Regression Analysis of Mileage Per Gallon Performance

Abstract: This project focuses on the mileage per gallon performance(MPG) for various vehicles. The multiple linear regression approach is introduced to fit the model. Before modelling data set cleaning is performed firstly, followed by statistical descriptions,correlation computation and feature selection. Eventually, the experimental data analysis for 392 samples indicated that the response(MPG) do have a linear relationship with these predictors, and the predictors(model.year and origin) do have significant influence on MPG.
Keywords: Mileage Per Gallon(MPG);

1. Introduction

The fuel economy has been a consistently significant indicator for automobiles, indicating the relationship between the distance traveled by a vehicle and fuel consumed. Generally, it can be expressed in two ways: one is the unites of fuel per fixed distance, usually shown as liters per 100 kilometers and used in most European countries, another is miles or mileage per gallon (MPG) which has been commonly used in the United States. Therefore, figuring out what groups of factors truly affect the MPG of a vehicle will be naturally considered about by automakers to manufacture high-qualified and energy-saving vehicles, and by consumers to make sense of how to maintain the cars' performance logically. In other words, what factors might significantly contribute to the MPG of a vehicle can be an interesting and worthy exploring work, involving the relationship between the response and the predictors.

From previous research outcomes related to the relations of the response and the predictors, it is pronounced that many statistical learning approaches can be highly efficient in solving these problems, where one of the most useful methods is the regression fitting and its various extensions, applied widely in linear and nonlinear fitting, quantitative and qualitative variables. Here in my case, the multiple linear regression method is introduced to fit the mathematical relationship between the response (MPG) and some quantitative predictors.

2. Question

As mentioned before, in this project the goal is to make sense two questions:

1) What is the relationship between the response (MPG) and the predictors (such as cylinders, displacement,horsepower,etc.) ?

2) and what predictors do significantly have influence on the response ?

3. Data set

3.1 Data source

The data set for all analysis in this project related to the MPG originally derives from 'Kaggle', a free dateset open source website. *https://www.kaggle.com/uciml/autompg-dataset*

3.2 Data properties

The basic information o this data set chosen is listed as below in Tab.1.

Tab.1 Basic information of data set

Basic information	
Title	Auto-Mpg Data
Sources	This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. (c) Date: July 7, 1993

Number of Instances	398
Number of Attributes	9 including 8 numerical and 1 string
Attribute Information	mpg: continuous cylinders: multi-valued discrete displacement: continuous horsepower: continuous weight: continuous acceleration: continuous model year: multi-valued discrete origin: multi-valued discrete car name: string (unique for each instance)

Th part of this data set is also displayed as in Tab.2.

Tab.2 Part of Auto-mpg data set

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
14	8	455	225	3086	10	70	1	buick estate wagon (sw)
24	4	113	95	2372	15	70	3	toyota corona mark ii
22	6	198	95	2833	15.5	70	1	plymouth duster
18	6	199	97	2774	15.5	70	1	amc hornet
21	6	200	85	2587	16	70	1	ford maverick
27	4	97	88	2130	14.5	70	3	datsun pl510
26	4	97	46	1835	20.5	70	2	volkswagen 1131 deluxe sedan
26	4	121	113	2234	12.5	70	2	bmw 2002
21	6	199	90	2648	15	70	1	amc gremlin

4. Modeling Methodology

4.1 Data Cleaning/Statistical descriptions

Firstly, importing the original data set into my workspace based on R language by ‘readcsv’ function is performed, followed by a missing data deletion using ‘na.omit’ function and a summary of statistical terms for each variables, represented as below in Tab.3 and Tab.4.

Tab.3 Code for Data cleaning

```
>Auto = read.csv("E:/STA5104/autompg-dataset/auto-mpg.csv", header=T, na.strings="?")
>dim(Auto)          # returen sample size and number of attributes
[1] 398    9
>Auto1 = na.omit(Auto)     # 6 samples are deleted
>dim(Auto1)
[1] 392    9
>summary(Auto1)   # statistical descriptions
```

Tab.4 Statistical terms

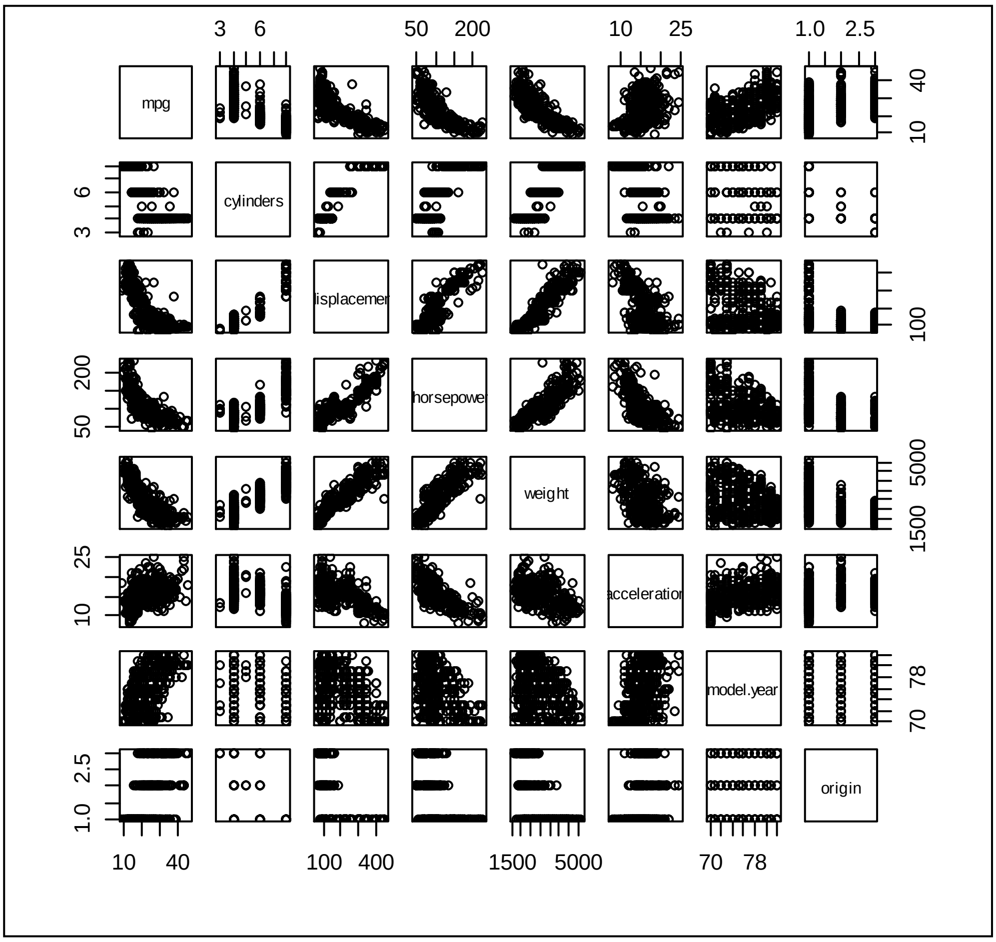
```
      mpg          cylinders      displacement     horsepower        weight
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140

  acceleration     model.year        origin                      car.name
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
 1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
 Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
 Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
 3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
 Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
                                                 (Other)           :365
```

4.2 Correlation coefficients

For graphically and mathematically identifying the correlations between variables, both ‘pairs’ and ‘cor’ functions are applied in this case, described like in Fig.1 and Tab.5.

Fig.1 pairs of scatterplots



Tab.5 Correlation coefficients between variables

	mpg	cylinders	displacement	horsepower	weight	acceleration	model.year	origin
mpg	1.0000000	-0.7776175	-0.8051269	-0.7784268	-0.8322442	0.4233285	0.5805410	0.5652088
cylinders	-0.7776175	1.0000000	0.9508233	0.8429834	0.8975273	-0.5046834	-0.3456474	-0.5689316
displacement	-0.8051269	0.9508233	1.0000000	0.8972570	0.9329944	-0.5438005	-0.3698552	-0.6145351
horsepower	-0.7784268	0.8429834	0.8972570	1.0000000	0.8645377	-0.6891955	-0.4163615	-0.4551715
weight	-0.8322442	0.8975273	0.9329944	0.8645377	1.0000000	-0.4168392	-0.3091199	-0.5850054
acceleration	0.4233285	-0.5046834	-0.5438005	-0.6891955	-0.4168392	1.0000000	0.2903161	0.2127458
model.year	0.5805410	-0.3456474	-0.3698552	-0.4163615	-0.3091199	0.2903161	1.0000000	0.1815277
origin	0.5652088	-0.5689316	-0.6145351	-0.4551715	-0.5850054	0.2127458	0.1815277	1.0000000

