Regression Models Course Project

Alexey Maranda

Thursday, April 28, 2016

Overview

This analysis has been performed to fulfill the requirements of the course project for the course Regression Models offered by the Johns Hopkins University on Coursera. In this project, we will analyze the mtcars data set and explore the relationship between a set of variables and miles per gallon (MPG) which will be our outcome.

The main objectives of this research are as follows

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

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(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.1.3
```

```
library(ggplot2)
library(datasets)
# clear environment variable
rm(list=ls(all=TRUE))
```

```
# Take a look at what the datasets consists of
dim(mtcars)
```

```
## [1] 32 11
```

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

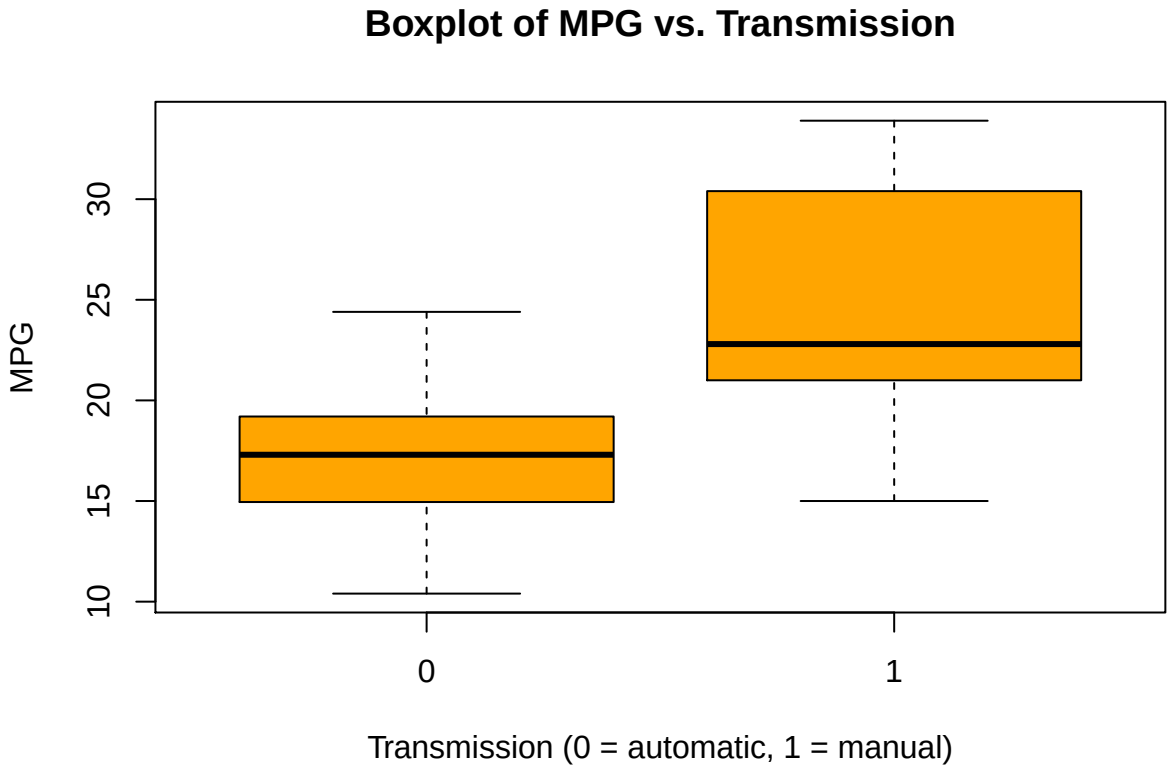
```
# Convert the categorical variables into factors
# We don't transform vs and am as factor because they only have a 2-state value
mtcars$cyl <- factor(mtcars$cyl)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

The dataset has 32 observations of 11 variables. We will do a quick analysis on the variables to gain some insight on the distribution of mpg and the two modes of transmission.

Exploratory Data Analysis

We dive deeper into our data and explore various relationships between variables of interest. Initially, 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. This plot clearly depicts an increase in the mpg when the transmission is Manual.

```
# We mainly focus on the relationship between variables mpg (Miles/(US) gallon) and am (Transmission). A
boxplot(mpg ~ am, data = mtcars, col = "orange",
        xlab = "Transmission (0 = automatic, 1 = manual)" ,
        ylab="MPG",
        main="Boxplot of MPG vs. Transmission")
```



Regression Analysis

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.

```
# Distribution of Manual and Automatic transmission Vehicles.
```

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

The best model obtained from the above computations consists of the variables, cyl, wt and hp as confounders and am as the independent variable. Details of the model are depicted below.

