Regression Model Project

Jason Liu

Oct 1, 2016

Statement This is the graduation project of Coursera course: Regression models, taught by JHU. The project can be viewed as a demonstration of regression techniques.

Load Data and Related Packages

```
library(datasets)
library(ggplot2)
library(GGally)
data(mtcars)
```

Data Exploration

First step of regression analysis is to explore the dataset. With the built-in function 'head' we quickly have a look into th dataset.

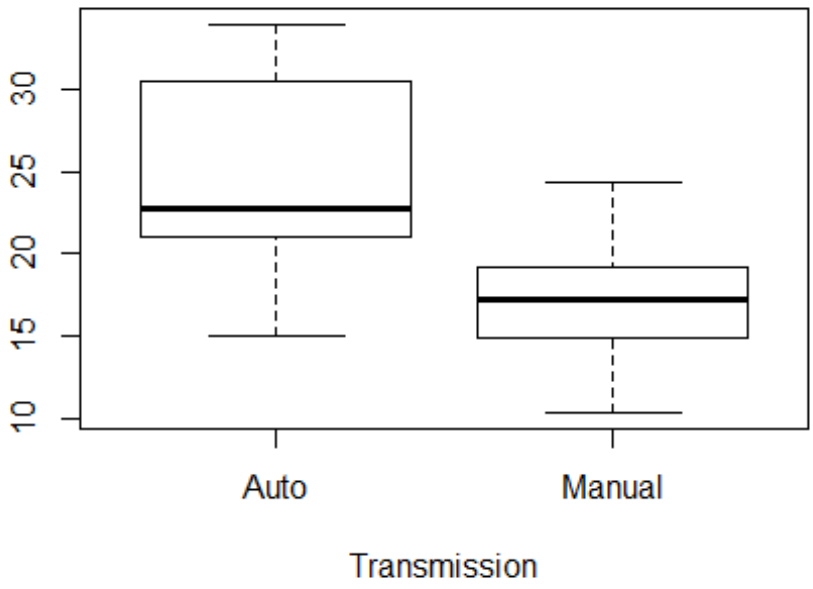
```
head(mtcars)
```

```
##                    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  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Subsquently, we plot the 'pair plots' of each variables in order to understand the correlations among different variables. (See Appendix) The variable,MPG, the abbreviation of 'Mile per Gallon', serves as the response variable of the model. Before ploting, we need to transfer two variables 'vs' and 'am' to the forms of factors as they do not have numeric meanings.

It is obvious that some variables are distributed discretely (i.e. cyl) while others are quite continous. The box plot is generated in order to briefly explore the oil consumptions of auto gear and mannual gear cars.

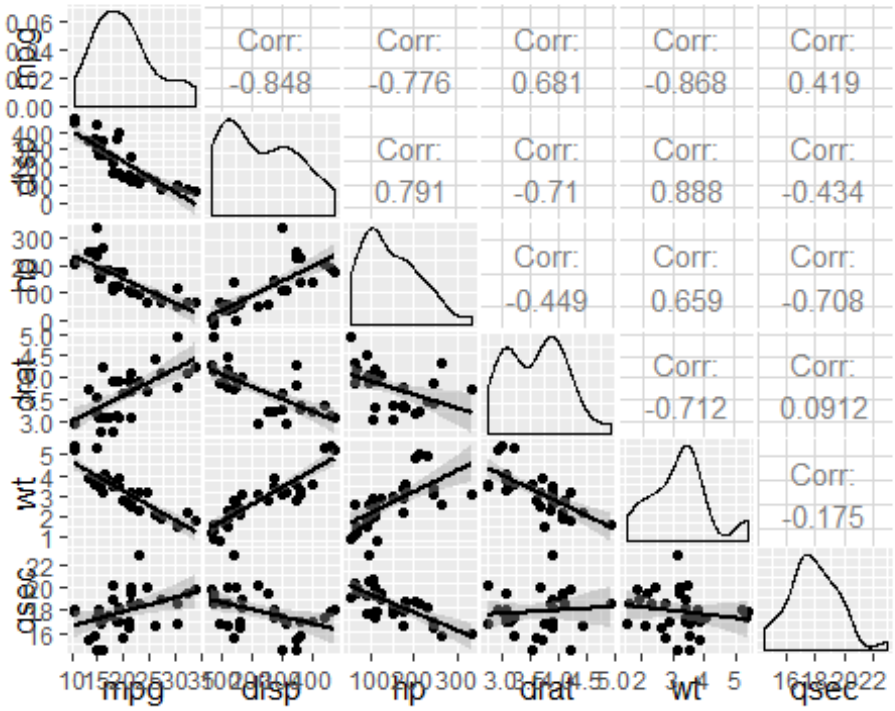
```
tmpData<-mtcars[,c(1,9)]
tmpData$am<-ifelse(tmpData$am=='1',tmpData$am<-'Auto',tmpData$am<-
'Manual')
boxplot(mpg ~ am, data = tmpData, xlab = "Transmission")
```



It is obvious that the auto gear car is always more gas-consuming. Before running the regression, we have already known the relationship between MPG and gear types to some extents.

Regression Analysis

The next step is to run regression analysis on the dataset. We first select several continous variables and see the regression plots beween the response variable and each of them.

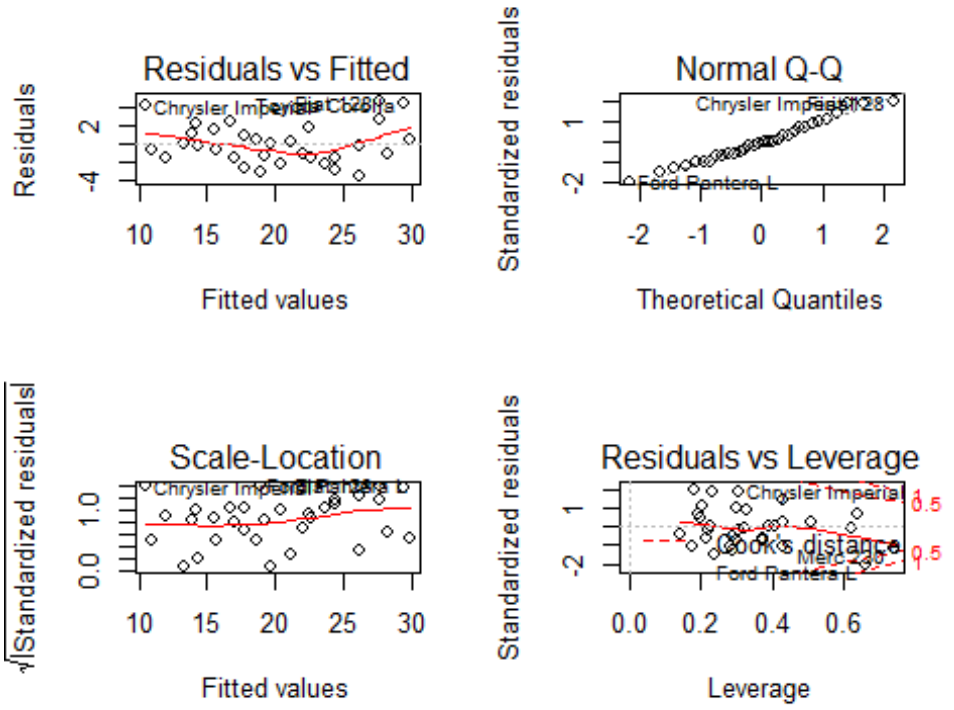


The first column of the plot shows the regression plots of the response variable and each individual variable. It is good to see that except for the variable 'qsec', the regession lines between 'mpg' and other variables look good. Visually judged, the outlier with huge leverage does not have profound impact on the results. However, through the other columns of the plot, we can find that there are strong correlations existing among some features, which could result in multicolinearity if all features are included in the regression function.

We first create a regression model with the inclusion of all variables. The summary of model is presented below. Also, we plot the diagnostic plots of the regression.

