Coursera Regression Models Project

Jose Antonio (joseantonio@me.com)

28 ago 2016

Motor Trend: Vehicle Fuel Consumption

[GitHub Project Link:]https://github.com/joseantonio11/coursera-regression-model-project

1-Executive Summary

In this project we will explore some features that affect fuel consumption in miles per gallon (MPG) answering some questions about automatic and manual transmissions (am).

We are looking a dataset of a collection of cars (mtcars - Motor Trend Car Road Tests), and are interested in exploring the relationship between a set of variables and Miles Per Gallon (MPG). In particularly we want answer two questions:

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

2-Describing the Data

The data of this project are extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973/74 models).

The data consists of 32 observations on 11 variables.

- **mpg**: Miles/(US) gallon
- **cyl**: Number of cylinders
- **disp**: Displacement (cu.in.)
- **hp**: Gross horsepower
- **drat**: Rear axle ratio
- **wt**: Weight (lb/1000)
- **qsec**: 1/4 mile time
- **vs**: V/S
- **am**: Transmission (0 = automatic, 1 = manual)
- **gear**: Number of forward gears
- **carb**: Number of carburetors

3-Data Loading

```
library(datasets)
data(mtcars)
```

4-Factor Data (wrangling)

Lets coerce the "cyl", "vs", "gear", "carb" and "am" variables into factor variables:

```
mtcars$cyl <- as.factor(mtcars$cyl); mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear); mtcars$carb <- as.factor(mtcars$carb)
mtcars$am <- as.factor(mtcars$am)
```

To better understanding will resume am levels to "Auto" and "Manual"

```
levels(mtcars$am) <- c("Auto", "Manual")
```

5-Preliminary EDA

Lets look at dimension, structure and head of mtcars after the coersion.

```
dim(mtcars)  ## 32 observations and 11 variables

## [1] 32 11

head(mtcars) ## some observations to better understand mtcars

##                   mpg cyl disp  hp drat    wt  qsec vs     am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0 Manual    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0 Manual    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1 Manual    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1   Auto    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0   Auto    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1   Auto    3    1

str(mtcars) ## variable types after coersion

## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Auto","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

To answer our questions we are interested initially in the relation of two parameters that are Transmission (am) and Miles per Gallon (MPG).

Lets plotting this relation, to evaluate it (GRAPHIC FIGURE 1 AT APPENDIX).

```
## PLEASE, REFERS TO GRAPHIC FIGURE 1 AT APPENDIX
```

The plot show that manual transmissions have higher MPG's, but may have other variables that can play other role in determination of MPG, as `cyl`, `disp`, `hp`, `wt` and others. For example, it is common sense that the heavier the car, more likely he will fuel consumption.

Lets plot MPG with other variables to identify correlations (GRAPHIC FIGURE 2 AT APPENDIX)

```
## PLEASE, REFERS TO GRAPHIC FIGURE 2 AT APPENDIX
```

The graph shows that MPG has correlations with other variables than just `am`. To obtain a more accurate model, we need predicting MPG in correlation with other variables than `am`. Lets use some models to evaluate the correlations.

6-Models of Regression Analysis

Lets run some tests to compare the MPG with correlate variables.

(a) Simple Linear Regression Model

Below is the model to explain the MPG variability only with the transmission type (am).

```
fit1 <- lm(mpg ~ am, mtcars)
summary(fit1)

##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## 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 p-value are low (0.000285) and R-Squared is 0.3385. Before making any conclusions on the effect of transmission type on fuel efficiency, we look at the variances between several variables in the dataset.

Lets fitting all parameters of mtcars.

```
fitall <- lm(mpg ~ ., mtcars)
summary(fitall)
```

```
##
## Call:
## lm(formula = mpg ~ ., 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
## 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
## amManual     1.21212    3.21355   0.377   0.7113
## 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
```

Reading data, there is no coefficient significant at 0.05 level. R-Squared value has improved, but is not able to describe the remaining variance of the MPG variable and p-value show any significance anymore. We have to meet somewhete in the middle.