From Tab.5, it is obviously seen that there indeed exists strong associations between the predictors(cylinders, displacement, horsepower, weight, acceleration, model.year, and origin) and the response(mpg). Moreover, the collinearity can be found in this case between some predictors like cylinders and displacement, horsepower,weight. Note that 'car.name' in the dataset is not considered into the mode as the MPG is more relevant to the numerical physical variables.

4.3 Feature selection

There are various feature selection methods, such as best subset selection, forward/backward stepwise selection, and ridge/lasso regression. All of them enable redundant features/variables/predictors to be reduced logically. Here in this project, backward stepwise selection is used to minimize the number of predictors in the fitting model,combined with decision rules like Cp, BIC and adjusted R2, demonstrated as below in Tab.6 and Fig.2.

Tab.6 Backward stepwise selection by various decision rules

```
library(dplyr)
Auto2=select(Auto1,-car.name)
library(leaps)
regfit.full=regsubsets(mpg~., data=Auto2,nvmax=7,method="backward")    #use backward stepwise
reg.summary= summary(regfit.full)        #obtain the selection results
# best model according to Cp,BIC, adjusted R2
which.min(reg.summary$cp)    # return 6 as the best number of predictors by Cp
which.min(reg.summary$bic)    #return 3 as the best number of predictors by BIC
which.max(reg.summary$adjr2) # return 6 as the best number of predictors by adjusted R2
# plot the selections of Cp,BIC and adjusted R2
par(mfrow=c(1,3))
# Cp
plot(reg.summary$cp,xlab="subset size",ylab="cp",type="l")
points(6,reg.summary$cp[6],col="red",cex=2,pch=20)
# BIC
plot(reg.summary$bic,xlab="subset size",ylab="BIC",type="l")
points(3,reg.summary$bic[3],col="red",cex=2,pch=20)
# adjusted R2
plot(reg.summary$adjr2,xlab="subset size",ylab="Adjusted R2",type="l")
points(6,reg.summary$adjr2[6],col="red",cex=2,pch=20)
```

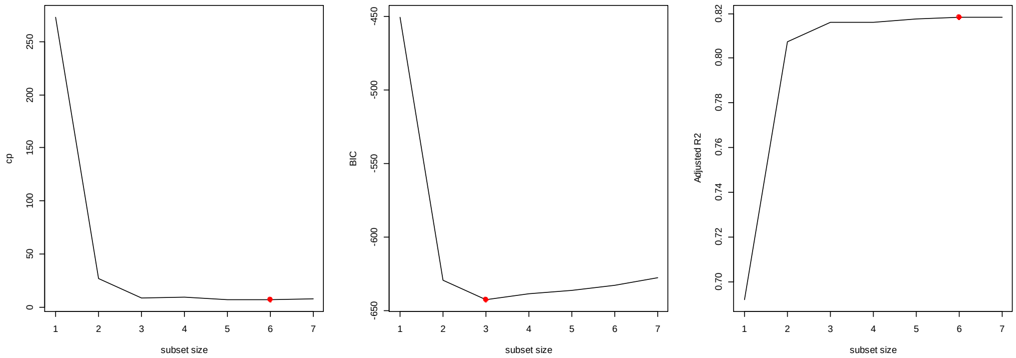


Fig.2 selection results by Cp, BIC and adjusted R2

From Tab.6 and Fig.2, it can been seen that backward stepwise selection outputs two types of feature selections, respectively, 6 predictors(cylinders, displacement, horsepower, weight, model.year, and origin) and 3 predictors(weight,model.year and origin). Meanwhile, both of them can be depicted by corresponding coefficients shown in Tab.7.

Tab.7 coefficients by two types of selections

```
>coef(regfit.full,6)    # 6 predictors
   (Intercept)      cylinders   displacement     horsepower          weight
-15.563492306   -0.506685137     0.019269286   -0.023895029   -0.006218311
    model.year           origin
   0.747515952     1.428241885
>coef(regfit.full,3)     #3 predictors
   (Intercept)           weight      model.year          origin
-18.045850149   -0.005994118     0.757126111     1.150390789
```

4.4 Multiple linear regression analysis

4.4.1. Methodology

In general, suppose that we have p distinct predictors. Then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ..\beta_p X_p + \varepsilon$$

where X_j represents the *j*th predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the average effect on *Y* of a one unit increase in X_j, holding all other predictors fixed.

4.4.2. Null hypothesis

Here in this case, the null hypothesis for this model is that whether all the regression coefficients are zero, β_j =0. The specific details about the test result will be illustrated as the following chapters through F -statistic value, t-statistic value, p-value, and R2, etc.

4.4.3 Regression without feature selection

For the later comparison analysis, the linear regression without reducing the redundant predictors is performed firstly, represented as below in Tab.8 and Tab.9.

Tab8. Linear regression without feature selection

```
lm.fit0=lm(mpg~.-car.name, data=Auto1)
summary(lm.fit0)
```

Tab.9 regression result without feature selection

```
Call:
lm(formula = mpg ~ . - car.name, data = Auto1)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5903 -2.1565 -0.1169  1.8690 13.0604 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780    
displacement   0.019896   0.007515   2.647  0.00844 ** 
horsepower    -0.016951   0.013787  -1.230  0.21963    
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548    
model.year     0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

From Tab.9, we can see that displacement, weight, model.year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not. Besides, the null hypothesis can be definitely rejected by above test results. The F-statistic is far from 1 (with a small p-value), indicating evidence against the null hypothesis.

As the collinearity in the model is implied within Tab.5, the variance inflation factor(VIF) described by 'vif' function is applied to identify whether there are predictors with VIF values greater than 5. If yes, then these predictors can be deleted in the model. The VIF values for regression result without feature selection is displayed in Tab.10, and in the model only model.year and origin are kept as the acceleration gets a high p-value in regression.

Tab.10 VIF values for each predictor in regression without feature selection

```
> vif(lm.fit0)
 Cylinders   displacement   horsepower    weight    acceleration   model.year   origin
10.737535    21.836792      9.943693     10.831260   2.625806      1.244952     1.772386
```

4.4.3 Regression with log transformation

Since the former regression model seems to be not good, here the log transfromation for the response(mpg) is conducted as below in Tab.11 and Tab.12, including the VIF values in this model. Similarly, the final model only includes model.year and origin in this model.

Tab.11 Code for regression with log transformation and VIF values for each predictor

```
>lm.fit_log=lm(log(mpg)~.-car.name, data=Auto1)
>summary(lm.fit_log)
> vif(lm.fit_log)
   cylinders   displacement   horsepower   weight     acceleration   model.year   origin
   10.737535    21.836792      9.943693    10.831260   2.625806      1.244952    1.772386
```

Tab.12 Regression results with log transformation

```
Call:
lm(formula = log(mpg) ~ . - car.name, data = Auto1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40955 -0.06533  0.00079  0.06785  0.33925

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.751e+00  1.662e-01  10.533  < 2e-16 ***
cylinders    -2.795e-02  1.157e-02  -2.415  0.01619 *
displacement  6.362e-04  2.690e-04   2.365  0.01852 *
horsepower   -1.475e-03  4.935e-04  -2.989  0.00298 **
weight       -2.551e-04  2.334e-05 -10.931  < 2e-16 ***
acceleration -1.348e-03  3.538e-03  -0.381  0.70339
model.year    2.958e-02  1.824e-03  16.211  < 2e-16 ***
origin        4.071e-02  9.955e-03   4.089 5.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1191 on 384 degrees of freedom
Multiple R-squared:  0.8795,    Adjusted R-squared:  0.8773
F-statistic: 400.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

4.4.5 Regression with feature selection

According to the outcomes of chapter4.3, both the 6-predictor model(lm.fit1) and the 3-predictor model(lm.fit2) can be separately fitted as below in Tab.13,Tab.14, and Tab.15, Tab.16.

Tab.13 Code for 6-predictor regression model

```
>lm.fit1 = lm(mpg~.-acceleration, data=Auto2)
>summary(lm.fit1)
>library(car)
>vif(lm.fit1)
 cylinders   displacement   horsepower    weight     model.year    origin
 10.710150    21.608513      6.147752    8.324047     1.237304    1.772234
```

Tab.14. regression result by 6-predictor model

```
Call:
lm(formula = mpg ~ . - acceleration, data = Auto2)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7604 -2.1791 -0.1535  1.8524 13.1209

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.556e+01  4.175e+00  -3.728 0.000222 ***
cylinders    -5.067e-01  3.227e-01  -1.570 0.117236
displacement  1.927e-02  7.472e-03   2.579 0.010287 *
horsepower   -2.389e-02  1.084e-02  -2.205 0.028031 *
weight       -6.218e-03  5.714e-04 -10.883  < 2e-16 ***
model.year    7.475e-01  5.079e-02  14.717  < 2e-16 ***
origin        1.428e+00  2.780e-01   5.138 4.43e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.326 on 385 degrees of freedom
Multiple R-squared:  0.8212,    Adjusted R-squared:  0.8184
F-statistic: 294.6 on 6 and 385 DF,  p-value: < 2.2e-16
```

Tab.15 Code for 3-predictor regression model

```
>lm.fit2=    lm(mpg~weight+model.year+origin, data=Auto2)
>summary(lm.fit2)
>vif(lm.fit2)
    weight      model.year       origin
  1.625522      1.105651      1.520292
```

Tab.16 regression result by 3-predictor model

```
Call:
lm(formula = mpg ~ weight + model.year + origin, data = Auto2)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.9440 -2.0948 -0.0389  1.7255 13.2722 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.805e+01  4.001e+00  -4.510 8.60e-06 ***
weight      -5.994e-03  2.541e-04 -23.588  < 2e-16 ***
model.year   7.571e-01  4.832e-02  15.668  < 2e-16 ***
origin       1.150e+00  2.591e-01   4.439 1.18e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.348 on 388 degrees of freedom
Multiple R-squared:  0.8175,	Adjusted R-squared:  0.816 
F-statistic: 579.2 on 3 and 388 DF,  p-value: < 2.2e-16
```

Above all, the lm.fit1 finally only contain model.year and origin as predictors with new fitting results shown in Tab.17 and Tab.18, while the lm.fit2 gets three predictors including weight, model.year and origin shown in Tab.16.

Tab.17 Code for the model with only model.year and origin

```
>lim.fitbest=lm(mpg~model.year+origin, data=Auto2)
>summary(lim.fitbest)
```

Tab.18 final model with only model.year and origin

```
Call:
lm(formula = mpg ~ model.year + origin, data = Auto2)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.3126  -3.7257  -0.4732   3.3893  15.5874 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -63.37982    5.46818  -11.59   <2e-16 ***
model.year    1.04715    0.07282   14.38   <2e-16 ***
origin        4.60725    0.33301   13.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.216 on 389 degrees of freedom
Multiple R-squared:  0.5557,	Adjusted R-squared:  0.5534 
F-statistic: 243.2 on 2 and 389 DF,  p-value: < 2.2e-16
```

5. Conclusion

Due to the above analysis, we conclude here for this project that the response(MPG) do have a linear relationship with some of these predictors with the following forms.
For the model with weight, model.year and origin:

$$y = -18.05 - 0.005994 weight + 0.7571 \bmod elyear + 1.15 origin$$

For the model with model.year and orgin:

$$y = -63.37982 + 1.04715 \bmod elyear + 4.60725 origin$$

Furthermore, predictors(model.year and origin) do have significant influence on the response.

6. Future Direction

Since there are two simplified models here to illustrate the linear relationship between the response(MPG) and the predictors, more works about which model can have a smaller test error will be my next research focus of interest.

7. Reference

[1] https://www.kaggle.com/uciml/autompg-dataset
[2] G. Casella, S. Fienberg,I. Olkin. An introduction to statistical learning with applications in R [M]. Springer,2017.