```
##
## 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
## vs1          0.31776    2.10451   0.151   0.8814
## am1          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
```



The influences of features on response variables are not statistically significiant if we have a look into the p values. However, considering we just have 32 rows of data for the regression, the p values do not mean everything. The adjusted R square is not bad, and the diagnostic of residuals reveals several facts: the residuals are patternless and normally distributed, which is a good sign of the model-fitting; the outliers with high leverage do not have a huge impact on the variance of residuals etc. Through the coefficients of features, we can find that 'drat', 'wt','aml' and 'gear' etc. have stong impacts on the dependent variable. Therefore, we use 'step' function to automatically choose the regressors that should be included in the model.

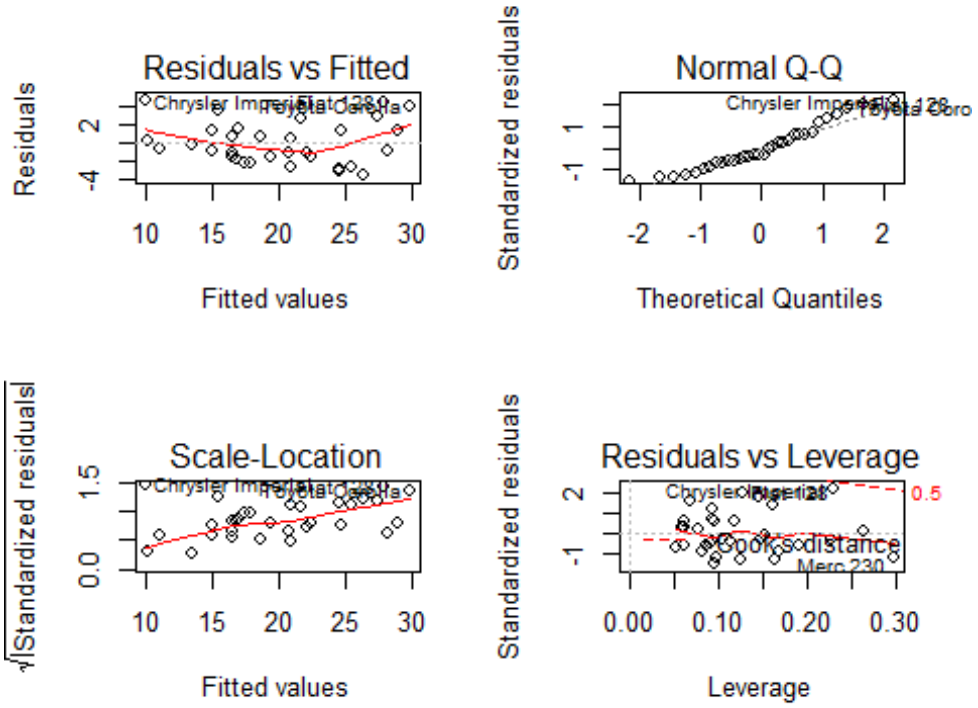
```
reg_adjusted<- step(reg,direction = 'both')

##
## 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 ***
## am1           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
```

The adjusted model incorporates three variables: 'wt','qsec' and 'am'. The summary of new model reveals a fact that the new model is way better than the previous one from the perspects of R-squared and p values. We also performs the diagnosis of residuals for the new model.

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



The diagnosis of residuals is quite similar with the previous one. But one obvious problem is identified that a outlier may exist in the dataset since the residual-leverage plot contains a point with both high leverage and influence. The problem becomes obvious as the quantity of variables shrinks.

Conclusion

In short, we can conclude that: 1. The type of transmission really has impact on the consumption of gasoline. Specifically, the auto transmission car will consume more gas. 2. Not all features in the dataset contributes a lot to the response variable. However, considering that the limited size of dataset, we cannot make a concrete conclusion regarding the inclusion of variables.

Appendix

Pairs plot of variables.