Coursera Regression Models Course Project

Author: Tham.Ng

Executive Summary

This report is a course project within the Regression Models course on the Data Science Specialization by Johns Hopkins University on Coursera.

We estimate the relationship between one variable (i.e. type of transmission (manual or automatic) and other independent variables along with miles per gallon (MPG).

Data Description

We analyze the '**mtcars**' data set through Regression Modelling and exploratory analysis to show how automatic (am = 0) and manual (am = 1) transmissions features affect the MPG feature. The dataset "mtcars" is located in the package "reshape2" first introduced in the Reshaping Data Course of the same Data Specialization Course.

The data was extracted from the 1974 Motor Trend US magazine, which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

The data set consists of a data frame with 32 observations (**nrow**) and 11 variables (**ncol**).

- mpg: Miles per US gallon
- cyl: Number of cylinders
- disp: Displacement (cubic inches)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (lb / 1000)
- qsec: 1 / 4 mile time
- vs: V/S
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

Data processing and transformation

We load in the data set, perform the necessary data transformations by factoring the necessary variables and look at the data, in the following section.

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

Data Preparation

We load in the data set, perform the necessary data transformations by factoring the necessary variables and look at the data, in this section.

```
data(mtcars)
mtcars <- mtcars %>%
  mutate(vs = as.factor(vs),
         am = as.factor(am, labels=c('Automatic','Manual')),
         cyl = as.ordered(cyl),
         gear = as.ordered(gear),
         carb = as.ordered(carb))
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 ...
```

Exploratory Data Analysis

In this section, we deep dive into our data and explore various relationships between variables of interest.

Initially, we plot the relationships between all the variables of the dataset in the appendix plot 2.

But we will use linear models to quantify that in the subsequent regression analysis section.

Since we are interested in the effects of car transmission type on mpg, we plot boxplots of the variable mpg when am is Automatic or Manual (see boxplot in the appendix plot 1). This plot clearly depicts an increase in the mpg when the transmission is Manual.

Regression Analysis

In this section, we start building linear regression models based on the different variables and try to find out the best model fit and compare it with the base model which we have using anova. After model selection, we also perform analysis of residuals.

Model building and selection

First, let start with simple regression model how am effect in mpg:

```
fit_base <- lm(mpg ~ am, data=mtcars)
summary(fit_base)
```

```
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 ***
amManual       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
```

This shows us that the average MPG for automatic is 17.1 MPG, while manual is 7.2 MPG higher. The R square value is 0.36 thus telling us this model only explains us 36% of the variance. As a result, significant effect on MPG otherwise it low bias, so we need more feature to complete the model.

Next: we will try **all feature** and check how it effect to **mpg.**

```
fit_all <- lm(mpg ~ ., data = mtcars)
summary(fit_all)
```

```
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
Call:
m(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5087 -1.3584 -0.0948  0.7745  4.6251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.57171   19.56616   1.358   0.1945
cyl.L       -0.23770    5.06256  -0.047   0.9632
cyl.Q        2.02541    2.14952   0.942   0.3610
disp         0.03555    0.03190   1.114   0.2827
hp          -0.07051    0.03943  -1.788   0.0939 .
drat         1.18283    2.48348   0.476   0.6407
wt          -4.52978    2.53875  -1.784   0.0946 .
qsec         0.36784    0.93540   0.393   0.6997
vs1          1.93085    2.87126   0.672   0.5115
amManual     1.21212    3.21355   0.377   0.7113
gear.L       1.78785    2.64200   0.677   0.5089
gear.Q       0.12235    2.40896   0.051   0.9602
carb.L       6.06156    6.72822   0.901   0.3819
carb.Q       1.78825    2.80043   0.639   0.5327
carb.C       0.42384    2.57389   0.165   0.8714
carb^4       0.93317    2.45041   0.381   0.7087
carb^5      -2.46410    2.90450  -0.848   0.4096
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 2.833 on 15 degrees of freedom
```

