Regression Models -

Jeroen Zonneveld, June 2015

Executive Summary

Motor Trend, an automobile trend magazine is interested in exploring the relationship between a set of variables and miles per gallon (MPG) outcome. In this project, we will analyze the `mtcars` dataset from the 1974 Motor Trend US magazine to answer the following questions:

- Is an automatic or manual transmission better for MPG?
  Manual transmissions achieve a higher value of MPG compared to automatic transmission.
- Quantify the MPG difference between automatic and manual transmissions.
  This increase is approximately 1.8 MPG when switching from an automatic transmission to a manual one, with all other variables held constant.

```
library(caret)
```

Exploratory Data Analysis

The data was extracted from the 1974 Motor Trend US magazine, and comprises 11 aspects of automobile design and performance for 32 automobiles (1973-1974 models). For a description of the data set please folow this link.

```
data(mtcars)
mtcarsRAW   <- mtcars
mtcars$am   <- factor(mtcars$am,labels=c("automatic","manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$carb <- factor(mtcars$carb)
mtcars$vs   <- factor(mtcars$vs)
```

To explore the relationship of the variable under study, `mpg`, with the other variables in the dataset, we plot `mpg` against each of the other variables (see appendix, figure 1). We try to find the highly correlated variables (>0.75), so we can excluded them from the model.

```
mtcarsCorrelation=cor(mtcarsRAW)
corIndices = findCorrelation(mtcarsCorrelation, cutoff = 0.75); print(corIndices)
```

```
## [1] 2 3 1 9
```

So the variables `cyl`, `disp`, `mpg` and `am` are highly correlated. As we are interested in the effect of `am` on `mpg` they will be part of the model, we will exclude `cyl` and `disp`. We will use the step function for the selection process.

Figure 2 in th appendix shows a first comparison of the effect of the type of transmission (`am`) on the fuel consumption (`mpg`),

Regression Model

First, we build a base regression model that includes all variables as predictors for `mpg`. Next we use the step function to refine our linear model, by selecting the most appropriate variables as predictors, and eliminating those that do not.

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

As we can see in the summary below, the algorithm has selected `cyl`, `hp`, `wt` and `am` as predictors. To check residuals for normality and homoskedasticity, we plot the residuals (see appendix, figure 3). We see that the residuals are normally distributed and homoskedastic.

```r
summary(bestModel)
```

```
##
## 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 **
## ammanual     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
```

Conclusions

bestModel, has successfully explained more than 84% of the variablilty of the data. The model coefficients provide the following insights:

- Miles per gallon increases by a factor of 1.8 (1.8) with manual transmission.
- Miles per gallon decreases by a factor of 0.03 (-0.03) as horspower increases.
- Miles per gallon decreases by a factor of 2.5 (-2.5) for every increase of 1000 lb in weight.
- Miles per gallon decreases by a factor of 3.03 (-3.03) for 6 cylinders and by a factor of 2.16 (-2.16) for 8 cylinders.
- The intercept is at 33.7 mpg.

The overall p-value is very small (1.506 x 10 ^ -10) which means that there is a very small uncertainty for the sign of the coefficients.

Appendix

Figure 1 - Relationships of the different variables

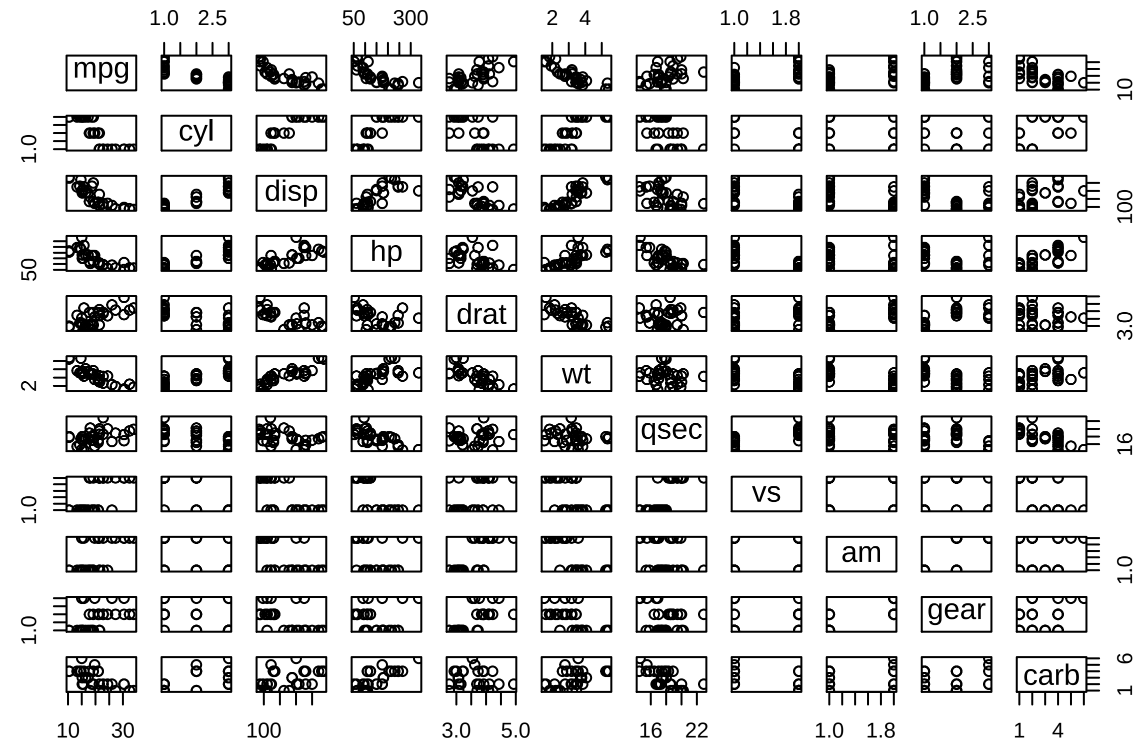

Figure 2 - MPG vs. Transmission Type

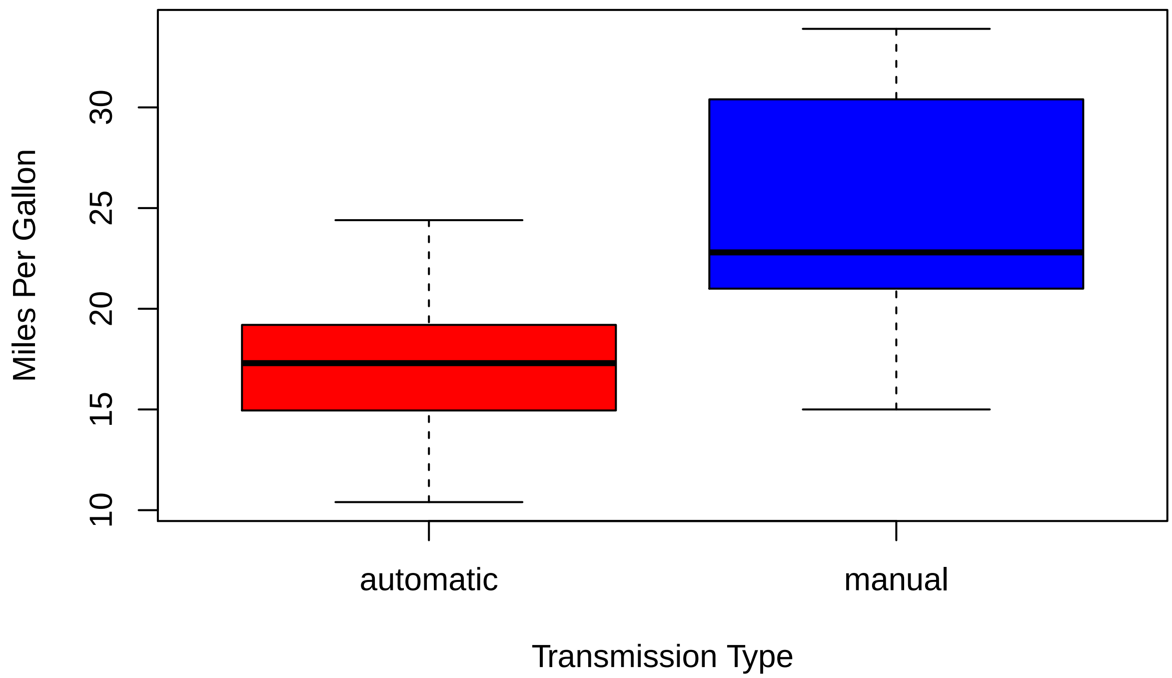

Figure 3 - Residuals