Regression Models Data Science Coursera R Project

Philippe-C - 22 March 2015 (using Rmd and Knitr)

Executive summary:

We have used the "mtcars" dataset from the 1974 "Motor Trend" US magazine to answer two questions:

- Is an automatic or manual transmission better for miles per gallon (MPG)?
- How different is the MPG between automatic and manual transmissions?

Based on a linear regression analysis, we show that there is a statistically significant difference between the mean MPG for automatic("am"=0) and manual transmission cars("am"=1). Manual transmissions allows a higher value of MPG compared to automatic transmission. This increase is approximately 7 MPG (+1.8 MPG when adjusted for other variables). The details of the complete analysis are presented below.

Exploratory Data analysis:

A) Dataprocessing:

It requires loading the dataset and transforming certain 'numeric' variables as 'factors' (see R code below).

```
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Looking at (**Appendix - Figure 1**), we see that the variables "cyl", "disp", "hp", "drat", "wt", "vs" and "am" have a strong correlation with "mpg". We are **mainly** interested in the effects of car transmission type on "mpg". Therefore, we look at the distribution of "mpg" for each level of "am" (Automatic or Manual) using a box plot (**Appendix - Figure 2**). This plot emphasizes that manual transmissions tend to have higher MPG.

B) Statistical inference:

We have performed a t-test making the null hypothesis that the MPG of the automatic and manual transmissions come from the same distribution (assuming the transmissions data have a normal distribution).

```
test<- t.test(mpg ~ am, data = mtcars)
test$p.value
```

```
## [1] 0.001373638
```

```
test$estimate
```

```
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Since the p-value (0.001374) is largely below 5%, we reject the null hypothesis. So, the automatic and manual transmissions MPG distributions are not the same. The difference in means is close to 7.24 MPG.

C) Regression Analysis, analysis of variance and model residuals:

We have used different models. The first model simply includes all variables as predictors. In order to build a "best model" (i.e. select the significant predictors of "mpg"), we use the R 'step()' function which calls 'lm()' repeatedly to build multiple regression models and select the best variables from them using both forward selection and backward elimination (AIC algorithm).

```
firstmodel <- lm(mpg ~ ., data = mtcars)
bestmodel <- step(firstmodel, direction = "both")
```

The "best model" shows that variables, "cyl", "wt" and "hp" are confounding variables and "am" the independent variable.

```
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
```

Looking at the adjusted R-squared and p-value (<5%), we can conclude with statistical significance that 84% of the variability in "mpg" is explained by the "best model" predictors. Then, we compare the "best model" with a "base model" including only the variable "am".

```
basemodel <- lm(mpg ~ am, data = mtcars)
anova(basemodel, bestmodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The critical value being highly significant, we reject the null hypothesis, in other words the variables "cyl", "hp", and "wt" contribute to the precision of the model. Looking at the residuals plots (**Appendix - Figure 3**), we can verify the independence conditions since the points are randomly scattered in the residuals vs fitted plot. The QQ plot indicates that the residuals are (at least approximately) normally distributed. Furthermore, the constant variance hypothesis is respected (scale-location plot). Finally, the Residuals vs. Leverage plot shows that no outliers are present, as all values fall well within the 0.5 bands. So, we can consider that the classical assumptions of the theory of linear regression are verified. Moreover, we can proceed with some regression diagnostics using the 'hatvalues()' and 'dfbetas()' functions.

```
leverage <- hatvalues(bestmodel)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```
influential <- dfbetas(bestmodel)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

As these results point out the same cars as in the residuals plots, they confirm that the analysis conducted in this project were correct.

Conclusions:

According to the "best model" results in section C (see also, *Appendix - Figures 1 and 4*), we can say with statistical significance (p-values<5%) that:

1. Manual transmission cars performed better in our dataset (1.8 more miles per gallon adjusted for "hp"", "cyl" and "wt").
2. MPG are negatively related to weight (-2.5 for every 1000 lbs in "wt").
3. MPG are negatively related to horse power (-0.32 for every increase of 10 in "hp").
4. Cylinders impact negatively MPG (adjusted for "hp", "wt" and "am", when "cyl" increase from 4 to 6, MPG decrease by a factor of 3).

