Regression Models Course Project

Sunday, October 26, 2014

Executive Summary

This analysis focuses on the effect of transmission on fuel efficiency.We will try to answer the following questions:

- Is an automatic or manual transmission better for miles per gallon (MPG)?
- How different is the MPG between automatic and manual transmissions?

We will use the dataset 'mtcars', which is taken from the 1974 Motor Trend US magazine. It contains infromation on fuel consumption and ten characteristics of 32 models of automobile.

Our results show that average MPG of cars with automatic transmission is different from the one of cars with manual transmission cars. Our analysis also includes the investingation (multiple linear regression) of the role of possible confounding variables (such as the weight and horsepower).

Reading and Exploring the Data

```
data(mtcars)
```

The main variable of interest is the one reporting the fuel consumption. As a matter of fact, we aim at exploring how this variable is affected by the other variables, i.e. by specific characteristics of the considered automobiles. More specifically, we are interested in studying the effect produced by the type of transmission on fuel consumption. The 'am' variable is a dummy variable: indeed, it takes value equal to 0 when transmission is automatic, and it takes value equal to 1 when transmission is manual.

Before starting our analysis, we can give a description of our dependent variable, at least to see whether it satisfies the main requirements for the application of a linear regression model. To do this, we can plot a histogram and a density plot (see Fig 1a and Fig 1b in the Appendix): we observe that the distribution of 'mpg' variable is approximately normal and there are no outlier. For this reasons, we can apply a linear regression model, whose parameters will not be influenced by outliers.

Is an automatic or manual transmission better for MPG?

In order to answer this first question, we can start observing the distribution of Miles per Gallon with respect to the two different levels of the factor variable of transmission we have defined before. By looking at the boxplot in Fig 2, we can suppose that Automatic transmission is better for fuel consumption.

However, in order to give a more precise answer we can apply a t-test on the difference between the means of the two groups, i.e. MPG of automobiles with Automatic transmission, and MPG of automobiles with Manual transmission.

```
t.test(mtcars$mpg~mtcars$am,conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  mtcars$mpg by mtcars$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
```

The p-value of the test is 0.001374, which is lower than 0.05. Hence, we can reject the null hypothesis of equal means. We can conclude that the two groups have a different mean, and that automatic transmission is related to a lower mpg with respect to manual transmission. However, in order to verify whether this conclusion holds also for different values of the other variables affecting mpg, it would be better to run a multiple linear regression.

How different is the MPG between automatic and manual transmissions?

We can start by the simplest resgression: mpg is the dependent variable and am is the explanatory variable.

```
lm<-lm(mpg~factor(am)-1,data=mtcars)
```

Looking at the estimated intercept and coefficient, we can say that automatic transmission cars have an average of 17.147 MPG (am=0); manual transmission cars an average of 24.392 MPG. Considering the goodness of fit of the model, we notice that the R^2 is euql to 0.3598. It order to try to get a better model, i.e. a model that is able to explain a higher proportion of the total variance, we need to add new variables. Therefore, we will use a multiple linear regression model.

In order to decide which variables can be relevant for explaining the variation of mpg, we can compute the correlation between mpg and the other potential regressors (cor(mtcars)[1,]). We immediately notice that wt, cyl, disp, and hp show a high correlation with mpg. However, since cyl and disp are highly correlated not only with each other, but also with the other variables, it would be better not to include them in the regression model (see the Appendix for regression results). The R^2 of this model is 0.84, and wt and hp can be considered as confounders (for this reason, see also the model with interactions in the Appendix). The estimations show that mpg increases by 2.08 when passing from a car with automatic transmission to a car with manual transmission.

After having defined the new model, we can perform an ANOVA test to verify which model performs better (see the Appendix).

```
mlm<-lm(mpg~factor(am)+wt+hp,data=mtcars)
anova(lm,mlm)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) - 1
## Model 2: mpg ~ factor(am) + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is 3.745e-09, we reject the null hypothesis: the two models are significantly different.

Finally, we can examine (see Fig 3) the residuals vs. fitted values: this would allow us to be sure that residuals are normal and homoskedastic.

Appendix

Fig 1a: Histogram of MPG

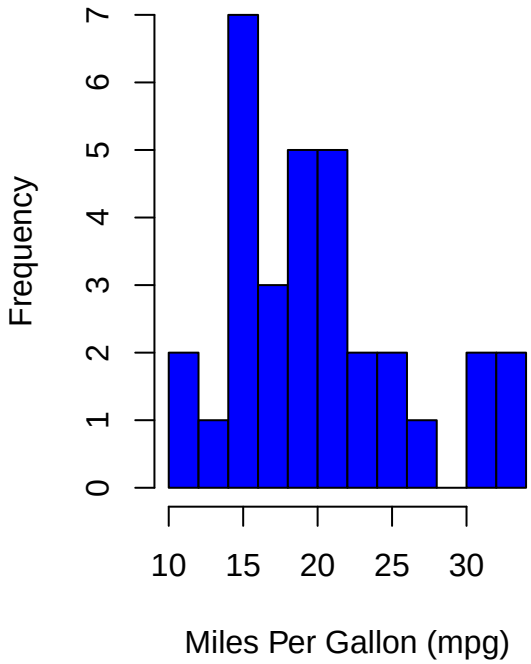


Fig 1b: Density Plot of MPG

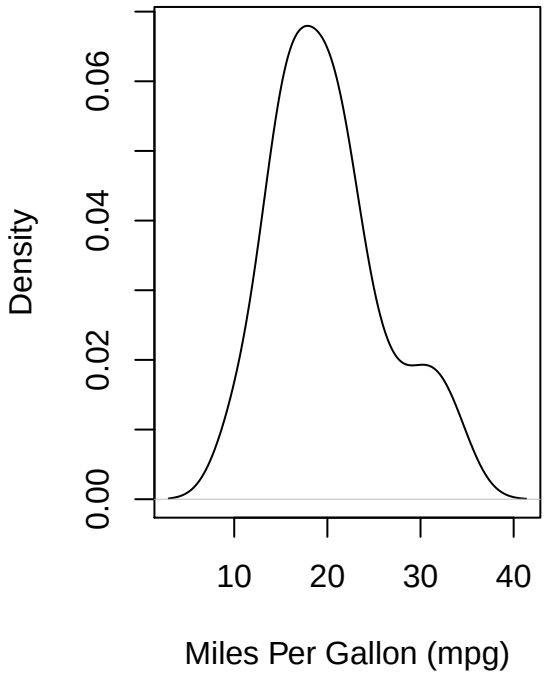
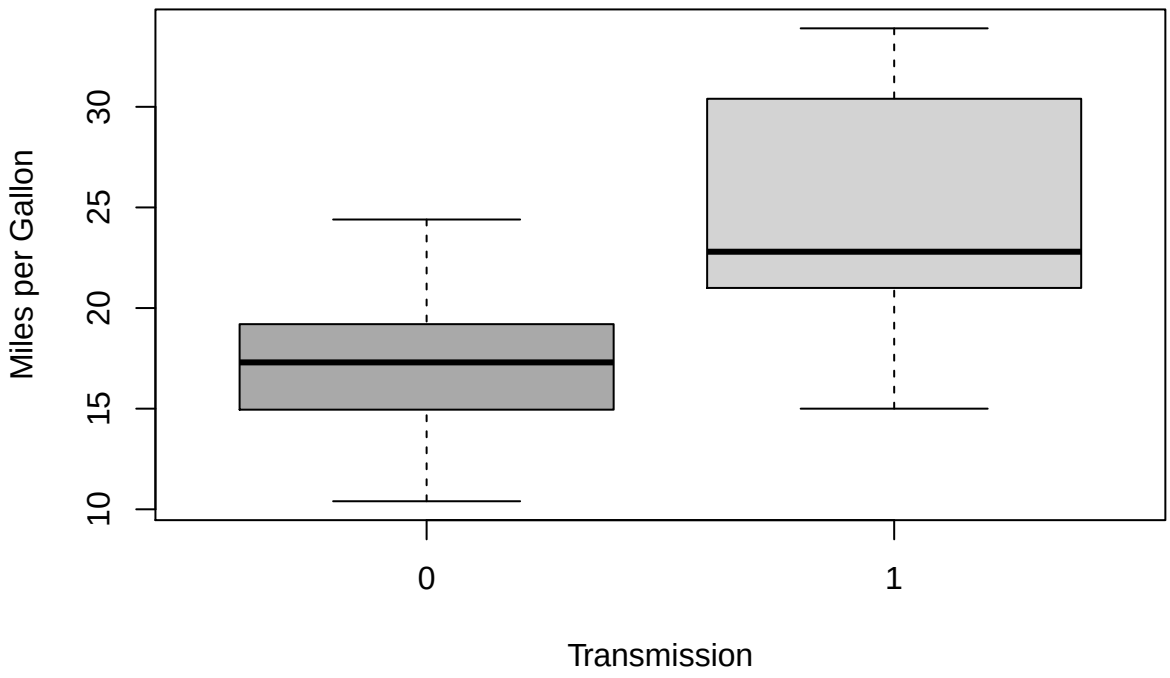


Fig 2: Miles Per Gallon by Transmission Type




