Regression Models - Course Project

Pier Lorenzo Paracchini, 18.12.2015

# Executive Summary

**Automatic** models are better for mpg. Actually **automatic** models have an estimated mpg of 26.585 (miles per gallon) with a 95% confidence interval included in [23.67 - 29.5], while **manual** models have an estimated increased mpg of 5.277 (miles per gallon) over the **automatic**.

# Data Exploration and Analysis

The mtcars dataset include 32 observations of 11 features. Each onservations comprises some information for specific automobiles (1973 - 1974 models). A scatterplot matrix for all features available in the datase is available in **Appendix, Figure 1**. The focus is on mpg (miles per gallon) and am (type of transmission - 0: automatic, 1: manual).

mtcars dataset contains 19 automatic car models and 13 manual car models. The observed mpg by type of transmission am can be seen in the provided hystogram (see **Appendix, Figure 2**). From the sample data, the am (predictor) is visibly related to mpg (outcome) as we can see from the sample mean of each group. Specifically the "automatic" group has a lower sample mean (17.147 miles per gallon) than the "manual" group (24.392 miles per gallon).

# Regression Model

## Simple Model

The simple model uses mpg (as outcome) using am as the only predictor (**mpg ~ am**). The coefficients of the fitted model are

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 17.147368 1.124603 15.247492 1.133983e-15  
## am1 7.244939 1.764422 4.106127 2.850207e-04

From the Residual vs Fitted plot (see **Appendix, Figure 3**) we can see that the model is just able to predict two possible values for the estimated mpg based on the value of the predictor - **17.147 mpg for automatic models** and **24.392 mpg for manual models** with an estimated residuals standard error of **4.902 mpg**. **R- squared** indicates that this model is able to explain only **35.98%** of the total variability of the data. The Q-Q plot (see **Appendix, Figure 3**) confirms the normality of the errors (assumption).

A more detailed interpration of the simple model can be found in **Appendix, Simple Model Interpretation**.

## Extending the Model adding new features

The simple model is quite limited. Looking at the available features, other possible features could be used to identify a "better" model. A possible feature that may be valuable investigating is hp (**gross horse power** in hp). The relationship between mpg vs. hp by am can be seen in **Appendix, Figure 4**.

Using the **nested model testing technique** for the following nested models we can see that, **added feature of model 2 are necessary over model 1** (P-value < 0.05).

## Analysis of Variance Table  
##   
## Model 1: mpg ~ am  
## Model 2: mpg ~ am + hp  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 30 720.90   
## 2 29 245.44 1 475.46 56.178 2.92e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Model 2 (**mpg ~ am + hp**) seems to provide a better "description" of the response and it will be used to answer the original questions. See the Residual vs Fitted plot and Q-Q plot (see **Appendix, Figure 5**). The estimated residuals standard error of **2.909 mpg** is decreased and **R- squared** of **78.2%** is increased (compared to model 1).

**Note!!** The same process can be executed adding other features to the model in a nested fashion - investigating the overall effect of the new features on the model in an incremental way.

## Findings & Interpretation

Model 2 (**mpg ~ am + hp**) has the following coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 26.5849137 1.425094292 18.654845 1.073954e-17  
## am1 5.2770853 1.079540576 4.888270 3.460318e-05  
## hp -0.0588878 0.007856745 -7.495191 2.920375e-08

* **automatic** models (reference group - am = 0) have an estimated mpg of 26.585 (miles per gallon) with a standard error of 1.425 (miles per gallon).
* **manual** models (am = 1) have an increased estimated mpg of 5.277 (miles per gallon) (over the reference group) with a standard error of 1.08 (miles per gallon). The P-value of 0 is statistically significant, reject the **null hypothersis** (having an increase/ decrease over the reference group null).

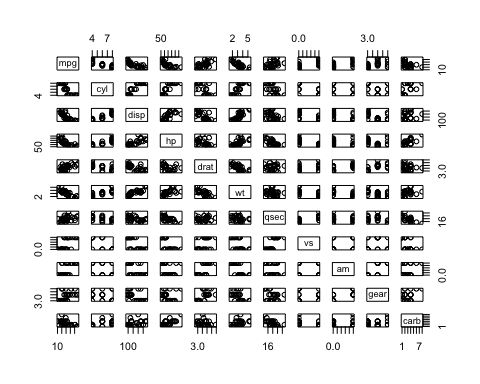
According to this very simple linear model, **automatic** transmission is better for **mpg** than **manual** transmission.

