Regression Models Course Project

Trieu Tran

October 17, 2015

Executive Summary

This is the course project for the Regression Models class. The goal of this project is to analyze the `mtcars` data set and the relationship between a set of variables within that data set and miles per gallon (MPG) as the dependent variable. Two particular questions of interest are:

```
    * Is an automatic or manual transmission better for MPG?
    * What is the quantified MPG difference between automatic and manual transmissions?
```

The analysis shows us that the average MPG of manual tranmission cars is slightly higher than automatic transmission cars. Holding other variables constant, manual transmission cars yield **0.18** higher MPG than automatic cars.

Data Exploratory

```
require(datasets); data(mtcars); require(GGally); require(ggplot2); require(car)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

First, we do some basic data exploration to see the relationship between variables of the data set. **Firgure 1-1** (Appendix) is a box plot that shows the different in MPG of two different groups of cars, based on their transmissions. From the plot, it appears that manual transmission cars have a better MPG than automatic transmission cars. Mean values are **24.392**, for manual cars and **17.147**, for automatic cars.

A pairs plot, which shows us the relationship between variables, (see Figure 1-2 in the Appendix) can tell us that variables like `disp` , `wt` and `cyl` are highly correlated with the dependent variable mpg. The variables `wt` and `disp` have high correlations. It makes common sense since a larger car (or a higher displacement car) will have a higher weight. Thus we just need to include one of these two variables in building our regression models; and in this case, we pick `wt` for its higher correlation to `mpg` . However, the questions of interest focus on how transmission type effects on the mpg. We include variable `am` in our model.

Inference analysis

We perform a t-test on two group of cars; one with manual transmissions and the other with automatic transmissions. In doing this test, we assume that the data is normally distributed and that the two groups of cars are unpaired and have different variances.

```
t.test(mpg ~ factor(am), data = mtcars, paired = FALSE, var.equal = FALSE, alternative="two.sided")
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mpg by factor(am)
## 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 in group 0 mean in group 1 
##        17.14737        24.39231
```

We gather from the test results that the p-value is **0.001374** which is very small; and the 95% confidence interval is from **-11.280194** to **-3.209684** which doesn't contain 0. So far, we may say manual transmission cars yield higher mpg than automatic transmission cars.

Regression analysis

```
# nesting models
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- update(fit1, mpg ~ factor(am) + wt)
fit3 <- update(fit1, mpg ~ factor(am) + wt + cyl)
anova(fit1, fit2, fit3)
```

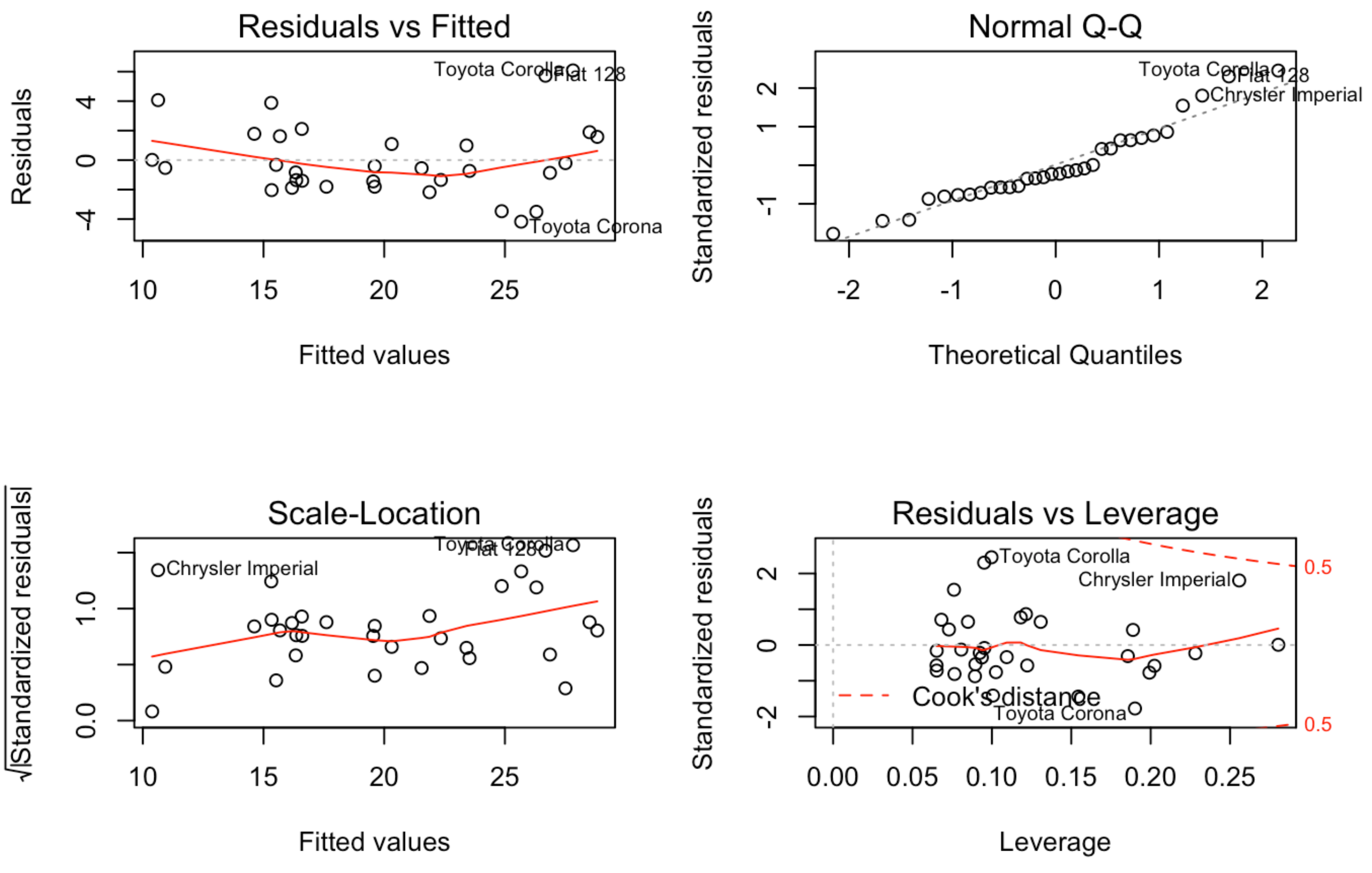
```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ factor(am) + wt + cyl
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 64.864 9.051e-09 ***
## 3     28 191.05  1     87.27 12.791  0.001292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `fit3` model which consists of the variables `wt` and `cyl` as cofounders and `am` as the the indepedent variable is the best model. The results from the above ANOVA test demonstrates `fit3` achieves the most significant p-value.

Adjusted R-squared of `fit3` is **0.8122** (see Output 3 - Appendix). This is almost as high as the adjusted R-squared of the model which includes all variables, **0.8066** (see Output 4 - Appendix). Therefore, we can confidently report that `fit3` model captures 82% of variability and it is the best model.

Based on the summary outpuf of `fit3` (Output 3 - Appendix), we conclude from our best fit model `fit3` that cars with manual transmission get slightly better **0.18** MPG than cars with automatic transmission.

```
par(mfrow = c(2, 2))
plot(fit3)
```



- In the upper left hand plot, Residuals Vs Fitted, data points lie under and above zero, excepts some outliers, and seem not to have any obvious pattern. It supports the indepedent condition of the data.
- Normal Q-Q plot indicates the data is normally distributed.

The above observations reaffirm our conclusion of the `fit3` model.

Appendix

Output 1

```
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Output 2

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## factor(am)1 -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

Output 3

```
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## factor(am)1   0.1765     1.3045   0.135  0.89334
## wt           -3.1251     0.9109  -3.431  0.00189 **
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

Output 4

```
newCars = mtcars
newCars$am <- factor(newCars$am)
fitAll <- lm(mpg ~ ., data = newCars)
summary(fitAll)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = newCars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4506 -1.6044 -0.1196  1.2193  4.6271 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 12.30337   18.71788   0.657   0.5181  
## cyl         -0.11144    1.04502  -0.107   0.9161  
## disp         0.01334    0.01786   0.747   0.4635  
## hp          -0.02148    0.02177  -0.987   0.3350  
## drat         0.78711    1.63537   0.481   0.6353  
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739  
## vs           0.31776    2.10451   0.151   0.8814  
## am1          2.52023    2.05665   1.225   0.2340  
## gear         0.65541    1.49326   0.439   0.6652  
## carb        -0.19942    0.82875  -0.241   0.8122  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066 
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Figure 1-1: Transmision Type Vs. MPG

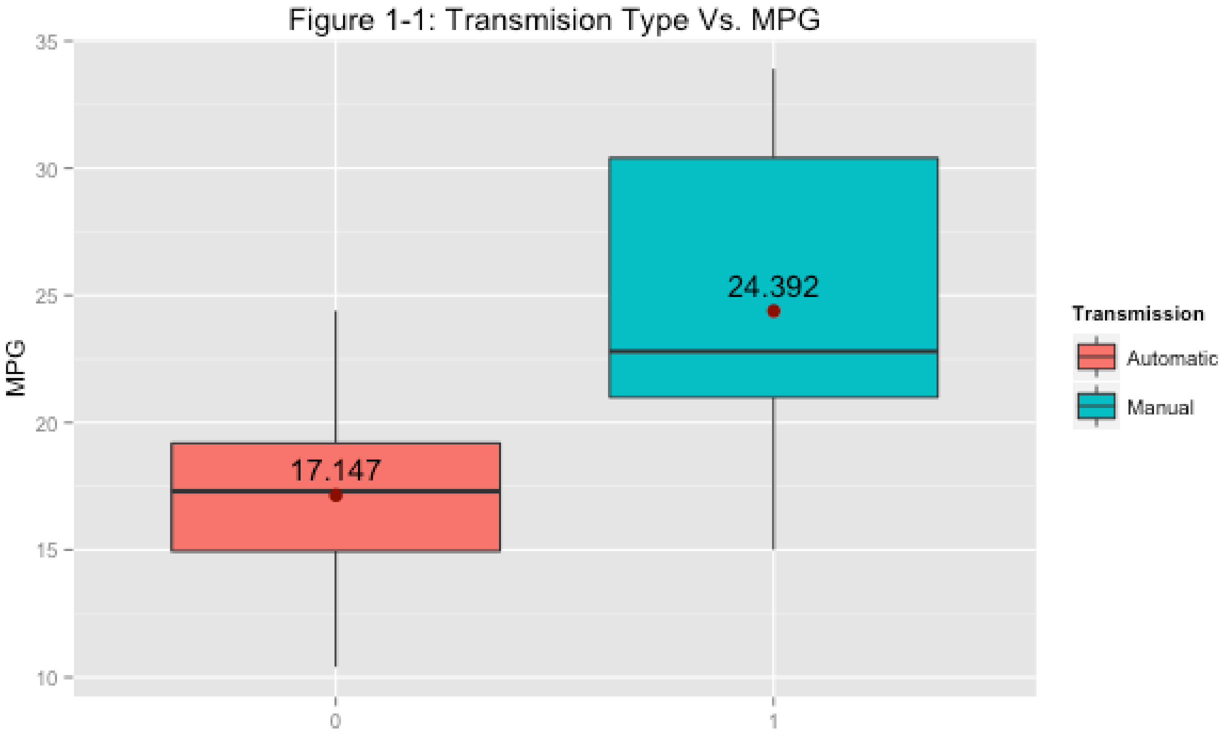
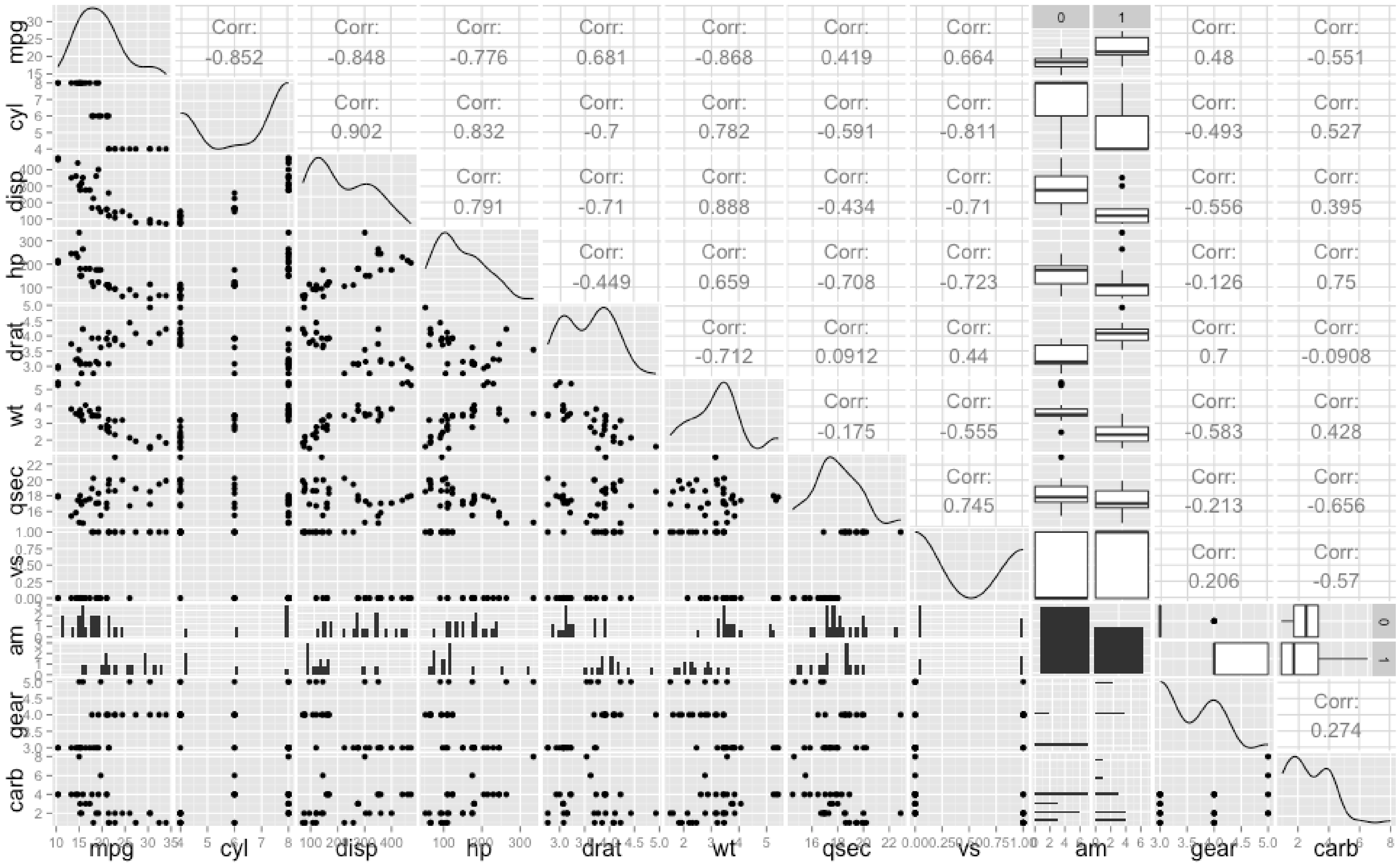


Figure 1-2: Mtcars


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
Corr: -0.852
Corr: -0.848
Corr: -0.776
Corr: 0.681
Corr: -0.868
Corr: 0.419
Corr: 0.664
Corr: 0.48
Corr: -0.551
Corr: 0.902
Corr: 0.832
Corr: -0.7
Corr: 0.782
Corr: -0.591
Corr: -0.811
Corr: -0.493
Corr: 0.527
Corr: 0.791
Corr: -0.71
Corr: 0.888
Corr: -0.434
Corr: -0.71
Corr: -0.556
Corr: 0.395
Corr: -0.449
Corr: 0.659
Corr: -0.708
Corr: -0.723
Corr: -0.126
Corr: 0.75
Corr: -0.712
Corr: 0.0912
Corr: 0.44
Corr: 0.7
Corr: -0.0908
Corr: -0.175
Corr: -0.555
Corr: -0.583
Corr: 0.428
Corr: 0.745
Corr: -0.213
Corr: -0.656
Corr: 0.206
Corr: -0.57
Corr: 0.274