Regression Models Course Project

Created by Eduardo Ferreira - 2014

Abstract

In this project, a data set of a collection of cars will be analysis, especially the relationship between a set of variables and miles per gallon. Particularly, following question will be answered by the data set:

We would like to know if automatic is better ou not than manual transmission for MPG. We would like to know how different is the MPG between automatc and manual transmission;

Briefly, cars with a manual transmission have a slightly better than automatic for MPG, but this different is statistical insignificant.

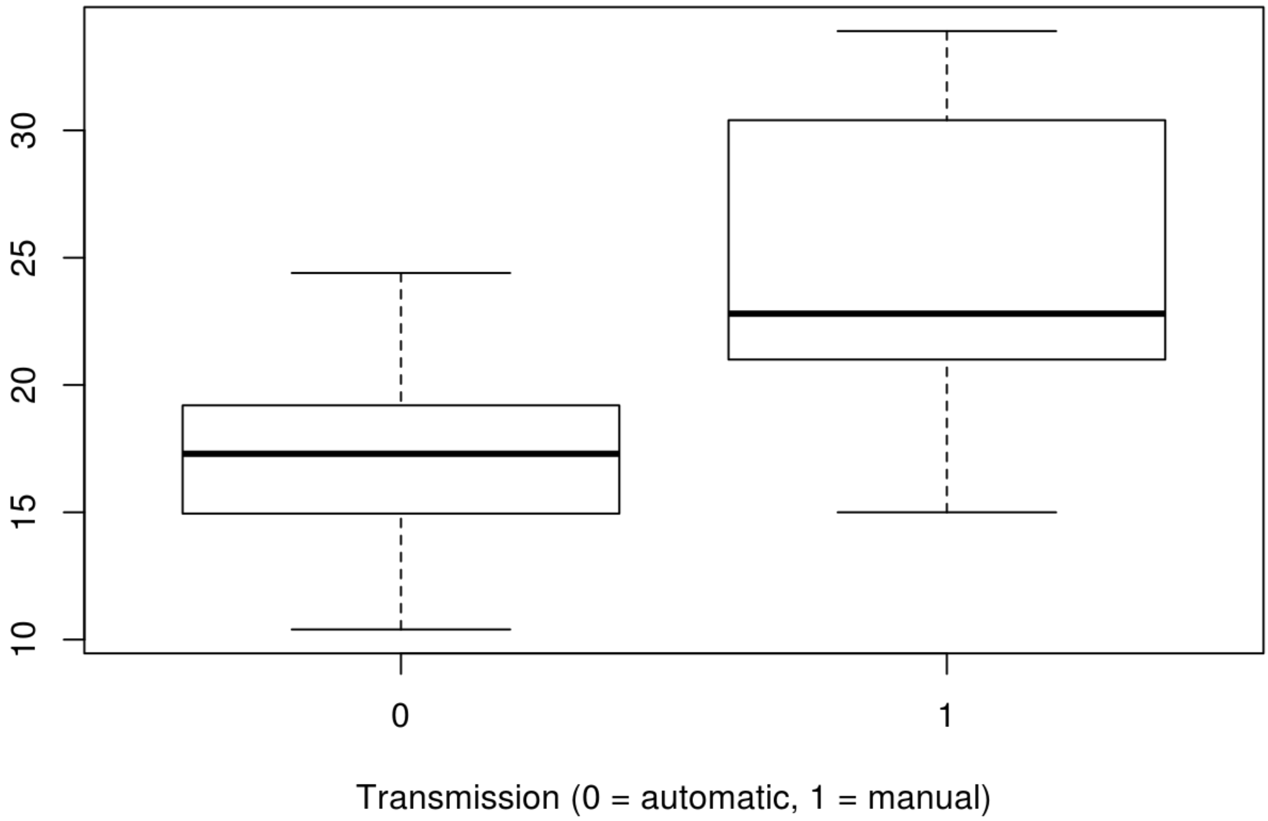
Analysis

mtcars is a data frame with 32 observations on 11 variables. Frist, using scatter plot matrices to show the data, as Figure 1 in appendix. Here you can see the bivariate relationship among all the variables. And the variable of mtcars$vsandmtcars$am should be category variables. So fix it.

```
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
```

