Coursera Data Science Specialisation: Regression Models Course Project

Danny Chan

April 2015

Executive Summary

We use the `mtcars` dataset from R to explore two questions, namely:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

We use data visualisation, hypothesis testing and regression analysis to shed light into these two questions. Our anlaysis shows us that:

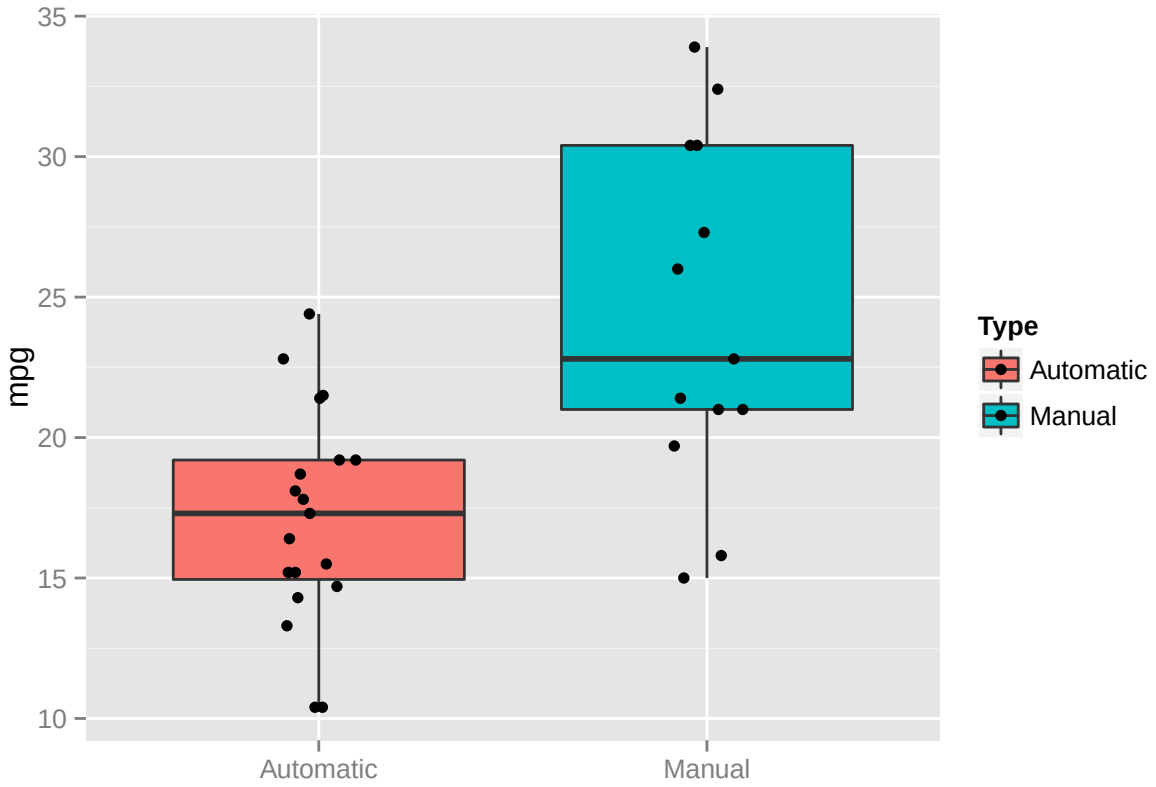
- Manual transmission is better for MPG;
- Cars with manual transmission have a MPG of 2.9 units higher than the MPG of cars with automatic transmission.

1. Exploratory Data Analysis & Statistical Inference

After loading in the dataset `mtcars`, let us first note what variables the dataset contains, and what they measure by using R's handy `?mtcars` help function.

Variable name	Data Type	Definition
mpg	numeric	Miles/(US) gallon
cyl	numeric	Number of cylinders
disp	numeric	Displacement (cu. in.)
hp	numeric	Gross horsepower
drat	numeric	Rear axle ratio
wt	numeric	Weight (lb / 1000)
qsec	numeric	1/4 mile time
vs	numeric	V/S
am	numeric	Transmission (0 = automatic, 1 = manual)
gear	numeric	Number of forward gears
carb	numeric	Number of carburetors

Now that we know what the variables mean, let's check out whether the mean of mpg varies by am. We can do a box plot using the very useful `ggplot2` library.



From the plot, it would seem that cars with manual transmissions are more efficient than cars with automatic transmissions (i.e. manual transmission cars have higher mpg on average).

Let's do a formal statistical inference test, specifically a one-sided unpaired t-test, to test whether the mpg of manual transmission cars are statistically higher than the mpg of automatic transmission cars. We can use R's `t.test` function to do this.

```
library(ggplot2)
g1 <- subset(mtcars, mtcars$am==0)
g2 <- subset(mtcars, mtcars$am==1)
t1 <- t.test(g1$mpg, g2$mpg, alternative="less", paired=F)
t1.summary <- data.frame("p-value" = c(t1$p.value), "CI-Lower" = c(t1$conf[1]), "CI-Upper" = c(t1$conf[
round(t1.summary, 3)
```

```
##                         p.value CI.Lower CI.Upper
## Automatic vs. Manual:     0.001     -Inf   -3.913
```

The null hypothesis for the one sided unpaired t-test is that "mpg of manual transmission cars is not different than the mpg of automatic transmission cars", and the alternative hypothesis is that "mpg of manual transmission cars is greater than mpg of authomatic transmission cars". As the p-value for the one sided unpaired t-test is less than 0.05, we can reject the null hypothesis and accept the alternative hypothesis at the 95% significance level.

2. Regression Analysis

In the previous section, we found through visualising the data and hypothesis testing that manual transmission cars have migher mpg than automatic transmission cars. In this section, we will use regression analysis to explore this further, and quantify the mpg difference between automatic and manual transmission cars.

Regression Analysis

After processing the data, we build an two models:

1. An initial model called `base_model` where we regress the variable `am` against `mpg`; and
2. A second model where we initially build a modelwith all the variables as predictors, and perfom 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.

We then perform an `anova` test to compare the two models.

The code for building the two models are presented below:

```
base_model <- lm(data = dB, mpg ~ am)
step_model <- step(lm(data = dB, mpg ~ .), direction = "both", trace = 0)
```

The results of the base model are:

```
summary(base_model)
```

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

The coefficient on `am` tells us that the mpg on manual transmission cars are on average 7.2 units higher than that of the mpg on automatic transmission cars.

The results of the model built using `step` method are:

```
summary(step_model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = dB)
##
```

```
## 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 ***
## am            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
```

In addition to the `am` variable, the model obtained from the `step` method computations consists of the variables `wt` and `qsec` as additional explanatory variables. The interpretation of the coefficients are:

- `wt`: An increase in the car's weight by 1000lbs decreases `mpg` by 3.9 units (i.e. lighter cars have better `mpg`)
- `qsec`: A 1 second increase in the 1/4 time increases `mpg` by 1.2 units (i.e., cars with quicker `qsec` have better `mpg`)
- `am`: A car with manual transmission increases `mpg` by 2.9 units.

A quick `anova` test shows us that the base model and the model estimated using the `step` are significantly different (o-value is smaller than 0.01, i.e. we reject the null hypothesis that the two models are identical at the 99% significance level) - i.e. the addition of the `wt` and `qsec` improves the model.

