Regression Models Project

Rustam Mansyrov

Thursday, September 24, 2015

Introduction

The principal objective of the project is to answer the posed question using data that was extracted from the 1974 *Motor Trend* US magazine. It comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. The question of interest is the difference between the consumption of fuel for vehicles with automatic and manual transmissions, respectively. Firstly, we will try to answer the question using Exploratory Data Analysis (EDA). Then, using modeling techniques, we will attempt to prove the results found by EDA.

Summary of Data

Quick and concise summary of the data is as follows:

```
      mpg             cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
      drat             wt             qsec       vs             am         
 Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   Automatic:19  
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13  
 Median :3.695   Median :3.325   Median :17.71                        
 Mean   :3.597   Mean   :3.217   Mean   :17.85                        
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90                        
 Max.   :4.930   Max.   :5.424   Max.   :22.90                        
```

We can see that average mpg is 20.09 and median of mpg is 19.20, that is quite close to the mean. So, mpg variable might be said to be approximately normally distributed. The question of interest is the difference of mpg between automatic and manual transmissions. Generally, manual transmission consumes more miles/gallon than automatic one. But, this is just an observation based on no facts. However, using this data, we might get an insight whether aforementioned is true or not.

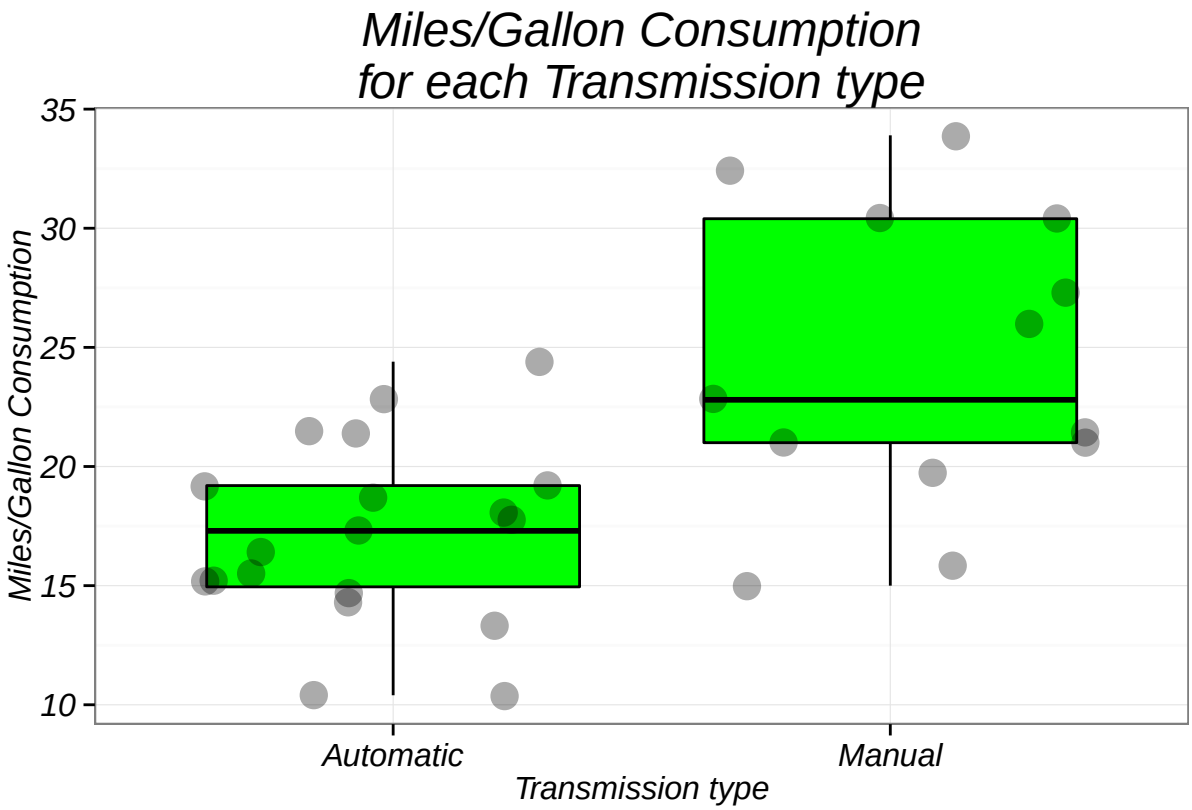
Exploratory Data Analysis

After getting an idea about the summary of the data, let‘s look at specific characteristics:

```
  Transmission Mean of mpg Median of mpg St. deviation of mpg
1    Automatic    17.14737          17.3             3.833966
2       Manual    24.39231          22.8             6.166504
```

