Regression Models - Course Project

Alberto A. Caeiro Jr

January 20, 2015

Executive Summary

This is the Regression Models Project Course. Project 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

Processing the data

Let's have a first look on the dataset and do some basic preparation

```
data(mtcars)
ds <- mtcars
ds$id <- seq(1:32)
ds$am <- factor(ds$am, labels=c("Auto", "Man"))
ds$cyl <- factor(ds$cyl)
head(ds)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs   am gear carb id
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  Man    4    4  1
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  Man    4    4  2
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  Man    4    1  3
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 Auto    3    1  4
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 Auto    3    2  5
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1 Auto    3    1  6
```

As one can see in the plots in the appendix, in a glace, the Automatic transmition seems to yields a lower mpg than the Manual transmition. And looking at the pairs plot, we can see that it seems that a lot of different variables influence the MPG measure. See appendix for the unadjusted mean for each group. Now, let's try some liner models to see what happens when we adjust the model considering the right predictors. So, first we have to find out what are the right predictors.

Selecting the best model.

As shown in the appendix, the best model (higher R^2) includes the following predictors and confounder: am, cyl, wt, hp. Adding carb and gear might suggest a better model, but the empirical selection do not sustain/confirm such beliefs (since R^2 did not increase, as a matter of fact, it decreases a little bit). This model accounts for aprox 84% of the variability of the data.

Residuals and Diagnostics

Now let's plot the fitted model to see both residuals and some diagnostics and to understand how good/poor is our fitted model.

```
fit3 <- lm(mpg ~ cyl + wt + am + hp, data=ds)
```

(Check appendix for the figures) We can see the residuals are normally distributed and that are some outliers (named at the plots). Looking for the most influencial points

```
infl <- dfbetas(fit3); infls <- sort(infl[,4]); tail(infls,2)
```

```
##     Toyota Corona Chrysler Imperial
##          0.3643262         0.9389082
```

Conclusion

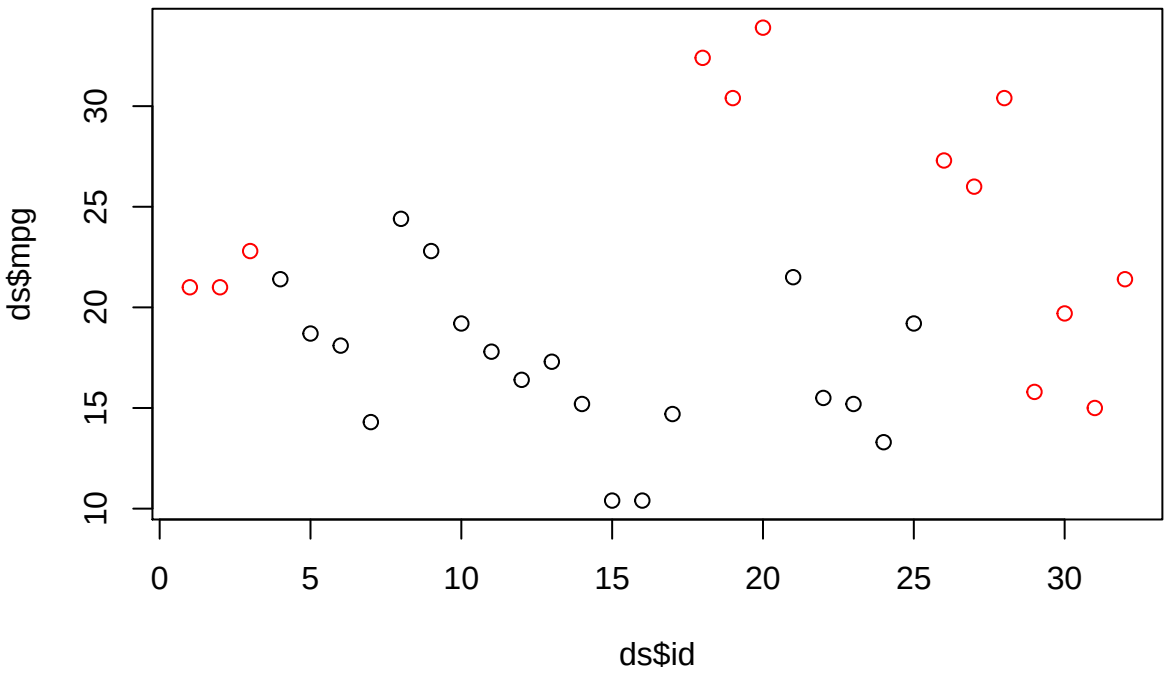
Now, we can anwser the 2 questions really based/backed up by the numbers. 1. Manual transmittion is really better for the MPG 2. Manual transmition is 1.4 times more economic (yielding a 1.4x higher mpg)

Appendix

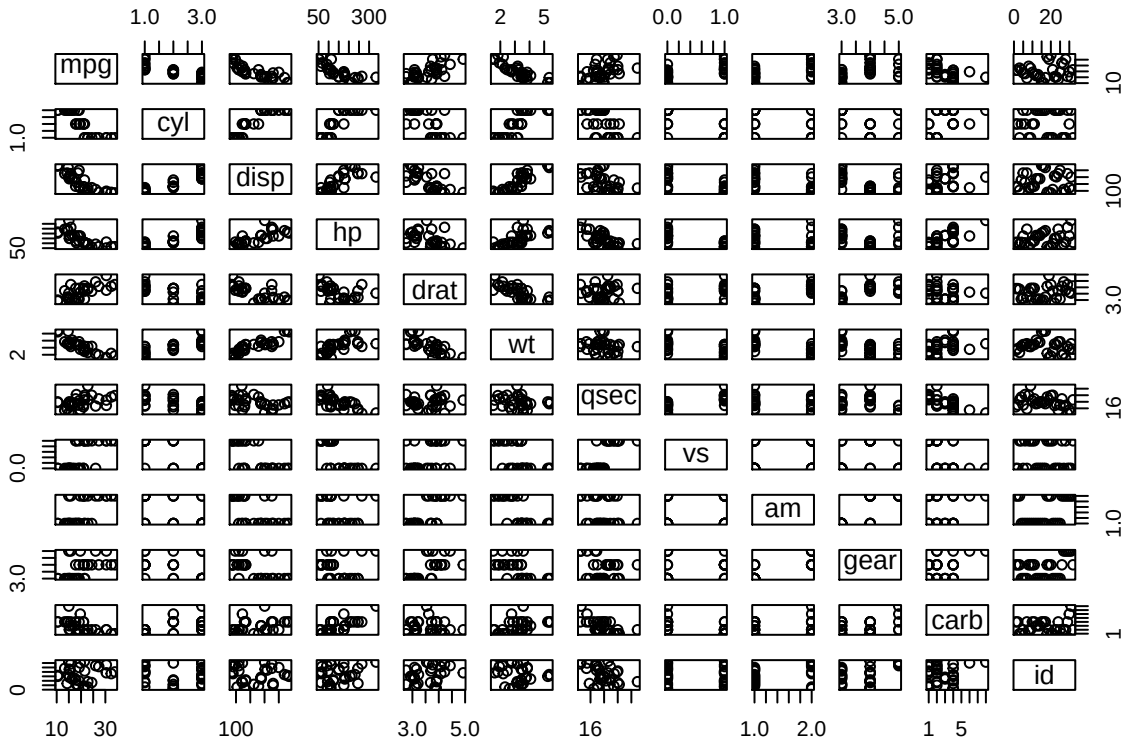
Looking the dataset

So first of all, let take a look at the dataset. In this first exploratory analysis, it seems that automatic transimission yields a lower MPG. And looking the the means in each group it seems to be something like 20 to 15% percent lower. Looking all cars and in the pairs info of the dataset

```
plot(ds$id, ds$mpg, col=factor(ds$am))
```

