Coursera Regression Models Peer Project

Akshay Amrit

5th December, 2019

Summary

This report is the final course project for the course **Regression Models** which is a part of **Data Science Specialization** by Johns Hopkins University on Coursera.
The objective of this course is to answer the questions:
- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions?
Simply put, we want to determine which mode of transmission consumes less fuel.
We will be using "**mtcars**" dataset which is available in "**datasets**" package. To find the answer to our questions,we will start with exploratory data analysis to get a better understanding of what our data looks like and which columns are correlated to MPG. To confirm our conclusion from exploratory data analysis, we will find the best model possible for our data using step function and conclude that **Switching to manual transmission will increase MPG by 1.8.** We will move to inference section from here and perform a t-test on our base model to conclude that **the difference in estimate between transmission is 7.24494 in favour of manual transmission.** We will plot some residual plots and perform diagnostic to confirm whether we made any wrong assumptions during the whole process.

Loading Libraries

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

Data Description and Processing

We analyse "mtcars" using exploratory data analysis techniques and regression models to compare the effect of different transmission techniques i.e. Automatic or Manual on MPG (Miles per Galon).

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).
This dataset is a data frame with 32 observations on 11 (numeric) variables.

[, 1] mpg: Miles/(US) gallon
[, 2] cyl: Number of cylinders
[, 3] disp: Displacement (cu.in.)
[, 4] hp: Gross horsepower
[, 5] drat: Rear axle ratio
[, 6] wt: Weight (1000 lbs)
[, 7] qsec: 1/4 mile time
[, 8] vs: Engine (0 = V-shaped, 1 = straight)
[, 9] am: Transmission (0 = automatic, 1 = manual)
[,10] gear: Number of forward gears
[,11] carb: Number of carburetors

We will now be performing necessary steps required to analyse data.

```r
data("mtcars")
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Exploratory Data Analysis

The objective of performing exploratory data analysis is to figure out what is important to answer our question and what isn't. To begin with, let us try to visualise the relation of MPG with Transmission. See **fig 1** in appendix for the boxplot. Looking at the figure, we can safely say that **MPG is higher for manual transmission**.

Our next step is to look at every column present in the data set and figure out which columns effect MPG and should be taken into account when we try to create our regression model. See **fig 2** in appendix for scatterplot for the entire dataset. From teh scatterplot, we can see that **cyl, disp, hp, drat, wt and am are strongly correlated to mpg**. We will double check it and qantify this relation in regression section.

Regression

Model Selection

We will perform stepwise model selection to get the best model out of the data available to us. We will start with the model which has all columns as predictors and reduce the predictors one at a time to find out the model which best satisfies our needs. This process is handled by **step** function which runs linear models repeatedly to find the best model using forward and backward elimination.

```r
start_model <- lm(mpg~., data = mtcars)
best_model <- step(start_model, direction = "both")
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## 
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - 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 ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
## + carb  5   13.5989 120.40 76.403
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## + gear  2    5.0215 134.00 69.828
## - wt    1   31.1896 170.21 71.482
## + carb  5   14.6475 124.38 73.442
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## + drat  1    0.9672 139.02 67.005
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## + gear  2    4.8033 135.19 68.110
## - wt    1   30.7151 170.71 69.574
## + carb  5   13.0509 126.94 72.095
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## + disp  1     1.247 139.99 65.227
## + drat  1     0.666 140.57 65.359
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## + gear  2     4.684 136.55 66.431
## - wt    1    39.645 180.88 69.428
## + carb  5     2.331 138.91 72.978
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
## - cyl   2    25.284 168.96 63.246
## + qsec  1     2.442 141.24 63.511
## - am    1    16.443 160.12 63.527
## + disp  1     0.589 143.09 63.928
## + drat  1     0.330 143.35 63.986
## + gear  2     3.437 140.24 65.284
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
## + carb  5     3.480 140.20 71.275
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## + vs    1     7.346 143.68 62.059
## + qsec  1     7.044 143.98 62.126
## - cyl   2    29.265 180.29 63.323
## + disp  1     0.617 150.41 63.524
## + drat  1     0.220 150.81 63.608
## + gear  2     1.361 149.66 65.365
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
## + carb  5     5.633 145.39 70.438
```

The best model obtained using the above method has cyl, hp, wt, am which is close to what we had guessed from our exploratory analysis. To analyse the properties of the model obtained:

