# R_Project_RegModel

*LI CHENYANG*

*June 4, 2017*

## Executive Summary

In this report I study the variables that may have an impact on MPG (miles per gallon). In particular, I am interested in the relationship between Auto/Manual transmissions and MPG. A quick exploratory analysis is carried out, before linear regression modelling is done to the data. From the study we can conclude that:
1) Keeping all other things constant, Manual has higher MPG output than Auto.
2) Auto has a MPG that is 14.079 + (-4.14)*Unit_Weight different from Manual.

The data used in this report was extracted from 1974 Motor Trend US magazine, as stated in R documentation.

## Exploratory Analysis

For starter, an exploratory analysis is done to explore a general relationship between MPG and other variables in the dataset. Please find the graphs in **Appendix**.
**1) Box Plot shows that Auto Transmission has a lower MPG output than Manual Translation,** if no other criteria is taken into consideration.
**2) MPG v.s Auto/Manual (with cyl) shows that for cars with the same cylinder, Auto SEEMS to have lower average MPG output than Manul.** However, the sample size is not large and the observation might be biased.
**3) MPG v.s Weight indicates that MPG is negatively related to weight.** We could also see most of the manual cars studied are on average lighter than those of auto cars. Therefore we need to adjust for weight when we do regression modelling on the MPG v.s. AM.

## Regression Analysis

### Strategy

1) In this case I am going to try three linear models, each with different predictors.

2) To select the best among the three, I am going to examine **Pr(>t)**, which indicate significance of the predictor,
3) and also **Adjusted R-Squared Value**, which indicates the amount of variance explained by the model.

At the end, we will do residual analysis to double check whether the model is a good fit.

### Model 1: Consider just AM inside mtcars

```
fit1 <- lm(mpg ~ am, mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

As we can see, the model is not very good in this case. The adjusted R-squared value indicates that the model can only explain 34% of the variance for the variables. This indicates that other variable must be added into the model, such weight and number of cylinders.

### Model 2: Consider AM, Weight and number of cylinders

```
fit2 <- lm(mpg ~ am + wt + cyl, mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4898 -1.3116 -0.5039  1.4162  5.7758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7536     2.8135  11.997  2.5e-12 ***
## am1           0.1501     1.3002   0.115  0.90895
## wt           -3.1496     0.9080  -3.469  0.00177 **
## cyl6         -4.2573     1.4112  -3.017  0.00551 **
## cyl8         -6.0791     1.6837  -3.611  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.603 on 27 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8134
## F-statistic: 34.79 on 4 and 27 DF,  p-value: 2.73e-10
```

This model is better than the previous one, in that the Adjusted R-squared value is 0.8134, which indicates that model explains 81% of the variance for the variables. However, the Pr(>t) value for am1 is 0.909, which is abover 5% threshold. This renders a variable of interest non-insiginifcant, and thus may not be an optimal choice.

Also note for cyl, there are only two factors present in summary. This indicates that cyl4 is collinear with one of the variables. Therefore, in the next case, we are taking out cyl variable, and replace it with qsec.

**Model 3: Adding intercation term, taking out cyliner, add horse power.**

```
fit3 <- lm(mpg ~ am + wt+ am*wt + qsec, mtcars)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + am * wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## am1           14.079      3.435   4.099 0.000341 ***
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am1:wt        -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

In this case we can tell that this model is a much better model because,
- First, the adjusted R-squared value increases from 81% to 88%.
- secondly, all coefficients are significant.

**Now we analyse the residual**
According to the residual plot shown in the appendix, we can observe that:
1) The residual v.s. fitted plot seems to be random, thereby indicating there is no other hidden relationship.
2) QQ-Q plot indicates strong normalities;
3) The scale-location plots contains totally random points, which is a sign for constant variance.
4) Last but not least residual v.s. leverage indicates there are not outliers.

## Conclusion

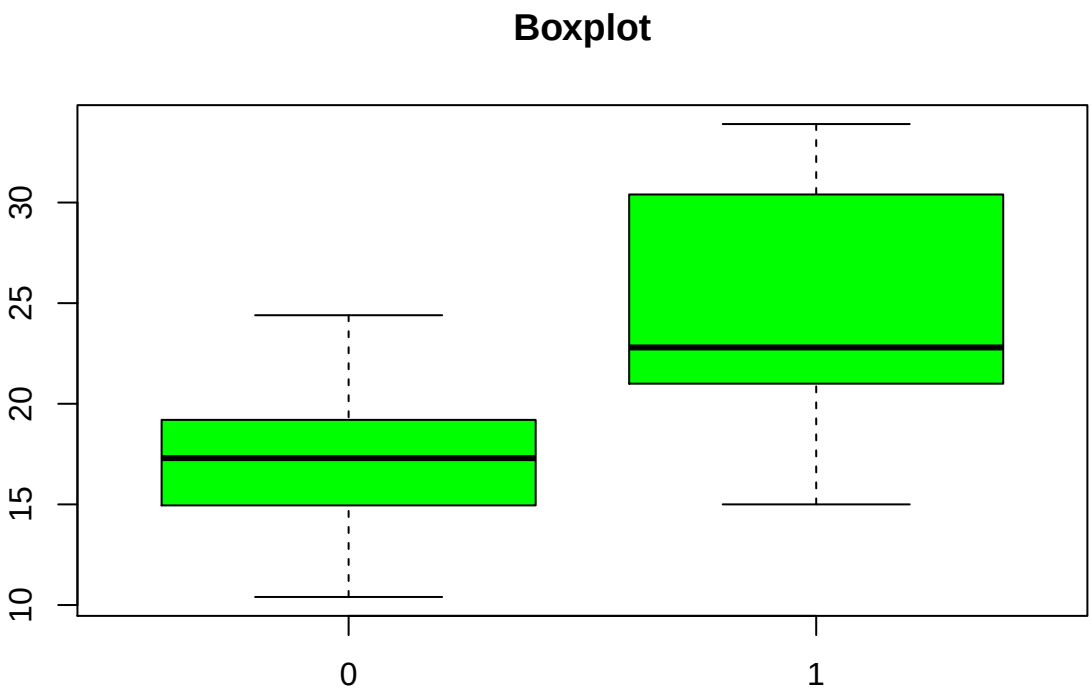
In this case we will select model 3. Using it to answer the questions in the executive summary:
1) Auto translation has a lower miles per gallon out put than Manual translation.
2) Keeping other variable constant, Manual Transmission will have 14.079 + (-4.14)*Unit_Weight MPG output difference from Auto Translation.
3) The confidence interval of this model is

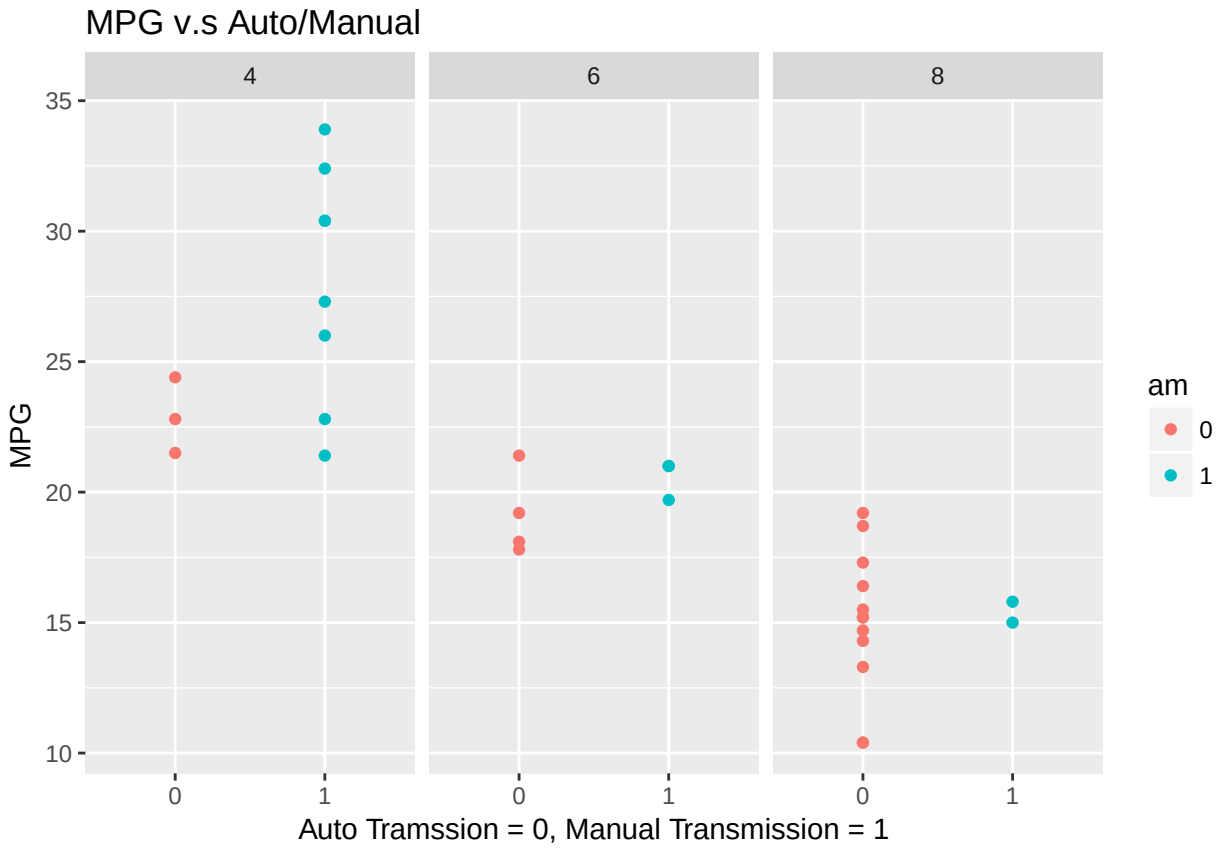
```
confint(fit3)
```

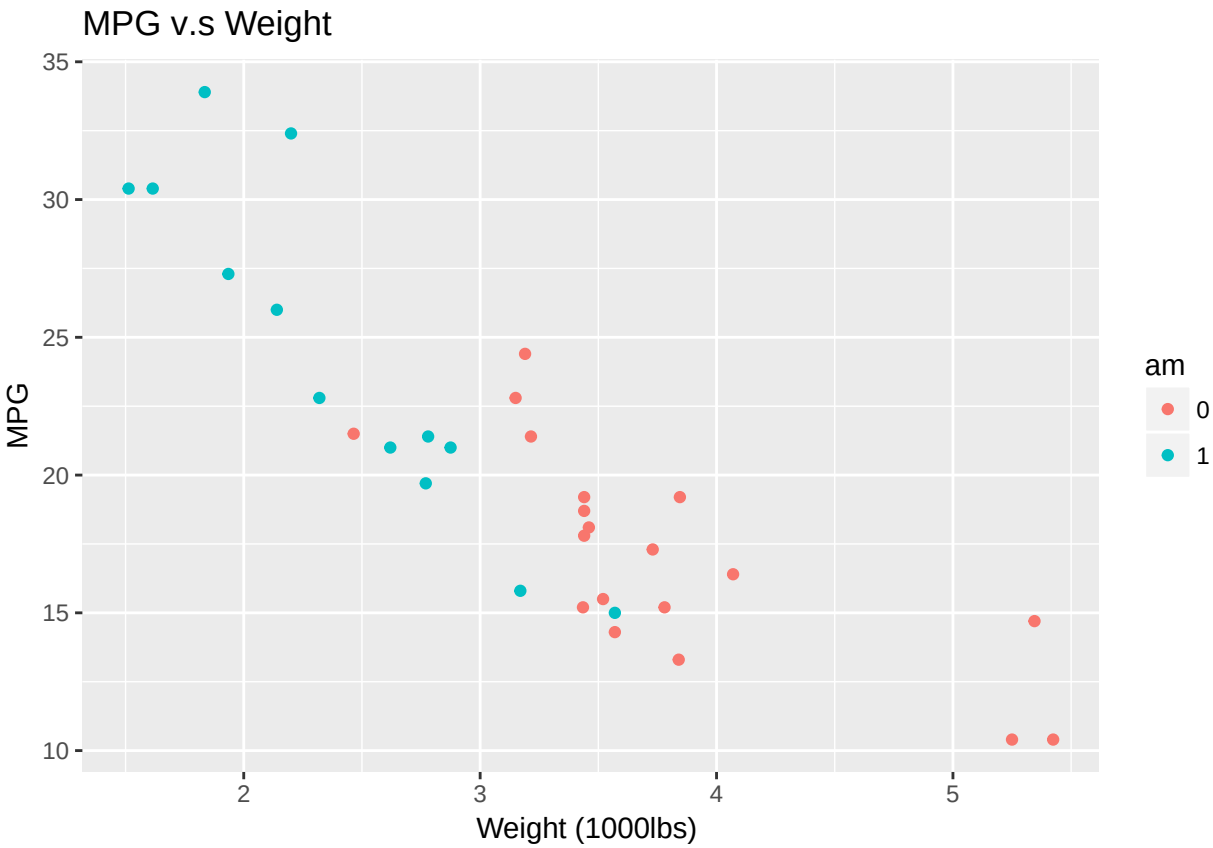
```
##                  2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## am1          7.0308746 21.127981
## wt          -4.3031019 -1.569960
## qsec         0.4998811  1.534066
## am1:wt      -6.5970316 -1.685721
```

At 95% confidence level, The difference in MPG ranges from 7.03 +(-6.59) x Weight to 21.12 + (-1.68) x Weight.

## Appendix






**Residual Plots: Best Model**