This result show that fit_all much better than fit_base. But we need to find the best model that fix higher variation. As mentioned, based on the pairs plot where several variables has high correlation with mpg, We build an initial model with all the variables as predictors, and stepwise model selection to select significant predictors for the final model which is the best model. This is taken care by the step method which runs **lm** multiple times to build multiple regression models and select the best variables from them using both **forward selection** and **backward elimination** methods by the **AIC** algorithm. The code is depicted in the section below, you can run it to see the detailed computations if required.

```
fit_final <- step(fit_all, direction = "both", k = log(nrow(mtcars))
```

```
Start:  AIC=101.32
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb


       Df Sum of Sq    RSS     AIC
- carb  5   13.5989 134.00  87.417
- gear  2    3.9729 124.38  95.428
- cyl   2   10.9314 131.33  97.170
- am    1    1.1420 121.55  98.157
- qsec  1    1.2413 121.64  98.183
- drat  1    1.8208 122.22  98.335
- vs    1    3.6299 124.03  98.806
- disp  1    9.9672 130.37 100.400
<none>              120.40 101.321
- wt    1   25.5541 145.96 104.014
- hp    1   25.6715 146.07 104.040
**********************************************
Step:  AIC=68.84
mpg ~ hp + wt + qsec + am


       Df Sum of Sq    RSS    AIC
- hp    1     9.219 169.29 67.170
<none>              160.07 68.844
```

```
- qsec  1    20.225 180.29 69.186
- am    1    25.993 186.06 70.193
+ disp  1     6.629 153.44 70.956
+ drat  1     1.428 158.64 72.023
+ vs    1     0.249 159.82 72.260
+ cyl   2    16.085 143.98 72.387
+ gear  2     1.764 158.30 75.421
- wt    1    78.494 238.56 78.147
+ carb  5     6.393 153.67 84.868

Step:  AIC=67.17
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              169.29 67.170
- am    1    26.178 195.46 68.306
+ hp    1     9.219 160.07 68.844
+ disp  1     3.276 166.01 70.011
+ drat  1     1.400 167.89 70.370
+ vs    1     0.000 169.29 70.636
+ cyl   2     9.862 159.42 72.181
+ gear  2     0.185 169.10 74.067
- qsec  1   109.034 278.32 79.614
+ carb  5    10.999 158.29 82.349
- wt    1   183.347 352.63 87.187
```

The final model obtained from the above computations consists of the variables **wt, qsec** and **am** the independent variable. Details of the model are depicted below.

```
summary(fit_final)
```

```
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
```

From the result above, we observe that the “**Adjusted R square** value”" is **0.83** which is the maximum obtained considering all combinations of variables. Thus, we can conclude that more than **83%** of the variability is explained by the above model.
Next, we use **anova** to compare aganist our base model and final model that was found through performing stepwise selection.

```
anova(base_model, best_model)
```

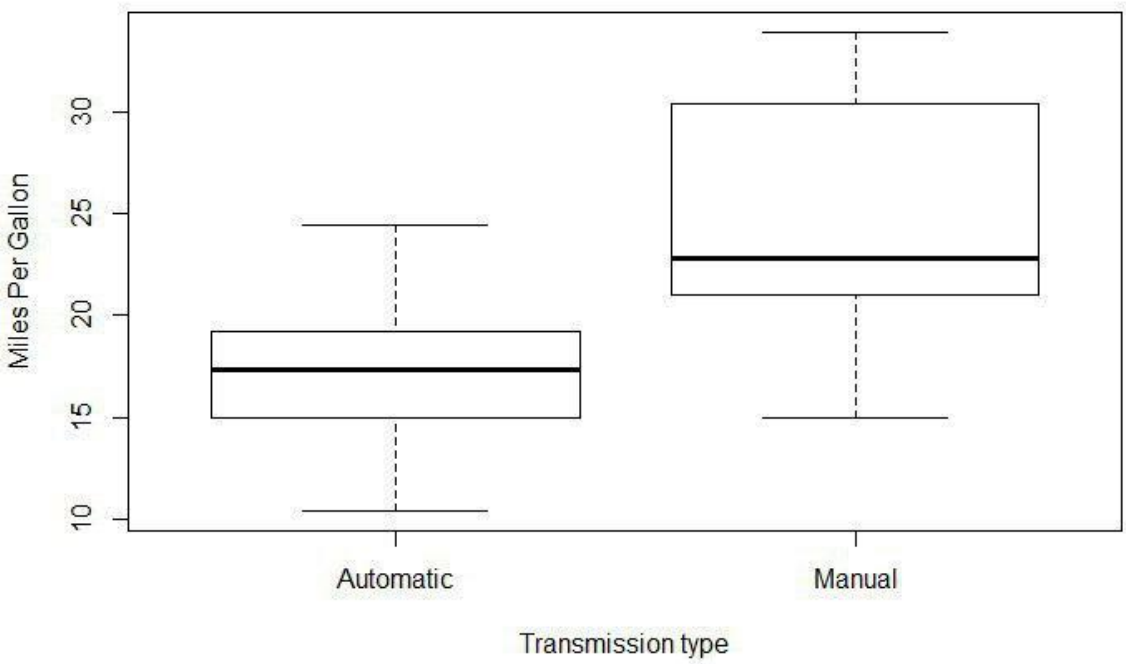
```
## Analysis of Variance Table
Model 1: mpg ~ am
Model 2: mpg ~ wt + qsec + am
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     30 720.90
2     28 169.29  2    551.61 45.618 1.55e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This results in a p-value of **1.55e-09**, and we can claim the fit_final model is significantly better than our fit_base simple model. We double-check the residuals for non-normality (Appendix - Plot 3) and can see they are all normally distributed.

Appendix

Plot 1: mpg along with am

```
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission type", ylab = "Miles
Per Gallon")
```



Plot 2: plot bettet_fit

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