```
summary(best_model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + 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 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -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 **
## amManual     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
```

The R^2 for **best_model** is 0.8659 which confirms that model confirms 87% variance observed in MPG. Using the coefficients we can infer that increasing number of cylinders from 4 to 6 will lead to a drop of 3.03 MPG but increasing it to 8 will lead to a drop of 2.16 MPG. Per unit increase in horsepower will lead to a drop of 0.03 MPG. Per unit increase in weight of the motor vehicle will lead to a drop 2.49 MPG. Changing the mode of Transmission to manual has a positive effect on MPG. **Switching to manual transmission will increase MPG by 1.8.**

Let us compare the **best_model** with **base_model** which only contains am as predictor to test the importance of predictors cyl, hp and wt. We define our **null hypothesis** as "**cyl, hp and wt do not contribute towards accuracy of the model.**"

```
base_model <- lm(mpg~am, data = mtcars)
anova(base_model, best_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using **anova** function, we can reject our null hypothesis as the p-value is significant.

Inference

We will perform t-test of the data assuming that Transmission has a normal distribution.

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

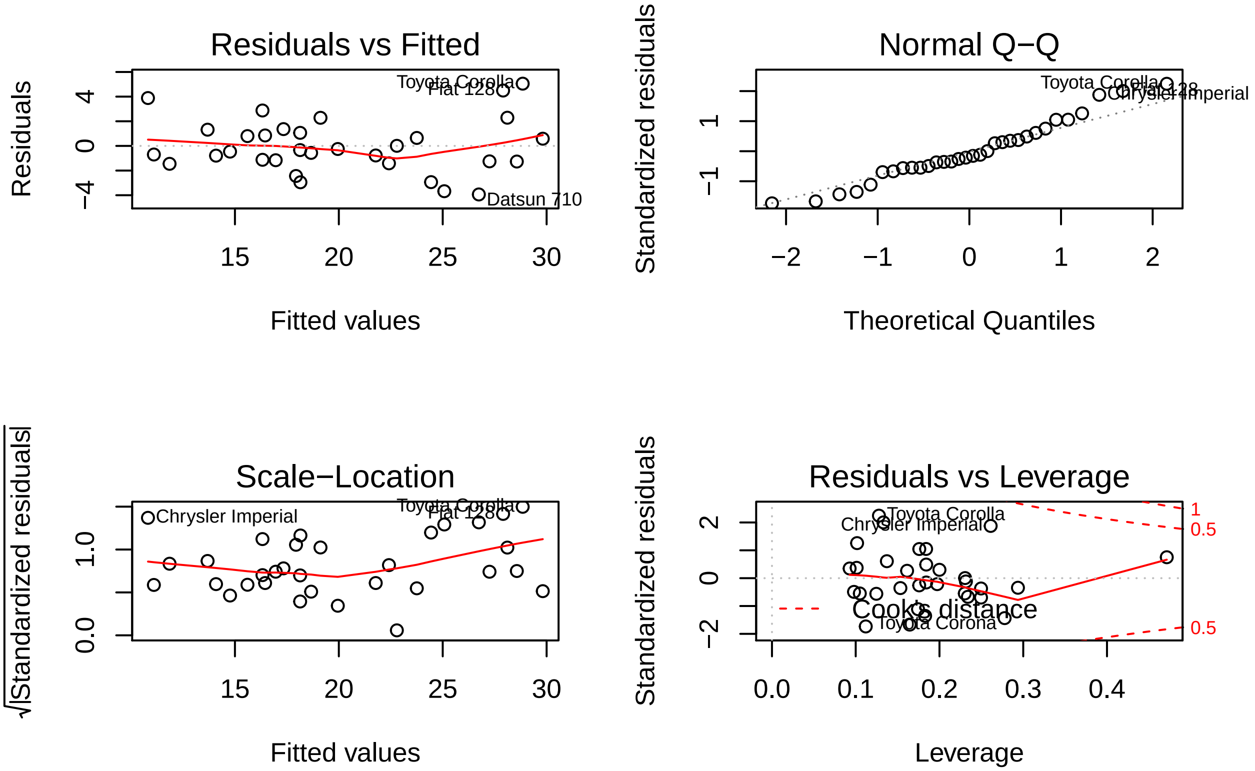
```
## 
##  Welch Two Sample t-test
## 
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group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

From the result, we can conclude that **the difference in estimate between transmission is 7.24494 in favour of manual transmission.**

Residual Plot and Diagnostic

To obtain multivariate regression residuals:

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



From the plots, we can observe that:
- The points in **Residuals vs Fitted** plot seems to be randomly scattered which supports independence condition.
- Most of the points on **Normal Q-Q** plot lie on the line which indicates that the points are normally distributed.
- The **Scale-Location** plot consists of points scattered in a constant band pattern, indicating constant variance.
- Since all points are within the 0.05 lines, the **Residuals vs. Leverage** concludes that there are no outliers.

