Summary statistics & Simple Linear Regression

Charlie Qu

Monday September 11, 2017

This writing involves the Auto data set from ISLR. I explore several practivcs in basic data analytics

Exploratory Data Analysis

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

The orginal data contained 408 observations but 16 observations with missing values were removed.Generally, the dataset contains 392 observations on the following 9 variables.

```
library(ISLR)
data(Auto)
colnames(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

The data dimision is

```
d = dim(Auto)
d
```

```
## [1] 392   9
```

Here is a summary of the 9 variables:

```
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
```

2. Based on the summary demonstrate which predictors are quantitative, and which are qualitative.

By viewing the dataset (check the code following), we find that (1) mpg(miles per gallon), (2) cylinders(Number of cylinders between 4 and 8), (3) displacement(Engine displacement (cu. inches)), (4) horsepower (Engine horsepower), (5) weight(Vehicle weight (lbs.)), (6) acceleration(Time to accelerate from 0 to 60 mph (sec.)), (7) year(Model year (modulo 100)) are quantitative predictors; (8) origin(Origin of car (1. American, 2. European, 3. Japanese)), (9) name(Vehicle name) are qulitative predictors

3. (1)Find the range of each quantitative predictor, with using the `range()` function. (2)Create a table with variable name, min, max with one row per variable. `kable` from the package `knitr` can display tables nicely.

The ranges as below

```
c("range of mpg", range(Auto$mpg)[2]-range(Auto$mpg)[1])
```

```
## [1] "range of mpg" "37.6"
```

```
c("range of cylinders", range(Auto$cylinders)[2]-range(Auto$cylinders)[1])
```

```
## [1] "range of cylinders" "5"
```

```
c("range of displacement", range(Auto$displacement)[2]-range(Auto$displacement)[1])
```

```
## [1] "range of displacement" "387"
```

```
c("range of horsepower", range(Auto$horsepower)[2]-range(Auto$horsepower)[1])
```

```
## [1] "range of horsepower" "184"
```

```
c("range of weight", range(Auto$weight)[2]-range(Auto$weight)[1])
```

```
## [1] "range of weight" "3527"
```

```
c("range of acceleration", range(Auto$acceleration)[2]-range(Auto$acceleration)[1])
```

```
## [1] "range of acceleration" "16.8"
```

```
c("range of year", range(Auto$year)[2]-range(Auto$year)[1])
```

```
## [1] "range of year" "12"
```

The min and max summary are as below, using the kable(), which is consistent with the range(), and the summary in question 1.

```
library(knitr)
colMax <- function(data) {sapply(data, max)}
colMin <- function(data) {sapply(data, min)}
max=colMax(Auto[, 1:7])
min=colMin(Auto[, 1:7])
Qualitative=data.frame(min,max)
kable(Qualitative)
```

	min	max
mpg	9	46.6
cylinders	3	8.0
displacement	68	455.0
horsepower	46	230.0
weight	1613	5140.0

	min	max
acceleration	8	24.8
year	70	82.0

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

```
colSD<- function(data) {sapply(data, sd, na.rm = TRUE)}
means=colMeans(Auto[, 1:7])
sds=colSD(Auto[, 1:7])
Qms=data.frame(means,sds)
kable(Qms)
```

	means	sds
mpg	23.445918	7.805008
cylinders	5.471939	1.705783
displacement	194.411990	104.644004
horsepower	104.469388	38.491160
weight	2977.584184	849.402560
acceleration	15.541327	2.758864
year	75.979592	3.683737

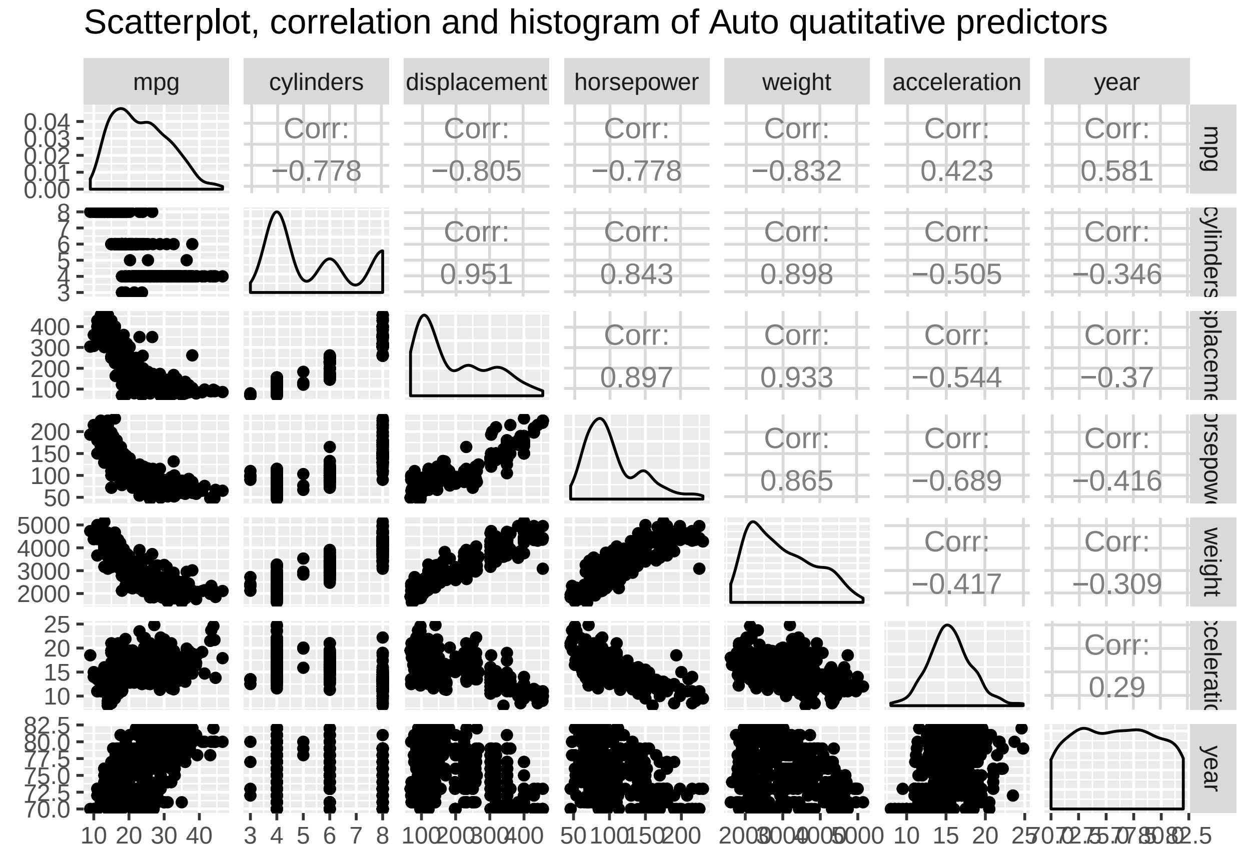
5. Now remove the 10th through 85th observations (try this with `filter` from the `dplyr` package). Find the range, mean, and standard deviation of each predictor in the subset of the remaining data. *Again, present the output as a nicely formated table*

```
library(dplyr)
N=1:nrow(Auto)
No.Auto=data.frame(Auto[,1:7],N)
Filterd=dplyr::filter(No.Auto,!(N>=10 & N<=85))
means.filter=colMeans(Filterd[, 1:7])
sds.filter=colSD(Filterd[, 1:7])
range.filter=colMax(Filterd[, 1:7])-colMin(Filterd[, 1:7])
Filter.Summary=data.frame(means.filter,sds.filter,range.filter)
kable(Filter.Summary)
```

	means.filter	sds.filter	range.filter
mpg	24.404430	7.867283	35.6
cylinders	5.373418	1.654179	5.0
displacement	187.240506	99.678367	387.0
horsepower	100.721519	35.708853	184.0
weight	2935.971519	811.300208	3348.0
acceleration	15.726899	2.693721	16.3
year	77.145570	3.106217	12.0

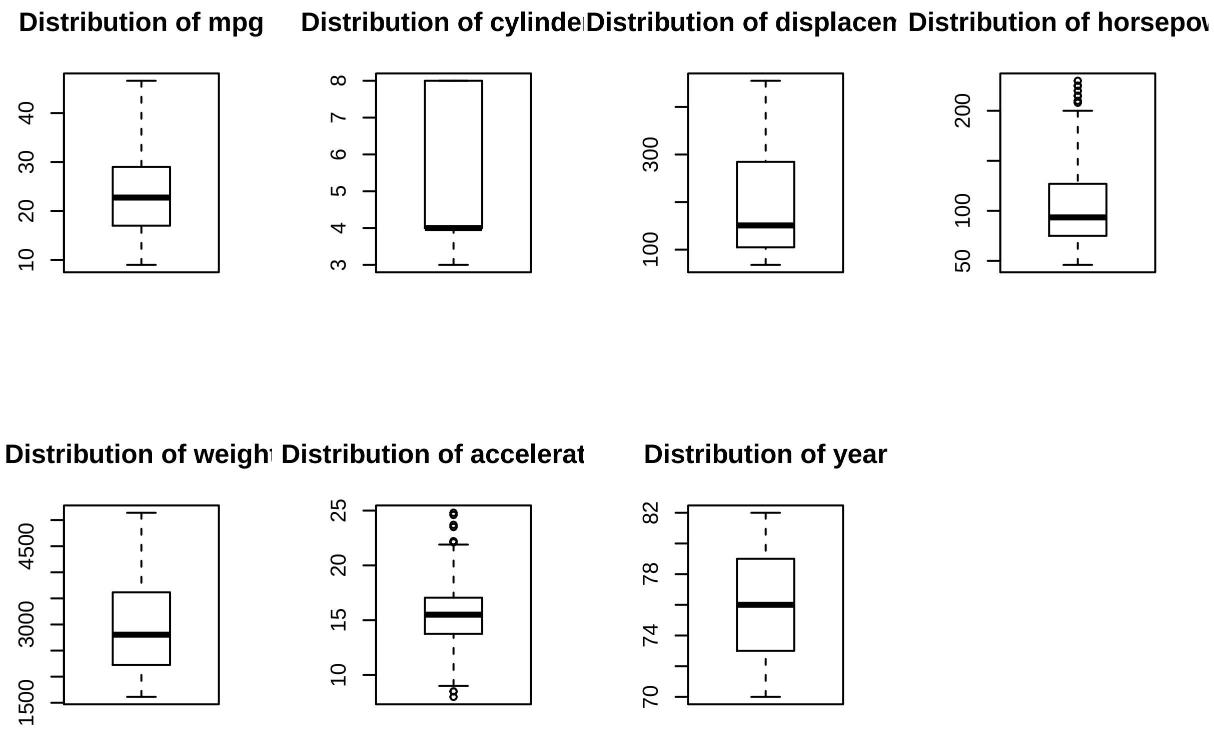
6. Investigate the predictors graphically, using scatterplot matrices (`ggpairs`) and other tools of your choice. Create some plots highlighting the relationships among the predictors. *Try adding a caption to your figure*

```
library(GGally)
library(ggplot2)
ggp=ggpairs(Auto, columns= 1:7)
print(ggp + ggtitle("Scatterplot, correlation and histogram of Auto quatitative predictors"))
```



By running the ggpairs() we find that: (1)For the relaltionships (a)very strongly correlated: displacement and cylinders(0.951), weiht and displacement(0933); (b)strongly correlated: cylinders and mpg (-0.778),displacement and mpg(-0.805),horsepower and mpg(-0.778),weight and mpg(-0.832),horsepower and cylinders(0.843),weight and cylinders(0.898),horspower and displacement(0.897), weight and horsepower(0.865); (c)moderdately strong correlated:accelaration and mpg(0.423),year and mpg(0.581),accelaration and cylinders(-0.505),acceleration and displacement(-0.544),accelerationo and horspower(-0.689),year and horspower(-0.416),acceleration and weight(-0.417); (d)weakly correlated: year and cylinders(-0.346), year and displacement(-0.37),year and weight(-0.309). (2) For the distributions: mpg, displacement, horsepower, weight are generally right skewed, acceleration is nearly prefect bell-shaped, year is symmetric closed to uniform distribution, and cylinders is of no particular pattern.

```
par(mfrow=c(2,4))
boxplot(Auto$mpg)
title("Distribution of mpg")
boxplot(Auto[,2])
title("Distribution of cylinders")
boxplot(Auto[,3])
title("Distribution of displacement")
boxplot(Auto[,4])
title("Distribution of horsepower")
boxplot(Auto[,5])
title("Distribution of weight")
boxplot(Auto[,6])
title("Distribution of acceleration")
boxplot(Auto[,7])
title("Distribution of year")
```



By the boxplots we can verify the skewness and potentially outliers of each predictors, as shown above.

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.

Based on the scattorplots and correlation coefficient, cylinders, displacement, horsepower and weight are strongly correlated to mpg, so that they can potentially be chosed as predictors of mpg.

```
summary(lm(mpg ~ cylinders, data=Auto))
```

```
##
## Call:
## lm(formula = mpg ~ cylinders, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2413  -3.1832  -0.6332   2.5491  17.9168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.9155     0.8349   51.40   <2e-16 ***
## cylinders    -3.5581     0.1457  -24.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.914 on 390 degrees of freedom
## Multiple R-squared:  0.6047, Adjusted R-squared:  0.6037
## F-statistic: 596.6 on 1 and 390 DF,  p-value: < 2.2e-16
```

```r
summary(lm(mpg ~ displacement, data=Auto))
```

```
## 
## Call:
## lm(formula = mpg ~ displacement, data = Auto)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.9170  -3.0243  -0.5021   2.3512  18.6128 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  35.12064    0.49443   71.03   <2e-16 ***
## displacement -0.06005    0.00224  -26.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.635 on 390 degrees of freedom
## Multiple R-squared:  0.6482, Adjusted R-squared:  0.6473 
## F-statistic: 718.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

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

```
## 
## 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
summary(lm(mpg ~ weight, data=Auto))
```

```
## 
## Call:
## lm(formula = mpg ~ weight, data = Auto)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.9736  -2.7556  -0.3358   2.1379  16.5194 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 46.216524   0.798673   57.87   <2e-16 ***
## weight      -0.007647   0.000258  -29.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.333 on 390 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6918
## F-statistic: 878.8 on 1 and 390 DF,  p-value: < 2.2e-16
```

As the summary displays the R-square values range from to 60.37% to 69.18%. Since there are also strong correlation between each pair of the for predictors, multicollinearity should be carefully considered, and with transformation the model can be fitted much better. The significance of the linear relationship are also validated by the small p-values of the intercepts, indicated the null hypothesis of non-related should be reject. For further example

```r
summary(lm(log(mpg) ~ log(weight)+log(cylinders), data=Auto))
```

```
##
## Call:
## lm(formula = log(mpg) ~ log(weight) + log(cylinders), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59128 -0.10270 -0.00582  0.09948  0.61682
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.03408    0.41151  24.383  < 2e-16 ***
## log(weight)    -0.81932    0.06196 -13.222  < 2e-16 ***
## log(cylinders) -0.25085    0.05765  -4.351 1.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1615 on 389 degrees of freedom
## Multiple R-squared:  0.7757, Adjusted R-squared:  0.7745
## F-statistic: 672.6 on 2 and 389 DF,  p-value: < 2.2e-16
```

This helps improve R-squre to 77.45%

Simple Linear Regression

8. Use the `lm()` function to perform a simple linearregression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results.

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

```
Yes. Based on the test of significance of regression, $H_0:\beta_1=0$ is rejected since p-value<0.05.
```

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

```
##
## 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
```

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

Since the t-statistics is fairly small, and the p-value is nearly 0, we can say there is strong linear relationship between mpg and horsepowre.

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

It is negative since β_1 =-0.1578.

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

Mpg is closely negatively related to the weight, the horsepower, the number of cylinders and the engine displacement of the car. For example, a car is of very high horsepower, or weight, the mpg will very likely be low. Therefore, if a customer looks for a economic car model, we may use low levels of weight, horsepower, number of cylinders and engine displacement as parameters to estimate. Besides, these factors also positively related to each other pairwise, that is, if we only have one or two of the four factors, we can still estimate the mpg as well as the other factors briefly.

(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.

We calculate prediction invertal using predict(), and calculate CI by the following definition

$$\hat{\mu}(x_0)) \pm t_{\alpha/2,n-2} \times \sqrt{MSE(1/n + (x_0 - \bar{x})^2/S_{xx})}$$

```r
pred <- predict(lm(mpg ~ horsepower, data=Auto),newdata=Auto[178,],se=T,interval = "prediction")
pred
```

```
## $fit
##          fit     lwr      upr
## 180 24.46708 14.8094 34.12476
##
## $se.fit
## [1] 0.2512623
##
## $df
## [1] 390
##
## $residual.scale
## [1] 4.905757
```

```r
data.lm <- lm(mpg ~ horsepower, data=Auto)
tval <- qt((1-0.95)/2, df=392-2)
Sxx <- sum((Auto$horsepower - mean(Auto$horsepower))^2)
MSres <- sum(data.lm$residuals^2)/392
lb=24.4671+tval*sqrt(MSres*(1/392+(24.4671-mean(Auto$horsepower))^2/Sxx))
```

```
ub=24.4671-tval*sqrt(MSres*(1/392+(24.4671-mean(Auto$horsepower))^2/Sxx))
c(lb,ub)
```

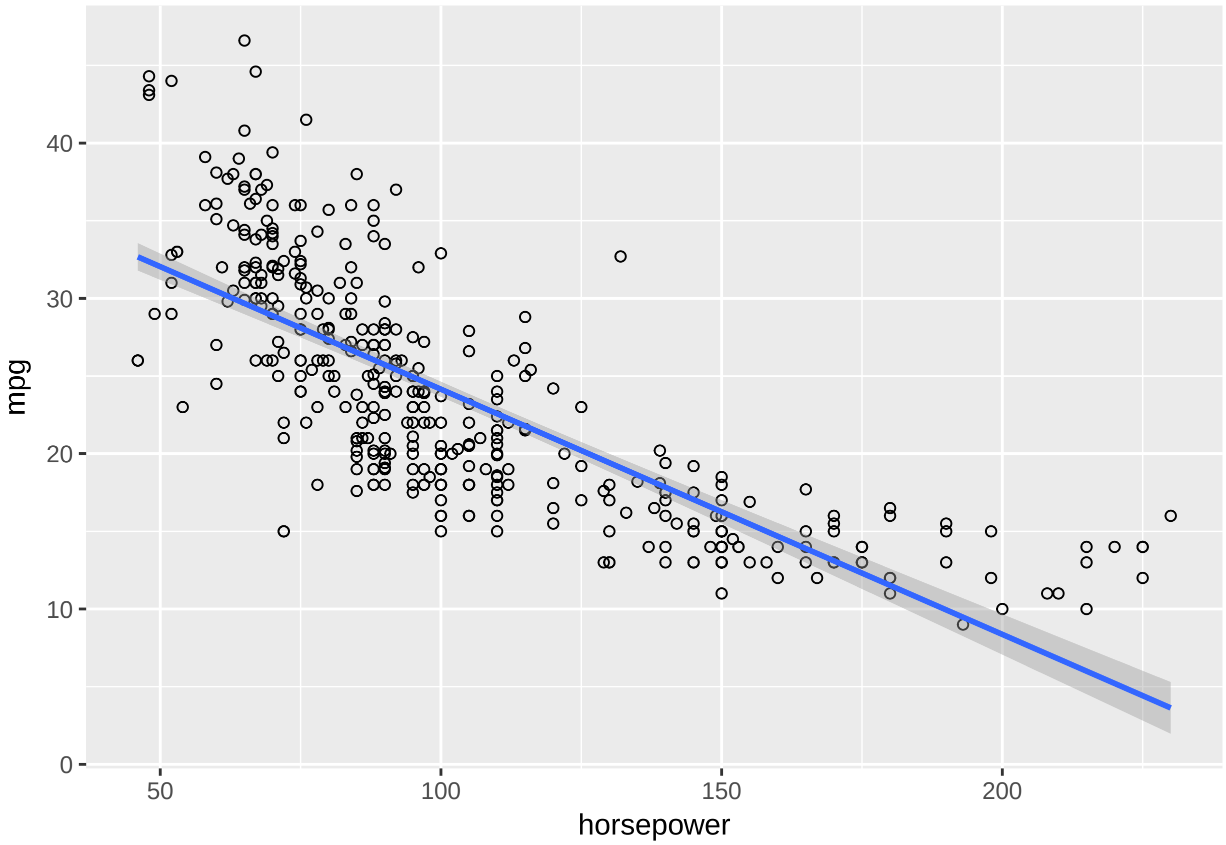
```
## [1] 23.34519 25.58901
```

```
The estimated mgp is 39.935861+(-0.157845)*98=24.4671
The 95% CI: [23.3452, 25.5890]
The 95% PI: [14.8094, 34.1248]

For a car with horsepower 98, the estimated mpg for that is about 24.5. If we randomly choose 100 cars
```

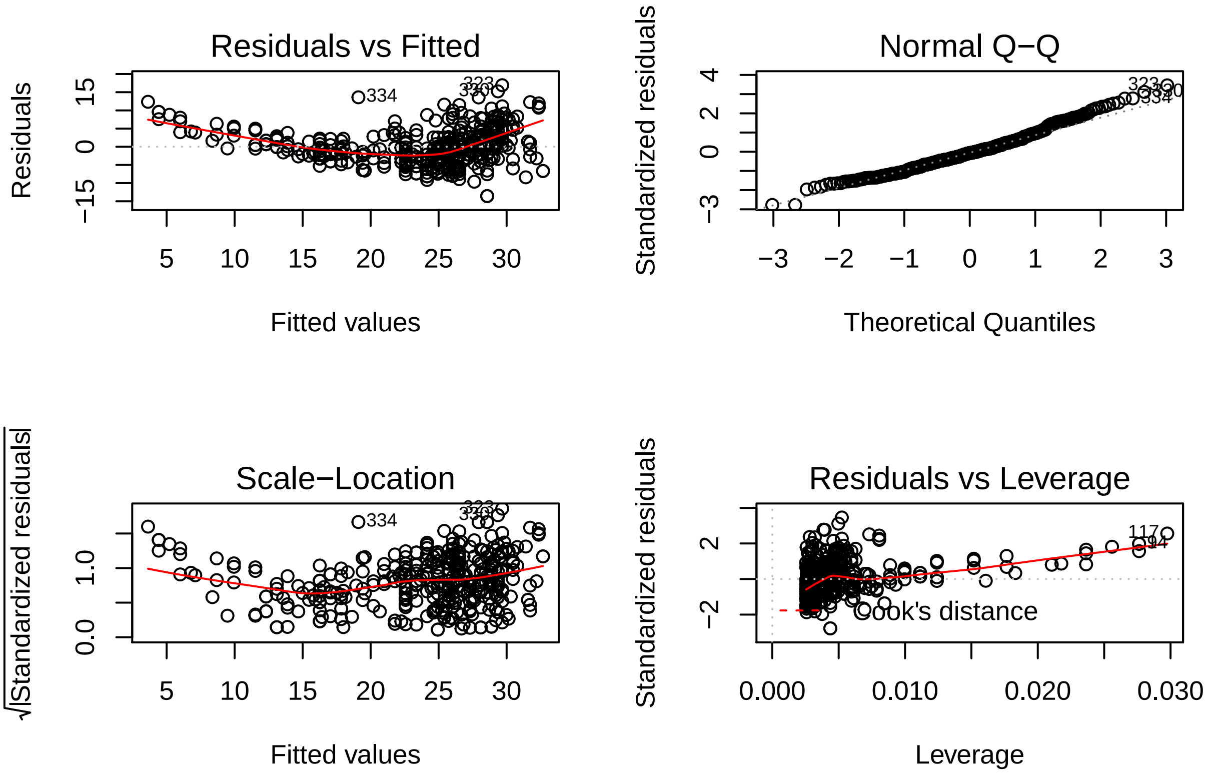
9. Plot the response and the predictor using `ggplot`. Add to the plot a line showing the least squares regression line.

```
library(ggplot2)
ggplot(Auto, aes(horsepower, mpg)) +
geom_point(shape=1) +
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(mpg ~ horsepower, data=Auto),ask=F)
```



(1) There is a slightly curved pattern for the residual plot, indicating that the variance is not contant and nonlinearity of the model. This could mean that other regressor variables are needed in the model.

(2) Normal Q-Q plots is nearly ideal since most of the points lie along the straight line. The head and tail of the points are a little bit above the line indicate that the left tail of the distribution is a bit heavier and the distribution is slightly positive skewed.

(3) Similarly as (1), the curved plot indicates the variance is not ideally contant.

(4) There are a bunch of influential points for this model, which indicates the model may be fitted better with transformations. For example, if we take log on the predictors and regressor, the influecial points issue may be improved significantly.

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 $X = x$.

12. Irreducible error:

(a) show that for any estimator $\hat{f}(x)$ that

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

Proof

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

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

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

$$= (f(x) - \hat{f(x)})^2 + E(\epsilon^2)$$

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

$$SinceE(\epsilon) = 0, thus$$

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

(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.
$proof$

$$Since(f(x) - \hat{f(x)})^2 \geq 0, thusE[(Y - \hat{f}(x))^2 \mid X = x] = \hat{f(x)})^2 + Var(\epsilon) \geq Var(\epsilon)$$