Coursera DS Regression Models Project

Raju Varghese

05/23/2015

Executive Summary

This is the detailed report of the Coursera Regression Models project. The standard report on Github is to be used for project evaluation. As there are page length constraints on the report that version is a shortened of this detailed report. In particular, the R code and results are displayed in their entirety in the detailed report.

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

Figure 1 (in the appendix) clearly shows that manual cars give better mileage than automatics. 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. Unfortunately, space does not permit the reproduction of that analysis here.

```
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)
```

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)
```

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.

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

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## ammanual     7.244939   1.764422  4.106127 2.850207e-04
```

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.

```
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. Please refer to Figure 2 in the appendix to view the above data graphically. We therefore do multivariable regression on those 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. Diagnostics and residual plots are in figure 3 in the appendix.

Appendix

This part contains graphic plots as permitted in the project specification of the course.

Figure 1: Boxplot of mileage by transmission type

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

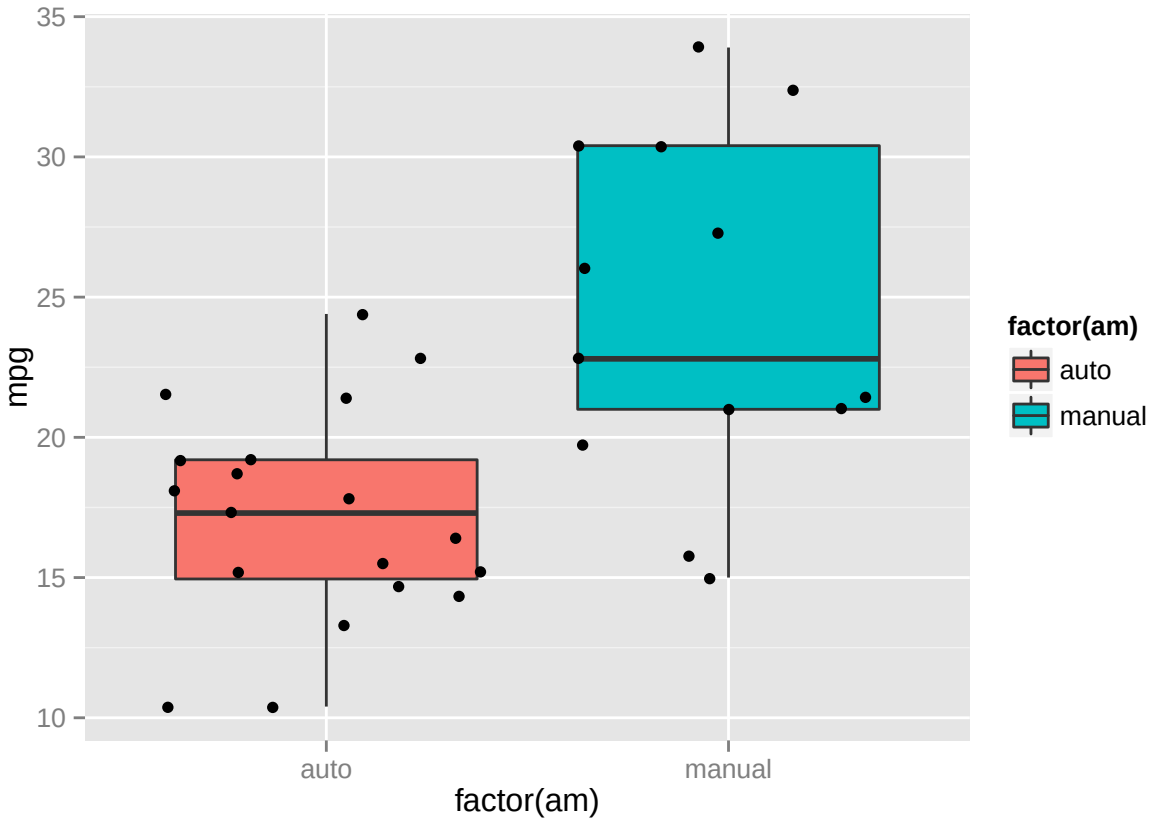


Figure 2: Correlation plot of all parameters

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

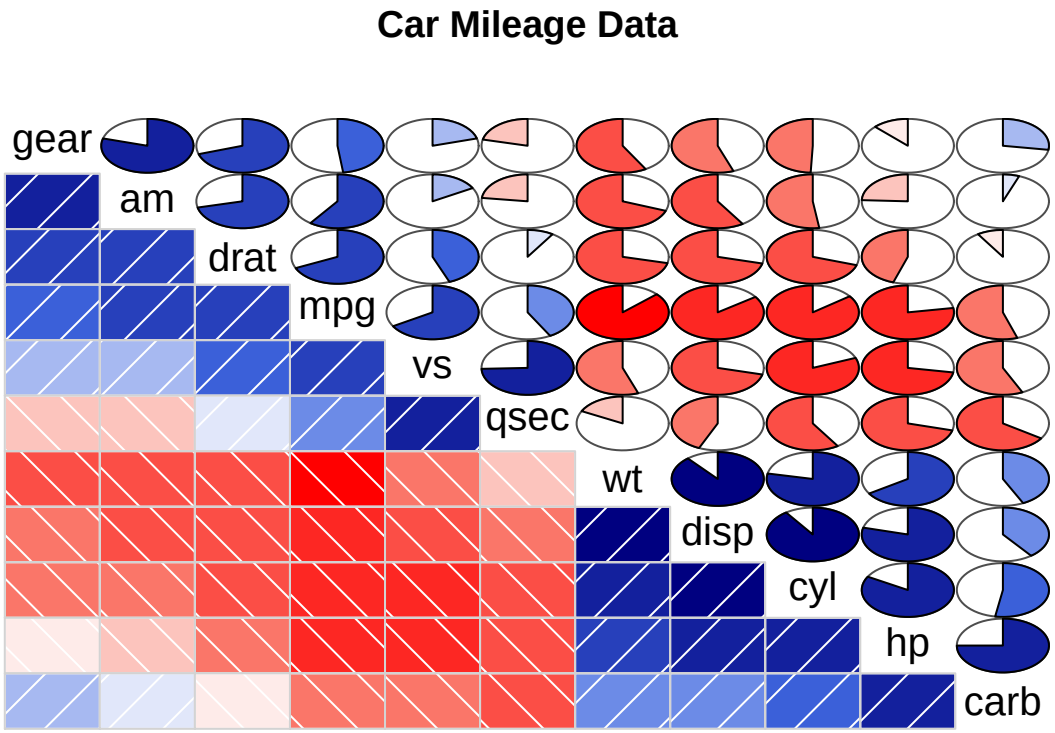
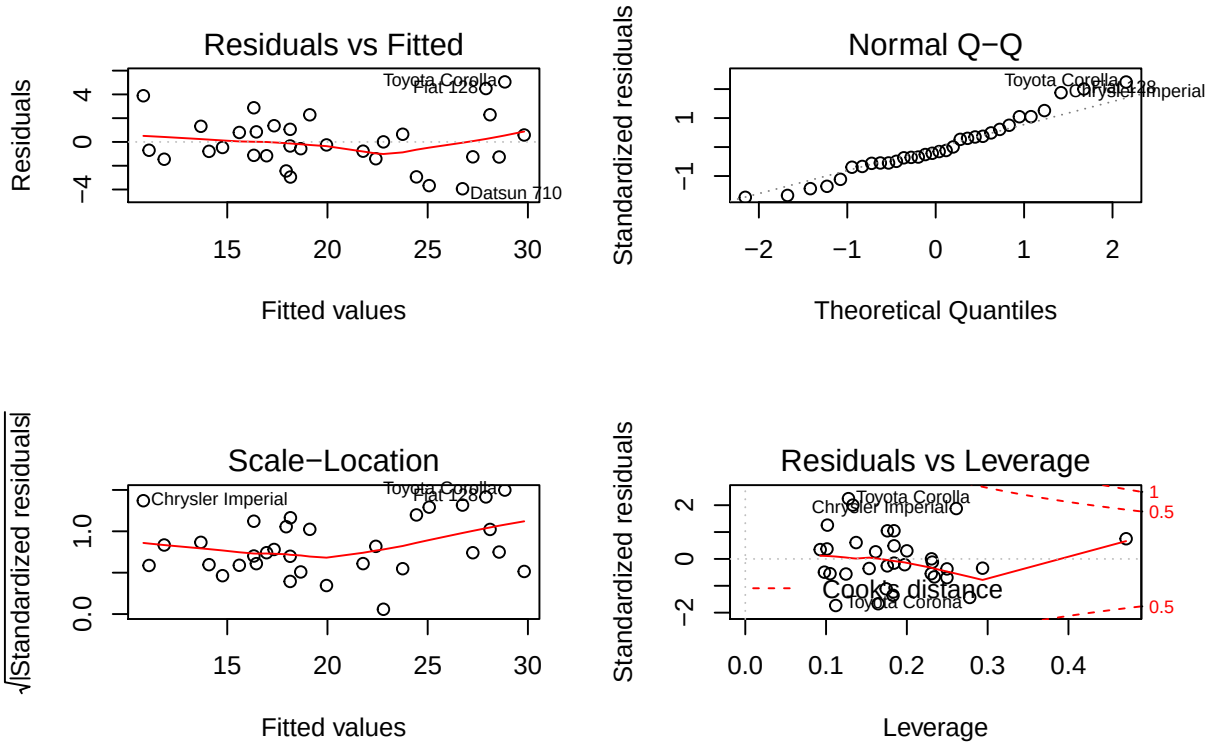


Figure 3: Diagnostics and Residual Data

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona