Regression Models Course Project

FChemin

November 25, 2016

This analysis will try to clarify if a manual transmission is better than an automatic one. We will look at a data set of a collection of cars (mtcars, from 1974 Motor Trend magazine) which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models), and explore the relationship between variables and miles per gallon.

Loading dataset and necessary libraries

```
data(mtcars)
attach(mtcars)
library(corrplot)
```

Data processing

We first need to change some variables classes from numeric to factor in order to be able to make our analysis:

```
mtcars$cyl <- as.factor(cyl)
mtcars$vs <- as.factor(vs)
mtcars$am <- as.factor(am)
mtcars$gear <- as.factor(gear)
mtcars$carb <- as.factor(carb)
```

Exploratory Data Analysis

```
## Showing the structure and summary of the data set
str(mtcars)
summary(mtcars)

## Showing first 6 lines of the data set
head(mtcars)
```

All the figures needed can be found in the **Appendix section** at the end of this document. According to our boxplot, given the median of each transmission, we could say that driving an automatic car seems to be more expensive than a manual one. The fuel economy for automatic cars seems lower. We need to investigate this assumption further and in order to do so, we make the null hypothesis that there's no difference between automatic and manual transmissions. In the plot of variable correlations, we can see there are strong correlations between **cyl**, **disp**, **hp** and **wt**.

Performing a t.test for our null hypothesis

Using a two sample t-test:

```r
tt <- t.test(mpg ~ am)
tt
```

```
##
##  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 0 mean in group 1
##        17.14737        24.39231
```

We can **reject** our null hypothesis because the p-value of our t.test is equal to 0.0013736. The difference between the mean for MPG of manual and automatic transmissions is about 7.

Regression models

We begin by using all the variables in our model:

```r
fit1 <- lm(mpg ~., data = mtcars)
summary(fit1)
```

No coefficient seems significant at 0.05 level in this model. Our residual standard error is equal to 2.65 on 21 degrees of freedom and the Adjusted R-squared, equal to 0.8066.

Now, we fit another model with the most significant variable:

```r
fit2 <- lm(mpg ~ wt + qsec + am, data = mtcars)
summary(fit2)
```

That's an improvement: our residual standard error is down to 2.459 on 28 degrees of freedom and our adjusted R-squared is now at 0.8336. Automatic transmission cars tend to be heavier so we need to add the interaction between **wt** and **am** in our model.

So, for this new model:

```r
fit3 <- lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)
summary(fit3)
```

We continue to improve. What about a simpler model?

```r
fit4 <- lm(mpg ~ am, data = mtcars)
summary(fit4)
```

We perform an analysis of variance (ANOVA) to find the best model and we calculate our confidence interval:

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

Integreting an interaction between wt and am since to have the most impact on our model so we choose **fit3**:

```r
summary(fit3)$coef
confint(fit3)
```

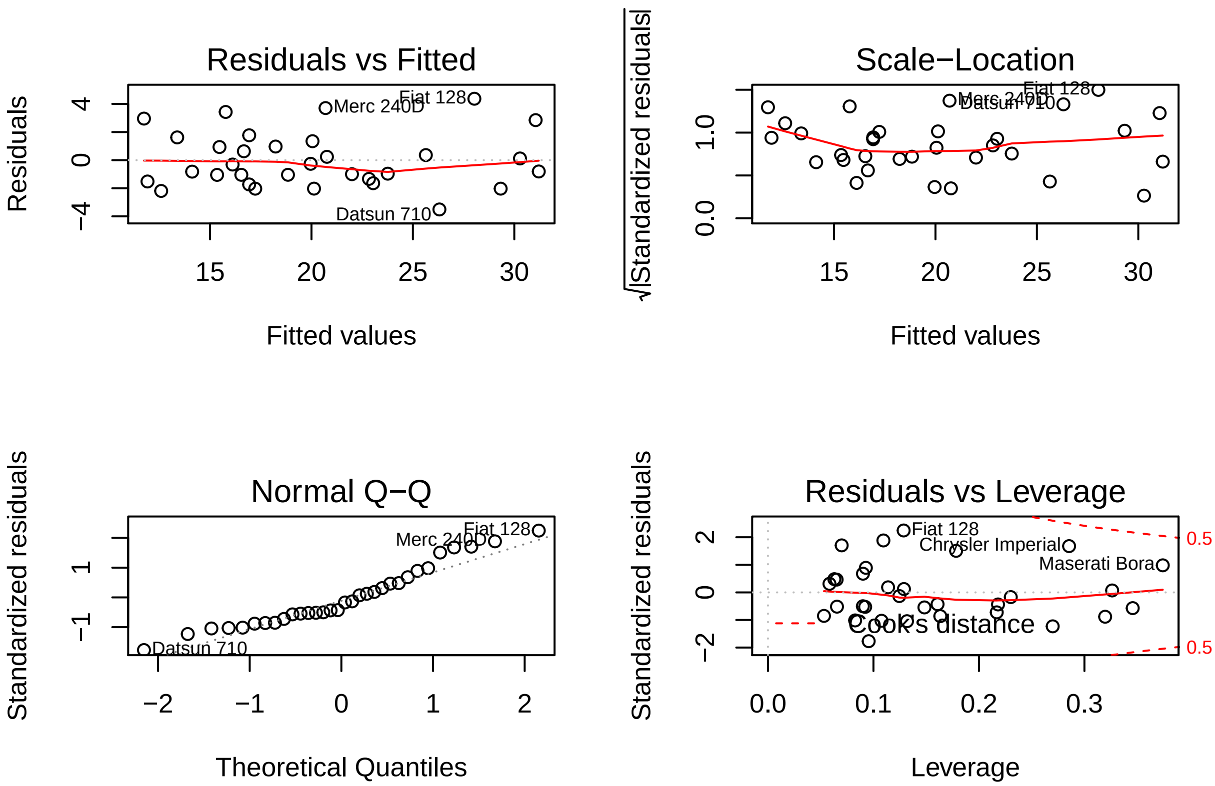
We can compute the difference in MPG between two identicaly weighting and 1/4 mile timing car with different transmissions with this formula:

MPG manual = 14.079428 + (-4.141376)*wt

That's 3.73 more MPG for a 2500lbs manual transmission car than an equivalent automatic transmission car.

Residual Plots

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



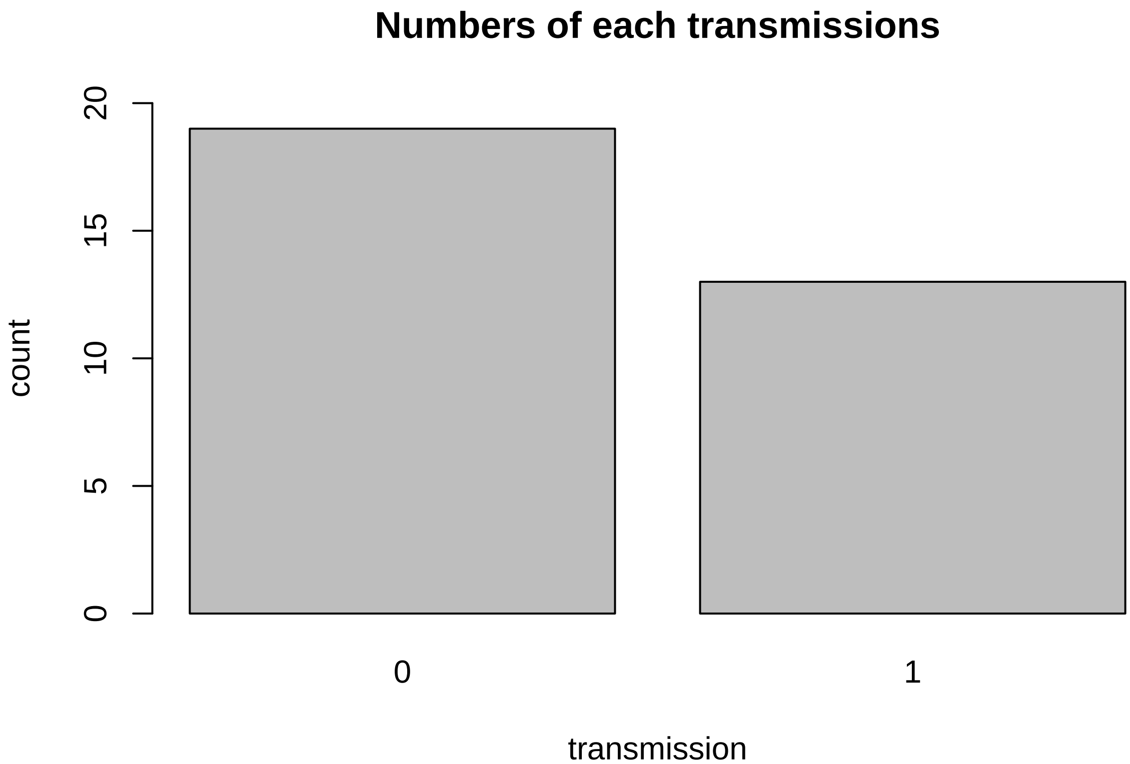
All assumptions are met.

Appendix

Figures

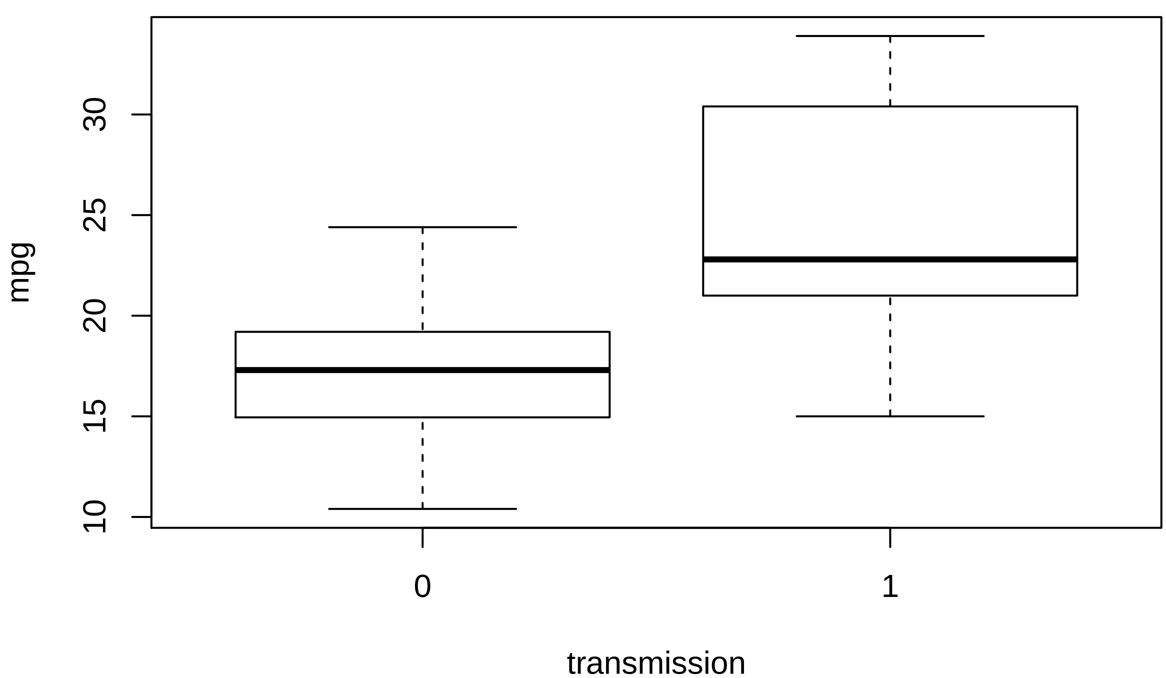
1. Distribution of transmissions

```r
barplot(table(am), xlab = "transmission", ylab = "count", ylim = c(0,20), main = "Numbers of each transm
```



2. Boxplot of MPG against transmission type

```r
boxplot(mpg ~ am, xlab = "transmission", ylab = "mpg")
```



3. Plot of the variable correlations

```
data(mtcars)
m <- mtcars[-1]
m <- cor(m)
corrplot(m, method = "circle")
```