Final Project, Regression Model

Executive Summary:

This project is set to explore the relationship between a set of variables and miles per gallon. We need to answer:

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

After analyzing the `mtcars` data, we can conclude that manual transmission produces more mpg compared to auto transmission. And according to our best fitted model, manual transmission achieve 2.936 more mpg than auto transmission

Analysis:

Loading libraries and datasets

```
library(datasets)
library(ggplot2)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.1
```

```
##
## 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
```

```
data(mtcars)
```

Transformation:

```
mtcars<- mutate(mtcars, am=factor(mtcars$am,labels=c("Auto","Manual")), vs=factor(vs),gear=factor(gear)
```

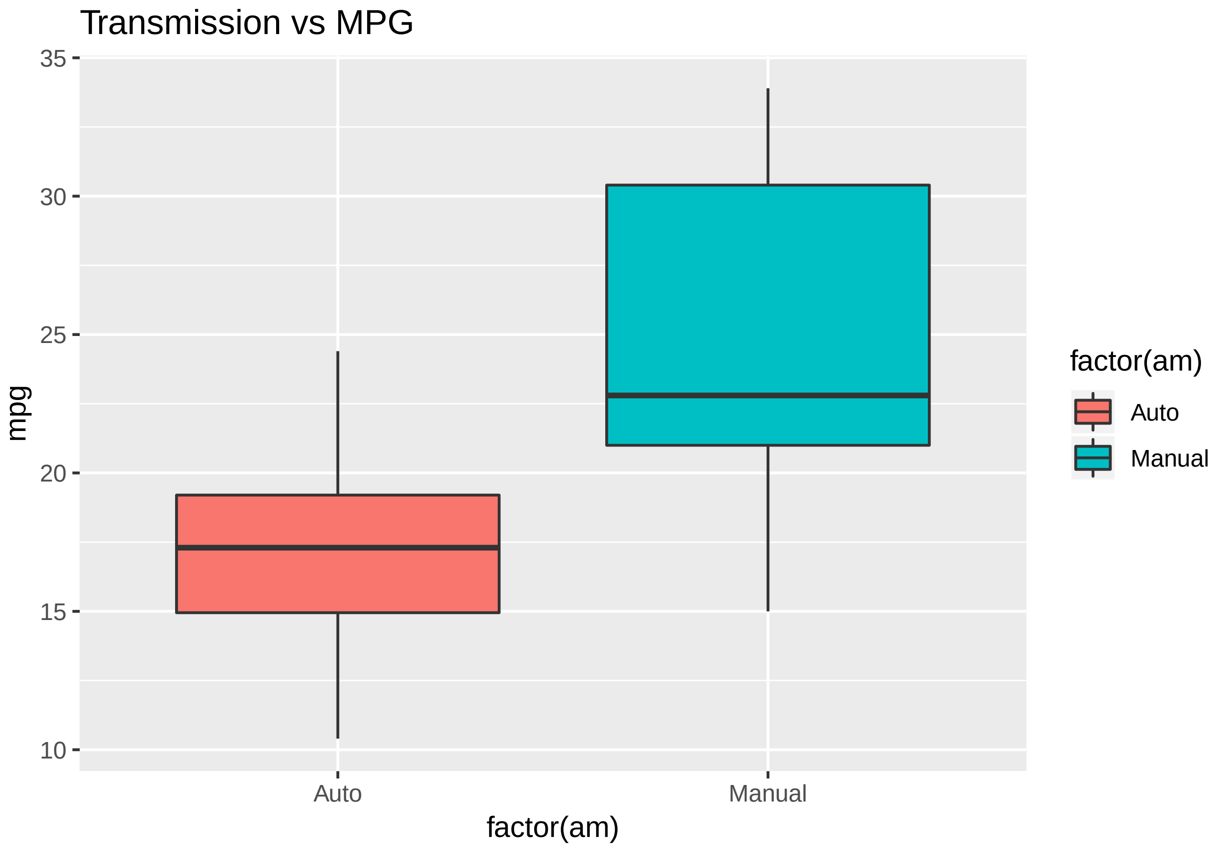
Basic preview:

```
head(mtcars)
```

```
##    mpg cyl disp  hp drat    wt  qsec vs     am gear carb
## 1 21.0   6  160 110 3.90 2.620 16.46  0 Manual    4    4
## 2 21.0   6  160 110 3.90 2.875 17.02  0 Manual    4    4
## 3 22.8   4  108  93 3.85 2.320 18.61  1 Manual    4    1
## 4 21.4   6  258 110 3.08 3.215 19.44  1   Auto    3    1
## 5 18.7   8  360 175 3.15 3.440 17.02  0   Auto    3    2
## 6 18.1   6  225 105 2.76 3.460 20.22  1   Auto    3    1
```

Exploratory analysis (Boxplot for Transmission method vs. MPG)

```r
g <- ggplot(mtcars, aes(x=factor(am), y= mpg))
g + geom_boxplot(aes(fill=factor(am))) + ggtitle("Transmission vs MPG")
```



Building models:

Method 1: (reference: how to build nested model in R)

```r
fit <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- update(fit, mpg ~ factor(am) + wt)
fit3 <- update(fit, mpg ~ factor(am)+ wt + hp)
fit4 <- update(fit, mpg ~ factor(am)+ wt + hp + qsec)
fit5<- update(fit, mpg ~ factor(am)+ wt + hp + qsec + cyl)
anova(fit, fit2, fit3, fit4, fit5)#use anova table to test whether you should include certain variables
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ factor(am) + wt + hp
## Model 4: mpg ~ factor(am) + wt + hp + qsec
## Model 5: mpg ~ factor(am) + wt + hp + qsec + cyl
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 72.0009  5.76e-09 ***
## 3     28 180.29  1     98.03 15.9478 0.0004755 ***
## 4     27 160.07  1     20.22  3.2903 0.0812504 .
## 5     26 159.82  1      0.25  0.0405 0.8420621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Method 2: Stepwise

```
fit_step <- step(fit, direction="both")
```

Comparison:

```r
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6533 -1.3325 -0.5166  0.7643  4.7284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.31994   23.88164   1.060   0.3048
## cyl         -1.02343    1.48131  -0.691   0.4995
## disp         0.04377    0.03058   1.431   0.1716
## hp          -0.04881    0.03189  -1.531   0.1454
## drat         1.82084    2.38101   0.765   0.4556
## wt          -4.63540    2.52737  -1.834   0.0853 .
## qsec         0.26967    0.92631   0.291   0.7747
## vs1          1.04908    2.70495   0.388   0.7032
## amManual     0.96265    3.19138   0.302   0.7668
## gear4        1.75360    3.72534   0.471   0.6442
## gear5        1.87899    3.65935   0.513   0.6146
## carb2       -0.93427    2.30934  -0.405   0.6912
## carb3        3.42169    4.25513   0.804   0.4331
## carb4       -0.99364    3.84683  -0.258   0.7995
## carb6        1.94389    5.76983   0.337   0.7406
## carb8        4.36998    7.75434   0.564   0.5809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.823 on 16 degrees of freedom
## Multiple R-squared:  0.8867, Adjusted R-squared:  0.7806
## F-statistic: 8.352 on 15 and 16 DF,  p-value: 6.044e-05
```

```r
summary(fit_step)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915    
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336 
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Conclusion :

Comparing the initial model (putting all the variables) and our best model(fit_step), we can conclude that the original model has a 0.78 Adjusted R square, meaning that there is only 78% of the variables is explained by this model. However, we have a higher Adjusted R square, 0.834, in our best fitted model.

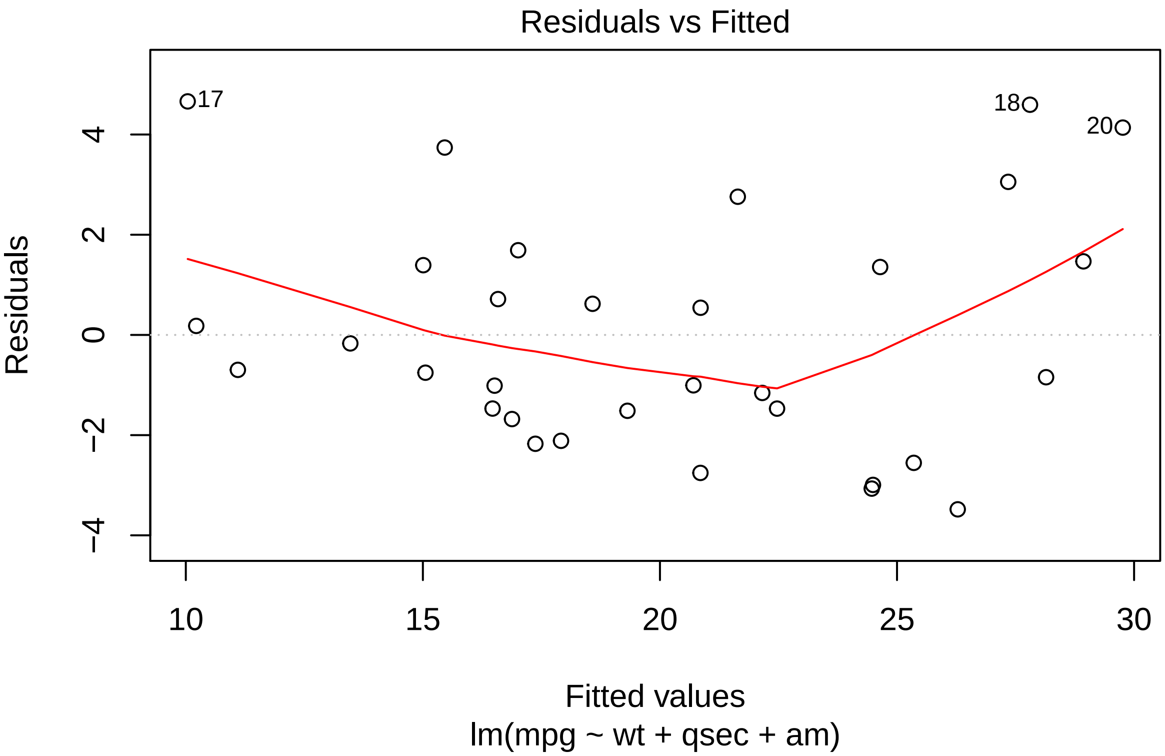
** Coefficient interpretation **

For every other variables stay the same, manual transmission will increase 2.936 more mpg, compared to auto transmission

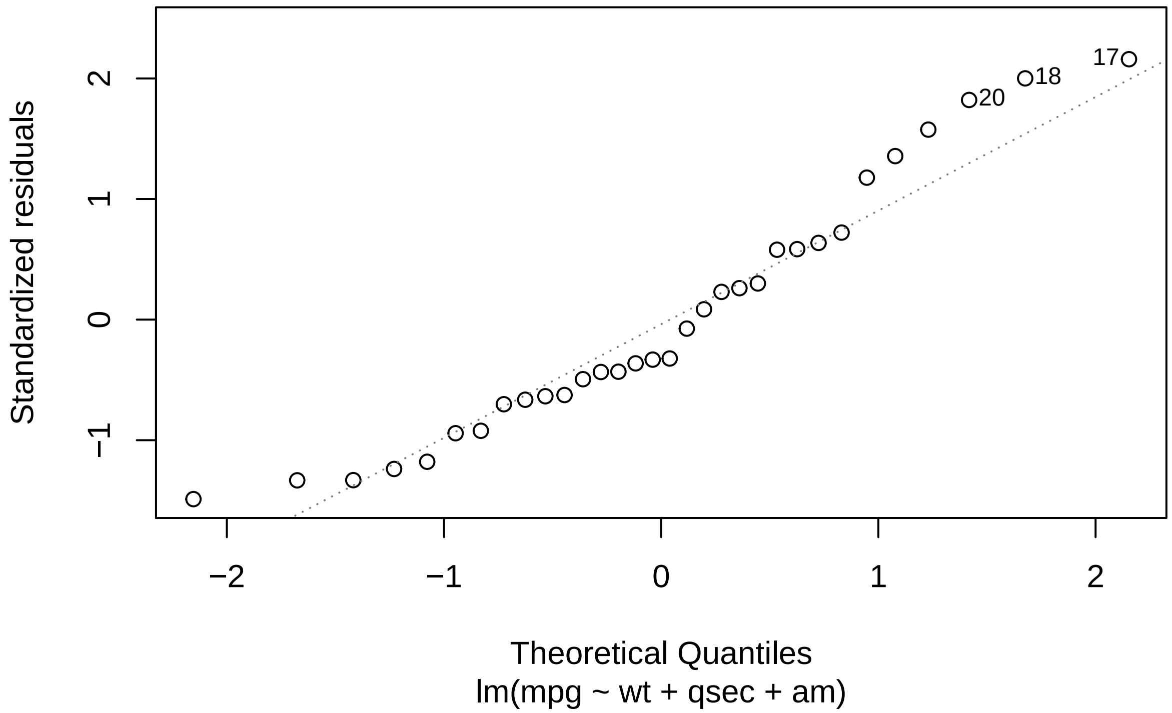
Model diagnosis

Residual analysis:

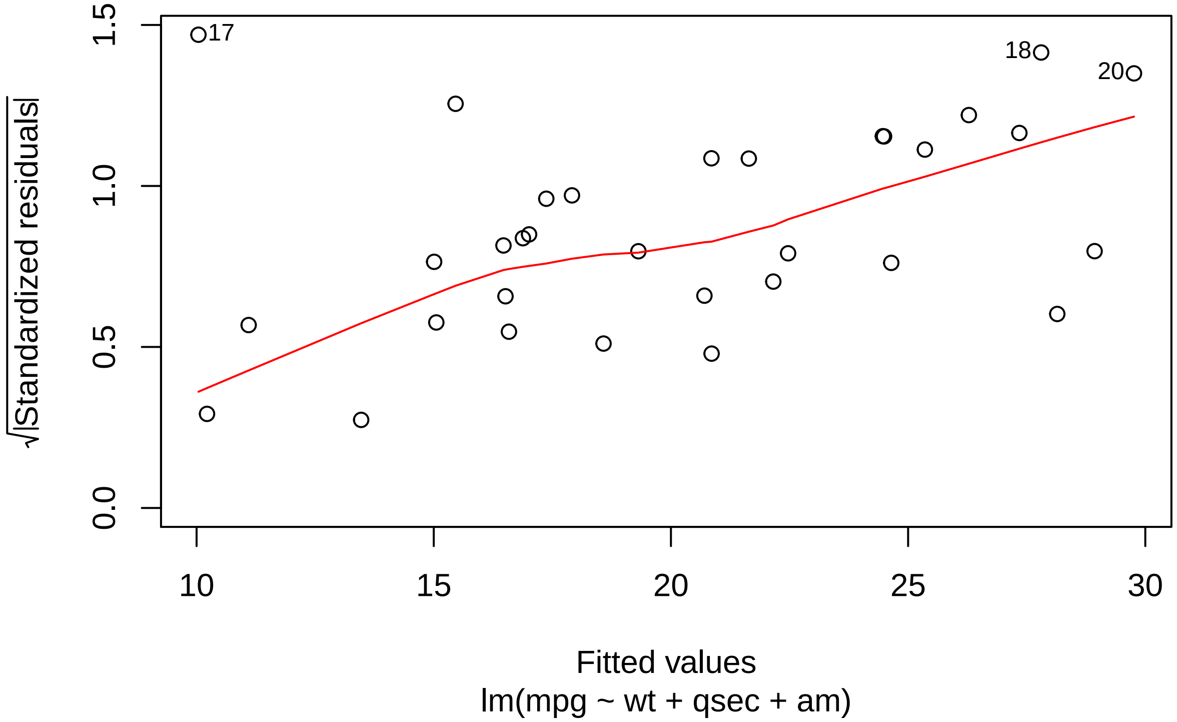
```r
plot(fit_step)
```

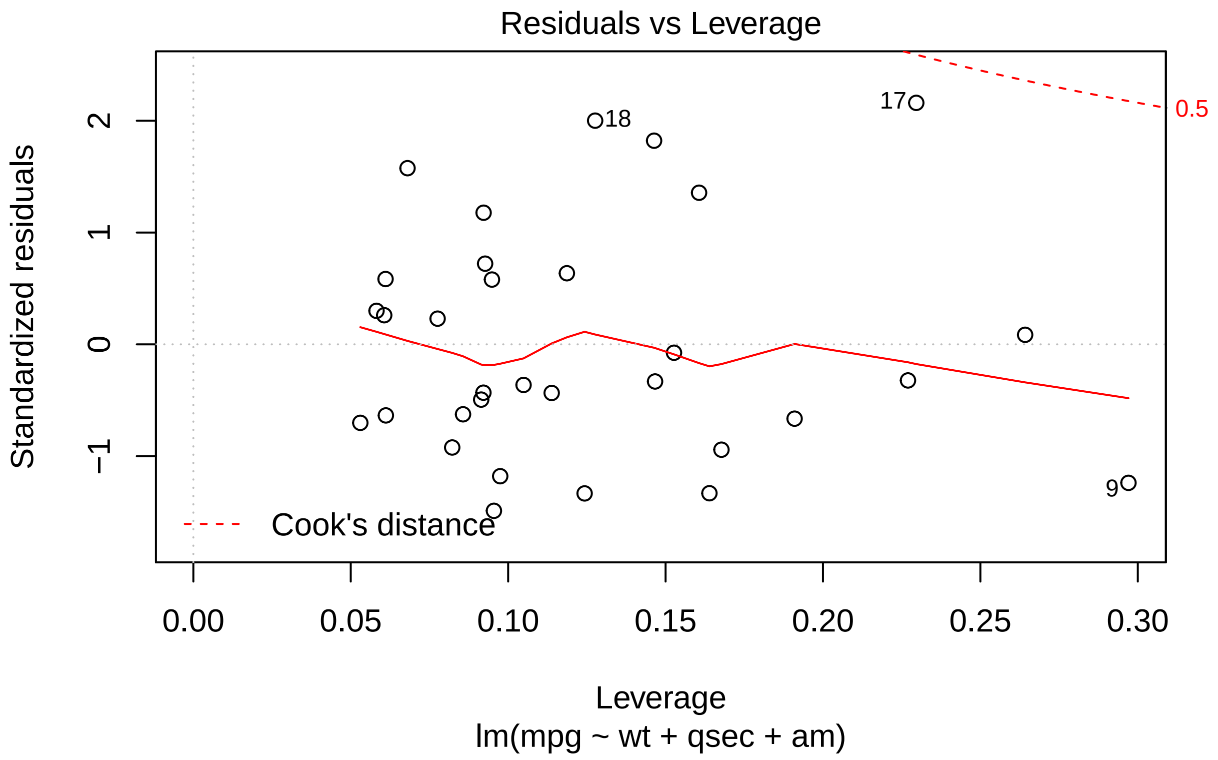


Normal Q−Q



Scale−Location





Residuals vs. Fitted :

The residuals are scattered, ensuring the independence between fitted values and residuals. If there is any pattern, then we should change the model.

Q-Q Plot:

The points are mostly closed with the line. Hence, we can suggest that the residuals are nornally distributed.

Statistical inference:

T-test

```
t.test(mpg~am, mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
##   mean in group Auto mean in group Manual
##             17.14737             24.39231
```

Conclusion

According to the t-test, we can reject the null hypothesis that the transmission method will not have an impact on the mpg.