```
sum((abs(dfbetas(best_model)))>1)
```

```
## [1] 0
```

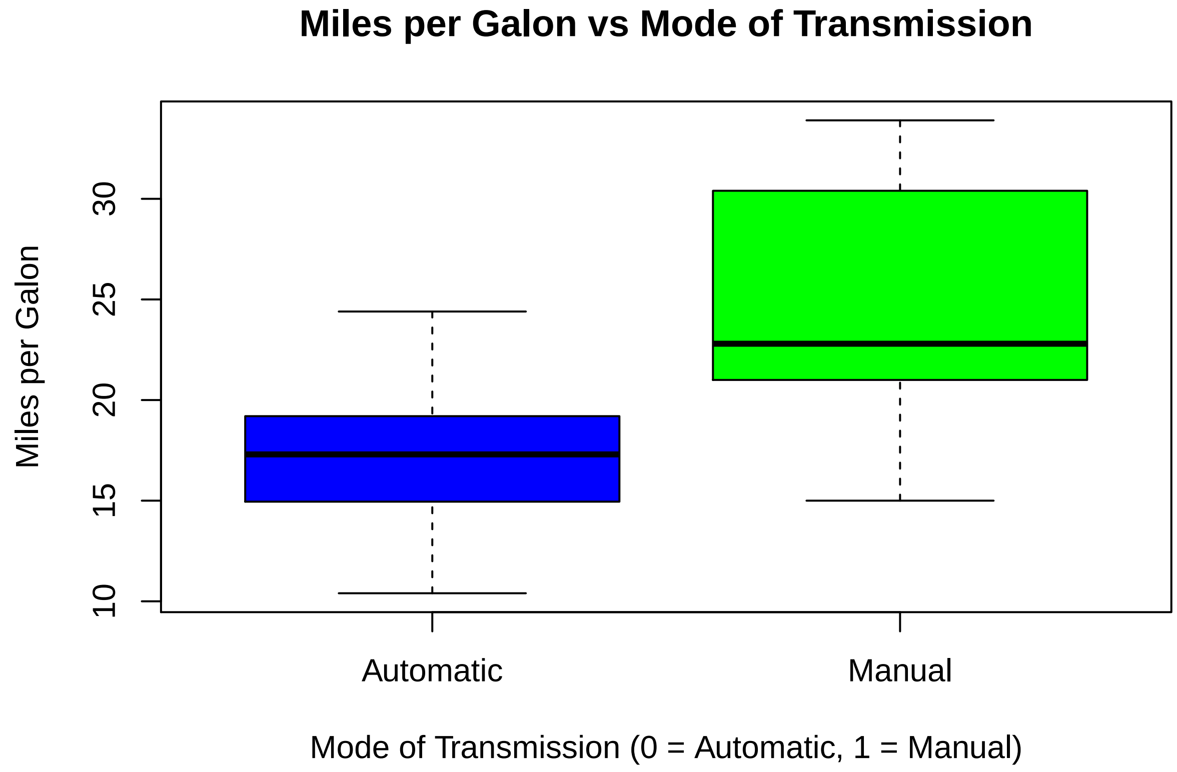
Conclusion

From our analysis, we can conclude that cars with manual transmission get more mileage compared to automatic variant. When adjusted by other factors like number of cylinders, horsepower and weight, we can observe that there is a rise of 1.8 MPG.

