Regression Models Course Project

Jure Bordon

Sunday, April 26, 2015

Executive Summary

This report is made for Regression Models Coursera course from the Data Science signature track by Johns Hopkins University. We analyze the mtcars data set and try to determine the relationship of various variables and miles per gallon (MPG). The main questions are:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions!

Our findings show that **manual transmission is** statistically significantly **better than automatic transmission** when it comes to MPG. In addition, we compared several other regression models with different combinations of variables. The conclusion that manual transmission is better holds even for the new models which we tested.

Analysis

Is an automatic or manual transmission better for MPG?

First load the data and since we sill be testing different models, we will transform all variables we use to factors:

```
data(mtcars)
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'))
```

We explore our data by first plotting a boxplot that shows MPG for automatic and manual tranismission types (see Figure 1 in appendix). This plot shows that manual transmission is more efficient than automatic. In order to quantify our assumption we will construct different linear models.

Before going deeper into analysis, we also check the relationship between all the variables of the dataset (see Figure 2 in appendix). From the plot it is clear that variables `cyl`, `disp`, `hp`, `drat`, `wt`, `vs` and `am` have a strong correlation with `mpg`. We will explore this relationship in the next section.

Quantifying the MPG difference between automatic and manual transmissions

Finding the best model

Our base model is constructed using all the variables as predictors. To extract the important vairables we will use the `step` function, which runs `lm` several times to build regression models and select the best variables to build a representative model.

```r
base_model <- lm(mpg ~ ., data = mtcars)
best_model <- step(base_model, direction = "both")
```

The best model obtained from `step` function includes variables `cyl`, `wt` and `hp` as confounders, while `am` is the independent variable:

```r
summary(best_model)
```

```
##
## 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
```

The adjusted R^2 value is 0.8401, which means that more than 84% of the variability is explained by the model which includes the above mentioned variables. We can now compare this model to the base model which is constructed using just `am`, to see if these additional variables contribute anything to the model.

```r
simple_model <- lm(mpg ~ am, data = mtcars)
anova(simple_model, best_model)
```

```
## 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 p-value suggests that the model using `am` and three confounder variables is significantly better than the model without the confounder variables. We can conclude that using `cyl`, `wt` and `hp` as confounder variables improves the accuracy of the model which we constructed constructed.

Inference

In order to confirm that manual (Group 1) and automatic (Group 2) transmissions are different we perform a t-test.

```
group1 <- mtcars[mtcars$am == "Manual",];group2 <- mtcars[mtcars$am == "Automatic",];
t.test(group1$mpg,group2$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  group1$mpg and group2$mpg
## 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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

With a p-value of 0.001374, we can now be certain that there is a significant difference in the mean MPG between manual and automatic transmission cars.

Residual plots and diagnostics

In order to dive a bit deeper into our results we can do residual plots and check some of the regression diagnostics of our model (see Figure 3 for residual plots).

Residuals vs Fitted show that points are randomly scattered, which verifies the independence condition. In addition, the Normal Q-Q plot points are mostly on the line which means that residuals are normally distributed. Scale-location plot shows constant variance, since the points are scattered in an even pattern.

To show important leverage points and the most influential measures we compute the top four points:

```
leverage <- hatvalues(best_model)
tail(sort(leverage),4)
```

```
##   Chrysler Imperial       Toyota Corona Lincoln Continental
##           0.2611168           0.2777872           0.2936819
##       Maserati Bora
##           0.4713671
```

```
influential <- dfbetas(best_model)
tail(sort(influential[,6]),4)
```

```
##    Toyota Corolla Chrysler Imperial          Fiat 128     Toyota Corona
##         0.2885399         0.3507458         0.4292043         0.7305402
```

Conclusion

From `summary(best_model)` we executed before, we can see that cars that have manual transmission will get 1.8 more `mpg` compared to those with automatic transmission when adjusted by `cyl`, `wt` and `hp` (and 7.245 when not adjusted - obtained buy running `summary(simple_model)`). We first saw this by plotting a simple boxplot, confirmed with model building and then making sure that the difference exists by performing a t-test.

Appendix

Figure 1:

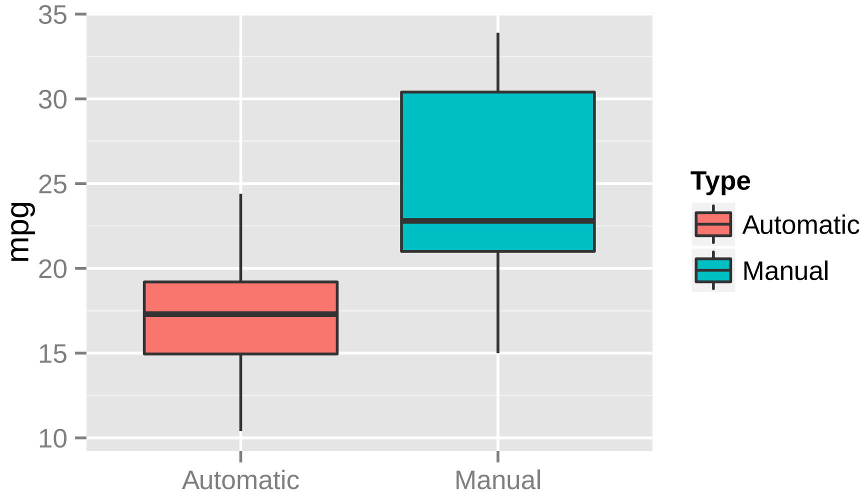


Figure 2:

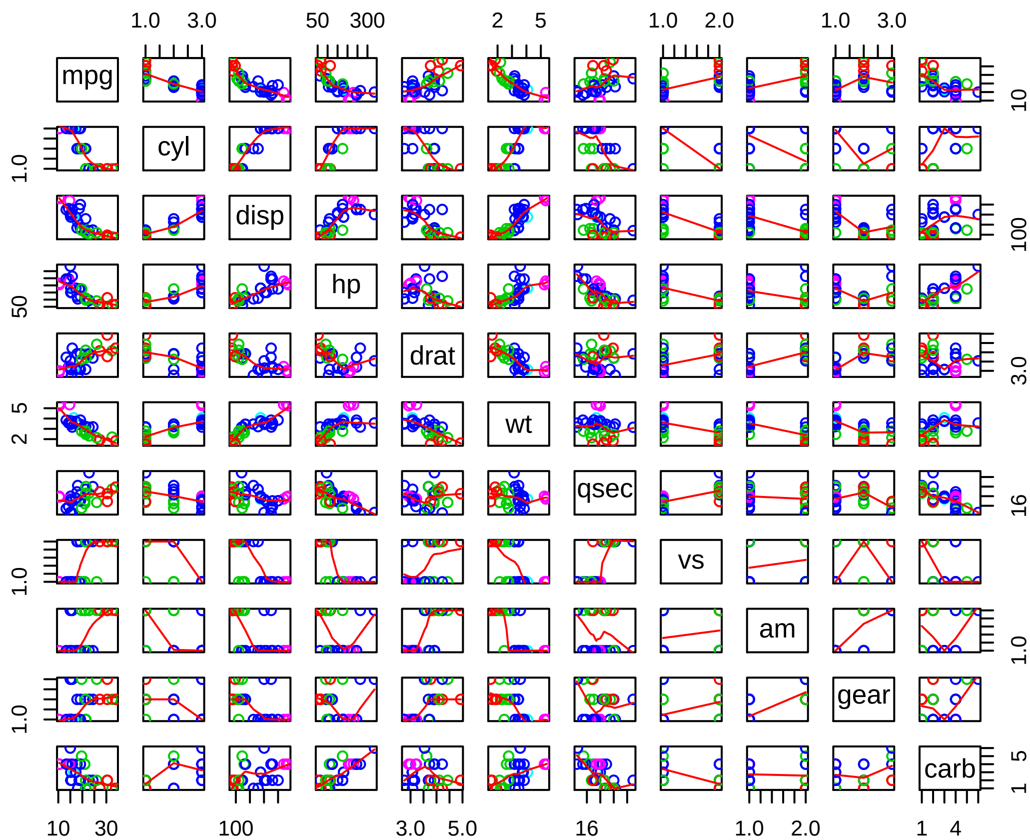
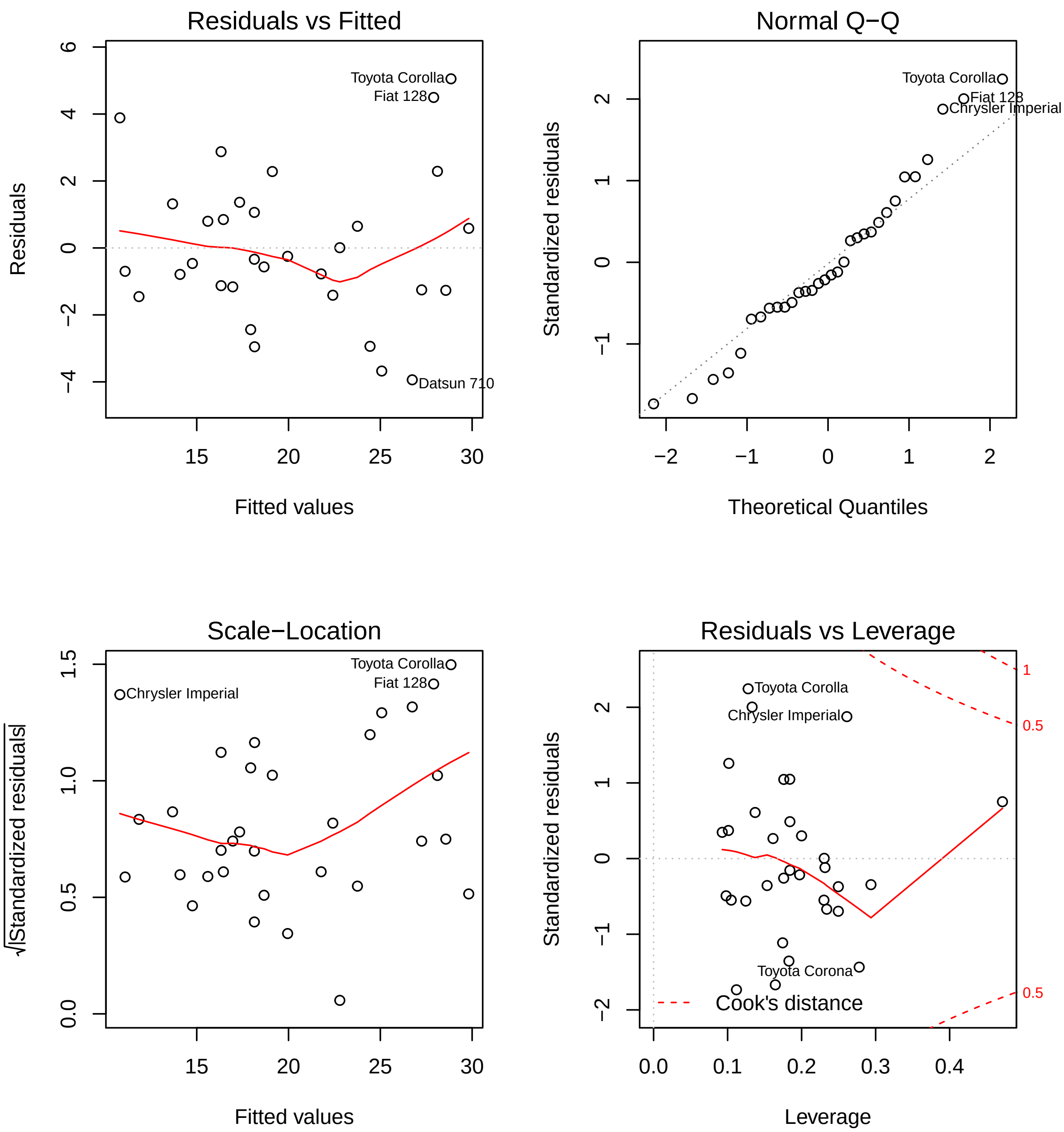


Figure 3:



Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Residuals
Fitted values
Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles
Scale−Location
Chrysler Imperial
Toyota Corolla
Fiat 128
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
Standardized residuals
Leverage