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

From the above model details, we observe that the adjusted (R-squared) value is 0.84 which is the maximum obtained considering all combinations of variables. Thus, we can conclude that more than 84% of the variability is explained by the above model.

In the following section, we compare the base model with only am as the predictor variable and the best model which we obtained earlier containing confounder variables also.

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

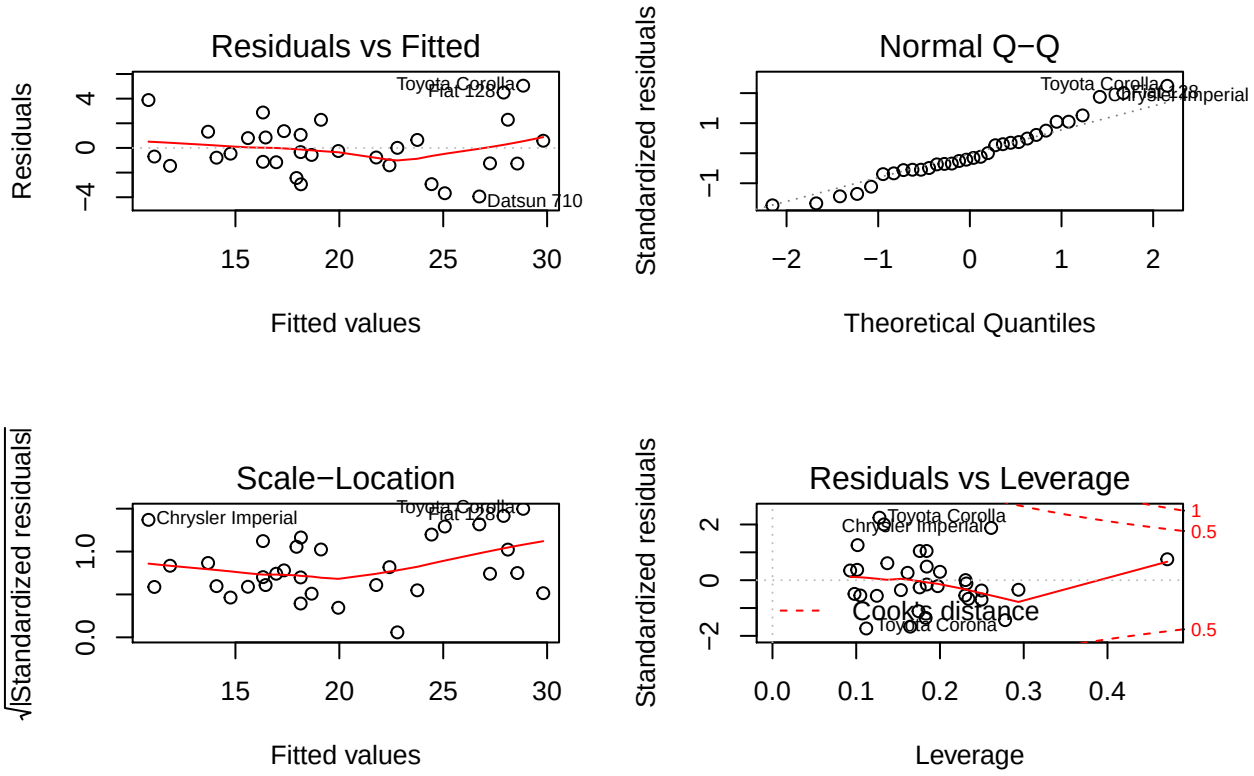
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the above results, the p-value obtained is highly significant and we reject the null hypothesis that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model.

Residuals

We study the residual plots of our regression model and also compute some of the regression diagnostics for our model to find out some interesting leverage points (often called as outliers) in the data set.

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



From the above plots, we can make the following observations:

- The points in the Residuals vs. Fitted plot seem to be randomly scattered on the plot and verify 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.
- There are some distinct points of interest (outliers or leverage points) in the top right of the plots.

Conclusion

We can conclude the following:

- Cars with Manual transmission get more miles per gallon mpg compared to cars with Automatic transmission.

- mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt.
- mpg decreases negligibly with increase of hp.