Based on the linear model previosly created we can state that

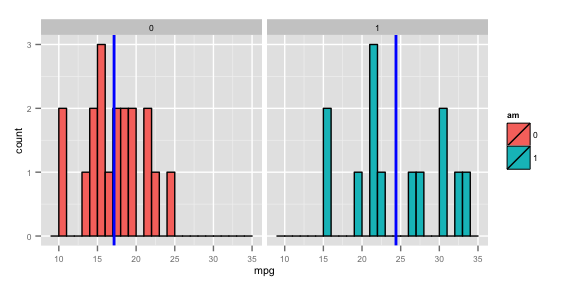
* **automatic** models use an estimated mpg of 26.585 (miles per gallon) with a 95% confidence interval included in [23.67 - 29.5] miles per gallon.
* **manual** models use an increased estimated mpg of 5.277 (miles per gallon) over the reference group with a 95% confidence interval included in [3.069 - 7.485] miles per gallon.

# Appendix

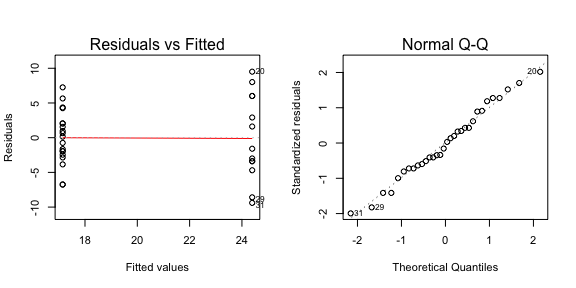
## Figure1: Scatterplot Matrix for features in mtcars dataset



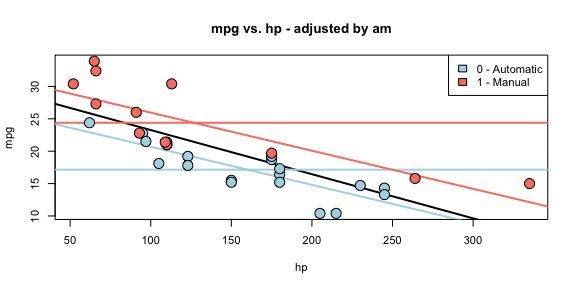
## Figure2: Hystogram of observed mpg by am (type of transmission)

Note the *blue* line represent the sample mean for each of the type of transmission (0: automatic, 1: manual).  


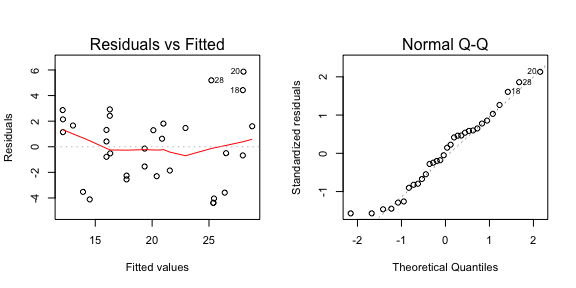
## Figure3: Residual Plots for (*mpg ~ am*) model



## Figure4: (*mpg vs. hp*) by *am* (transmission type) plots

For each type of transmission the **unadjusted** line (black), **adjusted** lines (lightblue for automatic, salmon for manual) and sample averages (horizontal) lines (lightblue for automatic, salmon for manual) are plot.  


## Figure5: Residual Plots for (*mpg ~ am + hp*) model



## Simple Model Interpretation

Looking at the coefficients of the simple model (mpg ~ am)

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 17.147368 1.124603 15.247492 1.133983e-15  
## am1 7.244939 1.764422 4.106127 2.850207e-04

* **automatic** models (reference group) have an estimated mpg of 17.147 (miles per gallon) with a standard error of 1.125 (miles per gallon).
* **manual** models have an increased estimated mpg of 7.245 (miles per gallon) (over the reference group) with a standard error of 1.764 (miles per gallon). The P-value of 0 (statistically significant) confirms that the **null hypothersis** (having an increase/ decrease over the reference group null) can be rejected.

Based on the linear model previosly created we can state that

* **automatic** models use an estimated mpg of 17.147 (miles per gallon) with a 95% confidence interval included in [14.851 - 19.444] miles per gallon.
* **manual** models use an increased estimated mpg of 7.245 (miles per gallon) over the reference group with a 95% confidence interval included in [3.642 - 10.848] miles per gallon.

Note that this model is quite limited, **R- squared** indicates that this model is able to explain only **35.98%** of the total variability of the data.