Lets use the R function STEP to do the variable selection.

(b) STEP function

```
bestFit <- step(fitall,direction="both",trace=FALSE)
summary(bestFit)
```

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

The Residual standard error of this model is 2.41 on 26 degrees of freedom. The Adjusted R-Squared value has increased to 0.8401 and the coefficients are significant at 0.05 significant level.

(d) Final Model Examination

Now we fit the model "mpg ~ wt + qsec + am" as final examination model.

```
lastModel <- lm(mpg ~ wt + qsec + am, data = mtcars)
summary(lastModel)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

(e) Residual Analisys

The resulting final model examination is dependant on the transmission (am), but also weight (wt) and 1/4 mile time (qsec). All have significant p-values and the R-squared is pretty good to (0.85)

Now let's look (amongst others) at the Residuals vs Fitted (GRAPHIC FIGURE 3 AT APPENDIX).

```
# PLEASE, REFERS TO GRAPHIC FIGURE 3 AT APPENDIX
```

The Normal Q-Q plot (GRAPHIC FIGURE 3 AT APPENDIX) show that residuals are normally distributed (points close to line). Scale-Location plot shows a constant variance due to a constant band pattern. Residuals and Leverage shows that most of the points are contained in the 0.5 bands

7 - CONCLUSION

All modes show that manual transmission will increase MPG, and this answer the first question (Is an automatic or manual transmission better for MPG?). Based on var lastModel using `mpg ~ wt + qsec + am` it is possible to conclude that manual transmission have more miles per gallon than automatic transmissions. With a `p < 0.05` confidence cars with manual transmission have near 3 more miles per gallon than automatic transmissions.

If we have more observations available, with same cars model, using manual and automatic transmission, could help us better answer the second question about (Quantify the MPG difference between automatica and manual transmissions?). With 32 observations it was not confidence conclude that this model fit all future observations.

APPENDIX - GRAPHICS

Figure 1 : Boxplots of "mpg" versus "am"

```
## GRAPHIC FIGURE 1
plot(mpg ~ am, data = mtcars, main = "MPG BY TRANSMISSION TYPE (AM)", xlab =
"Transmission Type (AM)", ylab = "Miles Per Gallon (MPG)")
```

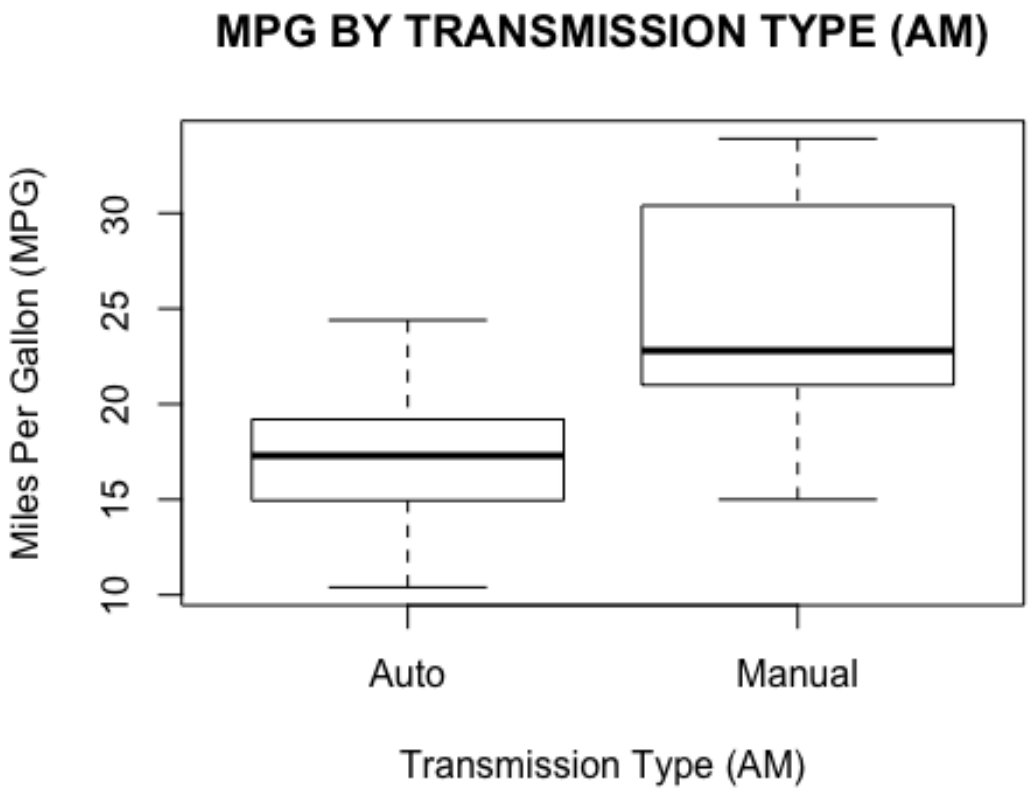


Figure 2 : Pairs graph

```
## GRAPHIC FIGURE 2
pairs(mtcars, panel = panel.smooth, main = "MTCARS PAIRS GRAPHS")
```

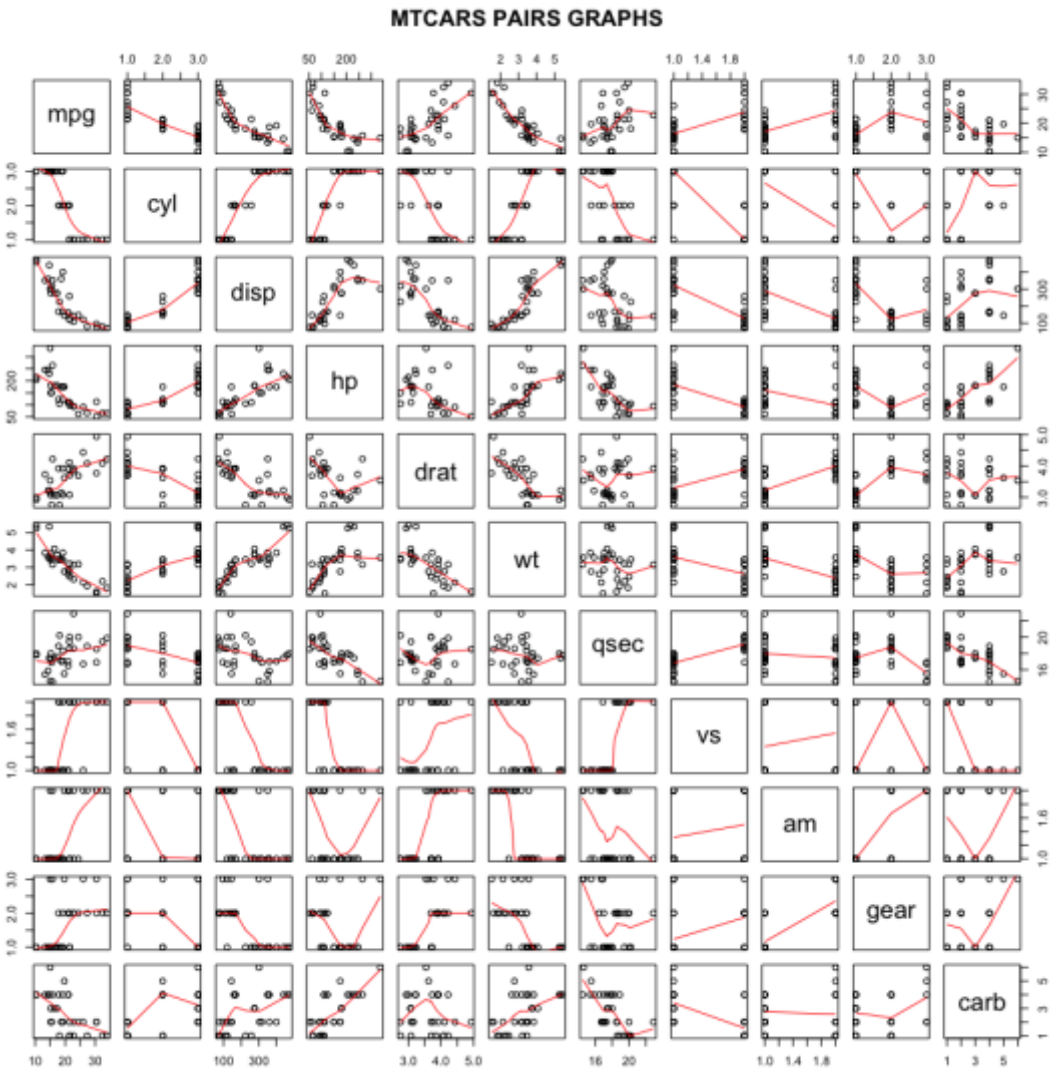
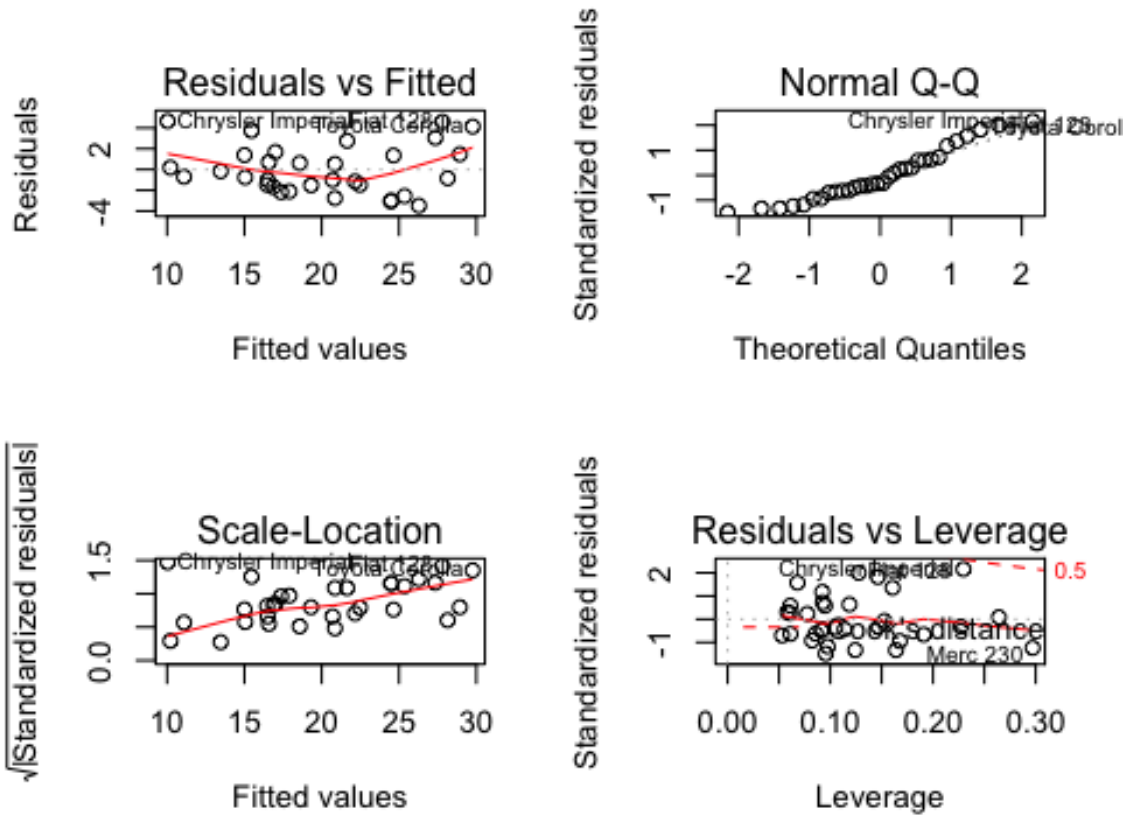


Figure 3 : Residual plots

```
# GRAPHIC FIGURE 3
par(mfrow = c(2, 2))
plot(lastModel)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
Merc 230