Regression Models Course Project

By: Misal6

January, 2015

Executive Summary

The focus of this analysis is the effect of car transmission on miles per gallon (from mtcars dataset). Both the requirements relate to this measurement. First to determine this relationship and second to quantify it. The approach I have taken to address thses can be summarized as follows.

Meaningfulness of Results (questions of interest)

1. Theoretical importance of measuring transmissions relation to mpg. Yes there is a relationship as demonstrated by the R-Squared values.
2. The statistical Effect size of this measure is the difference between the means of mpg determined by transmission type.
3. Can we rule OUT random chance? No, because of significant p-value
4. Can we rule out alternative explainations(LURKING variables). No, there is a stronger relationship on mpg in the data set relative of other variables as demonstrated by correlations.

Read in the data

```
mydata <- mtcars
mydata$am <- as.factor(mydata$am)
levels(mydata$am)[1] <- "Automatic"
levels(mydata$am)[2] <- "Manual"
```

Exploratory data analysis

```
summary(mydata$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

Means by transmission type.

```
aggregate(mpg~am, data = mydata, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

Hypothesis Testing

Null hypothesis “There is no significant difference in mpg by transmission type of a car”

```
am.manual <- mydata[mydata$am == "Manual",]
am.auto <- mydata[mydata$am == "Automatic",]
t.test(am.manual$mpg,am.auto$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  am.manual$mpg and am.auto$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:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

Since p-value is less than .005 we reject null hypothesis. MPG is effected by transmission type of cars.
To Quantify this effect, we fit a model with transmission as dependent variable

```
fit.am <- lm(mpg ~ am, data=mydata)
round(summary(fit.am)$r.squared,2)
```

```
## [1] 0.36
```

Only 36% of variance in mpg can be explained by difference in transmission type. (quantifying the uncertanity)
There must be other variables effecting mpg more then transmission type. The correlation table indicates the top choices.

```
sort(round(cor(mtcars,mtcars$mpg)[-1,],2))
```

```
##    wt   cyl  disp    hp  carb  qsec  gear    am    vs  drat
## -0.87 -0.85 -0.85 -0.78 -0.55  0.42  0.48  0.60  0.66  0.68
```

Taking the top correlated variables and getting the best fit model.(multiple models)

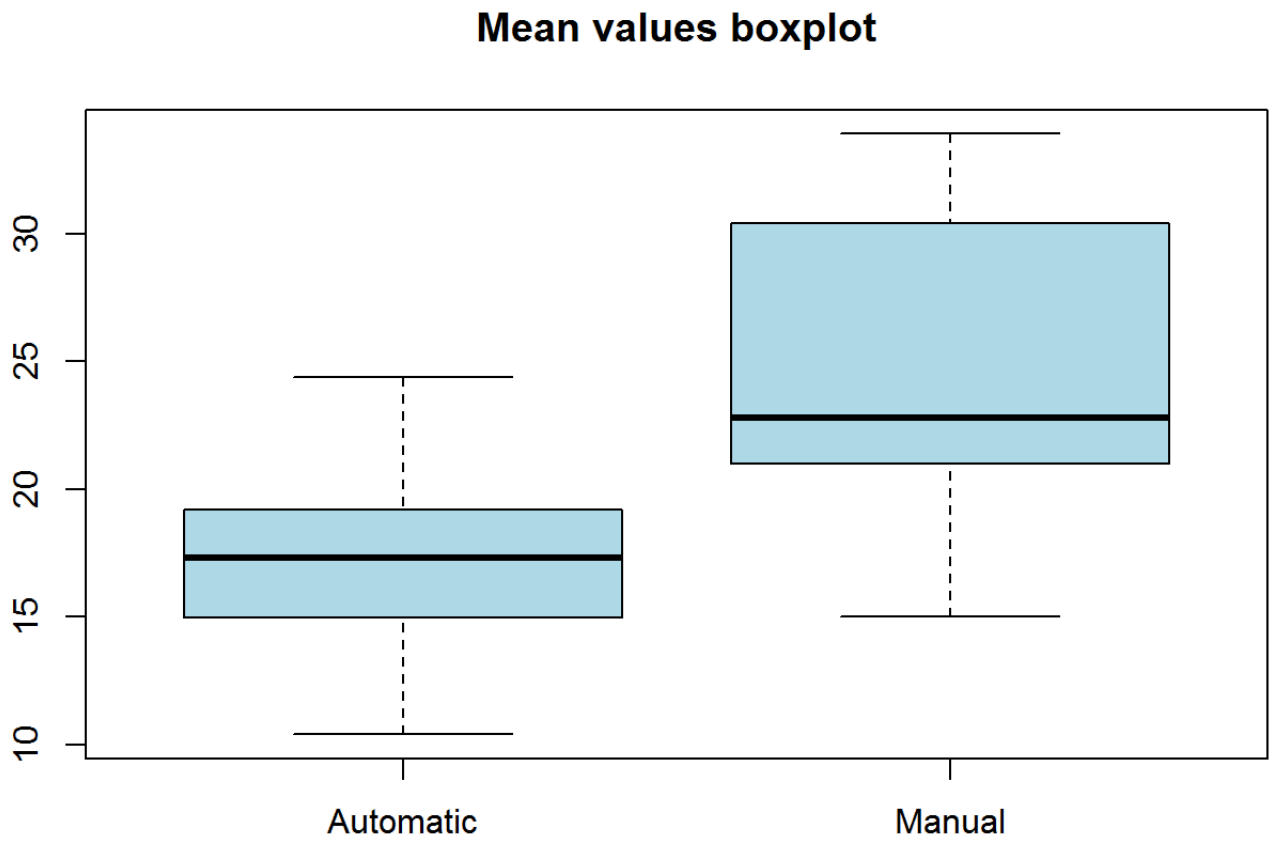
```
fit.best <-lm(mpg ~ am + wt + cyl + disp + hp, data=mydata)
round(summary(fit.best)$r.squared,2)
```

```
## [1] 0.86
```

These variables explain 86% variability on mpg, more than double of transmission type, **and is the best model representing effect of different variables on miles per gallon measure of cars**.

Appendix A

```
boxplot(mpg ~ am, data = mydata,col = "lightblue",main="Mean values boxplot")
```



Appendix B

Residual plot for the best fit model. (of particular interest is the Q-Q normal plot)

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + disp + hp, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## amManual     1.55649    1.44054   1.080  0.28984
## wt          -3.30262    1.13364  -2.913  0.00726 **
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## hp          -0.02796    0.01392  -2.008  0.05510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```