Regression models course project

Collin

Friday, July 25, 2014

Summary

This report shows that manual transmission is on average better for the miles per gallon (MPG) of a car than automatic transmission, but the result is not statistically significant. This research was done using the dataset mtcars in R. This type of data analysis is very difficult because there are many confounding variables, more than we can account for. Thus the result is not as definitive as it seems at first glance. T

Results

First we need to load the data, which will be the dataset mtcars.

```
data(mtcars)
```

The first question we want to answer, is whether an automatic or manual transmission is better for MPG? By splitting the cars into the two categories and looking at MPG, we can use a boxplot to compare the two sets. This boxplot is shown at the bottom of this report. The boxplot shows that the manual cars have much better MPG, since all quartiles are higher. We can perform a t-test with 95% confidence level on whether the two sample means are the same. We get a p-value of .0014, so we reject the null hypothesis that the means are the same. The full t-test is shown at the bottom of the report.

```
t.results <- t.test(mtcars$mpg[mtcars$am==0],mtcars$mpg[mtcars$am==1])
t.results$statistic
```

```
##      t
## -3.767
```

```
t.results$p.value
```

```
## [1] 0.001374
```

A linear regression is not too useful on this, but can do it. First we need to turn am into a factor because the numeric values of 0 and 1 are meaningless.

```
mtcars$am <- as.factor(mtcars$am)
lm(mpg~am,data=mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Coefficients:
## (Intercept)          am1
##       17.15         7.24
```

```
confint(lm(mpg~am,data=mtcars))
```

```
##              2.5 % 97.5 %
## (Intercept) 14.851  19.44
## am1          3.642  10.85
```

This shows that having manual transmission increases MPG by 7.245. Note that this is the difference between the means of the two groups MPG, 24.39-17.15. The 95% confidence interval of this variable is the range from 3.64 to 10.84. This range does not include zero, so there is definitely an effect from the type of transmission. This preliminary analysis tells us that manual is better for MPG than automatic transmission. However, we have only considered the MPG to be a function of the type of transmission, which is not true. There are many other factors to consider. We can see that there are 9 other variables in the mtcars dataset that we have not considered.

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

This is way too many variables to study, so we will only use mpg, am, the number of cylinders (cyl), horse power (hp), and weight (wt). A data frame with only these five variables will be created for ease.

```
car <- mtcars[,c('mpg','am','cyl','hp','wt')]
```

The pairs plot is much easier to read now that we have removed many variables, it can be found at the bottom of the report. Now we can try another regression using mpg as the outcome and the other four as the inputs.

```
lm(mpg~.,data=car)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = car)
## 
## Coefficients:
## (Intercept)          am1          cyl           hp           wt  
##      36.147        1.478       -0.745       -0.025       -2.606
```

```
confint(lm(mpg~.,data=car))
```

```
##                2.5 %    97.5 %
## (Intercept) 29.77605 42.517020
## am1         -1.47895  4.435042
## cyl         -1.94094  0.450624
## hp          -0.05295  0.003049
## wt          -4.49383 -0.719130
```

Now we see that a manual transmission now only accounts for 1.48 miles per gallon. Also, the 95% confidence interval now includes 0, going from -1.48 to 4.44, so we no longer are certain that there is a statistically significant result for the type of transmission affecting the MPG.

Extra data and plots

```
t.results
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean of x mean of y
##     17.15     24.39
```

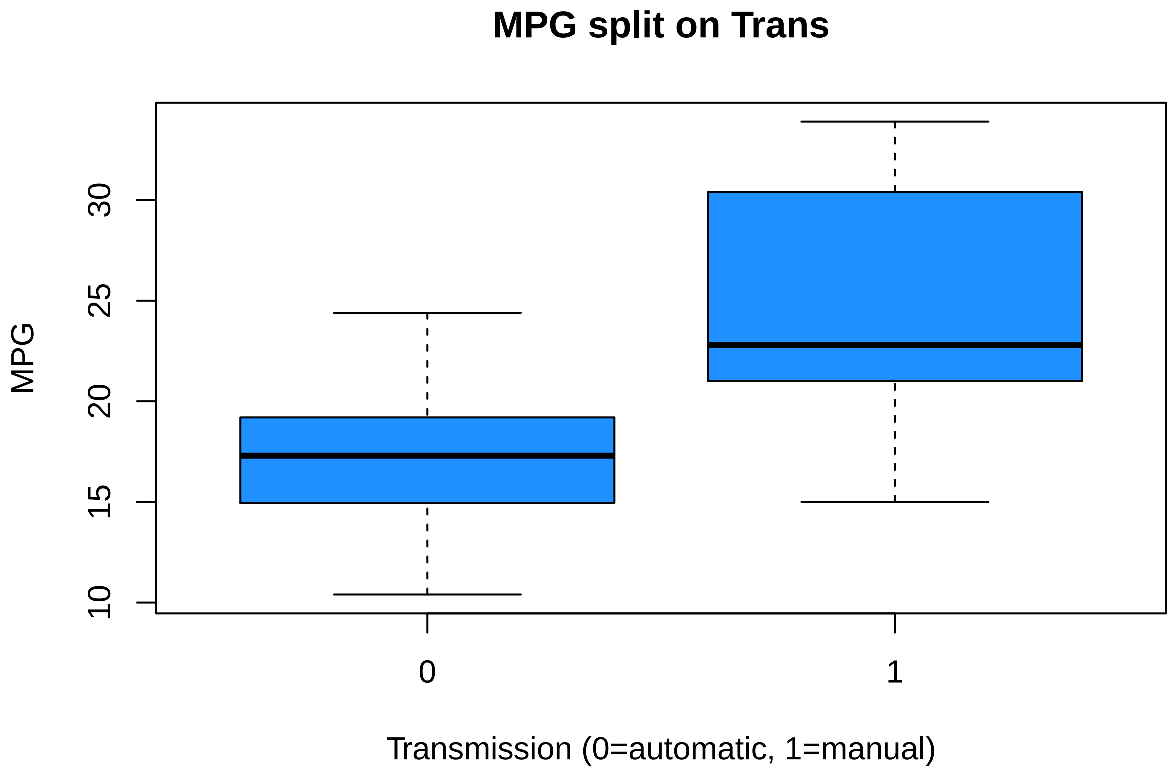
```
require(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.1.1
```

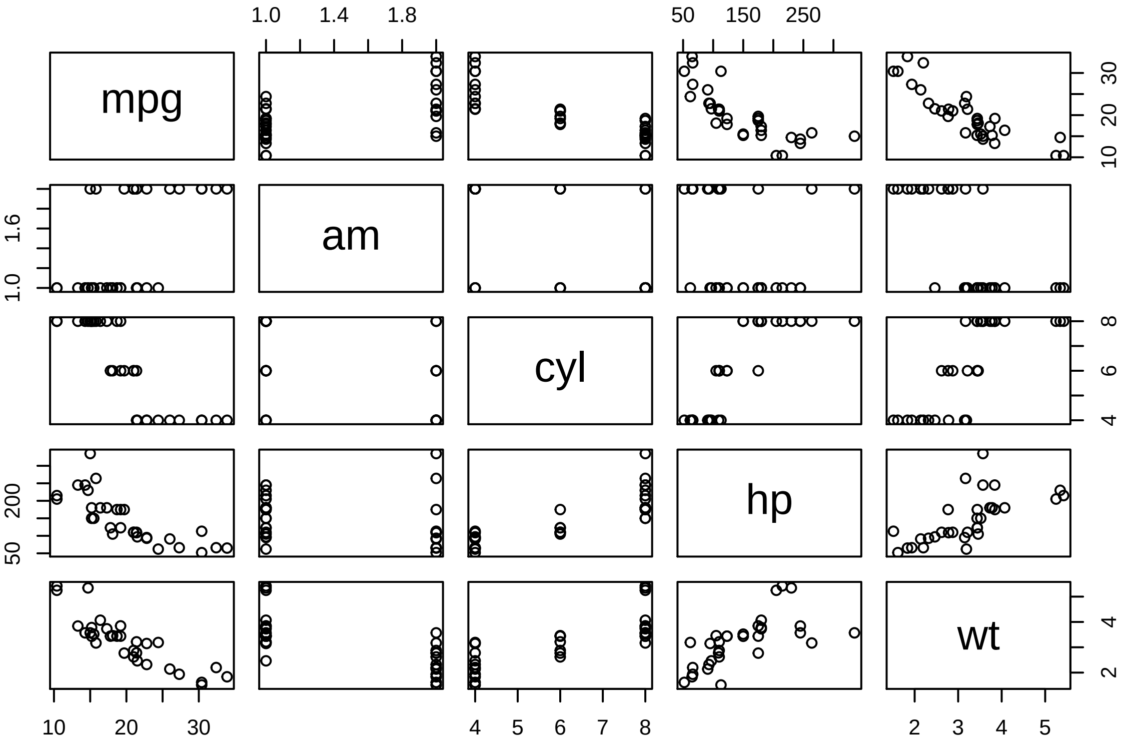
```
ddply(mtcars,'am',function(x){summary(x$mpg)})
```

```
##   am Min. 1st Qu. Median Mean 3rd Qu. Max.
## 1  0 10.4      15   17.3 17.1    19.2 24.4
## 2  1 15.0      21   22.8 24.4    30.4 33.9
```

```
boxplot(mpg~am,data=mtcars,xlab='Transmission (0=automatic, 1=manual)',ylab='MPG',main='MPG split on Tra
```


MPG split on Trans
MPG
10
15
20
25
30
0
1
Transmission (0=automatic, 1=manual)

```
pairs(car)
```


mpg
am
cyl
hp
wt
1.0
1.4
1.8
50
150
250
10
20
30
1.0
1.6
4
6
8
50
200
2
4
10
20
30
4
5
6
7
8
2
3
4
5