Dividing mpg variable into two groups yields the apparent difference in means, medians and standard deviations. We can see that mean consumption for the manual transmission is significantly larger than for the automatic transmission. The same might be stated for medians and standard deviations.

In order to observe the relationship better, let us use the boxplots and visually convince ourselves that the stated above is true:



Apparently from the boxplot above, we can see that miles/gallon consumption for manual transmission is larger than that for the automatic transmission.

Modelling

Using exploratory data analysis may not be necessarily effective, because using graphics may not always yield objective answer. However, modeling part always provides analysts with the most precise results. For this project, multivariable regression analysis is implemented and before selecting the best model, it is always useful to check underlying assumptions of all regression models. And, the first thing to do is getting rid of redundancy of information in the model. That is, let us check whether covariates are correlated with each other or not:

```
            mpg       disp         hp        drat         wt        qsec
mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403
disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
```

From the correlation table above, we see that weight correlated is the most correlated with mpg. So, it is relevant to drop all regressors higly correlated with weight variable. As a result, we get the following model:

```
                     Estimate Std. Error   t value     Pr(>|t|)
(Intercept)          9.617781  6.9595930  1.381946 1.779152e-01
wt                  -3.916504  0.7112016 -5.506882 6.952711e-06
qsec                 1.225886  0.2886696  4.246676 2.161737e-04
as.factor(am)Manual  2.935837  1.4109045  2.080819 4.671551e-02
```

We see that intercept's p - value is larger than 0.05, meaning that it is insignificant. Let us get rid of continuous regressors:

```
                     Estimate Std. Error   t value     Pr(>|t|)
(Intercept)         17.147368   1.124603 15.247492 1.133983e-15
as.factor(am)Manual  7.244939   1.764422  4.106127 2.850207e-04
```

After getting rid of weight variable, the model has finally all significant regression coefficients. But, this might be the case that interaction term is necessary. Let us check it using ANOVA test:

```
Analysis of Variance Table

Model 1: mpg ~ am
Model 2: mpg ~ am + wt * am
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     30 720.90
2     28 188.01  2    532.89 39.682 6.733e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

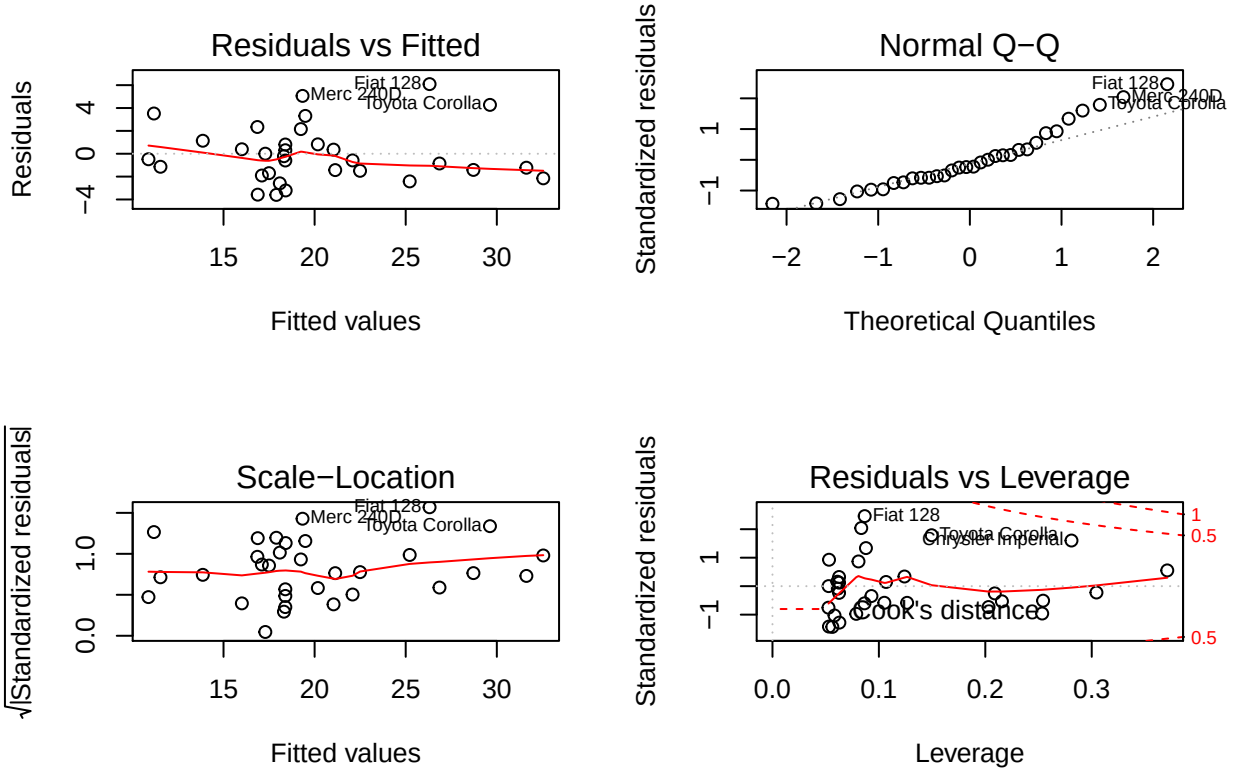
We see that p - value for second model is low and significant to reject null hypothesis, stating that the first model is sufficient. Therefore, the second model is better. The entire summary for the model is as follows:

```
Call:
lm(formula = mpg ~ am + wt * am, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6004 -1.5446 -0.5325  0.9012  6.0909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.4161     3.0201  10.402 4.00e-11 ***
amManual     14.8784     4.2640   3.489  0.00162 **
wt           -3.7859     0.7856  -4.819 4.55e-05 ***
amManual:wt  -5.2984     1.4447  -3.667  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.833, Adjusted R-squared:  0.8151
F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

Looking at the summary, the first thing to report is adjusted r-squared which is quite large. It means that our regressors explain 81 % of variation in expected miles/gallon consumption. As for regression coefficients, the following might be reported:

- If we use manual transmission, mpg = 45 - 8 * wt. This means that for every one unit increase in weight, the expected change in miles/gallon consumption is -8.
- If we use automatic transmission, mpg = 31 - 3 * wt. This means that for every one unit increase in weight, the expected change in miles/gallon consumption is -3.
- When the weight is zero(which is obviously impossible), the expected miles/gallon consumption 45 for manual transmission and 31 for automatic transmission, which implies that manual transmission always requires larger amount of fuel.

Diagnostics for the Best Model



- On residuals versus fitted values plot, no pattern is observed. So, homoscedasticity is obvious.
- The normality plot of residuals is shown on the upper right plot. The residuals seem to be normally distributed.
- Several residuals do surpass the cook's distance. Meaning that there might be slight evidence of outliers.
- Plot of standardized residuals versus fitted values depict no systematic pattern, meaning that no heteroscedasticity is possible here.