Appendix:

Figure 1 - Pairs plot for the "mtcars" dataset

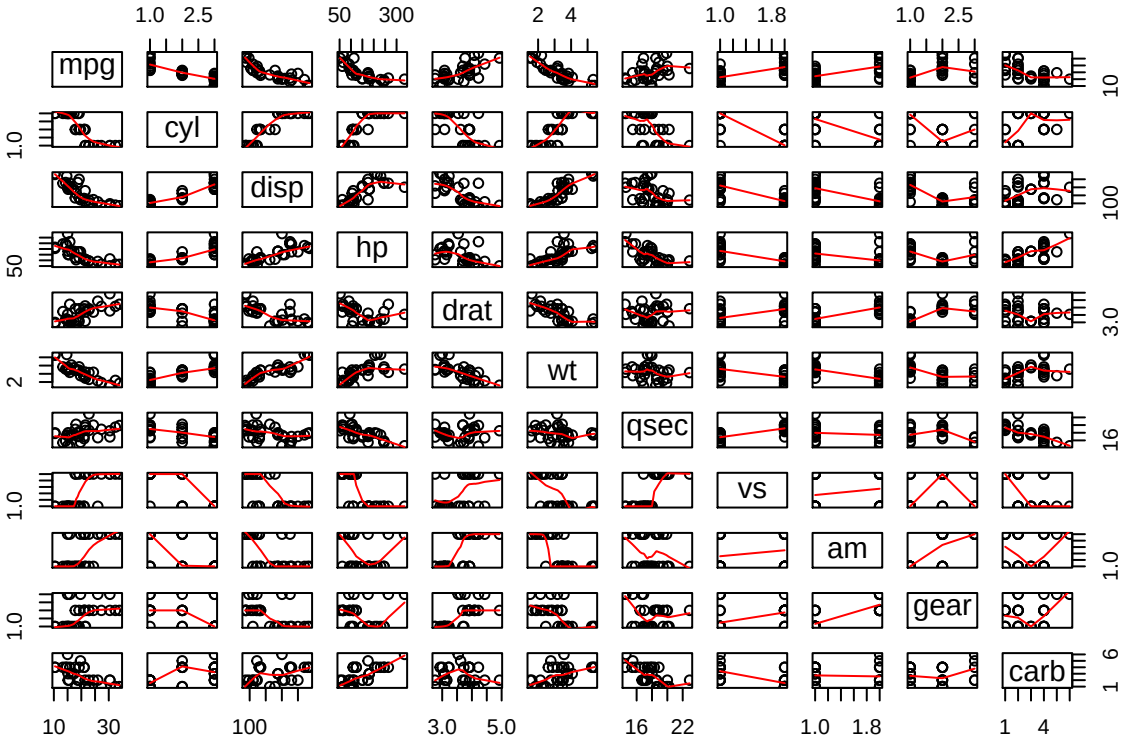

Figure 2 - Boxplot of MPG by transmission type