```
summary(lm)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) - 1, 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|)
## factor(am)0   17.147      1.125   15.25 1.13e-15 ***
## factor(am)1   24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

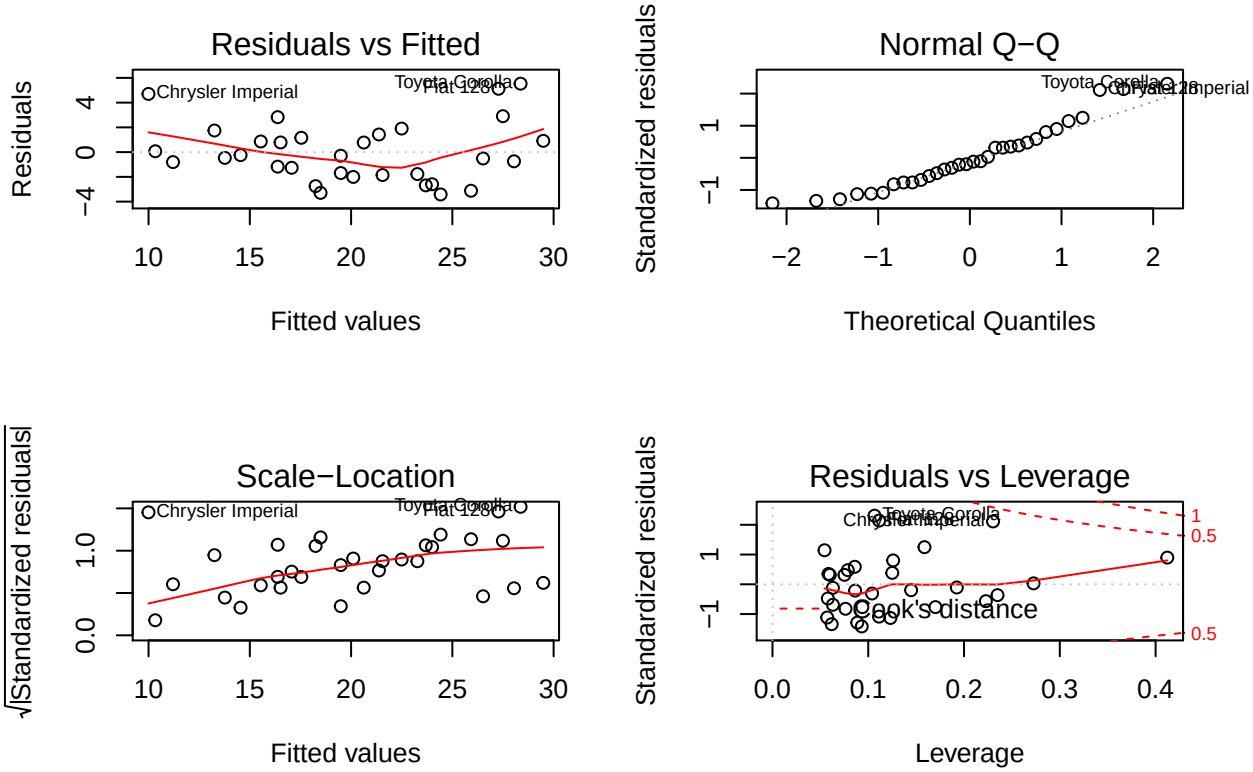
```
summary(mlm)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ factor(am) + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## factor(am)1  2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

```
imlm<-lm(mpg~factor(am)+wt+hp+factor(am):wt+factor(am):hp,data=mtcars)
summary(mlm)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## factor(am)1  2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1
0.5
0.5