Automatic or Manual transamission: A regression analysis on mtcars dataset

JL Siaw

Saturday, November 14, 2015

Executive summary

This analysis is focusing on data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) as outcome. We are particularly interested to explore:

- Is an automatic or manual transmission better for MPG?
- How difference of MPG between automatic and manual transmissions?

Statistic testing indicates that multivariate linear regression is better in getting the answer. Our multivariate linear regression model shows that manual transmission cars is more efficient, it gets **2.08** MPG more than automatic transmission cars on average.

Data Explorartory Analysis

First of all, let's label the predictor variable of interest `am` to value `Automatic` (0) and `Manual` (1) in a separate field for better interpretability.

In order to ensure data is suitable for the statistical and regression analysis, we need to inspect the scatter plot (figure 1) on how `mpg` varies by manual vs auntomatic transmission. simply looking at the plot, we know that `cyl`, `disp`, `hp`, `drat`, `wt`, `vs` and `am` has strong correlations with `mpg`.

First row of the plot shows that distribution of `mpg` (top left curve) of car dataset is quite normal and it is fit for statistical analysis. When we compared transmission types against `mpg` in figure 2, it looks like manual transmission cars have better `mpg` than automatic type. We will confirm this using hypothesis testing and regression model.

Statistics Testing

The distribution of boxplot in figure 2 shows that the mean MPG (red point) of manual transmission cars (**24.39**) is slightly better than that of automatic transmission cars (**17.15**). Is this significant?

We used $\alpha = 0.05$ to run t-test on dataset in hypothesis testing. We will test null hypothesis if there is no difference between the mean `mpg` of two transmission types.

With p-value of **0.0014** as shown in result 1, we reject the null hypothesis bacause p-value $\leq \alpha$. There is a significiant difference in the mean MPG between manual transmission cars and that of automatic transmission cars. Now we have to quantify the difference with regression model.

Regression Model

Correlation Analysis

To decide which predictor variables are suitable for the regression models, we look at the correlation matrix in figure 1. Data in `mpg` row shows the correlations with other variables. Since the categorical variable with multiple levels such as `cyl` has consistent trend, we interpret the correlation value same as other continuous variables.

Besides the mandatory `am` in this study, we see that the high correlation predictor variables with `mpg` are `cyl`, `disp`, `hp` and `wt`. Those variables can be the potential candidates in regression model. However, we found from figure 1 that there is high correlation between `cyl` and `disp` and also few other variables. We decided to drop `cyl` and `disp` from the model since predictors should not exhibit collinearity. In answering the questions with more evidence, We will build and compare two linear regression models. Model(1) mpg ~ am and Model(2) mpg ~ am + hp + wt

Regression Analysis

Simple Linear Regression

Data in `am` is dichotomous, it can be used in the regression model directly without extra dummy coding.

Simple linear regression analysis for `mpg` vs `am` is in Result 2. The coefficient and intercept show that the 'Automatic' cars (value=0) have **17.15** MPG on average, whereas 'Manual' cars (value=1) have **7.25** more than `Automatic`. The R^2 value with **0.36** indicates that only 36% variation in the estimated `mpg` variable that is explained by the regression model.

Multivariate Linear Regression

Multivariate linear regression analysis for `mpg` vs `am` + `hp` + `wt` is in Result 3. The R^2 appears that this model can explain about 84% variation in `mpg`. Coefficient summary also indicates that `am` has confound relationship with both `wt` and `hp` in predicting `mpg`. On our focus study, coefficient of `am` suggests that 'Manual' transmission cars have **2.08** more MPG than `Automatic` on average.

Residual Analysis

We now examine the residuals of model 2 to see there is any systematic pattern on fitted `mpg` value.

The residuals in figure 3 `Residuals vs Fitted` looks random and close to 0. `Normal Q-Q` plot shows points that fall on or close to the line, indicating the residuals are approximately normally distributed. So multivariate linear regression has no problem with the model too.

Model Selection

Analysis of Variance with ANOVA in Result 4 shows that p-value for model 2 is very small. We reject the null hypothesis and accept variables `hp` and `wt` as both of them appear necessary in the model.

Appendix

Figure 1: Matrix of Data Relationships

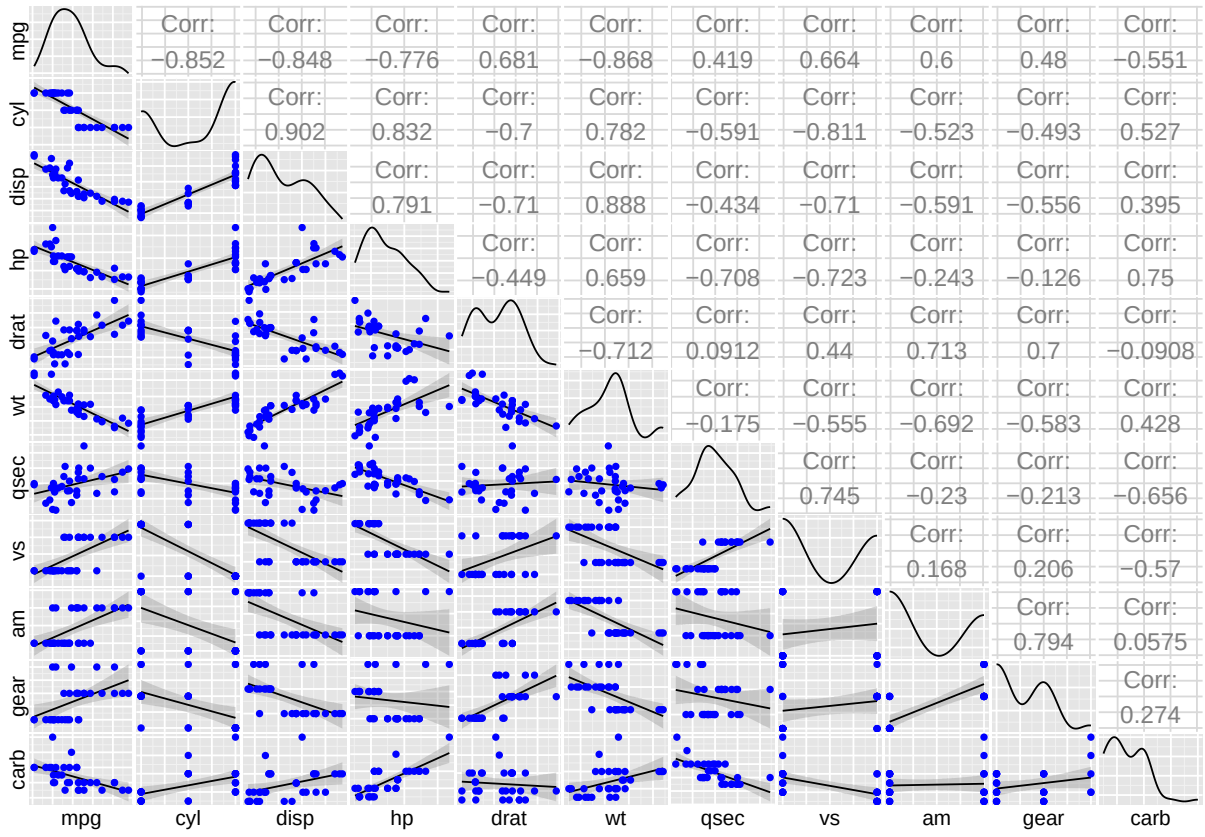
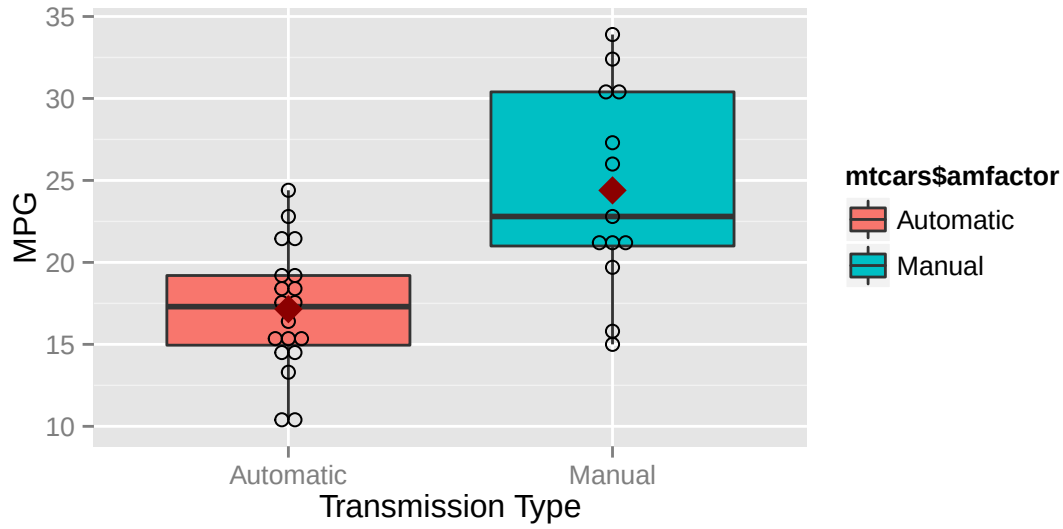


Figure 2: MPG by Transmission Type



Result 1: t-test on MPG by Transmission Types

```
##
##  Welch Two Sample t-test
##
## data:  auto$mpg and manual$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

Result 2: Simple Linear Regression Model

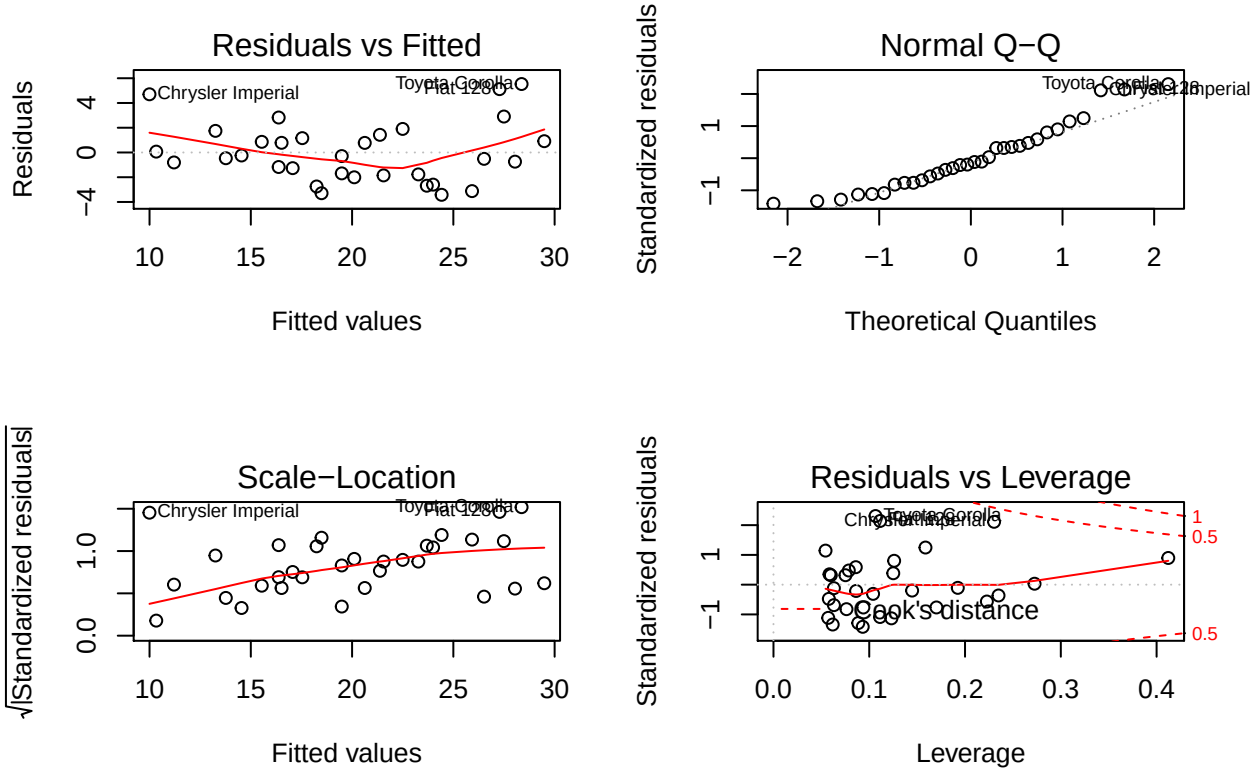
```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             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
```

Result 3: Multivariate Linear Regression Model

```
##
## Call:
## lm(formula = mpg ~ am + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## wt          -2.878575   0.904971  -3.181 0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Figure 3: Residuals Plot



Result 4: Model Selection

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```