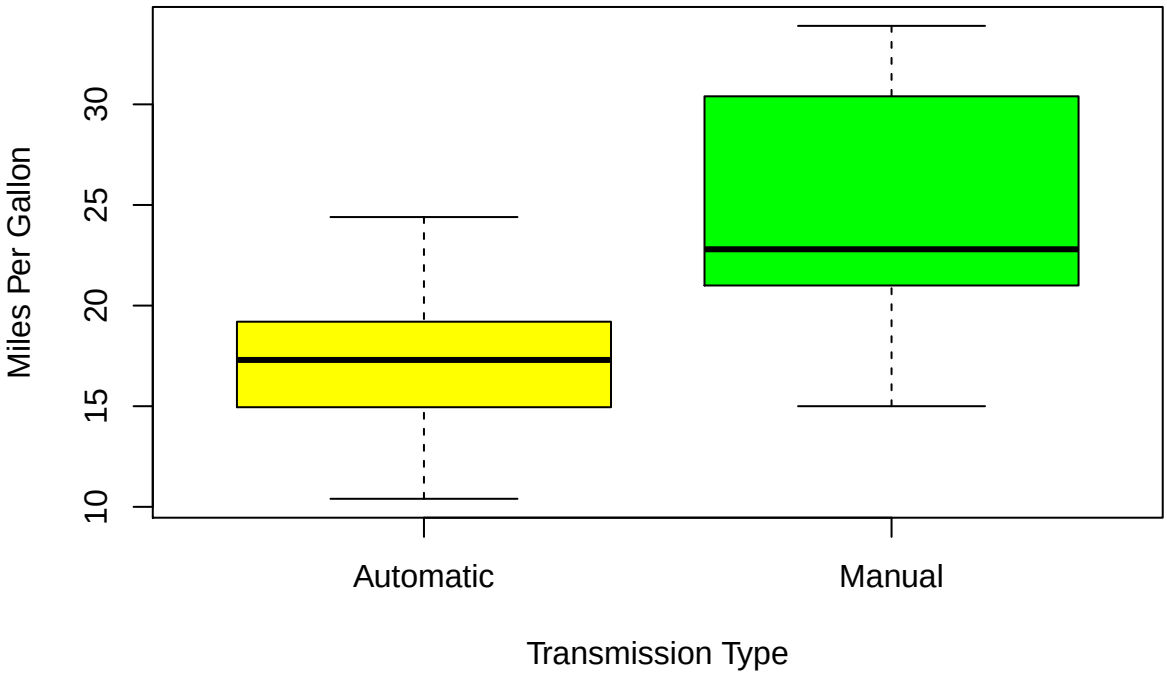

Figure 3 - "Best model" residuals plots

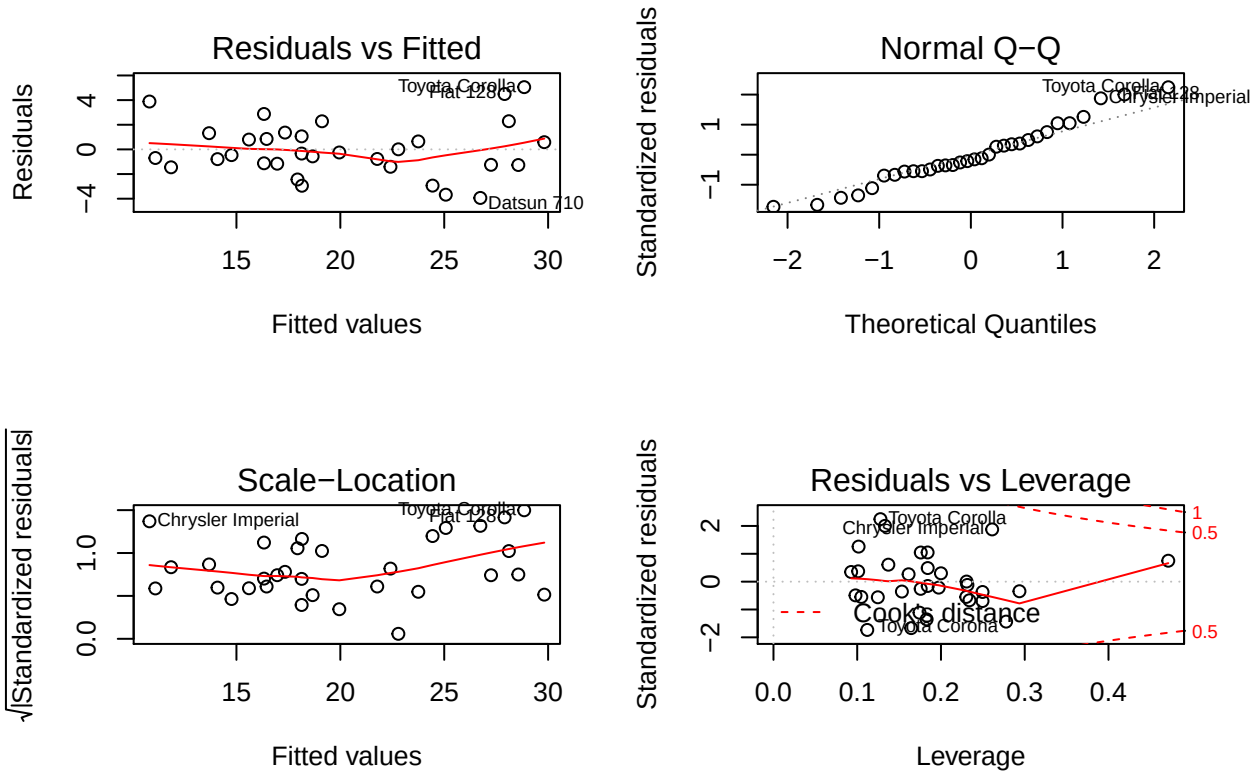

Figure 4 - Scatter Plot of MPG as a function of Weight