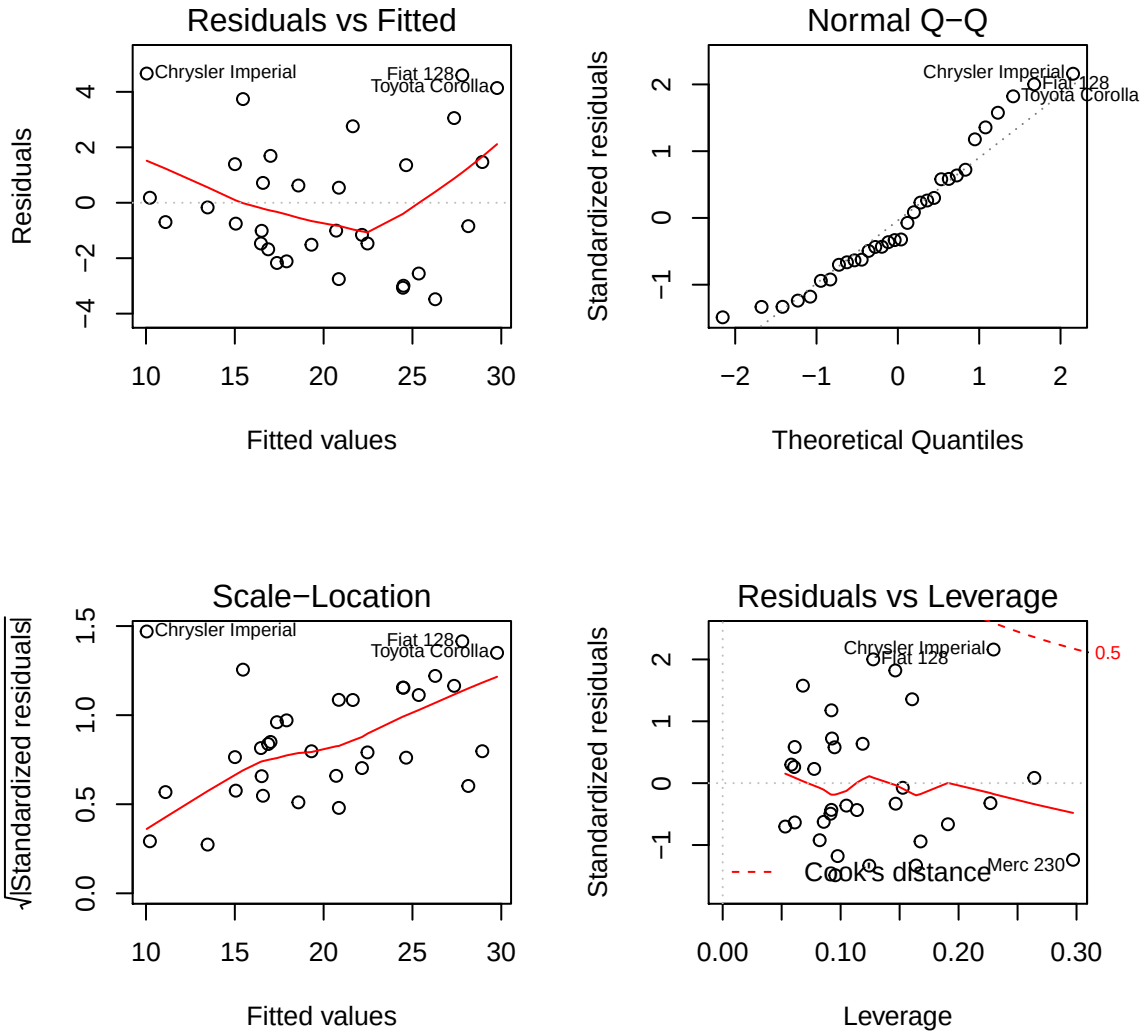
```
anova(base_model, step_model)
```

```
## 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
```

Residuals and Diagnostics

In this section, we shall study the residual plots of our regression model.

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



From the above plots, the following observations can be made:

- The points in the Residuals vs. Fitted plot seem to be randomly scattered on the plot, which verifies the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.

Conclusion

Through the use of data visualisation, hypothesis testing and regression analysis, our analysis shows us that:

- Manul transmission is better for MPG;
- Cars with manual transmission have a MPG of 2.9 units higher than the MPG of cars with automatic transmission.