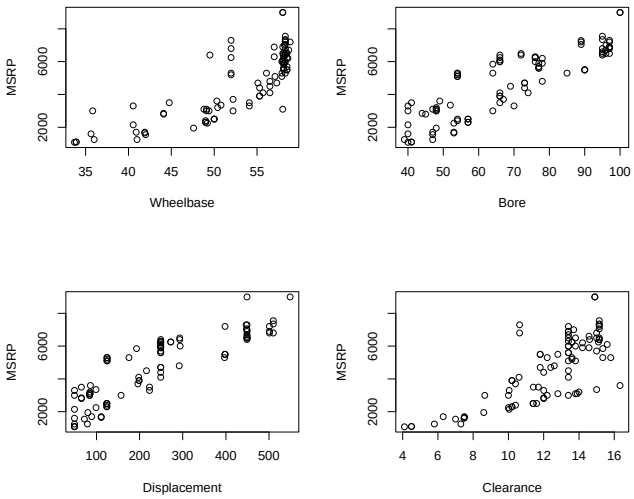
Marcus Crowder
Statistics 3154
R Homework 4

1. Let's take MSRP ($) as response variable and consider Wheelbase (in), Displacement (cu in), Bore (in) and Clearance (in) as potential predictors. Use scatterplots to see which variables can be appropriately used as predictors in simple linear regression.

To plot the MSRP as response variable towards Wheelbase, Displacement, Bore and Clearance. I used these functions

```
> motor <- read.table('~/Downloads/Motorcycles.txt', sep = '\t', header = TRUE)
> head(motor)
        Model MSRP Bore Displacement Clearance Engine.strokes Wheelbase
1   HondaCR85R 2999   48           84      12.2              2      49.1
2  HondaCR85RB 3099   48           84      13.9              2      49.1
3 HondaCRF100F 2249   53           99      10.0              4      49.2
4 HondaCRF150F 2999   64          157      10.0              4      52.2
5 HondaCRF230F 3499   66          223      11.7              4      54.1
6 HondaCRF250R 5899   78          249      14.2              4      58.2
> str(motor)
'data.frame':    93 obs. of  7 variables:
 $ Model         : Factor w/ 90 levels "GasGasPRO125",..: 4 5 6 7 8 9 10 11 12 13 ...
 $ MSRP          : int  2999 3099 2249 2999 3499 5899 6199 6699 6999 1249 ...
 $ Bore          : int  48 48 53 64 66 78 78 96 96 39 ...
 $ Displacement  : num  84 84 99 157 223 249 249 449 449 49 ...
 $ Clearance     : num  12.2 13.9 10 10 11.7 14.2 14.2 13.4 13.7 5.8 ...
 $ Engine.strokes: int  2 2 4 4 4 4 4 4 4 4 ...
 $ Wheelbase     : num  49.1 49.1 49.2 52.2 54.1 58.2 58.6 58.7 58.2 36 ...
> plot(motor$Bore, motor$MSRP, xlab="Bore", ylab="MSRP")
> par(mfrow = c(2,2))
> plot(motor$Bore, motor$MSRP, xlab="Bore", ylab="MSRP")
> plot(motor$Displacement, motor$MSRP, xlab="Displacement", ylab="MSRP")
> plot(motor$Clearance, motor$MSRP, xlab="Clearance", ylab="MSRP")
> plot(motor$Wheelbase, motor$MSRP, xlab="Clearance", ylab="Wheelbase")
> plot(motor$Wheelbase, motor$MSRP, xlab="Wheelbase", ylab="MSRP")
> plot(motor$Bore, motor$MSRP, xlab="Bore", ylab="MSRP")
> plot(motor$Displacement, motor$MSRP, xlab="Displacement", ylab="MSRP")
> plot(motor$Clearance, motor$MSRP, xlab="Clearance", ylab="MSRP")
```

```
> plot(motor$Wheelbase, motor$MSRP, xlab="Wheelbase", ylab="MSRP")
> plot(motor$Bore, motor$MSRP, xlab="Bore", ylab="MSRP")
> plot(motor$Displacement, motor$MSRP, xlab="Displacement", ylab="MSRP")
> plot(motor$Clearance, motor$MSRP, xlab="Clearance", ylab="MSRP")
```



The results were plotted on this graph. Wheelbase seems like it has a bend and could not be used as a predictor without a transformation. appropriately used as a predictor. Bore, Displacement and Clearance appear to be linearly associated with no distinguishable patterns.

2. Build a multiple regression model for MSRP using Displacement and Bore as predictors. Write down the fitted model. Report R^2 and adjusted R^2. Interpret the coefficients for Displacement and Bore.

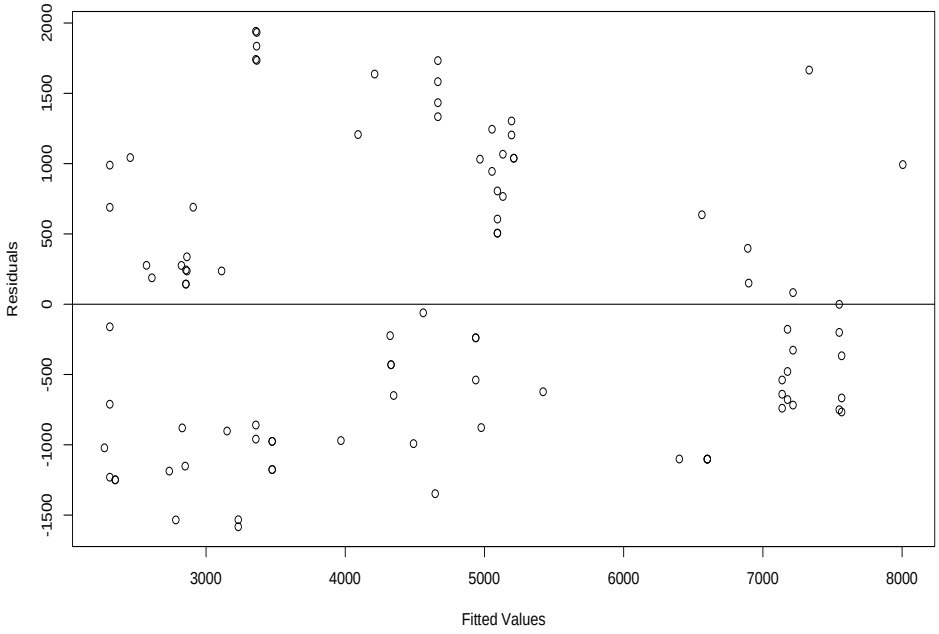
After Building a model we obtain the summary statistics of the Displacement and Bore Model.

```
> imod <- lm(MSRP ~ Displacement+Bore, data = motor)
> plot(imod$fitted.values, imod$residuals, xlab = 'Fitted Values', ylab = 'Residuals')
> abline(0,0)
> summary(imod)

Call:
lm(formula = MSRP ~ Displacement + Bore, data = motor)

Residuals:
    Min      1Q  Median      3Q     Max
-1582.7  -877.6  -178.2   805.6  1941.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   423.025   1036.588   0.408   0.6842
Displacement    6.722      3.324   2.022   0.0461 *
Bore           38.915     26.221   1.484   0.1413
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 998.6 on 90 degrees of freedom
Multiple R-squared:  0.7566,  Adjusted R-squared:  0.7512
F-statistic: 139.9 on 2 and 90 DF,  p-value: < 2.2e-16
```

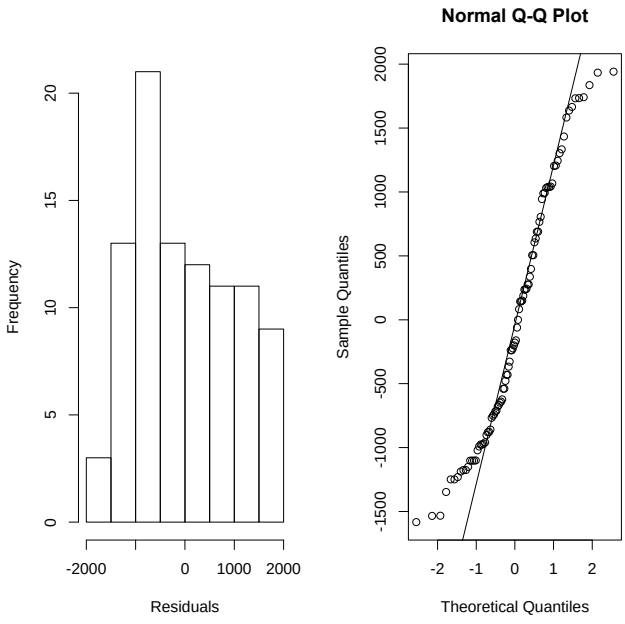
The R-squared for this model is .07566 which can be interpreted as: 75.66% of the variations in the multiple regression model can be explained by this model. On the Other hand the adjusted R squared is equal to .7512 which can be interpreted as: After taking account of all of the predictors and degrees of freedom 75.12% of the variations of this model can be explained by this model. The numbers are close together because of the high number of observations. The coefficients for displacement and bore cannot be interpreted by themselves. TO interpret the Displacement or Bore you would have to account for the other variable for instance: After allowing for Bore, each additional unit of Displacement is associated with a 6.722 increase in MSRP. Similarly for Bore if on Average Displacement stays the same then Bore will increase MSRP by 28.915.

3. Check the model you fitted in the previous question to see if it satisfies the assumptions as required in multiple regression.
   We assumed that the sample is random so we can say the independence assumption is met otherwise we are unable to check it.

The Residual plot of the multiple regression model shows no patterns or bends so we can say the linear assumption is met. We can also say the equal variance assumption is met because of the equal spread around 0.



To check the Normal Assumption we create a histogram and plot the residuals. The histogram does not appear to be unimodal or symmetrical. In addition the Normal Q-Q plot has slight bends on both sides. But it doesn't matter because there are over 90 observations so the rules of Central Limit Theorem can be applied. Because of the large sample size we can say the model tends to follow a normal distribution.



To produce these graphs I used these commands:

```
> plot(imod$fitted.values, imod$residuals, xlab = 'Fitted Values', ylab = 'Residuals')
> abline(0,0)
> par(mfrow = c(1,2))
> hist(imod$residuals, main = '', xlab = 'Residuals')
> qqnorm(imod$residuals)
> qqline(imod$residuals)
```

4. Conduct a test to see if the fitted multiple regression model is statistically useful. If useful, find the predictors that make significant contributions to the MSRP in the model. Explain.
   By looking at the F-stat of 139.9 and checking the F-stat table we can check that the F stat is significant. We can go even further by looking at the low P-Value associated with the F-stat to reject the null hypothesis and make the claim that the model is statistically useful. If we were to use a .05 level of significance we can say Displacement is useful given the other predictor Bore.

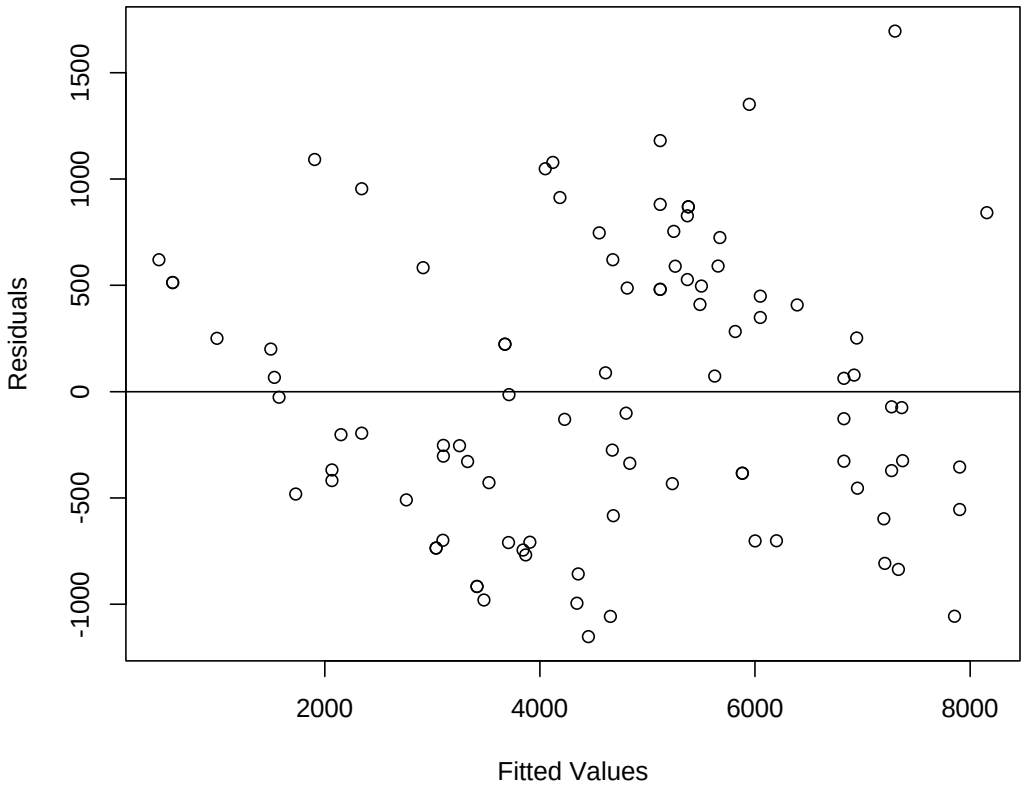
```
> summary(imod)

Call:
lm(formula = MSRP ~ Displacement + Bore, data = motor)

Residuals:
    Min      1Q  Median      3Q     Max
-1582.7  -877.6  -178.2   805.6  1941.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   423.025   1036.588   0.408   0.6842
Displacement    6.722      3.324   2.022   0.0461 *
Bore           38.915     26.221   1.484   0.1413
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 998.6 on 90 degrees of freedom
Multiple R-squared:  0.7566,  Adjusted R-squared:  0.7512
F-statistic: 139.9 on 2 and 90 DF,  p-value: < 2.2e-16
```

5. Suppose we are not satisfied with the R^2 given by the current model. Please propose a new multiple regression model in order to improve R^2. Compare the new model to the current one with respect to their R^2, coefficient estimates and hypothesis tests. Don't forget to check assumptions of the new model for its validity. (Hint: We have two potential predictors Wheelbase and Clearance in the pool. Think about how to use them to improve the model.)

Answer:

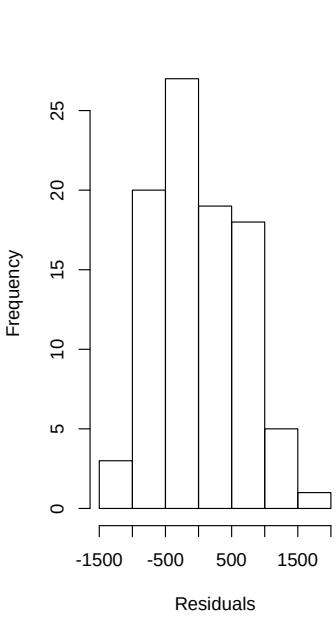
The Model I propose would be Displacement + Clearance. Again we check the assumptions. Because we assume these statistics are from a random sample we assume Independence.

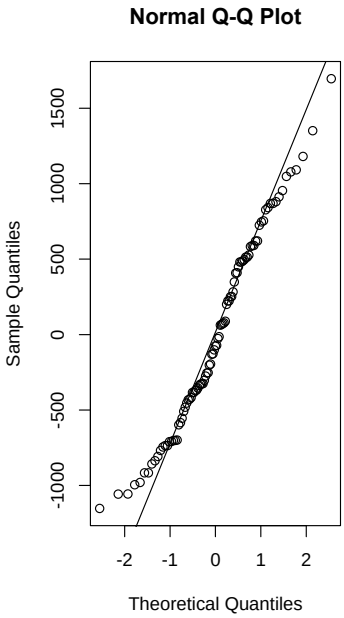
```
> imod <- lm(MSRP ~ Displacement+Clearance, data = motor)

> plot(imod$fitted.values, imod$residuals, xlab = 'Fitted Values', ylab = 'Residuals')

> abline(0,0)
```



After plotting the residual values we can see the plot appears to be linear since it does not seem to follow any patterns. It can also be said to have equal variance. We can again check for Normality but again because of The large number of samples we can say it doesn't matter the distribution since the Central Limit Theorem will be applied.

However the plot in the histogram appears to be normal and unimodal while the QQ plot also looks great with no skewness and the points tend to stick to the line.





Finally my final reason for selecting this model over the former is the higher R-squared value of 89.42% and adjusted r Squared of 89.18%. This model accounts for more variation than the former. Additionally the F statistics is higher with a similarly extremely low p value making it a useful model. Additionally both displacement and Clearance have low p-values suggesting that both predictors are significant given the other predictor. Because of all of this we can say this is a better model for rejecting H0.

```
> imod <- lm(MSRP ~ Displacement+Clearance, data = motor)
> summary(imod)

Call:
lm(formula = MSRP ~ Displacement + Clearance, data = motor)

Residuals:
    Min      1Q  Median      3Q     Max
-1152.7  -481.4   -75.4   513.3  1695.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1260.7470   316.0458  -3.989 0.000135 ***
Displacement     8.5412     0.5319  16.059  < 2e-16 ***
Clearance      317.3084    28.7170  11.049  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 658.4 on 90 degrees of freedom
Multiple R-squared:  0.8942,  Adjusted R-squared:  0.8918
F-statistic: 380.3 on 2 and 90 DF,  p-value: < 2.2e-16
```