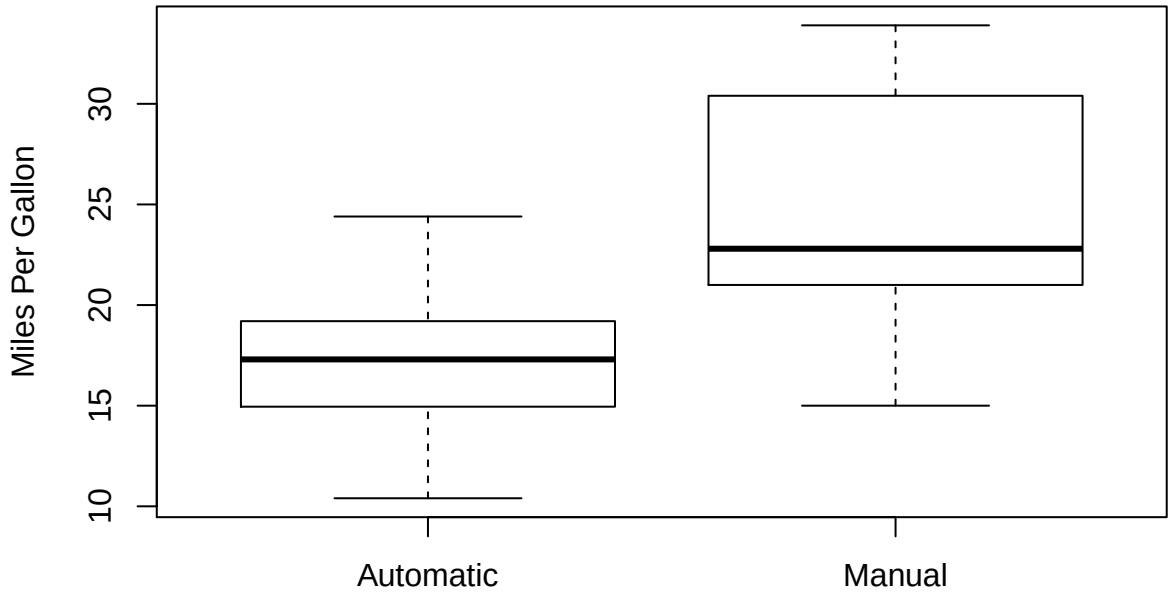


```
pairs(ds)
```



Unadjusted Means

```
boxplot(ds$mpg ~ ds$am, names=c("Automatic", "Manual"), ylab="Miles Per Gallon")
```



```r
manual_mean <- mean(ds$mpg[ds$am=="Man"])
autom_mean <- mean(ds$mpg[ds$am=="Auto"])
print(paste("Automatic Transmition - Unadjusted Mean: ", autom_mean))
```

```
## [1] "Automatic Transmition - Unadjusted Mean:  17.1473684210526"
```

```r
print(paste("Manual Transmition - Unadjusted Mean: ", manual_mean))
```

```
## [1] "Manual Transmition - Unadjusted Mean:  24.3923076923077"
```

Model Selection

Let's check some models, and then look at the summaries so we can try to figure out the best model. From the summaries below we can see the fit3 model is the best one.

```r
fit1 <- lm(mpg ~ ., data = ds)
summary(fit1)$adj.r.squared
```

```
## [1] 0.8068718
```

```r
fit2 <- lm(mpg ~ cyl + wt + am, data=ds)
summary(fit2)$adj.r.squared
```

```
## [1] 0.8134405
```

```r
fit3 <- lm(mpg ~ cyl + wt + am + hp, data=ds)
summary(fit3)$adj.r.squared
```

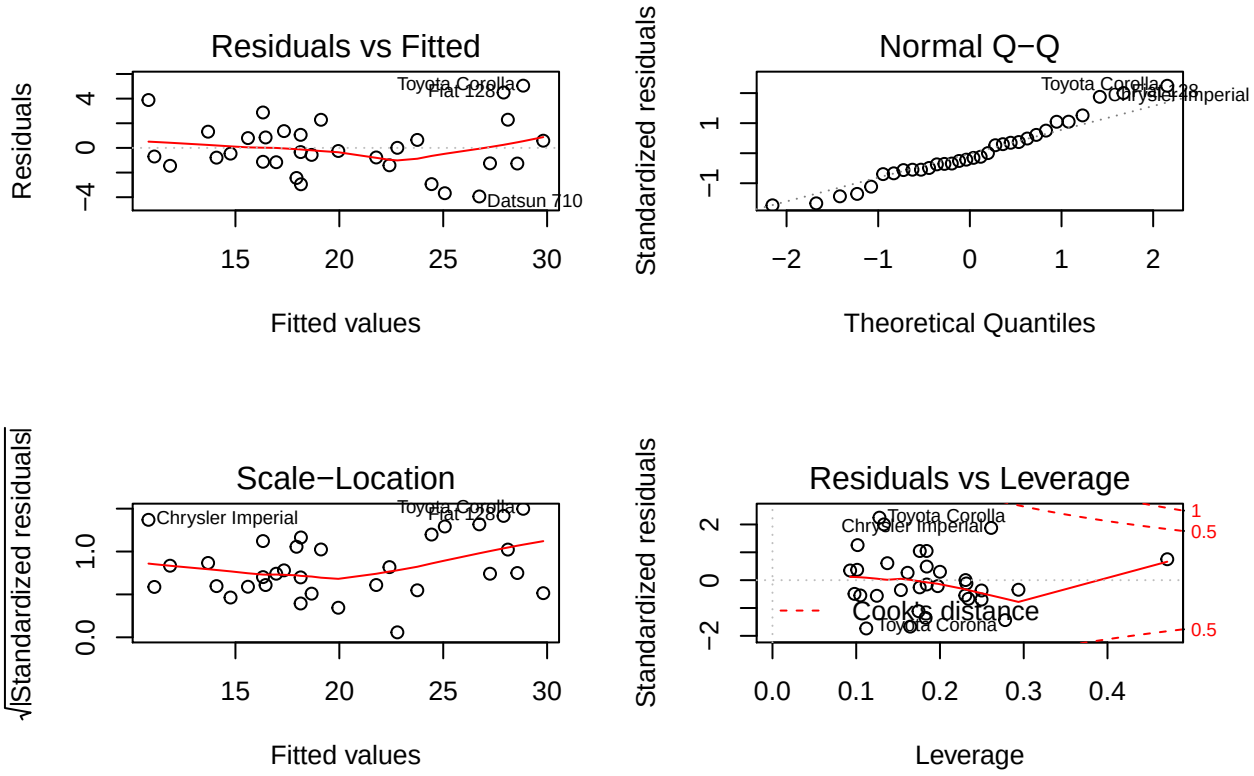
```
## [1] 0.8400875
```

```r
fit4 <- lm(mpg ~ cyl + wt + am + hp + gear + carb, data=ds)
summary(fit4)$adj.r.squared
```

```
## [1] 0.8286668
```

Residuals and Diagnostics

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



Some Basic Inteference

```
t.test(mpg ~ am , data = ds)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by 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 Auto  mean in group Man
##           17.14737           24.39231
```