We mainly focus on the relationship between variables mpg (Miles/(US) gallon) and am (Transmission). Box plot shows that there’s a good separation of groups based on gas mileage.

```
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission (0 = automatic, 1 = manual)")
```



But there are 10 predictor variables in the data set. Some may play bigger role to determination of mpg. An analysis of variance model is performed, as:

```
analysis <- aov(mpg ~ ., data = mtcars)
summary(analysis)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## cyl          1    818     818  116.42  5e-10 ***
## disp         1     38      38    5.35 0.0309 *
## hp           1      9       9    1.33 0.2610
## drat         1     16      16    2.34 0.1406
## wt           1     77      77   11.03 0.0032 **
## qsec         1      4       4    0.56 0.4617
## vs           1      0       0    0.02 0.8932
## am           1     14      14    2.06 0.1659
## gear         1      1       1    0.14 0.7137
## carb         1      0       0    0.06 0.8122
## Residuals   21    147       7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously, variables with p-value below 0.05 are more important. We choose cyl, disp, wt, drat, am as predictor variables for first model.

```
lm <- lm(mpg ~ cyl + disp + wt + drat + am, data = mtcars)
summary(lm)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + drat + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.383 -0.473  1.323  6.060
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.29638    7.53839    5.48  9.6e-06 ***
## cyl         -1.79400    0.65054   -2.76   0.0105 *
## disp         0.00737    0.01232    0.60   0.5546
## wt          -3.58704    1.21050   -2.96   0.0064 **
## drat        -0.09363    1.54878   -0.06   0.9523
## am1          0.17298    1.53004    0.11   0.9109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 26 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.801
## F-statistic: 25.9 on 5 and 26 DF,  p-value: 2.53e-09
```

Coefficient of drat has a p-value of 0.95226, refine the model as:

```
lm <- lm(mpg ~ cyl + disp + wt + am, data = mtcars)
summary(lm)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.058
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.8983     3.6015   11.36  8.7e-12 ***
## cyl          -1.7842     0.6182   -2.89   0.0076 **
## disp          0.0074     0.0121    0.61   0.5451
## wt           -3.5834     1.1865   -3.02   0.0055 **
## am1           0.1291     1.3215    0.10   0.9229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.64 on 27 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.808
## F-statistic: 33.6 on 4 and 27 DF,  p-value: 4.04e-10
```

And remove variable disp, now the model is:

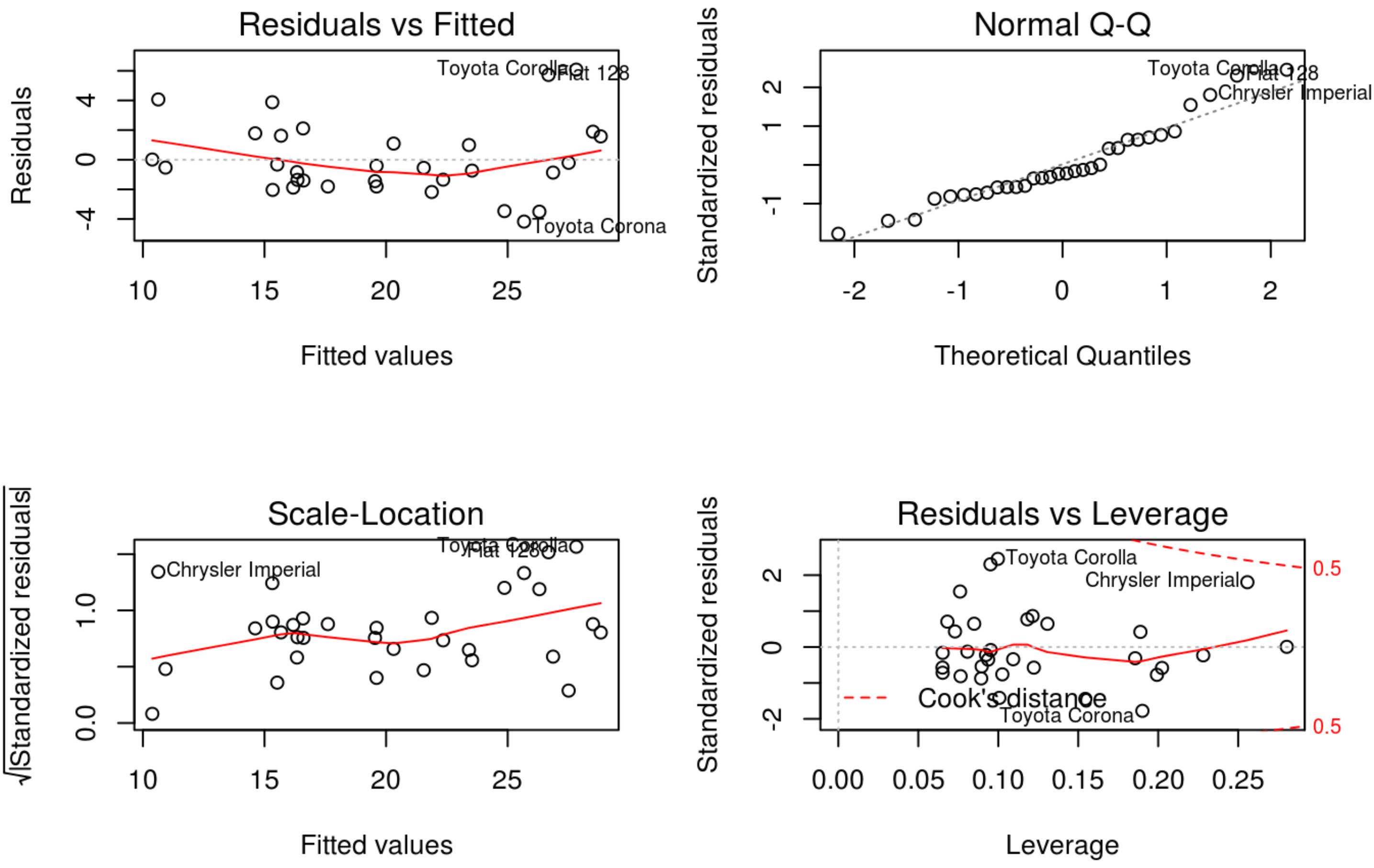
```
lm <- lm(mpg ~ cyl + wt + am, data = mtcars)
summary(lm)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.173 -1.534 -0.539  1.586  6.081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.418      2.641   14.92  7.4e-15 ***
## cyl           -1.510      0.422   -3.58   0.0013 **
## wt            -3.125      0.911   -3.43   0.0019 **
## am1            0.176      1.304    0.14   0.8933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 28 degrees of freedom
## Multiple R-squared:  0.83,   Adjusted R-squared:  0.812
## F-statistic: 45.7 on 3 and 28 DF,  p-value: 6.51e-11
```

The adjusted r-squared is 0.83 and this is our final model. Clearly, with cylinders and weight as confounding variables, the ceofficient of the am variable is small but has a large p-value. We cannot reject the hypothesis that the coefficient of am is 0.

To diagnostic the model, we apply the plot() to the object returned by the lm(). There is no discernible pattern found according to upper left graph. The normal Q-Q plot (upper right) indicates the model met the normality assumption. Scale-Location graph (bottom left) shows constant variance assumption are satisfied.

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

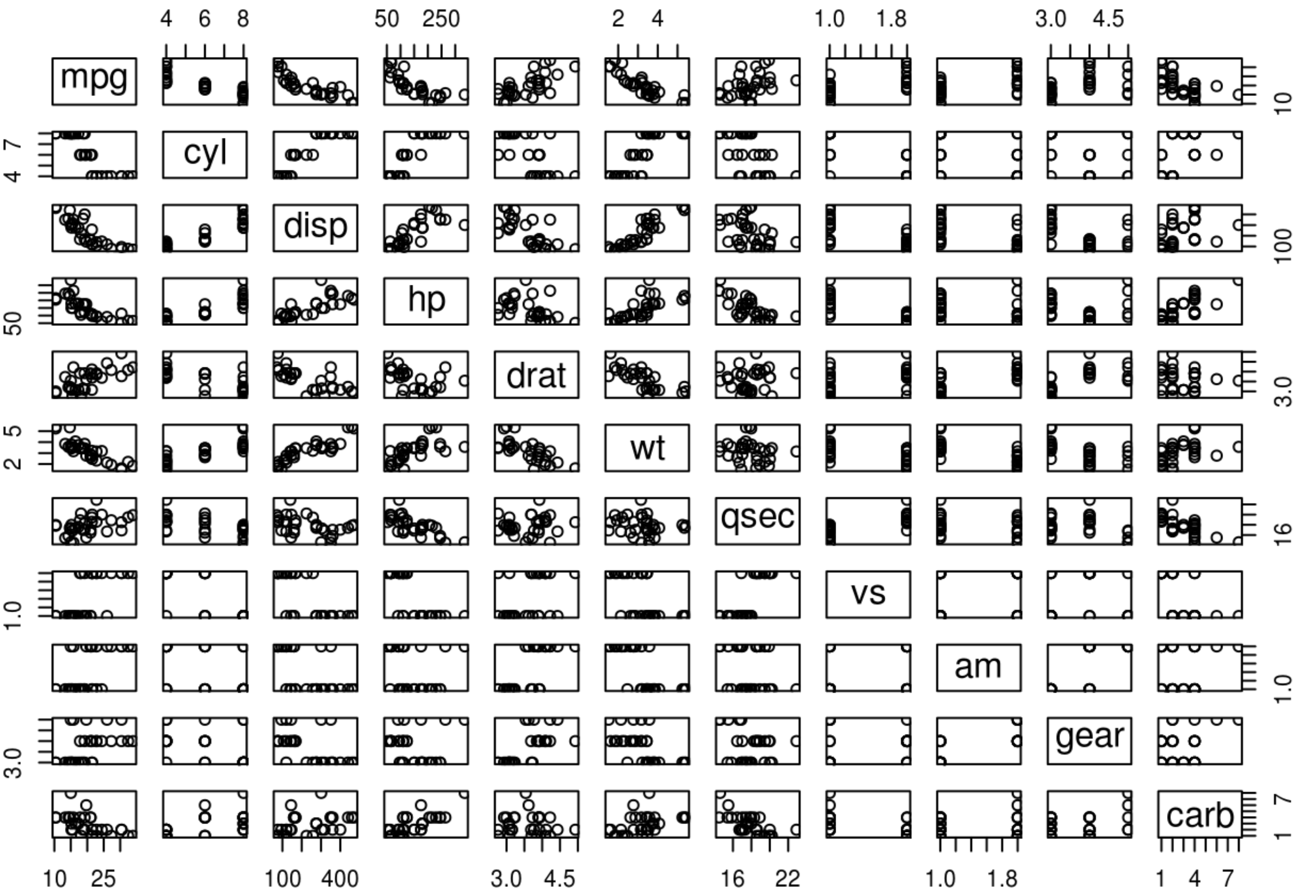


Now we can conclude that weight and number of cylinders play important role to determination of mpg.

Appendix

Figure 1

```
pairs(mtcars)
```


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb