Project-RegressionModels

Miguel Porro

January 28, 2019

Coursera : Regression Models - Course Project

Exploring Motor Trend Car Road Tests

Executive Summary

This last project, was made for working with Motor Trend (a magazine about the automobile industry).

We will be analyzing the mtcars data set. Looking at a data set of a collection of cars (which contains 32 observations), they are interested in exploring the relationship between a set of variables and miles per gallon mpg (outcome).They are interested in the following two questions:

- Is an automatic or manual transmission better for mpg?
- Quantifying how different is the mpg between 'Automatic' and 'Manual' transmissions?

Data Preprocessing and Transformations

The first step is loading the DATA SET, performing the necessary data transformations by factoring some variables and look the data:

```
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
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 : 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 "Automatic","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
...
```

Exploratory Analysis AT THIS POINT

It has been explored various relationships between variables of interest and the outcomes. First, it was plotted the relationships between all the variables of the dataset (see plot 2 in the Appendix). From the plot 1 we notice that variables like cyl, disp, hp, drat, wt, vs and am seem to have some strong correlation with mpg. We will use linear models to quantify that in the next section.

Additionally we plot a boxplot of the variable mpg when am is 'Automatic' or 'Manual' (see plot 3 in the Appendix). This plot shows that the mpg incresases when the transmission is 'Manual'.

Regression Analysis

In this section it was built a linear regression models based on the different variables of interest and I tried to find out the best model fit. It will be compared it with the base model which we have been using ANOVA. After model selection, we will perform an analysis of residuals.

Build the Model

Based on plot 2 there are several variables seem to have high correlation with mpg. We willl build an initial model with all the variables as predictors and perfom stepwise model selection to select significant predictors for the final model. This is taken by the step method, which runs lm multiple times to build multiple regression models and select the best variables from them, using both forward selection and backward elimination methods by the AIC algorithm:

```
mod_init <- lm(mpg ~ ., data = mtcars)
mod_best <- step(mod_init, direction = "both")
```

As we can see, the best model obtained from the above computations have cyl, wt, hp and am as relevant variables:

```
summary(mod_best)

##
## 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
```

We can see that the adjusted R2 value is equal to 0.84 which is the maximum obtained considering all combinations of variables. Therefore we can conclude that more than 84% of the variability is explained by this model.

Now, using ANOVA, we will compare the base model with only am as the predictor variable and the best model obtained above:

```
mod_base <- lm(mpg ~ am, data = mtcars)
anova(mod_best, mod_base)
```

