Regression Models Course Project

George Liu

September 24, 2015

Executive Summary

This report is part of the JHU Coursera Regression Models course project. In this project, regression and exploratory data analysis are done on the “mtcars”" data set to answer two questions:

1. “Is an automatic or manual transmission better for MPG”
2. “Quantify the MPG difference between automatic and manual transmissions”

Based on the analysis, automatic transmission provides a better mpg, on average, it’s 1.8 mpg.

Exploratory Data Analysis and Hypothesis Testing

We start by loading required packages and make the categorical variables factors:

```r
library(datasets); library(car); library(dplyr); library(caret); library(ggplot2); data("mtcars")
cars <- mtcars
cars$cyl <- factor(cars$cyl)
cars$vs <- factor(cars$vs)
cars$am <- factor(cars$am, labels = c("automatic", "manual"))
cars$gear <- factor(cars$gear)
cars$carb <- factor(cars$carb)
```

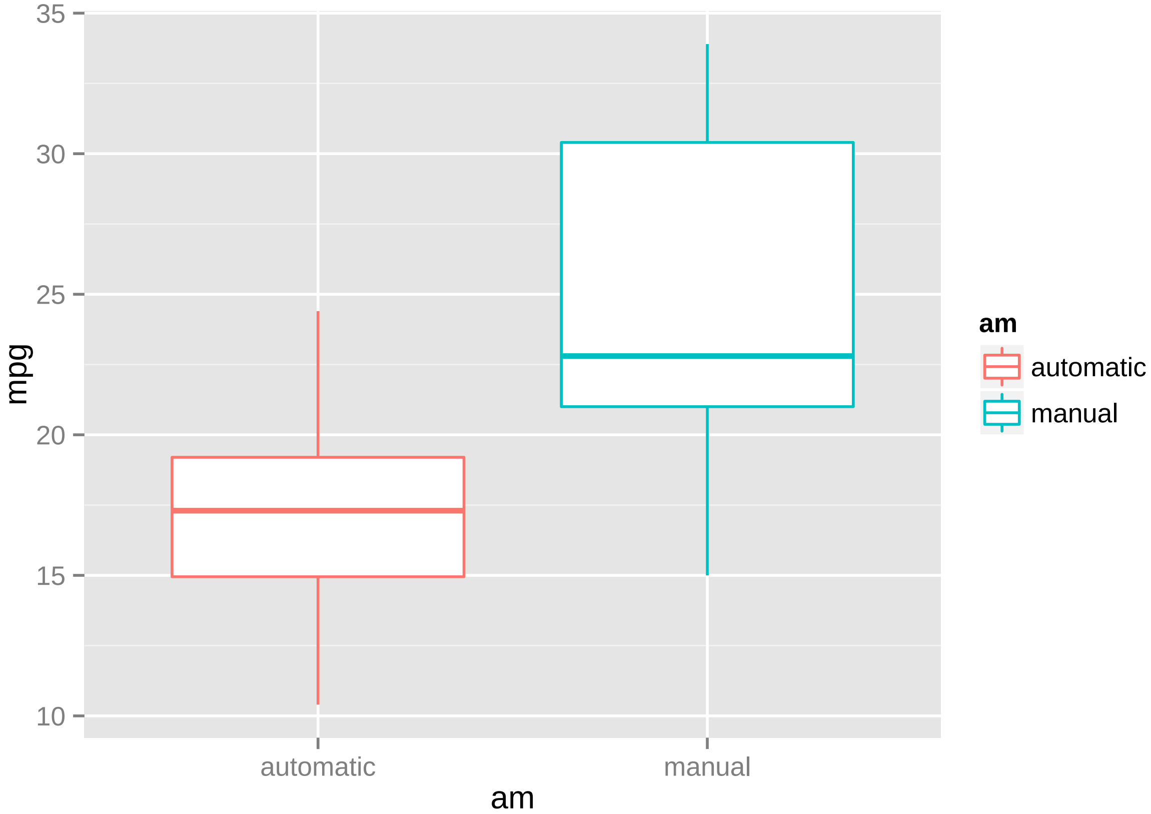
Now we do some exploratory data analysis. First, let’s look at the correlation among variables. This is important as highly correlated predictors will lead to multicollinearity in regression, which will affect the accuracy of coefficients. (See appendix)

```r
fit.all <- lm(mpg ~ ., data = cars)
vif(fit.all)
```

```
##            GVIF Df GVIF^(1/(2*Df))
## cyl  128.120962  2        3.364380
## disp  60.365687  1        7.769536
## hp    28.219577  1        5.312210
## drat   6.809663  1        2.609533
## wt    23.830830  1        4.881683
## qsec  10.790189  1        3.284842
## vs     8.088166  1        2.843970
## am     9.930495  1        3.151269
## gear  50.852311  2        2.670408
## carb 503.211851  5        1.862838
```

Second, we do some plotting here to explore the relationship between transmission type and mpg:

```
g <- ggplot(cars, aes(x = am, y = mpg, color = am))
g <- g + geom_boxplot(); g
```



Clearly, we see a difference in mpg between auto and manual transmissions. Below is a t-test for some inference:

```
t.test(cars$mpg[mtcars$am == 0], cars$mpg[mtcars$am == 1])
```

```
##
##  Welch Two Sample t-test
##
## data:  cars$mpg[mtcars$am == 0] and cars$mpg[mtcars$am == 1]
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

This tells us that the difference is statistically significant, and with 95% confidence, it's between -11.3 and -3.2, which is average auto mpg minus average manual mpg.

Modeling Building, Selection and Diagnotics

We now fit several models to quantify the difference using linear regression:

```
fit.am <- lm(mpg ~ am, data = cars)
fit.all <- lm(mpg ~ ., data = cars)
fit.step <- step(fit.all, direction = "both", trace = 0)
anova(fit.am, fit.step, fit.all)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 17.7489 1.476e-05 ***
## 3     15 120.40 11     30.62  0.3468    0.9588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

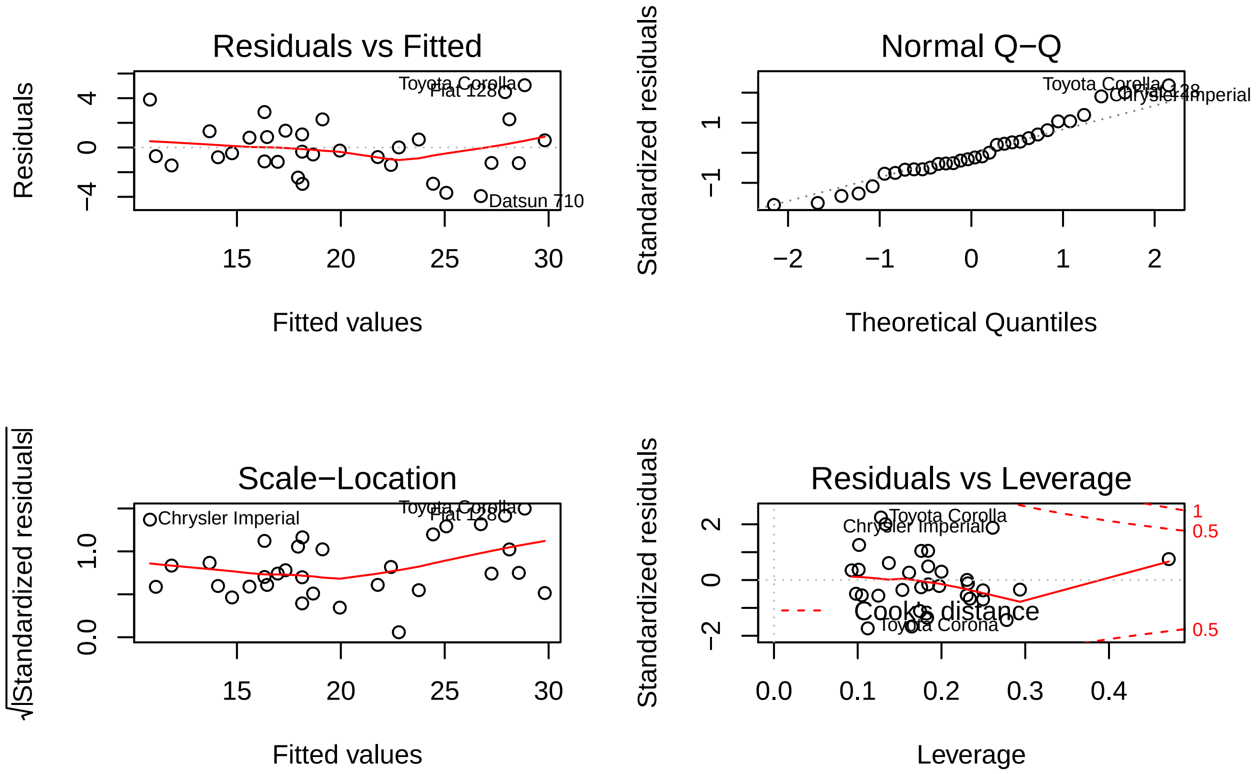
Here, it indicates the step method generated model is significantly different than the single variable model. However, using my automotive knowledge, I don't see the need to include the hp predictor, since like all the other predictors removed by the step method, hp is a result of the number of cylinders, not a cause. Therefore, I now construct my own model that only links the result mpg with the cause, and rules out any correlated variables that're equivalent to the cause.

```
fit.my <- lm(mpg ~ am + cyl + wt + carb, data = cars)
anova(fit.my, fit.step)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt + carb
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     22 163.89
## 2     26 151.03 -4    12.865
```

However, the adjusted R-Squared is smaller, so we'll keep the step method generated model. Below, a plot is shown to examine the conditions of the fit are met:

```
par(mfrow = c(2, 2))
plot(fit.step); summary(fit.step)
```

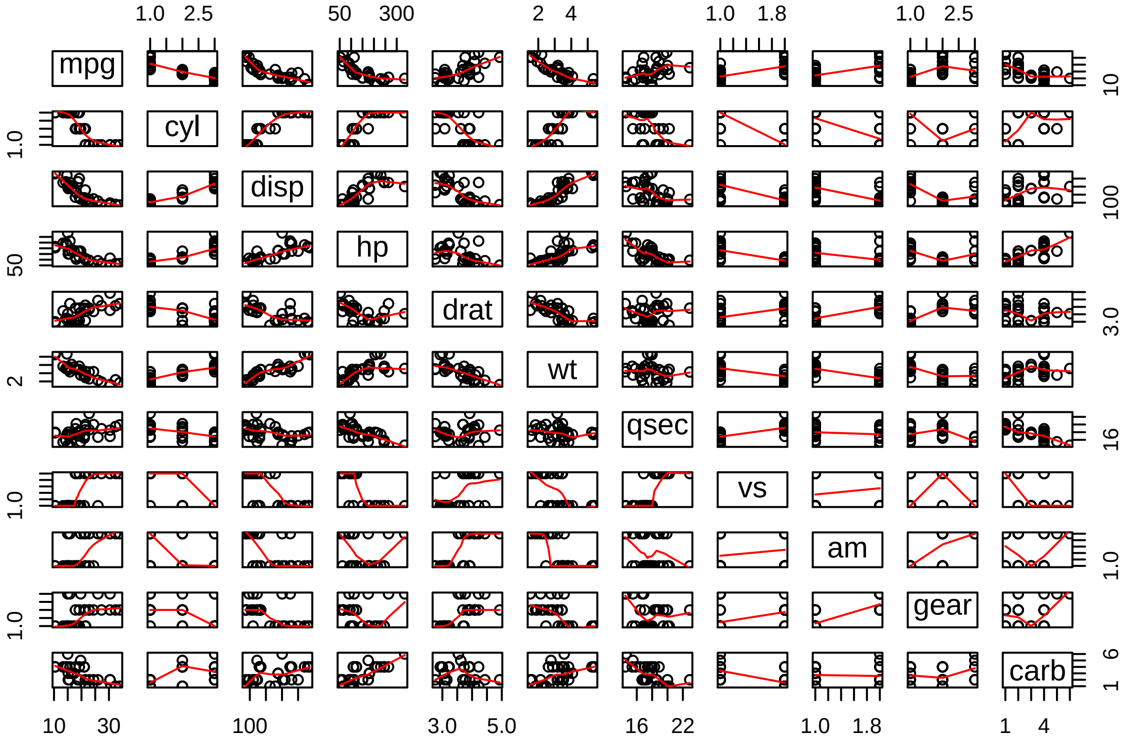


```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## 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
```

Conclusion

Based on the regression analysis, we can see that on average, manual transmission has a higher mpg than automatic which is about 1.8. Number of cylinders also have a strong impact on mpg, while hp only marginally influences mpg.

Appendix


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb