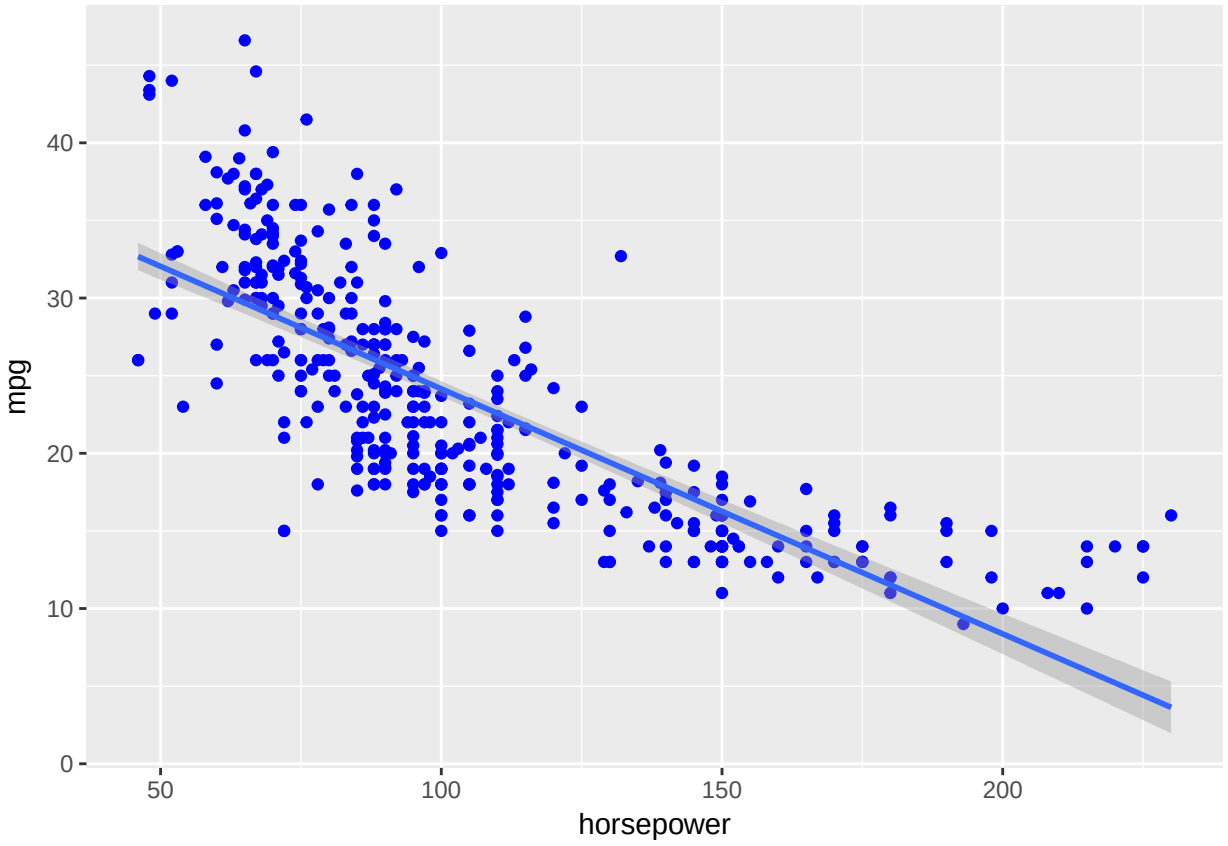
Table 2: confidence interval

fit	lwr	upr
24.46708	23.97308	24.96108

```
### calculate confidence interval
pred_clim <- predict(lm_reg, new, interval = "confidence")
kable(pred_clim,format="latex",caption = "confidence interval")
```

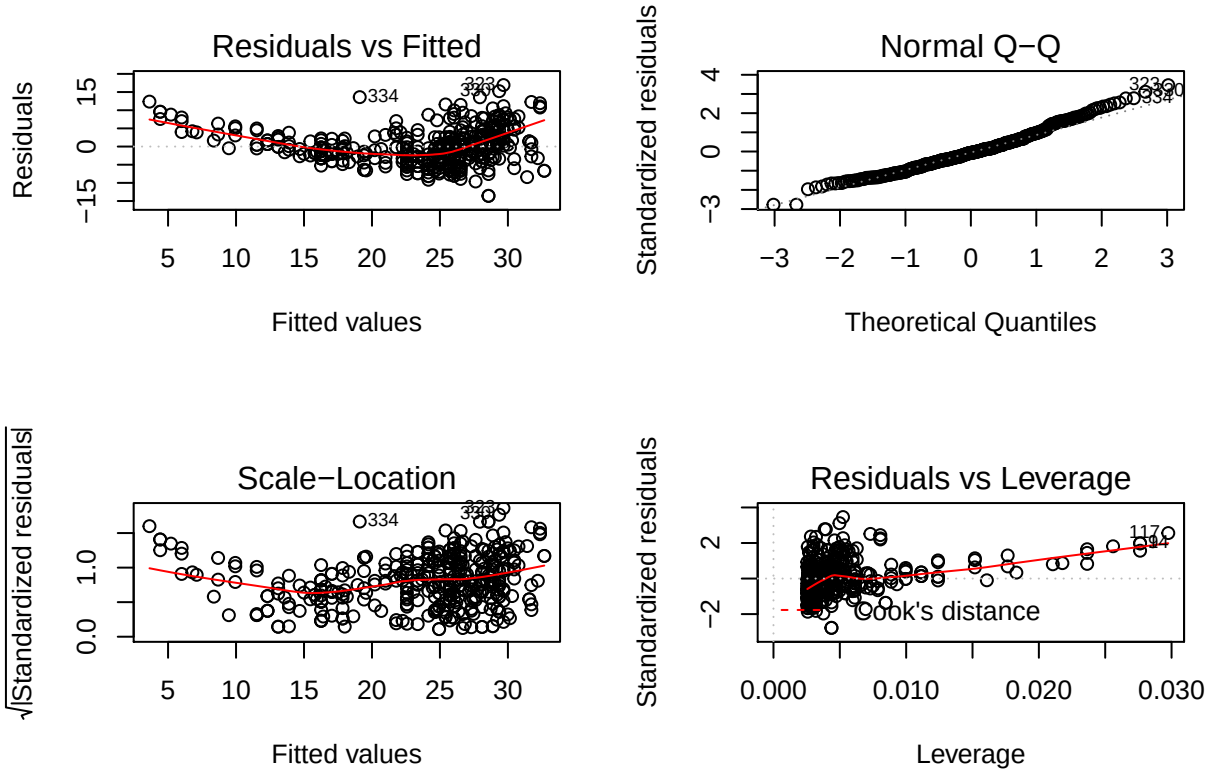
9. Plot the response and the predictor using `ggplot`. Add to the plot a line showing the least squares regression line.

```
library(ggplot2)
ggplot(data=Auto,aes(x=horsepower, y=mpg)) +
  geom_point(color="blue")+
  geom_smooth(method='lm')
```



10. 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(lm_reg,ask=F)
```



Answer to Question 10:

The frist Residuals-fitted figure and third figure both suggest curvature and nonconstant variance, which violate the assumption of zero mean and constant variance.

Theory

11. 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 $X = x$.

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

    In order to minimize $E[(Y - g(x))^2 \mid X = x]$

    $$\frac{d}{dg(x)} = 2g(x) - 2E(Y \mid X = x) = 0$$

    $$g(x) = E(Y \mid X = x)$$

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

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

We assume that there is some relationship between Y and X, which can be written in a general form

$$Y = f(x) + \epsilon$$

Our regression function is

$$E[Y \mid X = x] = f(x)$$

for any estimator $\hat{f}(x)$ it is easy to show that

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

We can optimize $\hat{f}(x)$ to minimize the reducibale error. Since ϵ is a random error term which is independent of X, the variance accosiated with the error term is irreducible.

(b) Show that the prediction error can never be smaller than

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

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

As we can see from part(a),

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

Since

$$(f(x) - \hat{f}(x))^2 \geqslant 0$$

We have

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