Regression Models - Course Project

Hasan Shojaei

26 August 2018

Executive Summary

In this report we investigate the relationship between MPG and a set of variables that describe a car using the mtcars dataset. We will address two basic questions:

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

The main findings from this analysis are:

- Cars with manual transmission perform better than automatic transmission cars
- Manual transmission appears to improve MPG by **7.25** when transmission type is the only parameter considered in the model, which explains only **34%** of variability in the data. When other important parameters (weight, horsepower, displacement) are also included in the model, the expected MPG improvement by manual transmission is only **2.16**. This model explains **82%** of variability in the data.

The *mtcars* Dataset

We first load the data from the mtcars dataset, and then show the first 6 rows of the data to get a sense of what parameters are included in the dataset.

```
data(mtcars)
head(mtcars)
```

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

Exploratory Data Analysis

We first use a boxplot (See Appendix A.1) to explore if transmission type has an impact on MPG. We observe that MPG is generally significanly higher for cars with manual transmission. To quantify this, we run a t-test (See Appendix A.2) and examine p-value and average MPG for each category. The p-value is < 0.05 which means the true difference between MPG of the two categories is statistically significant. The mean MPG for manual and automatic cars are **24.39** and **17.15** respectively (i.e. a difference of **7.24**).

Regression Models

To select model parameters, we first look at correlation between "mpg" and other parameters in the mtcars dataset. We then focus on the paramaters that have the largest correlation with "mpg": "cyl", "disp", "hp", "wt" (in addistion to "am" which is the main focus of this project) using the *pairs* plot (See Appendix A.3).

We observe that some of the parameters are highly correlated and therefore including all of them may not add muchg value to the model.

```
library(knitr)
cor_cars <- cor(mtcars)
kable(t(cor_cars[1,]), caption="Correlation Table", digits = 2)
```

Table 1: Correlation Table

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55

We build 6 different models for "mpg" as it can be seen below: The first model includes only "am" as the aexplanatory variable, while more and more explanatory variables are added for models 2 through 6. We then perform a test using the ANOVA function (See Appendix A.4) and find that "fit4" provides the best parsimonious fit of the data.

Last, we investigate the residual plot (See Appendix A.5) and observe that residuals do not show any particular pattern and therefore satisfy the basic requirements of a linear model.

```
fit1 <- lm(mpg ~ am , data = mtcars)
fit2 <- lm(mpg ~ am + wt, data = mtcars)
fit3 <- lm(mpg ~ am + wt + disp , data = mtcars)
fit4 <- lm(mpg ~ am + wt + disp + hp, data = mtcars)
fit5 <- lm(mpg ~ am + wt + disp + hp + cyl, data = mtcars)
fit6 <- lm(mpg ~ ., data = mtcars)
```

Conclusions

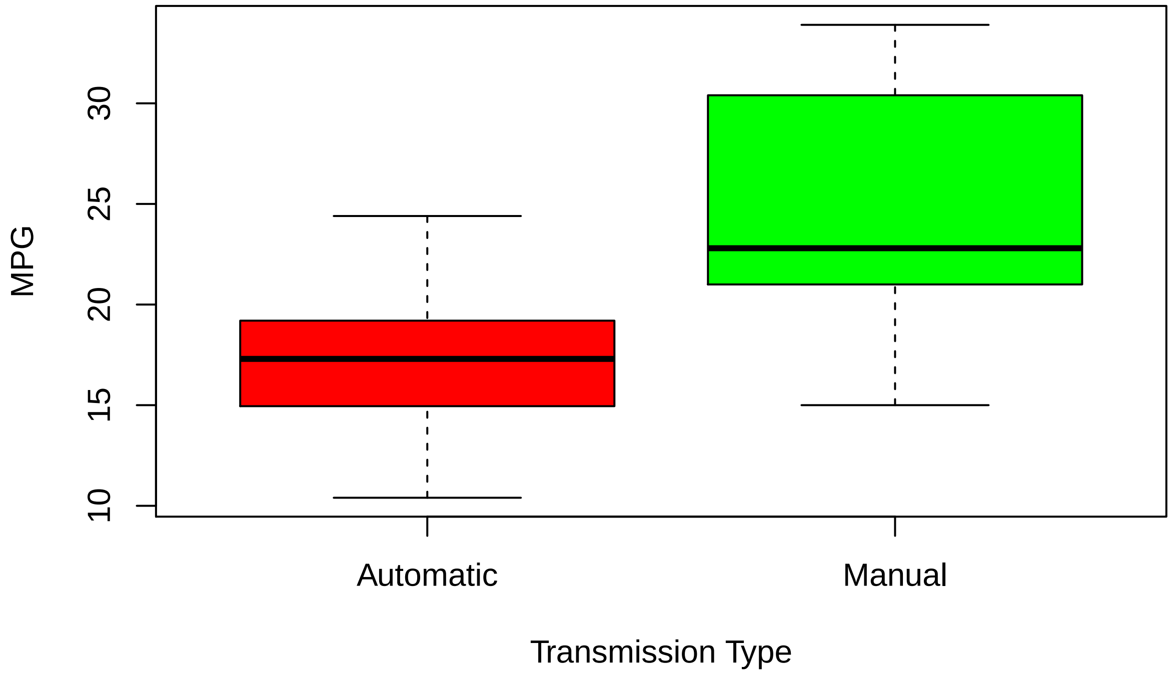
The adjusted R^2 for model "fit1" is 0.34 which means this model explains only 34% of variability in the data. On the other hand, the adjusted R^2 for model "fit4" is 0.82 meaning this model can explain 82% of variability in the data, which is a significant improvement over model "fit1".