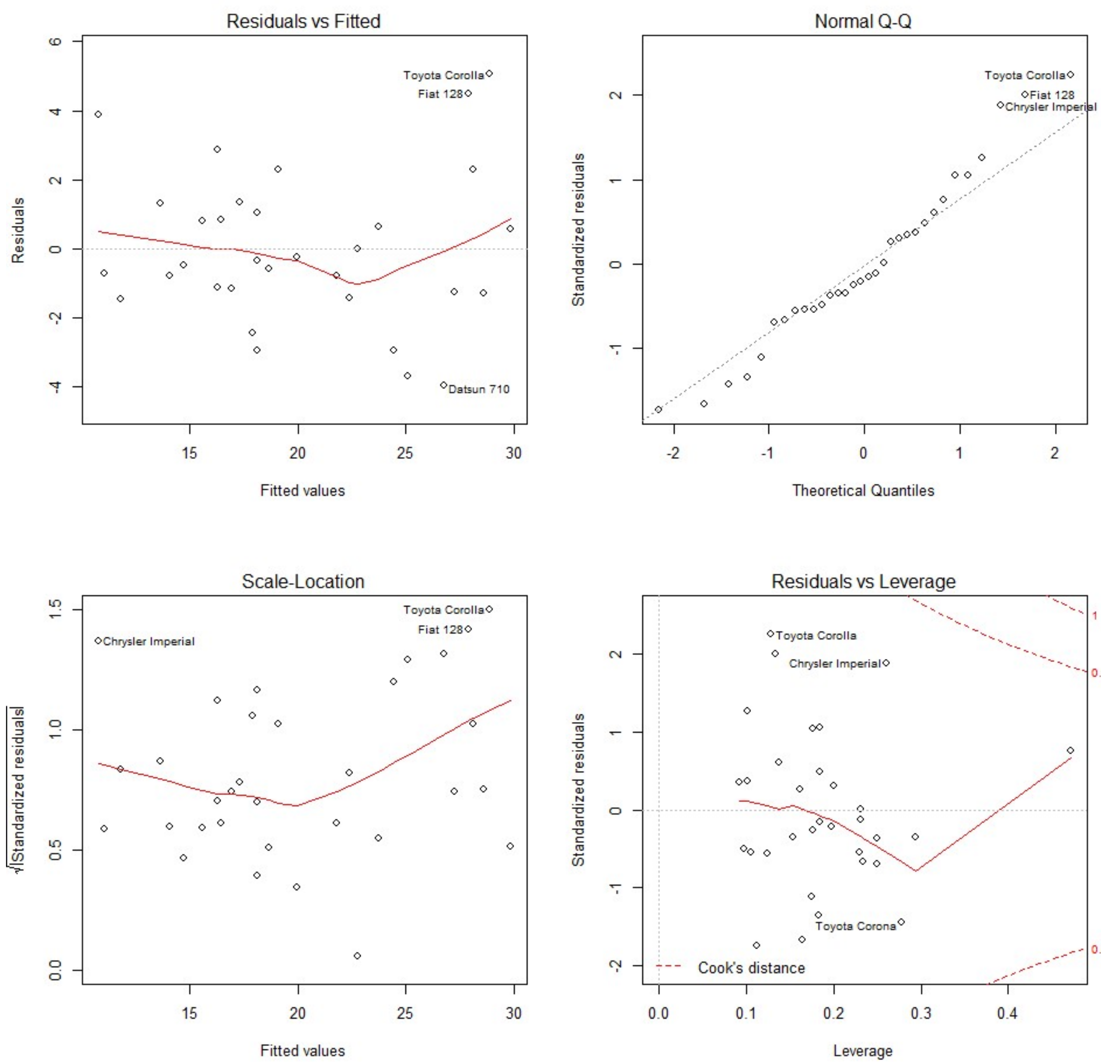
```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     26 151.03
## 2     30 720.90 -4   -569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at this result, the p-value obtained is highly significant, and we reject the null hypothesis that the confounder variables cyl, hp and wt do not contribute to the accuracy of the model.

Analysis of the Residuals and Diagnostics

Now we explore the residual plots of our regression model and also compute some of the regression diagnostics for our model to find out some interesting leverage points (often called as outliers) in the data set:

```
par(mfrow=c(2,2))
plot(mod_best, which=1)
plot(mod_best, which=2)
plot(mod_best, which=3)
plot(mod_best, which=5)
```



From the plots it can be concluded that:

- The Residuals vs Fitted plot shows random points on the plot that verifies the independence condition.
- In the Normal Q-Q plot the points mostly fall on the line indicating that the residuals are normally distributed.
- In the Scale-Location plot the points are in a constant band pattern, indicating constant variance.
- Finally, the Residuals vs Leverage plot shows some points of interest (outliers or leverage points) are in the top right corner.

Now we will compute some regression diagnostics of our model to find out these interesting leverage points. We compute top three points in each case of influence measures.

```
lev <- hatvalues(mod_best)
tail(sort(lev),3)

##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671

inf <- dfbetas(mod_best)
tail(sort(inf[,6]),3)

## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

Looking at this result we see that they the same cars shown in the residual plots.

Statistical Inference

In resume, it was performed a t-test assuming that the transmission data has a normal distribution and it will be seen that the manual and automatic transmissions are significatively different:

```
t.test(mpg ~ am, data = mtcars)

## 
##  Welch Two Sample t-test
## 
## data:  mpg by am
## 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:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Conclusions

From the summary(mod_best) it can be concluding that:

- Miles per gallon mpg will increase by 1.81 in cars with 'Manual' transmission compared to cars with 'Automatic' transmission (adjusted by hp, cyl, and wt). So, the conclusion for Motor Trend Magazine is: 'Manual' transmission is better for mpg.
- Miles per gallon mpg will decrease by 2.5 for every 1000 lb of increase in wt (adjusted by hp, cyl, and am).
- Miles per gallon mpg decreases with increase of hp.
- Miles per gallon mpg will decrease by a factor of 3 and 2.2 if number of cylinders cyl increases from 4 to 6 and 8, respectively (adjusted by hp, wt, and am).

Appendix

This is a boxplot of the variable mpg when am is ‘Automatic’ or ‘Manual’:

```
boxplot(mpg ~ am, data=mtcars, main="Plot 3: Miles per gallon by
Transmission type",
   xlab="Transmission type", ylab="Miles Per Gallon")
```