Project1

Tomohiro Justin Wada

2023-03-26

Background and problem definition

We want to determine the best multiple variable model to best predict the 1/4 mile time using Miles per gallon(mpg), Displacement(disp), gross horsepower(hp), rear axle ratio(drat), and weight(wt).

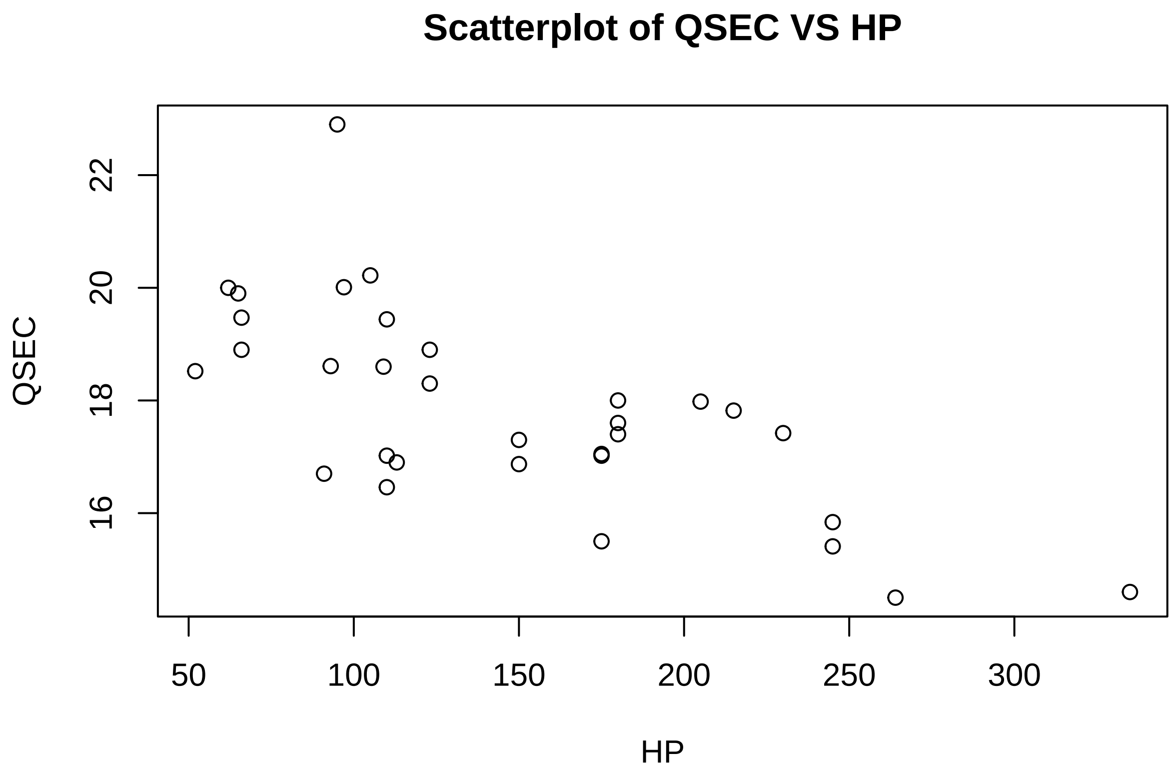
Setup

```
data(mtcars)
mydata <- mtcars[,c(7,1,3,4,5,6)]
```

No error or missing values within the data set.

Scatter Plots of QSEC VS HP

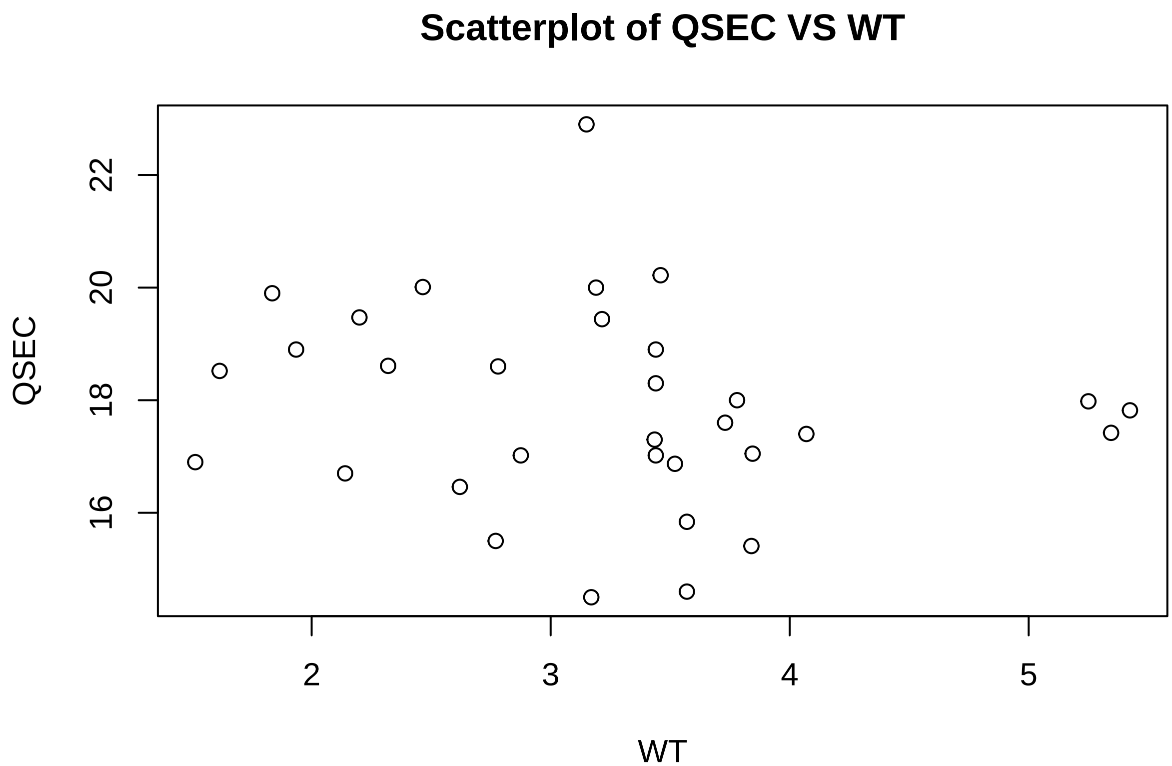
```
plot(x = mtcars$hp, y = mtcars$qsec, main = "Scatterplot of QSEC VS HP",
     xlab = "HP", ylab = "QSEC")
```



We see a moderate linear relationship between 1/4 mile time(qsec) and gross horsepower(hp) with moderate dispersion in the lower left region.

Scatter Plots of QSEC VS WT

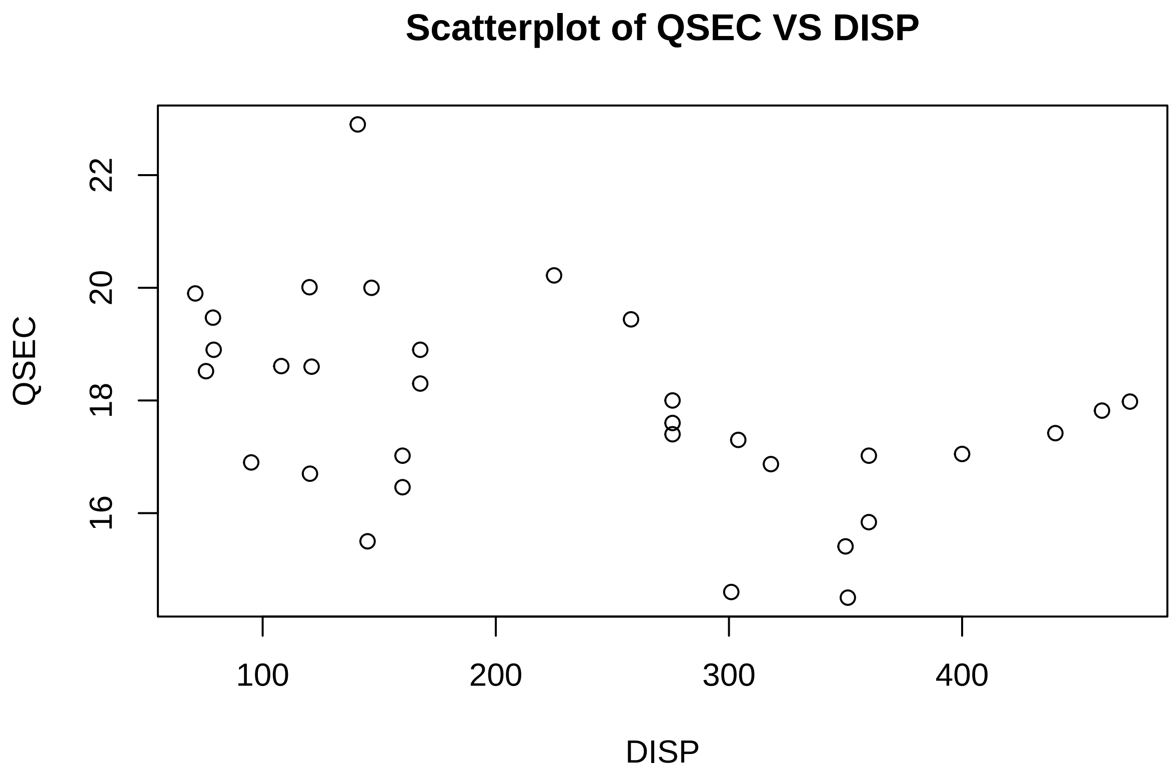
```
plot(x = mtcars$wt, y = mtcars$qsec, main = "Scatterplot of QSEC VS WT",
     xlab = "WT", ylab = "QSEC")
```



We don't see a clear linear relationship between 1/4 mile time(qsec) and Weight(wt) with moderate disperation on the center left region.

Scatter Plots of QSEC VS DISP

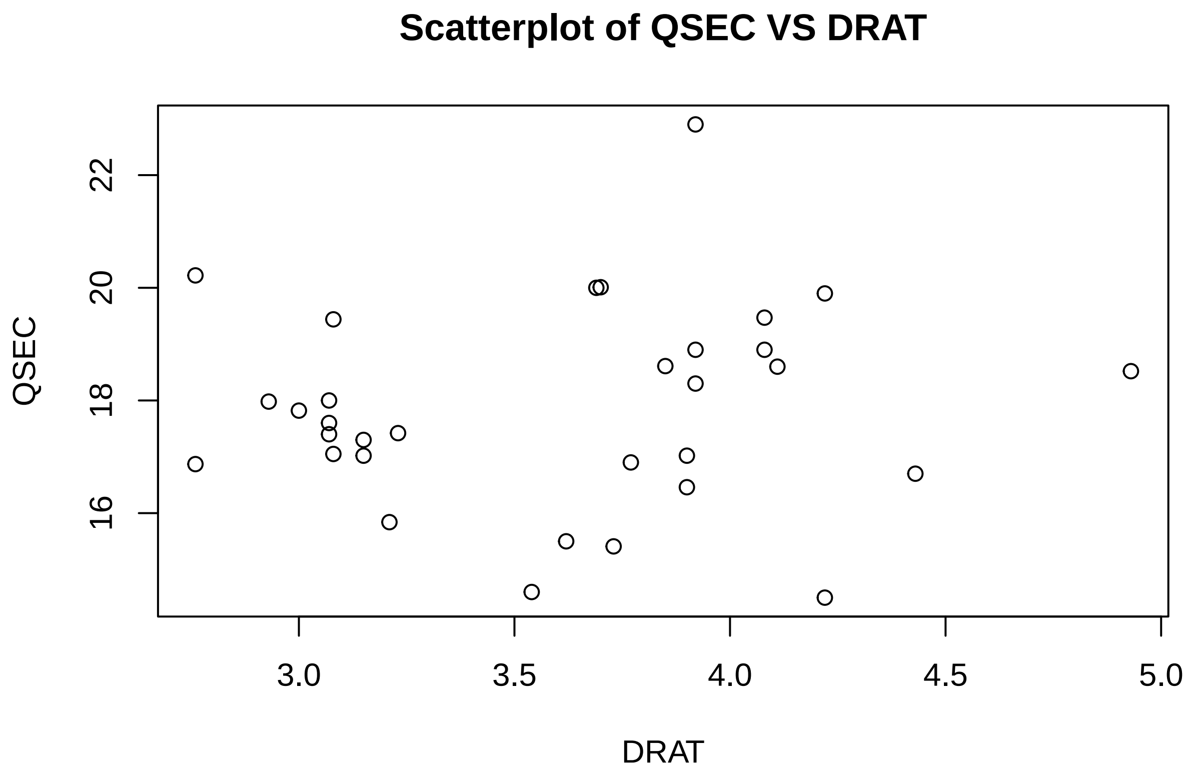
```
plot(x = mtcars$disp, y = mtcars$qsec, main = "Scatterplot of QSEC VS DISP",
     xlab = "DISP", ylab = "QSEC")
```



We see a moderate linear relationship between 1/4 mile time(qsec) and displacement with moderate dispersion in the left region.

Scatter Plots of QSEC VS DRAT

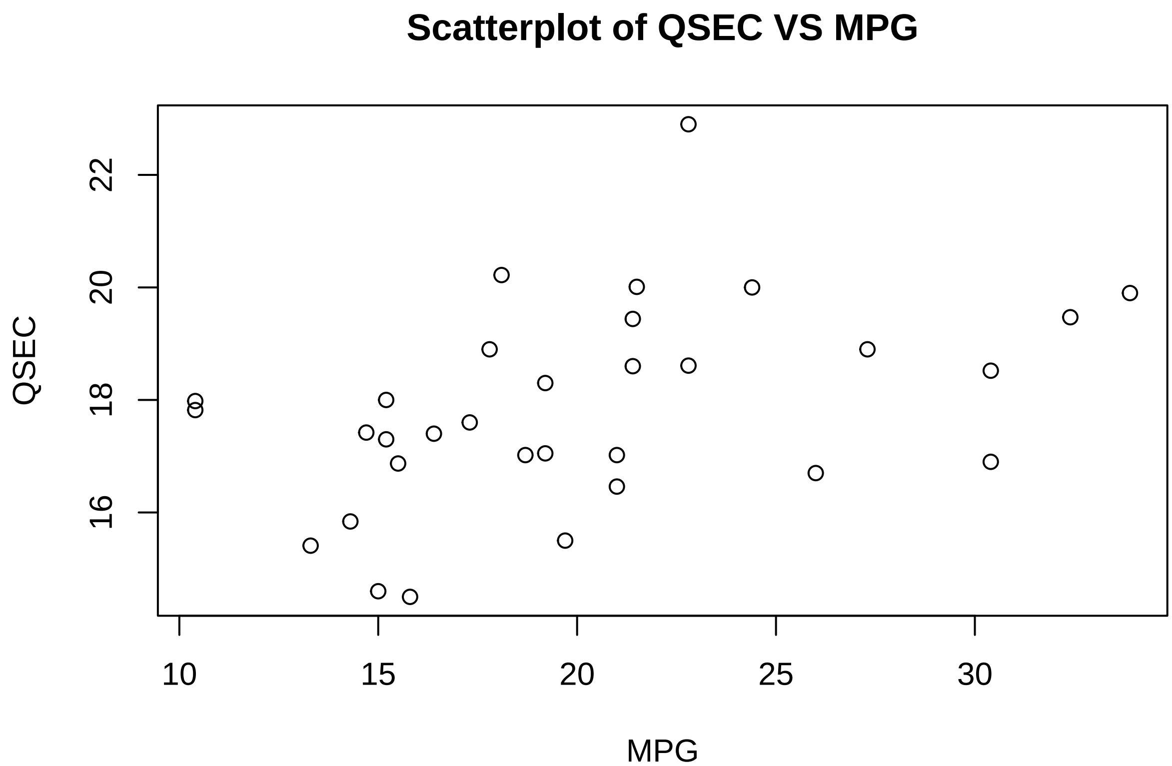
```
plot(x = mtcars$drat, y = mtcars$qsec, main = "Scatterplot of QSEC VS DRAT",
     xlab = "DRAT", ylab = "QSEC")
```



We see no linear relationship between 1/4 mile time(qsec) and rear axle ration (drat) with random dispersion.

Scatter Plots of QSEC VS MPG

```
plot(x = mtcars$mpg, y = mtcars$qsec, main = "Scatterplot of QSEC VS MPG",
     xlab = "MPG", ylab = "QSEC")
```



We can see a slight linear relationship between 1/4 mile time(qsec) and miles per gallon(mpg) with moderate dispersion in the center region.

Correlation Matrix

```
mydata.rcorr = rcorr(as.matrix(mydata))
mydata.rcorr
```

```
##       qsec   mpg  disp    hp  drat    wt
## qsec  1.00  0.42 -0.43 -0.71  0.09 -0.17
## mpg   0.42  1.00 -0.85 -0.78  0.68 -0.87
## disp -0.43 -0.85  1.00  0.79 -0.71  0.89
## hp   -0.71 -0.78  0.79  1.00 -0.45  0.66
## drat  0.09  0.68 -0.71 -0.45  1.00 -0.71
## wt   -0.17 -0.87  0.89  0.66 -0.71  1.00
##
## n= 32
##
##
## P
##      qsec   mpg    disp   hp     drat   wt
## qsec        0.0171 0.0131 0.0000 0.6196 0.3389
## mpg  0.0171        0.0000 0.0000 0.0000 0.0000
## disp 0.0131 0.0000        0.0000 0.0000 0.0000
```

```
## hp   0.0000 0.0000 0.0000        0.0100 0.0000
## drat 0.6196 0.0000 0.0000 0.0100        0.0000
## wt   0.3389 0.0000 0.0000 0.0000 0.0000
```

We can from the correlation matrix, that only hp is signicant. But we will see down on that wt(weight) and disp(displacement) is significant to the overall model.

Stepwise AIC Variable Selection

```r
model <- lm(mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp + mtcars$drat + mtcars$mpg
            , data = mtcars)
ols_step_forward_aic(model, details = TRUE)
```

```
## Forward Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1 . mtcars$hp
## 2 . mtcars$wt
## 3 . mtcars$disp
## 4 . mtcars$drat
## 5 . mtcars$mpg
##
##  Step 0: AIC = 130.9485
##  mtcars$qsec ~ 1
##
## -------------------------------------------------------------------------
## Variable       DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## mtcars$hp       1    110.667    49.651    49.338    0.502         0.485
## mtcars$disp     1    126.281    18.619    80.369    0.188         0.161
## mtcars$mpg      1    126.781    17.352    81.636    0.175         0.148
## mtcars$wt       1    131.956     3.022    95.966    0.031        -0.002
## mtcars$drat     1    132.681     0.823    98.165    0.008        -0.025
## -------------------------------------------------------------------------
##
##
## - mtcars$hp
##
##
##  Step 1 : AIC = 110.6665
##  mtcars$qsec ~ mtcars$hp
##
## -------------------------------------------------------------------------
## Variable       DF      AIC      Sum Sq      RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## mtcars$wt       1    101.168    14.893    34.445    0.652         0.628
## mtcars$drat     1    108.246     6.365    42.972    0.566         0.536
## mtcars$mpg      1    109.767     4.274    45.064    0.545         0.513
## mtcars$disp     1    109.799     4.229    45.109    0.544         0.513
```

```
## -------------------------------------------------------------------------
##
## - mtcars$wt
##
##
##  Step 2 : AIC = 101.1682
##  mtcars$qsec ~ mtcars$hp + mtcars$wt
##
## -------------------------------------------------------------------------
## Variable       DF      AIC      Sum Sq     RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## mtcars$disp     1    100.404     2.851    31.594    0.681        0.647
## mtcars$mpg      1    101.658     1.587    32.858    0.668        0.633
## mtcars$drat     1    103.142     0.028    34.417    0.652        0.615
## -------------------------------------------------------------------------
##
## - mtcars$disp
##
##
##  Step 3 : AIC = 100.4036
##  mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp
##
## -------------------------------------------------------------------------
## Variable       DF      AIC      Sum Sq     RSS      R-Sq     Adj. R-Sq
## -------------------------------------------------------------------------
## mtcars$mpg      1    100.830     1.516    30.078    0.696        0.651
## mtcars$drat     1    101.808     0.582    31.012    0.687        0.640
## -------------------------------------------------------------------------
##
##
## No more variables to be added.
##
## Variables Entered:
##
## - mtcars$hp
## - mtcars$wt
## - mtcars$disp
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ----------------------------------------------------------------
## R                       0.825       RMSE                 1.062
## R-Squared               0.681       Coef. Var            5.951
## Adj. R-Squared          0.647       MSE                  1.128
## Pred R-Squared          0.587       MAE                  0.687
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
```

```
## -------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F         Sig.
## -------------------------------------------------------------------
## Regression     67.394         3         22.465    19.909    0.0000
## Residual       31.594        28          1.128
## Total          98.988        31
## -------------------------------------------------------------------
##
##                                   Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    17.965         0.850                 21.144    0.000    16.225    19.706
##   mtcars$hp    -0.023         0.005       -0.881    -4.986    0.000    -0.032    -0.014
##   mtcars$wt     1.485         0.429        0.813     3.461    0.002     0.606     2.364
## mtcars$disp    -0.007         0.004       -0.459    -1.590    0.123    -0.015     0.002
## ----------------------------------------------------------------------------------------
```

```
##
##                         Selection Summary
## ------------------------------------------------------------------
## Variable          AIC      Sum Sq     RSS       R-Sq      Adj. R-Sq
## ------------------------------------------------------------------
## mtcars$hp      110.667    49.651    49.338    0.50158      0.48497
## mtcars$wt      101.168    64.543    34.445    0.65203      0.62803
## mtcars$disp    100.404    67.394    31.594    0.68083      0.64663
## ------------------------------------------------------------------
```

From the five variables, the selection process decides not to include mpg(miles per gallon) and drat(rear axle ratio) into the final model as it was not significant. But if we see the significance column we see that disp is not significant either.

Interaction Terms

```
InteractionModel <- lm(mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp  +
                         mtcars$hp*mtcars$wt + mtcars$hp*mtcars$disp + mtcars$wt*mtcars$disp,
                          data = mtcars)
summary(InteractionModel)
```

```
##
## Call:
## lm(formula = mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp +
##     mtcars$hp * mtcars$wt + mtcars$hp * mtcars$disp + mtcars$wt *
##     mtcars$disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6266 -0.5214 -0.1063  0.4593  2.9960
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             1.762e+01  2.308e+00   7.636 5.44e-08 ***
## mtcars$hp              -3.906e-02  3.242e-02  -1.205   0.2396    
## mtcars$wt               2.823e+00  1.279e+00   2.208   0.0366 *  
## mtcars$disp            -1.519e-02  1.815e-02  -0.837   0.4107    
## mtcars$hp:mtcars$wt    -4.072e-03  1.166e-02  -0.349   0.7299    
## mtcars$hp:mtcars$disp  1.082e-04  6.536e-05   1.655   0.1104    
## mtcars$wt:mtcars$disp -2.237e-03  3.682e-03  -0.608   0.5489    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 25 degrees of freedom
## Multiple R-squared:  0.7202, Adjusted R-squared:  0.6531 
## F-statistic: 10.73 on 6 and 25 DF,  p-value: 6.546e-06
```

From the summary, we see that the adjused R-squred is 0.6531 including disp. Let's see without disp.

Interation Terms test 2

```r
InteractionModel <- lm(mtcars$qsec ~ mtcars$hp + mtcars$wt +
                         mtcars$hp*mtcars$wt,
                          data = mtcars)
summary(InteractionModel)
```

```
##
## Call:
## lm(formula = mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$hp *
##     mtcars$wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.8264 -0.4046 -0.1506  0.3512  3.7076 
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         18.8766043  1.8574084  10.163 6.73e-11 ***
## mtcars$hp           -0.0276678  0.0127248  -2.174   0.0383 *  
## mtcars$wt            0.9239377  0.6541649   1.412   0.1689    
## mtcars$hp:mtcars$wt  0.0001129  0.0038226   0.030   0.9766    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.109 on 28 degrees of freedom
## Multiple R-squared:  0.652,  Adjusted R-squared:  0.6148 
## F-statistic: 17.49 on 3 and 28 DF,  p-value: 1.358e-06
```

From the summary, we see that the new adjusted R-squared is 0.6148, lower than previously. This shows that disp(displacement) is important to the overall model. # Quadratic Terms

```r
sqHP <- mtcars$hp^2
sqWT <- mtcars$wt^2
sqDISP <- mtcars$disp^2
QuadraticModel <- lm(mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp + sqHP + sqWT + sqDISP,
                          data = mtcars)
summary(QuadraticModel)
```

```
##
## Call:
## lm(formula = mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp +
##     sqHP + sqWT + sqDISP, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6191 -0.3553 -0.1129  0.4693  3.1099
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.764e+01  2.326e+00   7.583 6.15e-08 ***
## mtcars$hp   -5.394e-02  1.827e-02  -2.952  0.00678 **
## mtcars$wt    3.378e+00  2.081e+00   1.624  0.11698
## mtcars$disp -1.275e-02  1.754e-02  -0.727  0.47399
## sqHP         7.672e-05  4.271e-05   1.796  0.08458 .
## sqWT        -2.617e-01  2.976e-01  -0.880  0.38749
## sqDISP       1.684e-05  3.158e-05   0.533  0.59850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 25 degrees of freedom
## Multiple R-squared:  0.7281, Adjusted R-squared:  0.6628
## F-statistic: 11.16 on 6 and 25 DF,  p-value: 4.672e-06
```

From the summary, we acheive a adjusted R-squared value of 0.6628. Lets see without disp in the quadratic test. # Quadratic Test2

```r
sqHP <- mtcars$hp^2
sqWT <- mtcars$wt^2
QuadraticModel <- lm(mtcars$qsec ~ mtcars$hp + mtcars$wt + sqHP + sqWT,
                          data = mtcars)
summary(QuadraticModel)
```

```
##
## Call:
## lm(formula = mtcars$qsec ~ mtcars$hp + mtcars$wt + sqHP + sqWT,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6669 -0.4432 -0.0415  0.4366  3.3142
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.885e+01  1.680e+00  11.222 1.13e-11 ***
## mtcars$hp   -6.363e-02  1.523e-02  -4.177 0.000277 ***
## mtcars$wt    2.352e+00  1.083e+00   2.172 0.038767 *
## sqHP         9.358e-05  3.851e-05   2.430 0.022041 *
## sqWT        -1.486e-01  1.433e-01  -1.037 0.308924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 27 degrees of freedom
## Multiple R-squared:  0.7171, Adjusted R-squared:  0.6752
## F-statistic: 17.11 on 4 and 27 DF,  p-value: 4.222e-07
```

We can see that for the quadratic test, it does increase the adjusted R-squared value but we will see that for the best adjusted R-squared value includes disp.

Multicollinearity

```
vif(lm(mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp,
                          data = mtcars))
```

```
##   mtcars$hp   mtcars$wt mtcars$disp
##    2.736633    4.844618    7.324517
```

Based on the Variance Inflation factor, we see each predictor variable is less than 10. There is no multicollinearity.

Final Model

```
FinalModel <- lm(mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp + sqHP,
                         data = mtcars)
summary(FinalModel)
```

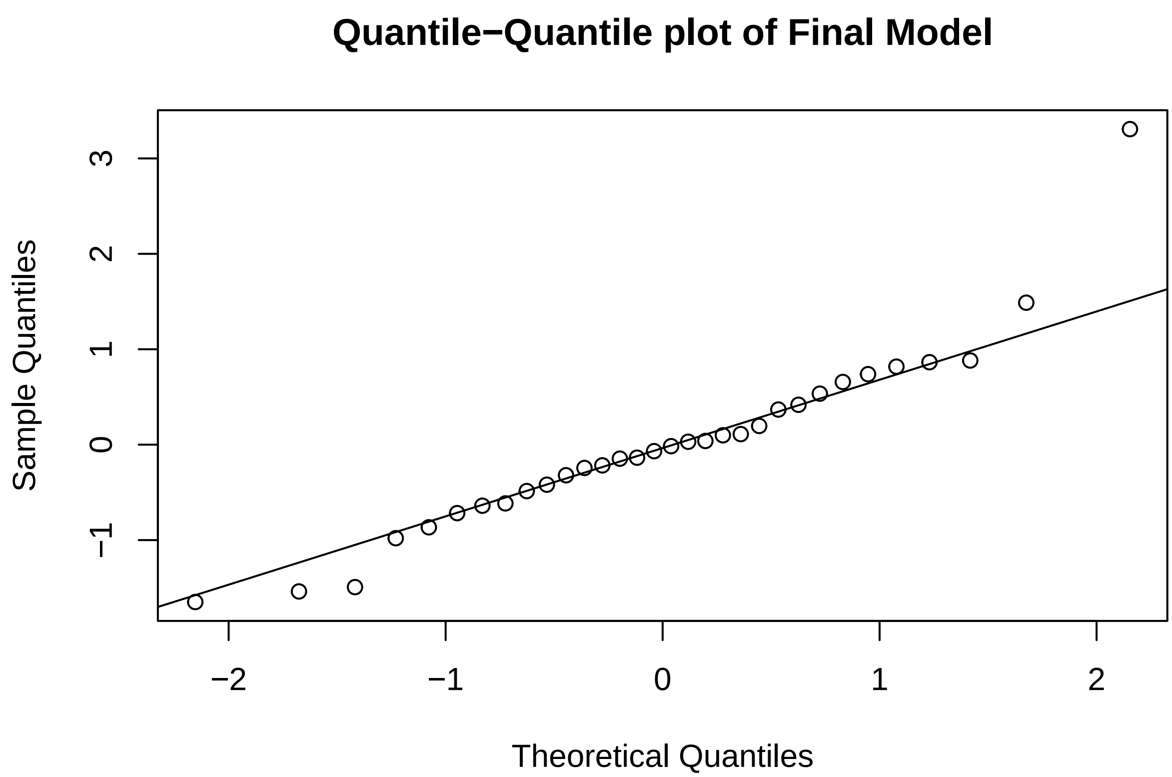
```
##
## Call:
## lm(formula = mtcars$qsec ~ mtcars$hp + mtcars$wt + mtcars$disp +
##     sqHP, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6490 -0.5184 -0.0416  0.4473  3.3070
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.954e+01  1.171e+00  16.680 9.59e-16 ***
## mtcars$hp   -5.245e-02  1.638e-02  -3.201 0.003489 **
## mtcars$wt    1.584e+00  4.146e-01   3.820 0.000711 ***
## mtcars$disp -4.380e-03  4.168e-03  -1.051 0.302627
## sqHP         7.356e-05  3.936e-05   1.869 0.072476 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 27 degrees of freedom
## Multiple R-squared:  0.7174, Adjusted R-squared:  0.6755
## F-statistic: 17.14 on 4 and 27 DF,  p-value: 4.165e-07
```

We achieve an final ajusted R-squared value of 0.6755.

$y = 1.954e + 01 - 5.245e - 02 * x1 + 1.584e + 00 * x2 - 4.380e - 03 * x3 + 7.356e - 05 * x4$

```r
resids <- FinalModel$residuals
qqnorm(resids,main="Quantile-Quantile plot of Final Model")
qqline(resids)
```



Based on the normal qqplot, we can see that the majority of the data points fall on the line with only 4 outliers on both ends of the graph showing the model is accurate predictor.