Regression Models Course Project

Sohail Munir Khan

26 July 2015

Context

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

"Is an automatic or manual transmission better for MPG" "Quantify the MPG difference between automatic and manual transmissions"

Executive Summary

This report provides an analysis and evaluation of the relationship between a set of variables and miles per gallon (MPG)
Method included exploratory data analysis and using linear regression to find the most important regressors for our conclusion
Findings reveal that along with am (transmisson type automatic or manual), these variables had significant affect: carb + wt + cyl
We have more than 85% conclusive results with hightly significant P-value

Exploratory Data Analysis

- See Appendix for a summary of mtcars data after all variables that should be factors have been updated
- Showing the histogram for the MPG variable in Appendix
- Showing the boxplot for difference between Automatic(am==0) vs Manual(am==1) Transmission in Appendix. The answer to our first question is that MPG for Manual transmission cars have better MPG

```
library(datasets)
data(mtcars)

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

library(ggplot2)
mpg_mean <- mean(mtcars$mpg)
ghist <- ggplot(mtcars, aes(x = mpg)) + geom_histogram(fill = "salmon", colour = "black",
                                                       binwidth = 1)
ghist <- ghist + geom_vline(xintercept = mpg_mean, size = 2)
ghist <- ghist + ggtitle(paste("mean = ", mpg_mean))
```

```
gtrans <- ggplot(mtcars, aes(am, mpg, fill=am))
gtrans <- gtrans + geom_boxplot()
gtrans <- gtrans + scale_fill_discrete(name="Transmission", breaks=c("0", "1"),
                                       labels=c("Automatic", "Manual"))
gtrans <- gtrans + scale_x_discrete(breaks=c("0", "1"), labels=c("Automatic", "Manual"))
```

Regression Models

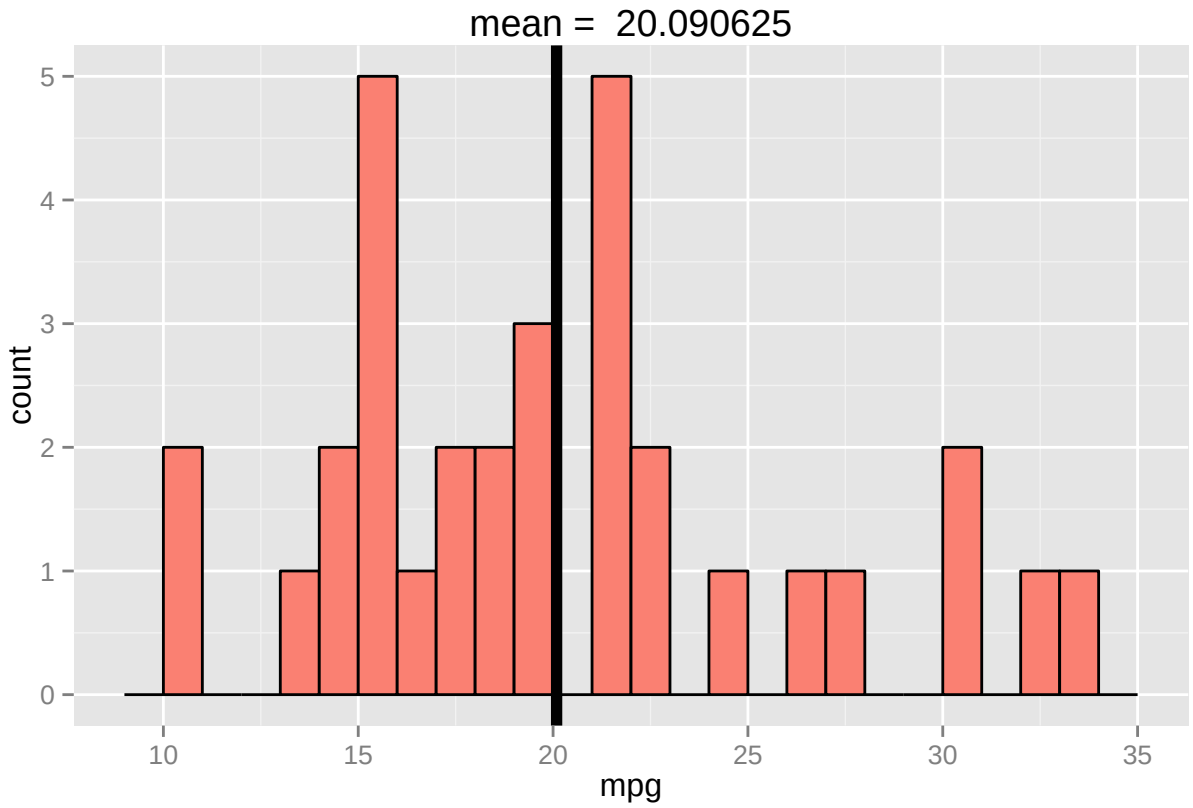
- See Appendix for summary: First we fit a linear model by using am as our first variable and then looking at the rest while interpreting our MPG prediction. We note that there is a 23.88 slope for am0 (intercept) implying that there is a positive relationship between am0 (Automatic) and MPG. am1(Manual) has even higher value since it increases the slope by 1.21. P values for both these are close to 0 but not that significant (still > 0.25). We also note that some other important slopes in our summary are carb then wt then cyl. We will be using only these variables to create another model

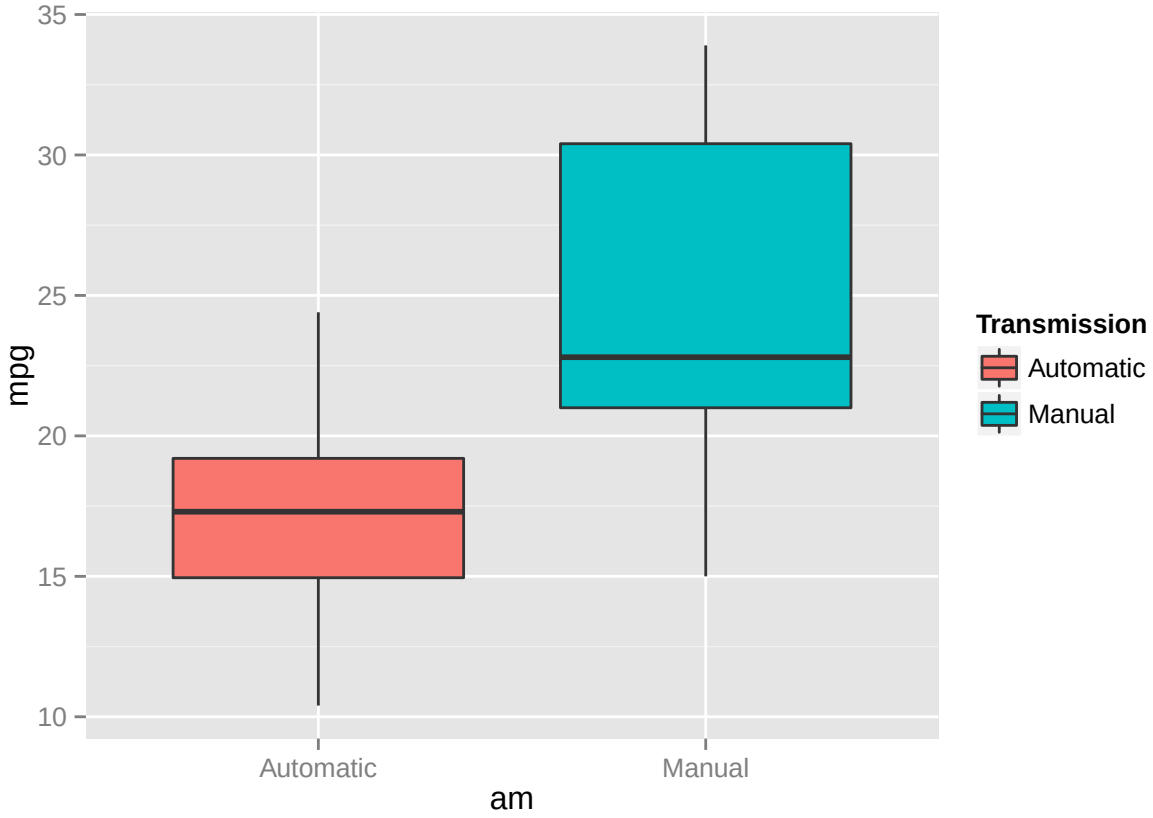
- See next Appendix summary: With a very high P-value significance (4.63e-08) we deduce that we have a very high correlation from am0 and MPG. And even higher for am1. All other variables are negatively impacting MPG. We have 0.8545 R-squared value for more than 85% significance and a very high P-value for the entire model (2,714e-07)

- Next we plot residual values and deduce there is no shape that we can see to go against this model.

Appendix

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs     am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   0:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   4:12   2:10
##  Median :3.325   Median :17.71                 5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                        4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                        6: 1
##  Max.   :5.424   Max.   :22.90                        8: 1
```






```
##
## Call:
## lm(formula = mpg ~ am + ., data = mtcars)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## am1          1.21212    3.21355   0.377   0.7113
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124


##
## Call:
## lm(formula = mpg ~ am + carb + wt + cyl, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.549 -1.291 -0.137  1.429  5.420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.8849     3.8047   8.118 4.63e-08 ***
## am1           1.8736     1.7793   1.053   0.3038
## carb2        -0.4656     1.4465  -0.322   0.7506
## carb3        -0.3510     2.2888  -0.153   0.8795
## carb4        -2.3001     1.9025  -1.209   0.2395
## carb6        -3.7769     3.4614  -1.091   0.2870
## carb8        -4.2009     3.6077  -1.164   0.2567
## wt           -2.3318     1.1300  -2.063   0.0511 .
## cyl6         -2.8225     1.8603  -1.517   0.1435
## cyl8         -5.2330     2.0141  -2.598   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.729 on 22 degrees of freedom
## Multiple R-squared:  0.8545, Adjusted R-squared:  0.7949
## F-statistic: 14.35 on 9 and 22 DF,  p-value: 2.714e-07

## Warning: not plotting observations with leverage one:
##   30, 31

## Warning: not plotting observations with leverage one:
##   30, 31
```