Regression Models: Course Project

Amin

Friday, October 23, 2015

Summary

The goal of this project is to investigate the relationship between a set of variables and miles per gallon (mpg) given a dataset of a collection of cars (mtcars). We are particularly interested in finding any relation between MPG and the type of transmission (automatic or manual), and quantifying this relation. We use exploratory data analysis and regression model to explore this relationship.

Global Setting

Here is the global setting for the code used throughout the report.

```r
echo = TRUE
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.2.2
```

Loading the data

Firstly, we load the data.

```r
data(mtcars)
head(mtcars, n = 2)
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

Exploratory Data Analysis

To do the analysis, we start with some basic exploratory data analysis. To explore the relation between mpg and transmisson, we plot the boxplot of these two variables. The plot shows that we expected value of mpg is higher for the manual transmission. To explore the relationship with other variables, we use the pairs function to explore the correlation between all variables. As the plot shows (see Appendix), there is high correlation between mpg and some other variables, especially with "wt".

Regression Analysis

In this section, we use the regression model to build our model of the data. In our initial model, we use all the variables as the predictor of "mpg".

```r
fit0 <- lm(mpg ~ ., data = mtcars)
summary(fit0)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

This model has the residual standard error of 2.65 on 21 degrees of freedom and adjusted R-square of 0.81 which means that 81% of the variability is explained by the model. To improve the model, we use the step function to perform the selection of the best variables.

```r
optimalfit <- step(fit0, direction= "both")
```

```r
summary(optimalfit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

This model suggests “am”" as independent variable and “wt” and “qsec” as confounders. The residual standard error is 2.459 on 28 degrees of freedom and the adjusted R-squared is 0.83 meaning that it explains 83% of variability. We can clearly see the improvement in this model comparing to the initial model.

Residual Analysis

In this part of the report, we use the residual plots to do regression diagnostics for our model. Please refer to the appendix for the plot. The scatter plot of the residuals vs fitted values shows a completely random pattern. It means that the error in the model is independent from the predictors, and this is the best linear model to predict the mpg. The Normal Q-Q plot shows that the points are on the or close the line meaning that the residual are approximately normally distributed. The Scale-Location plot has a constant width in the points, so we do not suffer from heteroscedasticity. There are some potential points of interest in the plots that may indicate values of increased leverage of outliers.

The data points with the most leverage in the model can be found by using the function hatvalues().