Another important observation from this study is that manual transmission appears to improve MPG by 7.25 when transmission type ("am") is the only explanatory parameter in the model, while expected MPG improvement by manual transmission is only 2.16 when three other important parameters (weight, horsepower, displacement) are also included in the model. Therefore transmission type does not have as significant of an impact on PMG as it first appeared.

Appendix

A.1 Exploratory Plot

```
mtcars$am <- factor(mtcars$am,labels=c("Automatic","Manual"))
with(mtcars, boxplot(mpg ~ am, col=c("red","green"),
                     xlab="Transmission Type", ylab="MPG"))
```



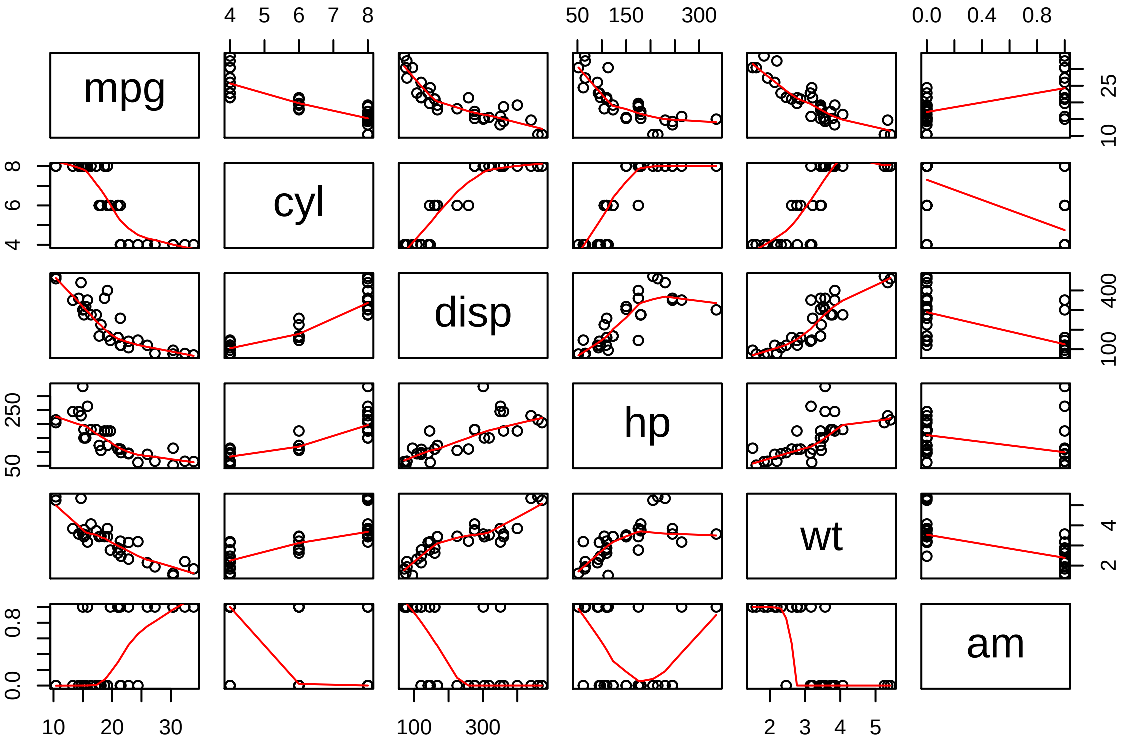
A.2 t-Test

```
cars_automatic <- mtcars[mtcars$am=="Automatic",]
cars_manual <- mtcars[mtcars$am=="Manual",]
t.test(cars_automatic$mpg, cars_manual$mpg)
```

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

A.3 Correlation Among Important Parameters

```
data(mtcars)
pairs(mtcars[,c(1:4,6,9)], panel = panel.smooth)
```



A.4 ANOVA and Summary of Models 1 & 4

```r
anova(fit1, fit2, fit3, fit4, fit5, fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + disp
## Model 4: mpg ~ am + wt + disp + hp
## Model 5: mpg ~ am + wt + disp + hp + cyl
## Model 6: 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     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 246.56  1     31.76  4.5224  0.045474 *
## 4     27 179.91  1     66.65  9.4893  0.005672 **
## 5     26 163.12  1     16.79  2.3902  0.137035
## 6     21 147.49  5     15.63  0.4449  0.812059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(fit1)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422   4.106127 2.850207e-04
```

```
summary(fit1)$adj.r.squared
```

```
## [1] 0.3384589
```

```
summary(fit4)$coeff
```

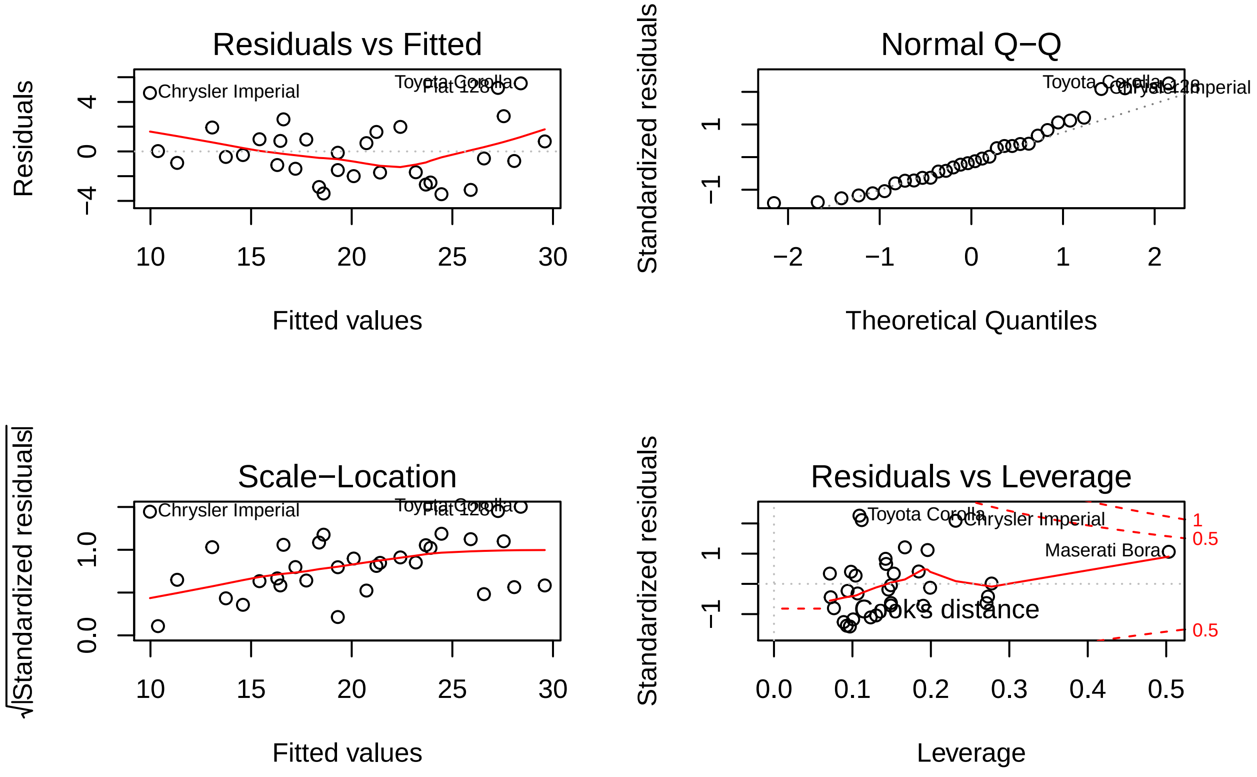
```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 34.209443370 2.82282610 12.1188632 1.979953e-12
## am           2.159270737 1.43517565  1.5045341 1.440531e-01
## wt          -3.046747000 1.15711931 -2.6330448 1.382936e-02
## disp         0.002489354 0.01037681  0.2398959 8.122229e-01
## hp          -0.039323213 0.01243358 -3.1626624 3.842032e-03
```

```
summary(fit4)$adj.r.squared
```

```
## [1] 0.8165613
```

A.5 ANOVA and Summary of Models 1 & 4

```
par(mfrow=c(2,2))
plot(fit4)
```


Residuals vs Fitted
Residuals
4
0
−4
10
15
20
25
30
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
1.0
0.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
1
−1
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
Toyota Corolla
Chrysler Imperial
Maserati Bora
Cook's distance
1
0.5
0.5