Appendix

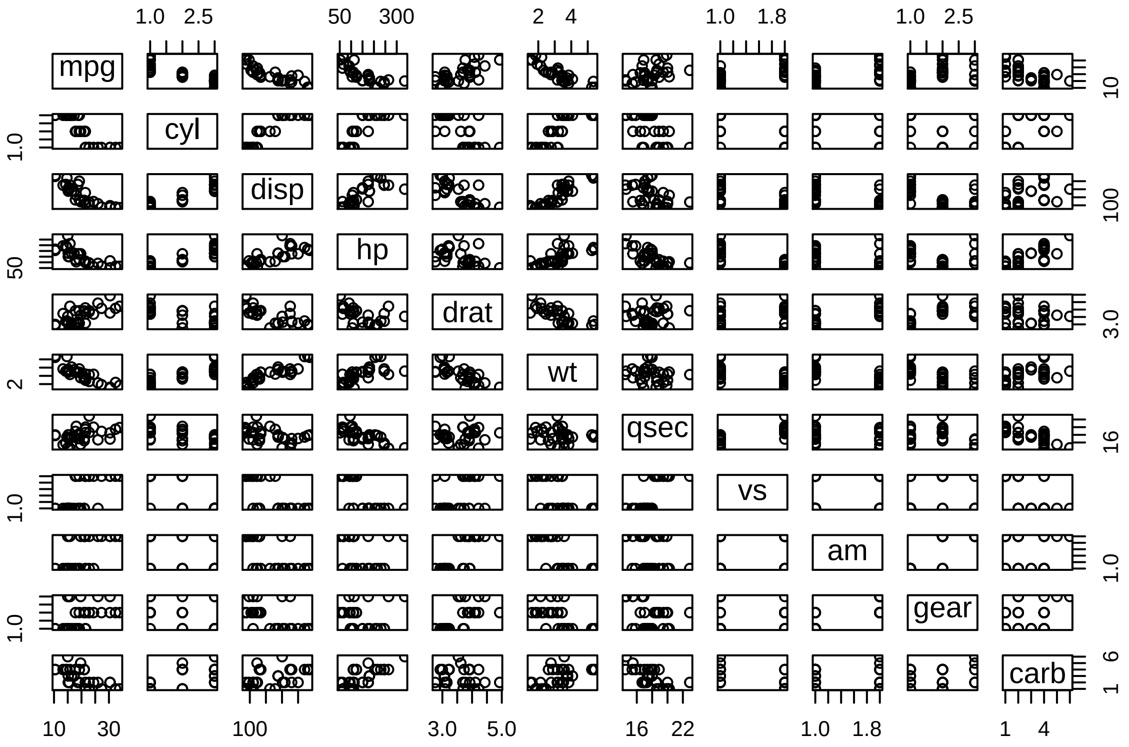
Boxplot of MPG vs Mode of Transmission

```
#Figure 1
boxplot(mpg~am, data = mtcars, col = c("blue", "green"), xlab = "Mode of Transmission (0 = Automatic, 1
```



Scatterplot for “mtcars” dataset

```r
#Figure 2
pairs(mpg~., data = mtcars)
```


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb