Coursera DS Regression Models Project

Raju Varghese

05/24/2015

Preface

This is the detailed report of the Coursera Regression Models project. The standard report on Github was submitted for project evaluation. As there are page length constraints on the report that is a shortened version of this detailed report. In particular, the R code and results are displayed in their entirety in this report. Figures are shown in the locations where they are discussed instead of relegating them to the appendix. Further, the result of the *anova* comparison could not be be shown (due to the length limitation) in the submitted report and that is included here.

Overview

This report of the Coursera Regression Model course project explores the relationship between the mileage (called mpg - miles per gallon - hereafter) and several other parameters of a collection of cars. In particular, the project expects to answer the question if automatic or manual transmission is better for mpg and to quantify it. The data is taken from the *mtcars* data set that is available in R. This report also looks at other parameters and shows that their contribution to mileage is just as important (or even more) as the transmission system.

```
library (datasets)
library (ggplot2)
```

Exploratory Analysis

To examine the contents of the mtcars data frame *str*, *summary* functions were invoked and plots were made.

```
str (mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
summary (mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
```

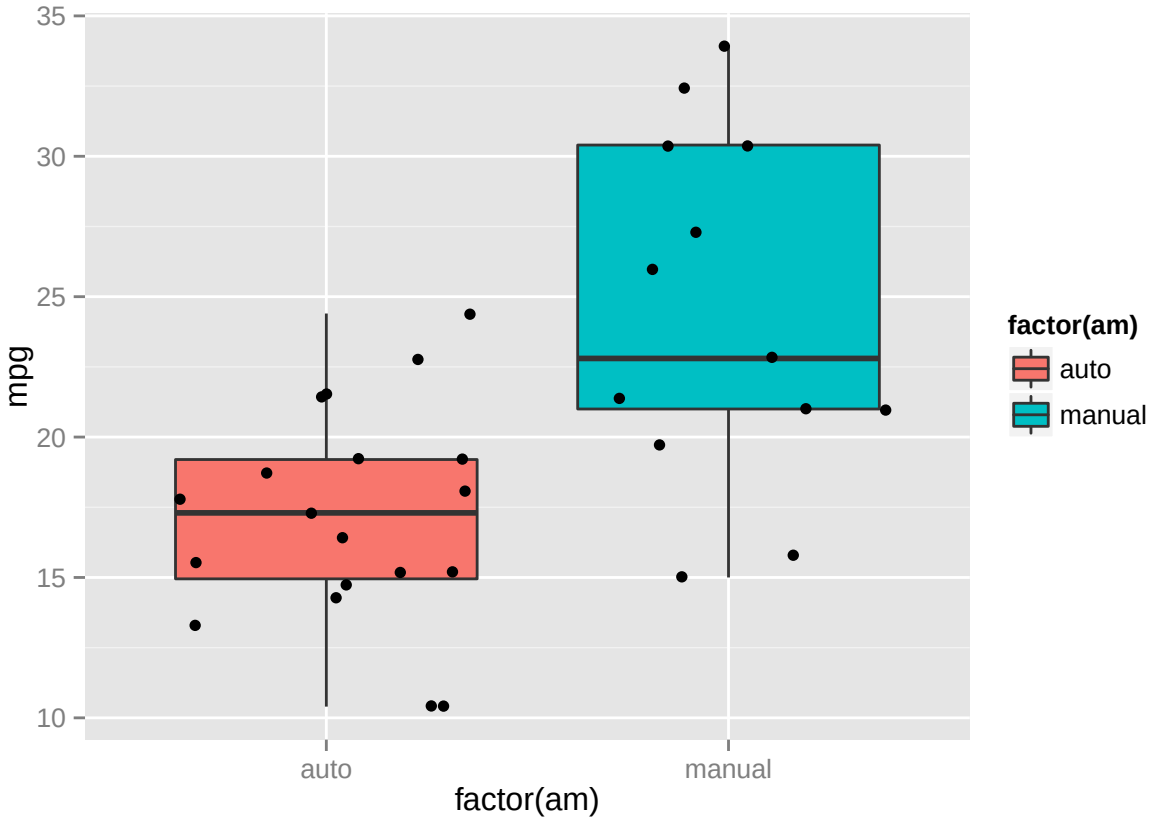
```
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

During exploratory analysis it was seen that making factors out of some of the parameters made it easier to see what the actual variables were. A new variable *mt* is used to make these changes instead of performing those actions on a built-in dataset.

```
data (mtcars)
mt <- mtcars
mt$am <- factor (mt$am, labels=c("auto", "manual"))
mt$carb <- factor (mt$carb)
mt$cyl <- factor (mt$cyl)
mt$gear <- factor (mt$gear)
mt$vs <- factor (mt$vs)
```

The figure below clearly shows that manual cars give better mileage than automatics.

```
p <- ggplot(mt, aes(factor (am), mpg))
p + geom_boxplot(aes (fill=factor (am))) + geom_jitter()
```



Statistical Inference

As the number of samples is low we shall do a t-test on the data. As we would like to be pretty sure of our conclusions we shall set the confidence interval to 99%; usually 95% is deemed good enough.

```
t <- t.test (mt [mt$am=="manual", "mpg"], mt [mt$am=="auto", "mpg"], conf.level=0.99)
t
```

```
##
##  Welch Two Sample t-test
##
## data:  mt[mt$am == "manual", "mpg"] and mt[mt$am == "auto", "mpg"]
## t = 3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   1.720751 12.769128
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

The confidence interval range of [1.7207509, 12.7691277] does not include 0 and the p-value of 0.0013736, gives a very strong indication of the effect of transmission system (automatic versus manual) on mileage.

Linear Regression

Doing a linear model fit of the transmission system gives us the result shown below.

```r
fit1 <- lm (mpg ~ am, data=mt)
summary (fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## ammanual       7.245      1.764   4.106 0.000285 ***
## ---
## 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:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The coefficients section above shows that the mean mileage of manually shifted cars are 7.25 miles per gallon higher than automatics. The signifiance codes (the 3 asterisks) at the end of the line confirms that this is an important factor. The R^2 value of 35.98% explains part of the variance. However, this does not say anything about the correlation - or lack thereof - of the other parameters. This prompts us to take a look at the other parameters and do a multiple regression analysis.

Multivariable Regression

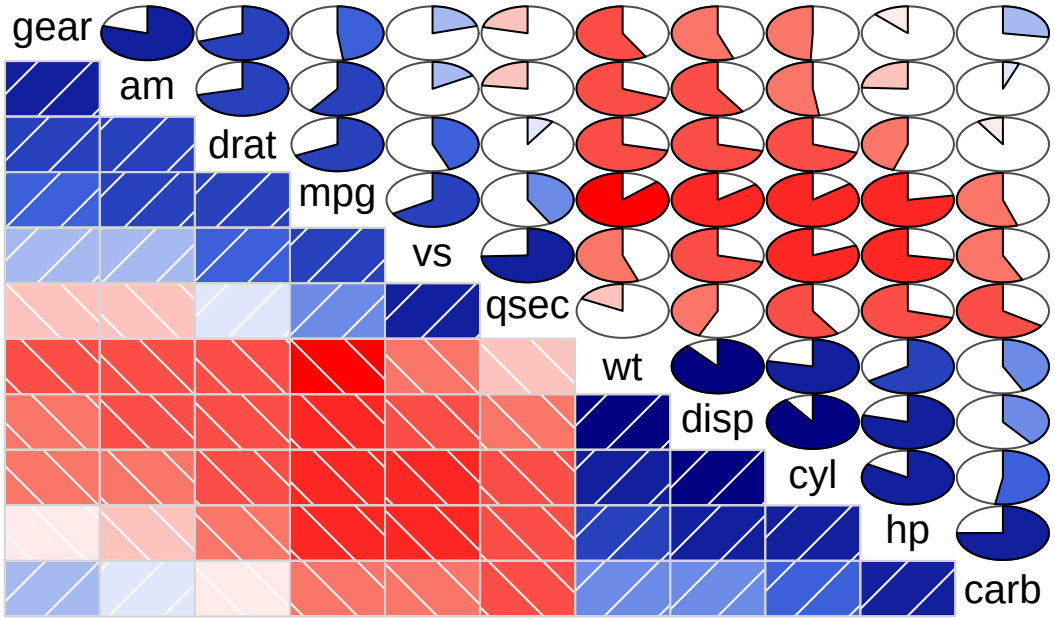
In order to determine which of the parameters are important in our further analysis thei correlation with mileage will be listed below with the help of the *cor* function. Please note that the absolute value of the correlation is sorted and hence it does not say whether there is a positive or negative correlation.

```r
sort (abs (cor (mtcars) [1,]), decreasing=T)
```

```
##       mpg        wt       cyl      disp        hp      drat        vs
## 1.0000000 0.8676594 0.8521620 0.8475514 0.7761684 0.6811719 0.6640389
##        am      carb      gear      qsec
## 0.5998324 0.5509251 0.4802848 0.4186840
```

The result shows that wt (weight), cyl (number of cylinders), disp (displacement), hp (horsepower), drat (rear axle ratio) and vs (v or straight engine) have a higher correlation than the transmission system. The figure below shows the above data graphically.

```r
library (corrgram)
corrgram(mtcars, order=TRUE, lower.panel=panel.shade,
         upper.panel=panel.pie, text.panel=panel.txt,
         main="Car Mileage Data")
```



We therefore perform multivariable regression to get a best fit. For that we take an initial model with all parameters and then *step* through them taking combinations in order to find the best one.

```
initialFit <- lm (mpg ~ ., data = mt)
bestFit <- step (initialFit, direction = "both", trace=0)
summary (bestFit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ammanual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

This shows that the cylinder (cyl), horsepower (hp), weight (wt) and transmission (am) give the best fit. With an R^2 value of 0.84 this looks a very good combination in the choice of a multivariable linear model.

```
anova (fit1, bestFit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics and residual plots are shown below.

```
par (mfrow=c(2,2))
plot (bestFit)
```