```
highlev <- hatvalues(optimalfit)
tail(sort(highlev),3)
```

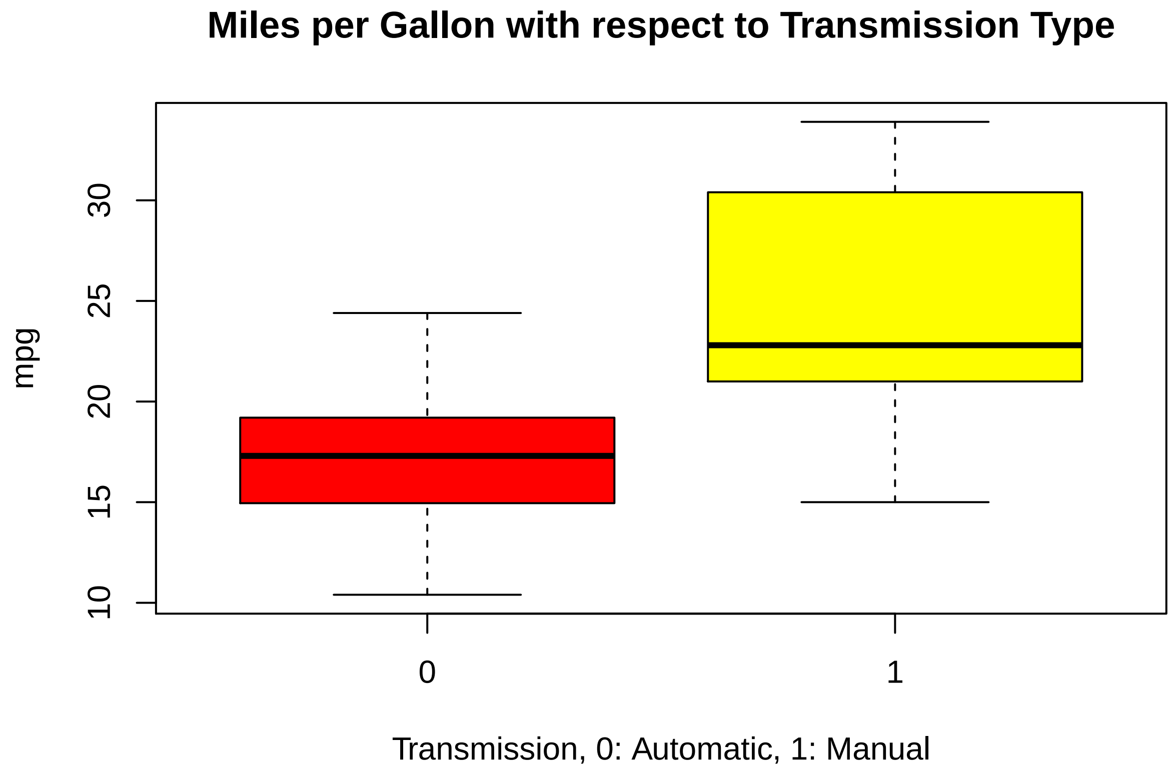
```
##   Chrysler Imperial Lincoln Continental            Merc 230
##           0.2296338           0.2642151           0.2970422
```

This means that our analysis was correct. These are the same cars as indicated in the residual plots.

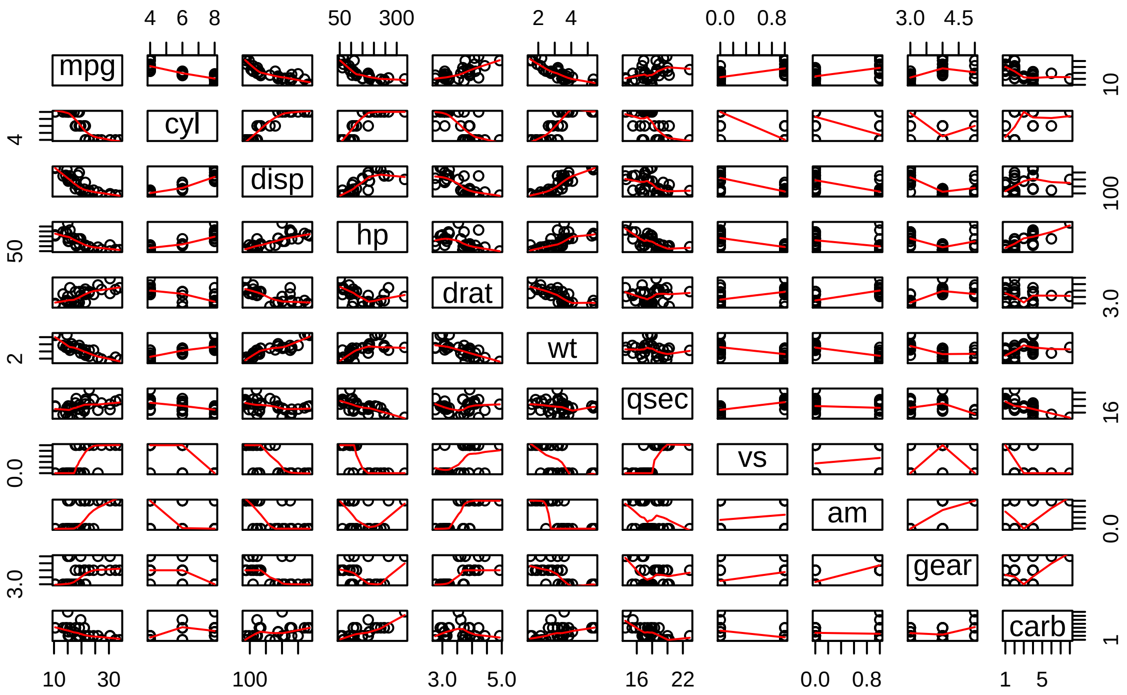
Appendix

This appendix contains all the plots for the report.

```
boxplot(mtcars$mpg ~ mtcars$am, xlab="Transmission, 0: Automatic, 1: Manual", col = (c("red","yellow"))
        main="Miles per Gallon with respect to Transmission Type")
```

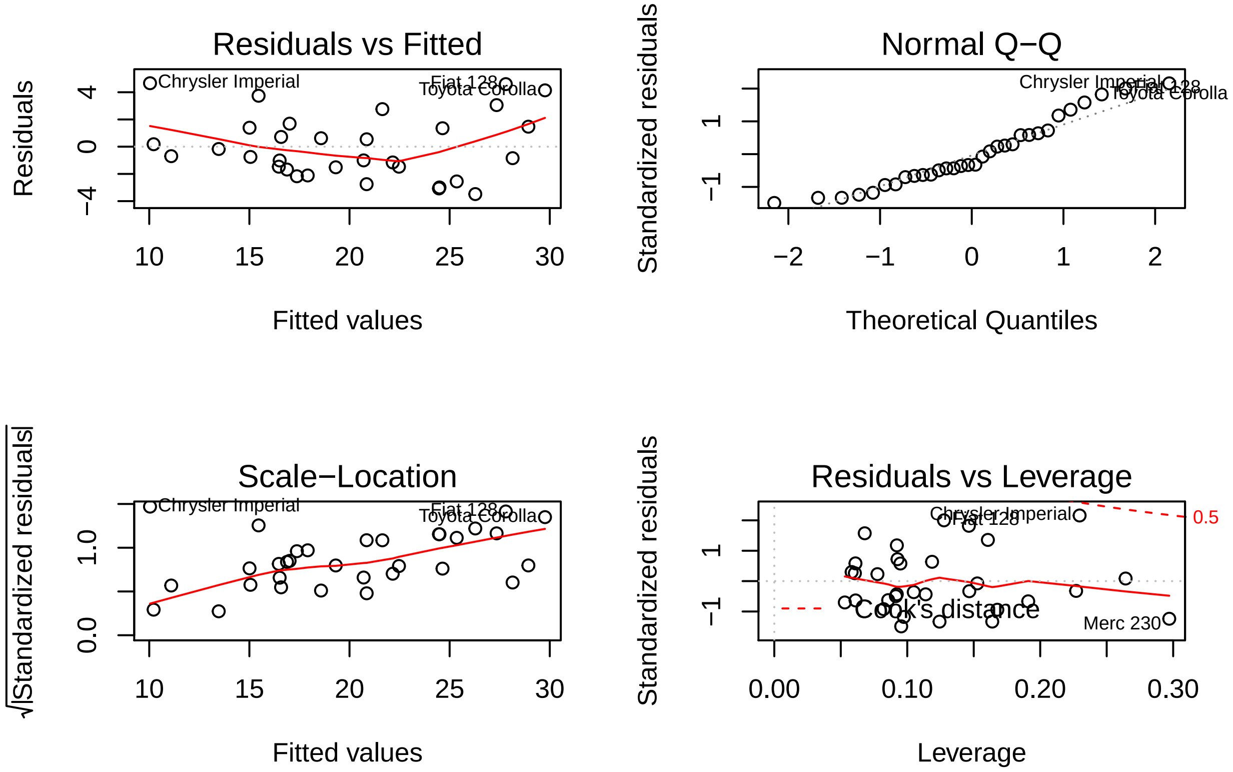


```r
pairs(mtcars, panel=panel.smooth, main="The correlation between all variables of mtcars dataset")
```



The correlation between all variables of mtcars dataset

```r
par(mfrow=c(2, 2))
plot(optimalfit)
```


Residuals vs Fitted
Residuals
4
0
−4
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale−Location
√|Standardized residuals|
0.0
1.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
1
−1
0.00
0.10
0.20
0.30
Leverage
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5