Regression Analysis on MTCARS

Executive Summary -

In this study, we develop a regression model for the MTCARS dataset to answer two questions: 1.Is an automatic or manual transmission better for MPG? 2.How do you quantify the MPG difference between automatic and manual transmissions?
The best fit model we come up with is this: mpg = 33.70832 - 3.03134x1(Cyl=6) - 2.16368(Cyl=8) - 0.03211xHp - 2.49683xWt + 1.80921x1(am="Manual"). Based on this model, we conclude that manual transmission is better than automatic transmission for MPG. There is 1.80921 MPG increase when a car is manual transmission with holding all of the other variables constant.

Exploratory Data Analysis on MTCARS -

```
library(datasets)
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Regression Modeling -

The modeling approach we use is **stepwise, backward elimination**, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model the most by being deleted, and repeating this process until no further improvement is possible.

```
## create an initial model with all variables
fit <- lm(mpg~as.factor(cyl)+disp+hp+drat+wt+qsec+as.factor(vs)+as.factor(am)+
          as.factor(gear)+as.factor(carb), data=mtcars)
```

```
## use the stepwise approach to come up with the best fit model
mymodel <- step(fit, direction="backward")
```

```
## Start:  AIC=76.4
## mpg ~ as.factor(cyl) + disp + hp + drat + wt + qsec + as.factor(vs) +
##     as.factor(am) + as.factor(gear) + as.factor(carb)
##
##                    Df Sum of Sq    RSS    AIC
## - as.factor(carb)  5   13.5989 134.00 69.828
## - as.factor(gear)  2    3.9729 124.38 73.442
## - as.factor(am)    1    1.1420 121.55 74.705
## - qsec             1    1.2413 121.64 74.732
## - drat             1    1.8208 122.22 74.884
## - as.factor(cyl)   2   10.9314 131.33 75.184
## - as.factor(vs)    1    3.6299 124.03 75.354
## <none>                         120.40 76.403
## - disp             1    9.9672 130.37 76.948
## - wt               1   25.5541 145.96 80.562
## - hp               1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ as.factor(cyl) + disp + hp + drat + wt + qsec + as.factor(vs) +
##     as.factor(am) + as.factor(gear)
##
##                   Df Sum of Sq    RSS    AIC
## - as.factor(gear)  2    5.0215 139.02 67.005
## - disp             1    0.9934 135.00 68.064
## - drat             1    1.1854 135.19 68.110
## - as.factor(vs)    1    3.6763 137.68 68.694
## - as.factor(cyl)   2   12.5642 146.57 68.696
## - qsec             1    5.2634 139.26 69.061
## <none>                         134.00 69.828
## - as.factor(am)    1   11.9255 145.93 70.556
## - wt               1   19.7963 153.80 72.237
## - hp               1   22.7935 156.79 72.855
##
## Step:  AIC=67
## mpg ~ as.factor(cyl) + disp + hp + drat + wt + qsec + as.factor(vs) +
##     as.factor(am)
##
##                  Df Sum of Sq    RSS    AIC
## - drat            1    0.9672 139.99 65.227
## - as.factor(cyl)  2   10.4247 149.45 65.319
## - disp            1    1.5483 140.57 65.359
## - as.factor(vs)   1    2.1829 141.21 65.503
## - qsec            1    3.6324 142.66 65.830
## <none>                        139.02 67.005
## - as.factor(am)   1   16.5665 155.59 68.608
## - hp              1   18.1768 157.20 68.937
## - wt              1   31.1896 170.21 71.482
##
## Step:  AIC=65.23
## mpg ~ as.factor(cyl) + disp + hp + wt + qsec + as.factor(vs) +
##     as.factor(am)
##
##                  Df Sum of Sq    RSS    AIC
## - disp            1    1.2474 141.24 63.511
## - as.factor(vs)   1    2.3403 142.33 63.757
## - as.factor(cyl)  2   12.3267 152.32 63.927
## - qsec            1    3.1000 143.09 63.928
## <none>                        139.99 65.227
## - hp              1   17.7382 157.73 67.044
## - as.factor(am)   1   19.4660 159.46 67.393
## - wt              1   30.7151 170.71 69.574
##
## Step:  AIC=63.51
## mpg ~ as.factor(cyl) + hp + wt + qsec + as.factor(vs) + as.factor(am)
##
##                   Df Sum of Sq    RSS    AIC
## - qsec             1     2.442 143.68 62.059
## - as.factor(vs)    1     2.744 143.98 62.126
## - as.factor(cyl)   2    18.580 159.82 63.466
## <none>                         141.24 63.511
## - hp               1    18.184 159.42 65.386
## - as.factor(am)    1    18.885 160.12 65.527
## - wt               1    39.645 180.88 69.428
##
## Step:  AIC=62.06
## mpg ~ as.factor(cyl) + hp + wt + as.factor(vs) + as.factor(am)
##
##                   Df Sum of Sq    RSS    AIC
## - as.factor(vs)    1     7.346 151.03 61.655
## <none>                         143.68 62.059
## - as.factor(cyl)   2    25.284 168.96 63.246
## - as.factor(am)    1    16.443 160.12 63.527
## - hp               1    36.344 180.02 67.275
## - wt               1    41.088 184.77 68.108
##
## Step:  AIC=61.65
## mpg ~ as.factor(cyl) + hp + wt + as.factor(am)
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         151.03 61.655
## - as.factor(am)    1     9.752 160.78 61.657
## - as.factor(cyl)   2    29.265 180.29 63.323
## - hp               1    31.943 182.97 65.794
## - wt               1    46.173 197.20 68.191
```

```r
summary(mymodel)
```

```
##
## Call:
## lm(formula = mpg ~ as.factor(cyl) + hp + wt + as.factor(am),
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     33.70832    2.60489  12.940 7.73e-13 ***
## as.factor(cyl)6 -3.03134    1.40728  -2.154  0.04068 *
## as.factor(cyl)8 -2.16368    2.28425  -0.947  0.35225
## hp              -0.03211    0.01369  -2.345  0.02693 *
## wt              -2.49683    0.88559  -2.819  0.00908 **
## as.factor(am)1   1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Confidence Limits on the Estimated Coefficients -

```r
confint(mymodel)
```

```
##                        2.5 %       97.5 %
## (Intercept)     28.35390366 39.062744138
## as.factor(cyl)6 -5.92405718 -0.138631806
## as.factor(cyl)8 -6.85902199  2.531671342
## hp              -0.06025492 -0.003963941
## wt              -4.31718120 -0.676477640
## as.factor(am)1  -1.06093363  4.679356394
```

Conclusion -

The best fit regression model:
mpg = b0 - b1x1(Cyl=6) + b2x1(Cyl=8) + b3xHp + b4xWt + b5x1(am="Manual") + ei
where b0 = 33.70832, b1 = -3.03134, b2 = -2.16368, b3 = -0.03211, b4 = -2.49683 and b5 = 1.80921

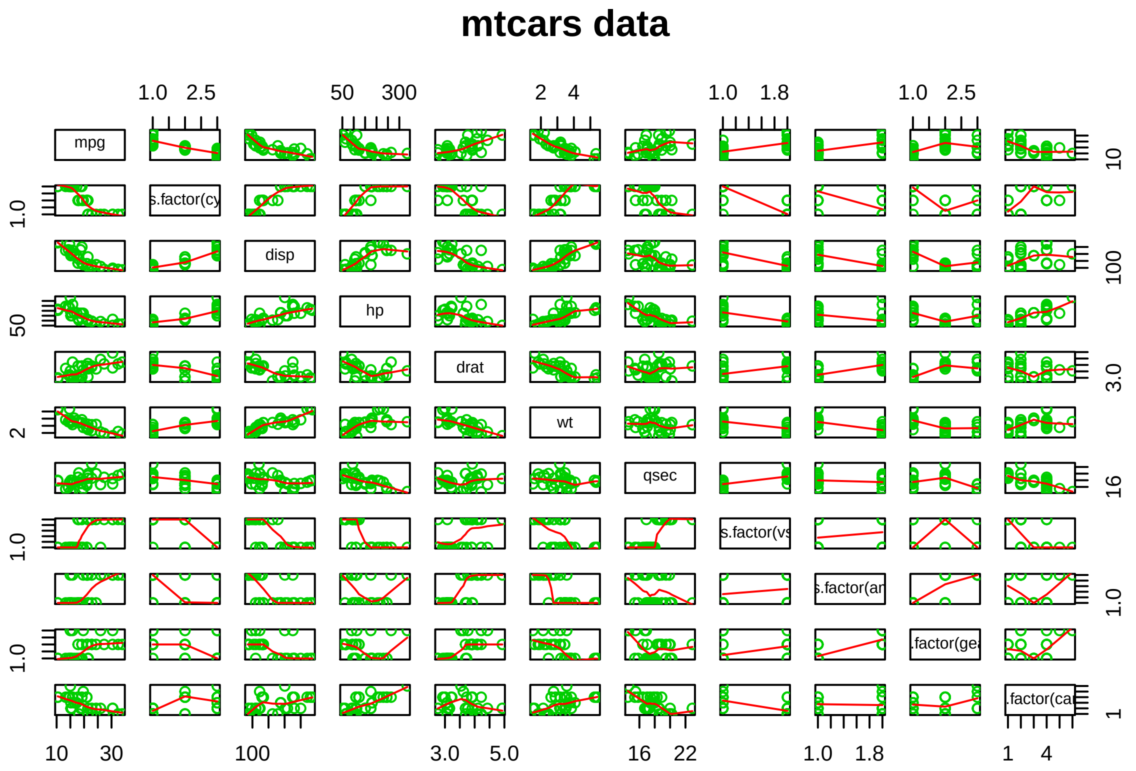
Coefficients interpretation:
b0 - mpg at 0 horse power, 0 weight and is automatic for 4 cylinders
b0+b1 - mpg at 0 horse power, 0 weight and is automatic for 6 cylinders
b0+b2 - mpg at 0 horse power, 0 weight and is automatic for 8 cylinders
b3 - change in mpg for each horse power at 0 weight, is automatic for 4 cylinders
b4 - change in mpg for each 1000 lbs of weight at 0 horse power and is automatic for 4 cylinders
b0+b5 - mpg at 0 horse power, 0 weight and is manual for 4 cylinders
ei - everything we don't measure

Questions:
1.Is an automatic or manual transmission better for MPG?
Answer: **Manual transmission is better for MPG based on the coefficient b5 which is positive.**

2.How do you quantify the MPG difference between automatic and manual transmissions?
Answer: **There is a 1.80921 increase of MPG (more efficient) for manual tranmission than automatic transmission, holding all of the other variables, such as weight fixed. The 95% confidence interval of b5 coefficient is [-1.06093363, 4.679356394] as shown above**

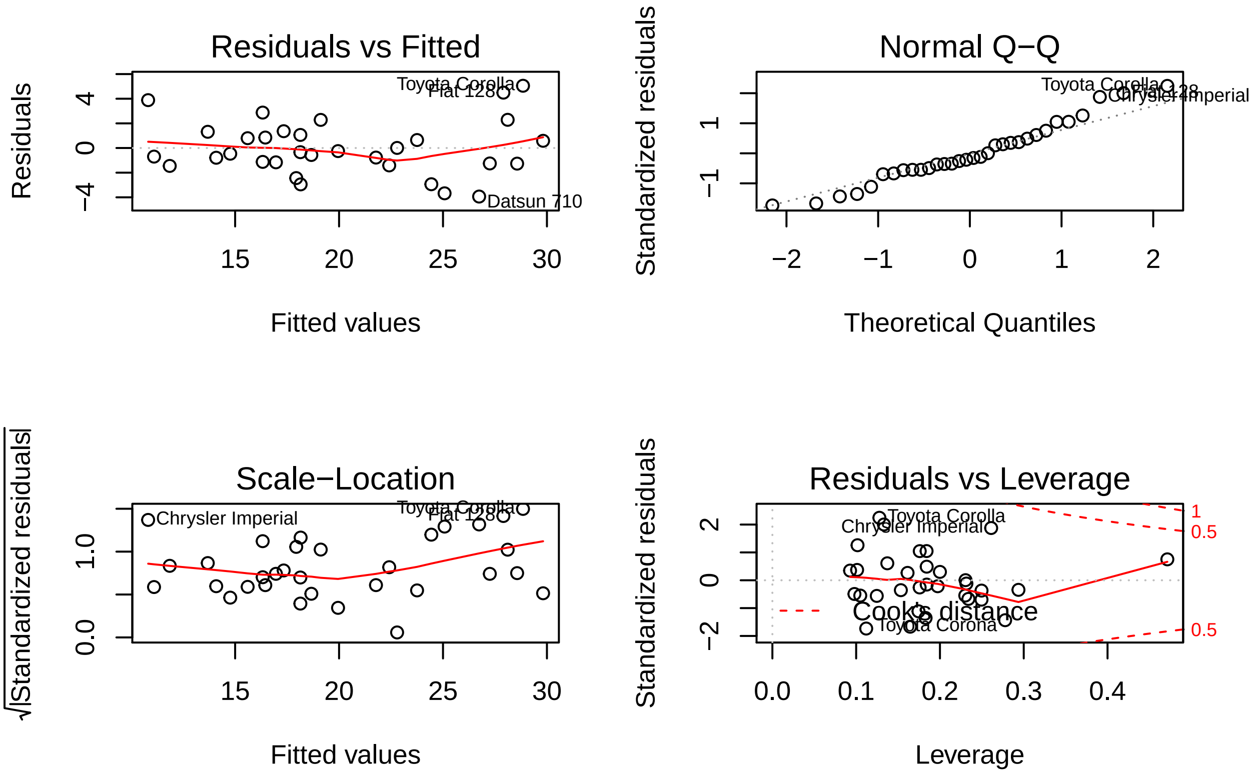
Appendix A - Scatterplot Matrices for Exploratory Data Analysis

```r
pairs(~mpg+as.factor(cyl)+disp+hp+drat+wt+qsec+as.factor(vs)+as.factor(am)+
        as.factor(gear)+as.factor(carb), panel = panel.smooth,
        main = "mtcars data", data=mtcars, col=3)
```

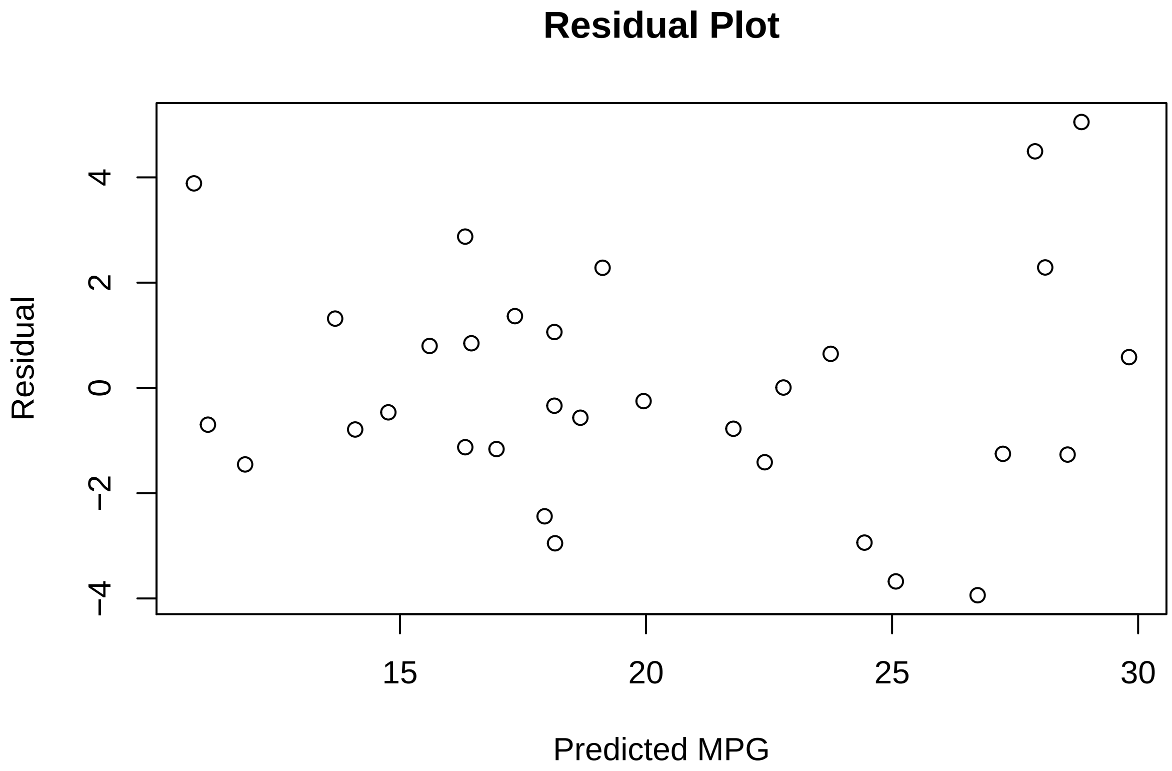


Appendix B - Model Dianostics and Residual Plot

```
par(mfrow=c(2,2))
plot(mymodel)
```



```
par(mfrow=c(1,1))
plot(predict(mymodel),resid(mymodel), main="Residual Plot", xlab="Predicted MPG", ylab="Residual")
```


Residual Plot
Residual
4
2
0
−2
−4
15
20
25
30
Predicted MPG