Regression Model | Course Project

Aaron Ran An

August 21, 2015

Executive Summary

Context:

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

"Is an automatic or manual transmission better for MPG" "Quantify the MPG difference between automatic and manual transmissions"

Questions:

Take the mtcars data set and write up an analysis to answer their question using regression models and exploratory data analyses.

Written as a PDF printout of a compiled (using knitr) R markdown document. Brief. Roughly the equivalent of 2 pages or less for the main text. Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures. Include a first paragraph executive summary.

Data Management and Exploratory Analysis

First step is to structure the data into a regression-friendly way. Therefore every categorical variable in the mtcars dataset is transformed into factor.

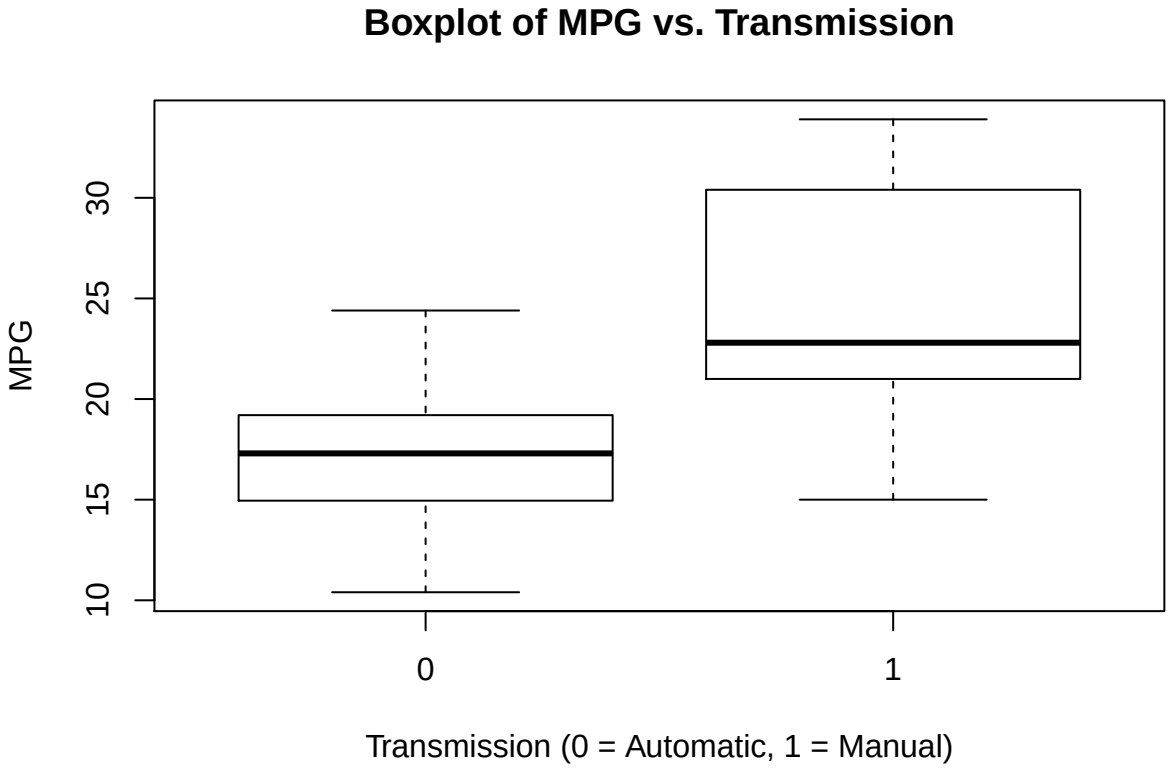
```
data("mtcars")

mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

Section 1: Inference and ANOVA

In order to explore if the automatic and manual transmission group have significantly different MPG value, our first attempt would be exploratory analysis and ANOVA.

In the exploratory analysis, a boxplot is used below to illustrate the difference of MPG between these two groups.



From the plot, we can visually infer that the Manual Transmission group has higher mean of MPG value than the Automatic group.

Next an ANOVA procedure is used to test if this visual different is really significantly different.

```
fit <- aov(mpg ~ factor(am), data=mtcars)

TukeyHSD(fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mpg ~ factor(am), data = mtcars)
##
## $`factor(am)`
##         diff     lwr      upr    p adj
## 1-0 7.244939 3.64151 10.84837 0.000285
```

The Tukey's HSD test yields a p-vaule of 0.000285, which suggests the group difference between two levels of transmission is significantly different from each other. Therefore, we can answer the first question.

Section 2: Regression Analysis

After we determined the group difference between two transmission levels is significant, we need to quatify that difference using regression models.

Step 1: simply fit a model on the mpg and transmission.

Our first model would be simply be fitted on the MPG and different transmission level.

```r
sim_mod <- lm(mpg ~ factor(am), data = mtcars); summary(sim_mod);
```

This simple model only achieved an R square of 0.3598, which means there are about 64% of variance is left unexplained or could be attributed to other variables, so we continue to fit the full model.

```r
full_mod <- lm(mpg ~ ., data = mtcars); summary(full_mod);
```

The full model achieved 0.779 adjusted R square but none of the predictors is significant. The reason could be the inclusion of too many un-important varibales. So next step we will use stepwise selection to select the relevant variables.

```r
step(full_mod)

step_mod <- lm(mpg ~ cyl +  hp + wt + am, data = mtcars); summary(step_mod);
```

Stepwise left us with four variables: cyl, hp, wt and am. The model is better this time, with three variables/levels being significant and adjusted R square of 0.8401.

In order to inspect the potential two-way interaction among the varaibles, we next fit an regression on all two-way interaction on the model, and use step wise again to select the relevant variables.

```r
int_step_mod <- lm(mpg ~ (cyl +  hp + wt + am)^2, data = mtcars)

step(int_step_mod)
```

The previous stepwise suggest there are only two interactions, between cyl / hp and between wt /am. We will add those two interactions to make the final models.

In order to make sure the inclusion of the two interation is really neccessary, I added a drop1 test to see if dropping any varibale would incur significant loss. The result sugguest that the interaction between cyl and hp could be dropped. Therefore we have our final model. The ANOVA test after that also suggest that including the interaction between cyl and hp is not necessary.

```r
sec_final_mod <- lm(mpg ~ cyl + hp + wt + am + cyl:hp + wt:am, data = mtcars)

drop1(sec_final_mod, test="F")
```

```
## Single term deletions
##
## Model:
## mpg ~ cyl + hp + wt + am + cyl:hp + wt:am
##        Df Sum of Sq    RSS    AIC F value  Pr(>F)
## <none>              108.53 57.082
## cyl:hp  2    21.939 130.47 58.973  2.3247 0.12036
## wt:am   1    19.049 127.58 60.257  4.0368 0.05639 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
final_mod <- lm(mpg ~ cyl + hp + wt + am + wt:am, data = mtcars)

anova(final_mod, sec_final_mod)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am + wt:am
## Model 2: mpg ~ cyl + hp + wt + am + cyl:hp + wt:am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     25 130.47
## 2     23 108.53  2    21.939 2.3247 0.1204
```

```
summary(final_mod)
```

The final model achieved 0.86 adjusted R square, which is better than the full model and the step model. Also the final model is simple and easy to understand so the parsimonious is prioritized so we can clearly quantify the effect on different levels of transmission.

The results sugguest