Motor Trend Car Road Test: Data Analysis Report

Coursera Regression Models Project

Giuseppe Di Bernardo

April 18, 2016

1. Executive Summary

This is the Assignment Project of the Coursera Regression Models course. We are going to analyze the `mtcars` data provided in the R dataset packages, and statistically infer conclusions about the relationship between the Miles per Gallon (MPG) and a set of variables. 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?

We address the above questions by performing regression models and exploratory data analysis. We fit several linear regression models and select the one with the highest Adjusted R-squared value. Moreover, in the Inference section we perform a statistical *t*-test showing difference between cars with Automatic and Manual transmission. We find that the data corresponding to the type transmissions are significantly different each others, and that the mean for MPG of manual transmitted cars is about 7 more than that of automatic transmitted cars.

2. Data Processing and Exploratory Data Analysis

We load in the `mtcars` data set to perform the necessary data transformations, changing the variables of interest from `numeric` class to `factor` class, e.g, :

```
data("mtcars")
mtcars$am <- factor(mtcars$am, labels = c('Automatic', 'Manual'))
mtcars$cyl <- as.factor(mtcars$cyl)
```

A basic exploratory data analysis - on the effects of car trasmission type - can be observed in the boxplot, reported for convenience in the **Appendix**, Figure 1. We may infer that `Manual` trasmission in general yields higher values of Miles per Gallon (`mpg`). Moreover, exploring the several relationships between the other quantieties present in the dataset, we notice that variables like `cyl`, `disp`, `hp`, `drat`, and `wt` all have some strong correlation with the `mpg`, as we can observe in the pair graph present in the **Appendix** Figure 2.

3. Regression Analysis

The next step is to perform multiple linear regression models based on all the variables seem to have high correlation with the `mpg` quantity. Through a stepwise regression - **forward selection** and **backward elimination** methods modeled by the `AIC` algorithm - we find out the best model fit, where a subset of predictor variables is selected from a larger set. The final model will be chosen comparing this latter one with the initial model, using the `ANOVA` analysis.
First, we fit the simple model with `mpg` as the outcome variable and `am` as the predictor variable.

```
transmission_model<-lm(mpg ~ am, data = mtcars)
summary(transmission_model) # hide results
```

It shows that on average, a car has 17.147 `mpg` with automatic transmission, and if it is manual transmission, 7.245 `mpg` is increased. This model has the `Residual standard error` as 4.902 on 30 degrees of freedom. And the `Adjusted R-squared value` is 0.3385, which means that the model can explain about 34% of the variance of the `mpg` variable. The low `Adjusted R-squared value` also indicates that we need to add other variables to the model. Next, we fit the full model as it follows:

```
full_model <- lm(mpg ~ ., data = mtcars)  # multiple regression analysis
summary(full_model) # hide results
```

A model built in this way has a `Residual standard error:2.833` on 15 degrees of freedom, and an `Adjusted R-squared:0.779` which means that it can explain about 78% of the variance of the `mpg` variable. However, none of the coefficients are significant at 0.05 significant level.
To select some statistically significant variables we use both the forward and the backward stepwise selection:

```
best_model <- step(full_model, direction = 'both')
```

The `best_model` has `cyl`, `wt` and `hp` as confounders and `am` as the indipendent variable, a `Residual standard error:  2.41` on 26 degrees of freedom, and `Adjusted R-squared:  0.8401`, which means that the model can explain about 84% of the variance of the `mpg` variable. All of the coefficients are significant at 0.05 significant level. This is a pretty good one. Finally, we select the ultimate model, by comparing all the models considered so far with the `anova()` function:

```
anova(transmission_model, full_model,best_model)
```

Looking at the above results, we reject the null hypothesis that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model, and end up selecting the model with the highest Adjusted R-squared value: `mpg ~ cyl + hp + wt + am`:

```
summary(best_model)$coef
```

```
##              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
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

5. Inference Analysis

The question "*How different is the MPG between automatic manual transmission?*" needs to be quantitatively addressed, in order to correctly support our conclusions. We perform an inference analysis, and at this step we make the null hypothesis H_0 as the `mpg` of both the automatic and manual transmissions are from the same population (assuming the data have a gaussian distribution). Then, from the results of a two samples t-test

```
##
##  Welch Two Sample t-test
##
```

```
## data:  mpg by am
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

we can affirm that the two transmissions type data are significantly different, at 95% of confidence level. In fact, the very low `p-value` (about 0.1%) of the statistical test tells us to reject H_0. Moreover, the mean value for `mpg` of manual transmitted cars is about 7 more than that of automatic transmitted cars, confirming so the conclusion drawn in section 2.

6. Regression Diagnostic

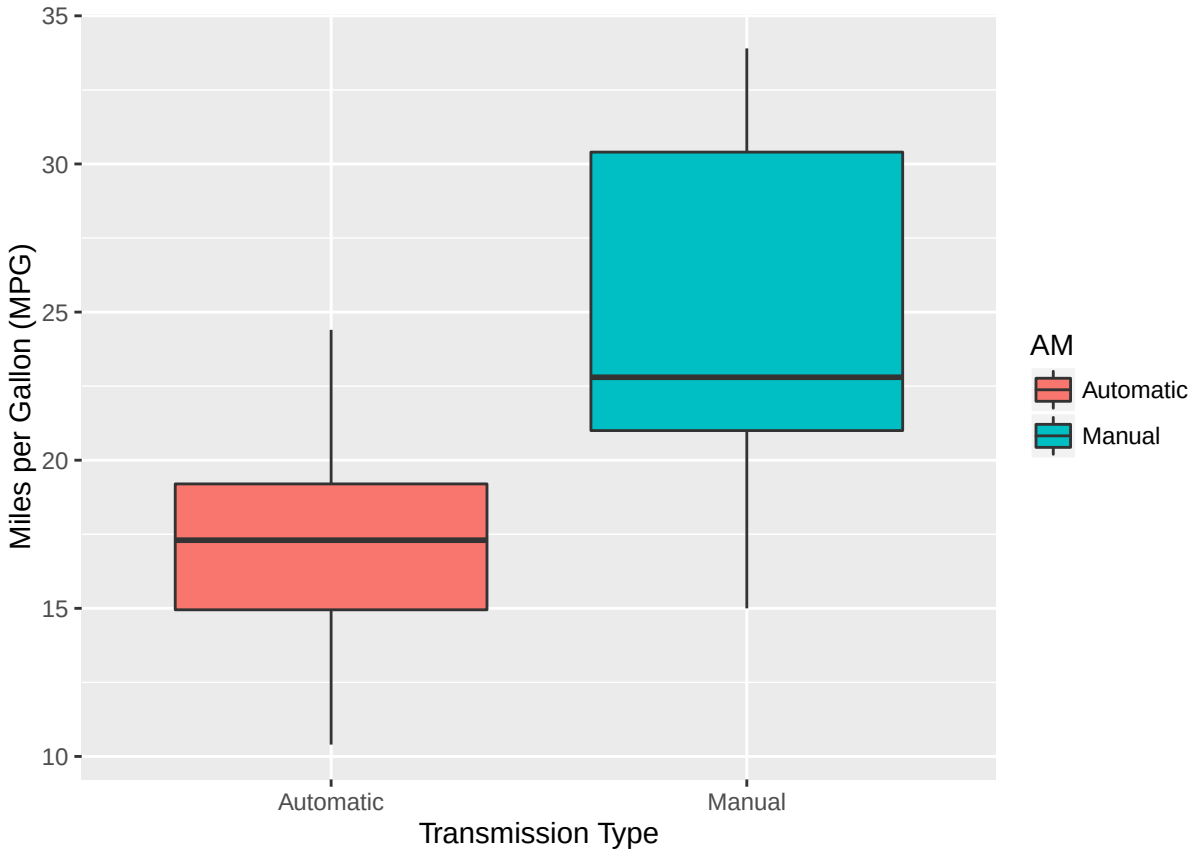
The model fitting is just the first part of the story for regression analysis, since this is all based on certain assumptions. Regression diagnostics are used to evaluate the model assumptions and investigate wheter or not there are observations with a large, undue influence on the analysis. With the reference to the plots shown in **Appendix** Figure 3., we use the `plot()` function to check:

- Linearity and Homoscedasticity: the 1*st* of the four plots ensure that the residuals are not too far away from zero. Then, they are equally spread araound the $y = 0$ line;
- Normality: the assumption is evaluated using the *QQ-plot* (plot 2) by comparing the residuals to "ideal" normal observations. These lie well along the 45-degree line;
- Homoscedasticity: in the *scale-location plot* we don't see any particular pattern in the residuals;
- Indipendence: the fourh plot is of *Residuals vs. Leverage* which argues that no outliers are present, as all values fall well within the 0.5 bands.

Appendix: Figures

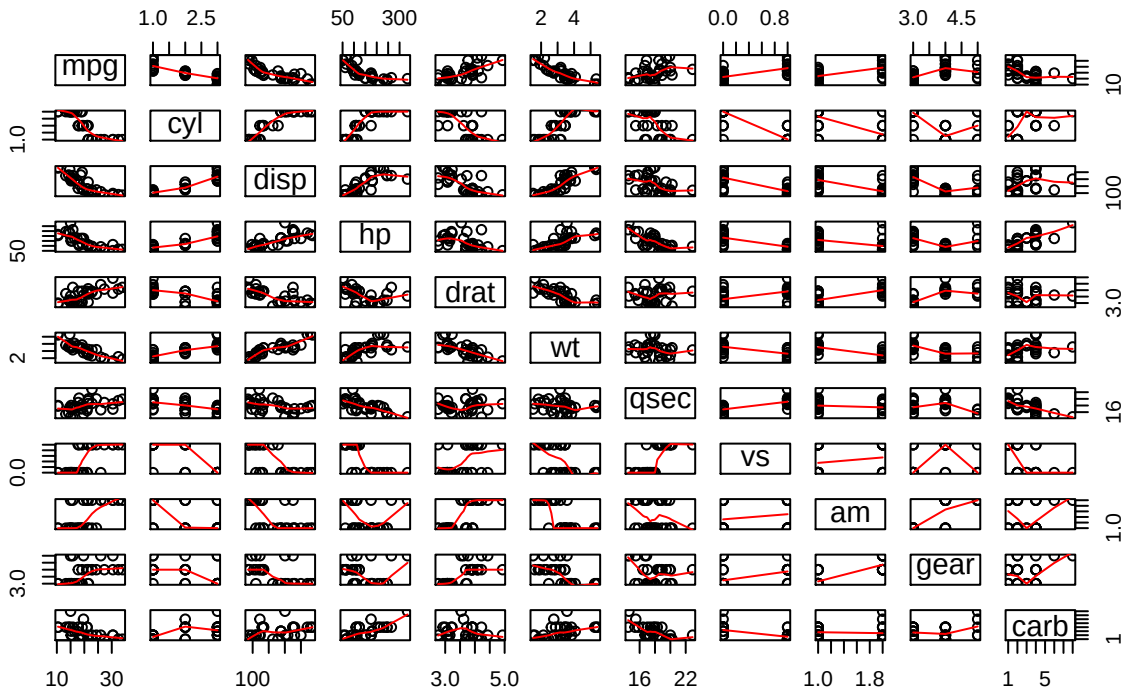
1. **Boxplot** between the `mpg` variable and the transmission type variables, `Automatic` and `Manual`:

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```



2. **Pair Graph** between all the quantities present in the `mtcars` dataset:

```
pairs(mtcars,panel = panel.smooth, main= "Pair Graph for the Motor Trend Car Road Tests")
```



3. **Diagnostic Plots**

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