Regression Model | Course Project

Aaron Ran An

August 21, 2015

Executive Summary

In this study, we aim to explore the relatonship between various variables and miles per gallon (MPG), using the mtcars dataset in the R default. The question we want to answer is whether different tansmission levels, automatic or manual, has an effect on the miles per gallon for a car. Furthermore, we want to quantify that difference using regression models.

Specifically, we began analysis by using exploratory analysis and ANOVA to determine there is a statistically significant difference in MPG for different tranmission levels. Then we fit several multiple linear regression and select the best one to quatify that diiferent. In conclusion, holding the number of cylinders, gross horsepower and weight constant, cars with manual transmission add 9.89860 + (-3.14499)*wt more MPG on the average than cars with automatic transmission. Thus, cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values. A series of model diagnostics plots are produced at the end of this report.

Section 1: 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.

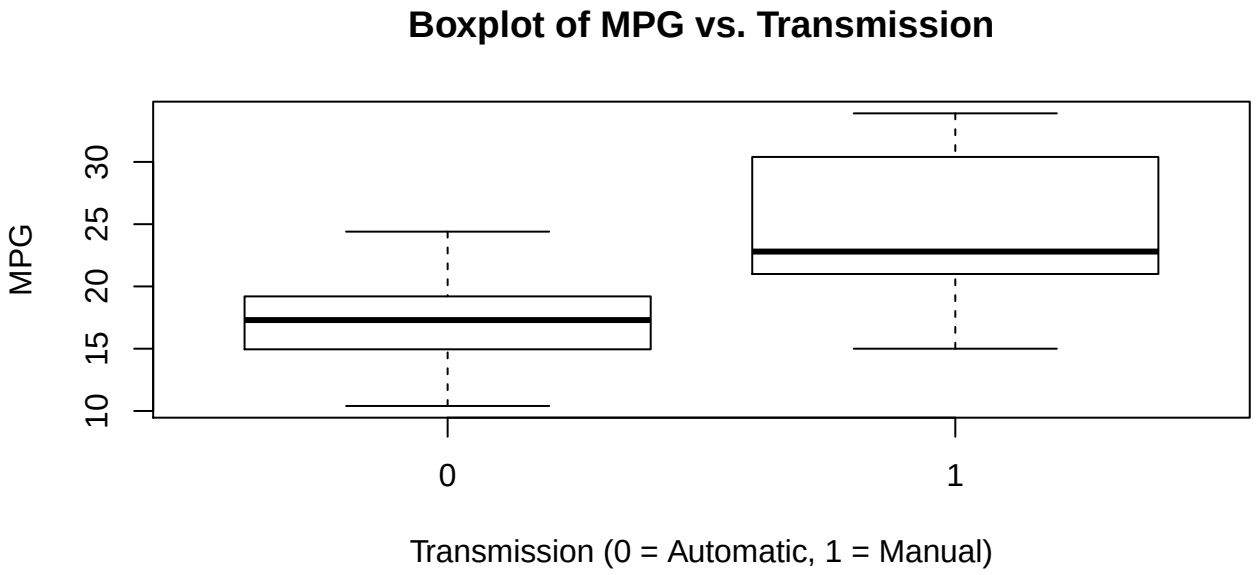
```
data("mtcars")

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 2: 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 boxplot, 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 determine the MPG difference between different tranmission levels is significant.

Section 3: Regression Analysis

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

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

```
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.

```
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.

```
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.

```
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.

```
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
```

```
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
```

```r
summary(final_mod)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 30.65246564 2.90990761 10.533828 1.111073e-10
## cyl6        -2.38062017 1.37363506 -1.733081 9.540117e-02
## cyl8        -2.89910832 2.19666356 -1.319778 1.988677e-01
## hp          -0.01781723 0.01484303 -1.200377 2.412443e-01
## wt          -2.20686780 0.85204530 -2.590083 1.577795e-02
## am1          9.89860309 4.28571455  2.309674 2.944922e-02
## wt:am1      -3.14498752 1.58475347 -1.984528 5.827576e-02
```

```r
confint(final_mod)
```

```
##                   2.5 %      97.5 %
## (Intercept) 24.65939873 36.64553256
## cyl6        -5.20967454  0.44843419
## cyl8        -7.42322161  1.62500497
## hp          -0.04838703  0.01275257
## wt          -3.96168794 -0.45204766
## am1          1.07200875 18.72519743
## wt:am1      -6.40884840  0.11887335
```

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 sugguests that holding cyl, hp and wt constant, cars with manual transmission add 9.89860 + (-3.14499)*wt more MPG on the average than cars with automatic transmission.

The aforementioned interaction also suggest the difference between MPG that due to transimission difference diminish as the weight of car gets bigger. That is, the heavier the car, the less different their MPG level due to different transmission. This could be revealed from the scatter plot in next section.

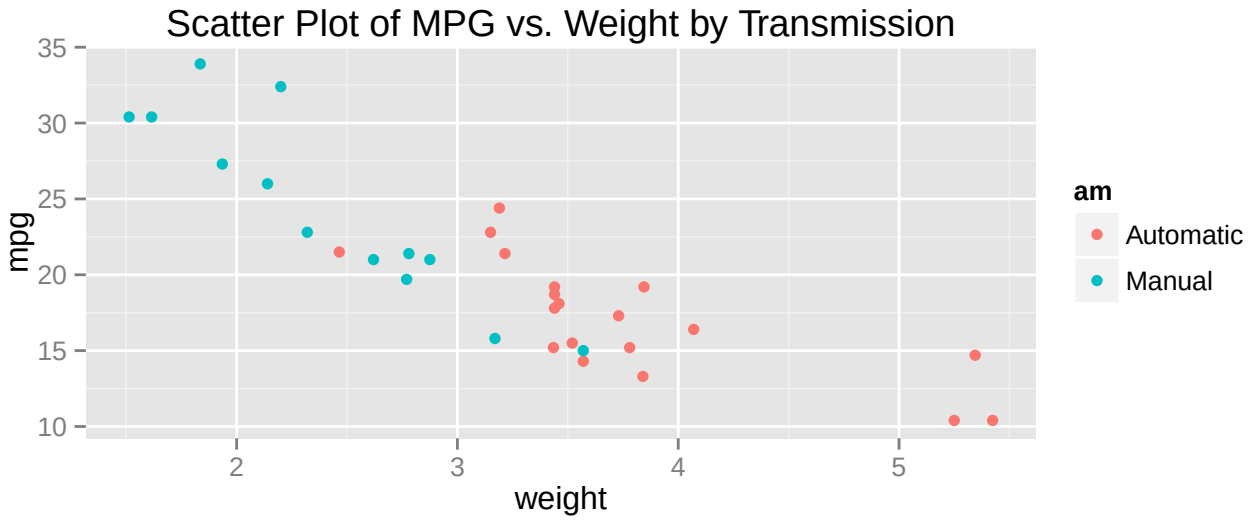
For instance, holding other variables (cyl and hp) constant, a manual transmitted car that weighs 2000 lbs have 3.60862 more MPG than an automatic transmitted car that has both the same weight.

Section 4: Model Dignostics and Supporting Figures

1. Scatter Plot

```
##
## Attaching package: 'ggplot2'
##
## The following object is masked from 'mtcars':
##
##     mpg
```

```r
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
```



2. Residual Plot

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