Multiple Linear Regression

Objective:

Objective of the project is to establish a multiple linear regression model between the dependent variable CO2 emissions and explanatory variables engine size, cylinders, fuel consumption in city roads, fuel consumption in highways, fuel consumption combined (L/100km) and fuel consumption combined (mpg).

Introduction:

This dataset captures the details of how CO2 emissions by a vehicle can vary with the different features. The dataset has been taken from kaggle where the reference of the data was from Canada Government official open data website. This contains data over a period of 7 years.

Data Description:

There are total 7385 rows and 12 columns.

- **Make:** Company of the vehicle
- **Model:** Car Model
- **Vehicle Class:** Class of vehicle depending on their utility, capacity and weight
- **Engine Size(L):** Size of engine used in Litre
- **Cylinders:** Number of cylinders
- **Transmission**: Transmission type with number of gears.
- **Fuel Type:** Type of Fuel used
- **Fuel consumption city(L/100km):** Fuel consumption in city roads (L/100 km)
- **Fuel consumption Hwy(L/100km):** Fuel consumption in highways (L/100 km)
- **Fuel Consumption Comb (L/100 km):** The combined fuel consumption (55% city, 45% highway) is shown in L/100 km

- **Fuel Consumption Comb (mpg):** The combined fuel consumption in both city and highway is shown in mile per gallon(mpg)
- **CO2 Emissions(g/km):** The tailpipe emissions of carbon dioxide (in grams per kilometre) for combined city and highway driving. It is the dependent variable in the model.

Methodology:

- Firstly, the data was cleaned by removing outliers and categorical variables.
- It is a secondary source of data and statistical concepts of multiple linear regression was used. A multiple linear model was fitted taking Y as a dependent variable and the Xi's (i=1 to 6) as explanatory variables.
- The model is further tested for multicollinearity, heteroscedasticity and auto correlation and treated accordingly.
- The entire project and the statistical tests in it are carried using the R software.

Data Cleaning:

```
vehicles<-read.csv(file.choose())
head(vehicles)
  Make       Model Vehicle.Class Engine.Size.L. Cylinders Transmission Fuel.Type
ACURA          ILX       COMPACT            2.0         4          AS5         Z
ACURA          ILX       COMPACT            2.4         4           M6         Z
ACURA ILX HYBRID         COMPACT            1.5         4          AV7         Z
ACURA      MDX 4WD   SUV - SMALL            3.5         6          AS6         Z
ACURA      RDX AWD   SUV - SMALL            3.5         6          AS6         Z
ACURA          RLX      MID-SIZE            3.5         6          AS6         Z
Fuel.Consumption.City..L.100.km. Fuel.Consumption.Hwy..L.100.km. Fuel.Consumption.Comb..L.100.km.
                             9.9                             6.7                              8.5
                            11.2                             7.7                              9.6
                             6.0                             5.8                              5.9
                            12.7                             9.1                             11.1
                            12.1                             8.7                             10.6
                            11.9                             7.7                             10.0
Fuel.Consumption.Comb..mpg. CO2.Emissions.g.km.
                         33                 196
                         29                 221
                         48                 136
                         25                 255
                         27                 244
                         28                 230
```

```
> #remove outliers
> outliers <- function(x) {
+
+   Q1 <- quantile(x, probs=.25)
+   Q3 <- quantile(x, probs=.75)
+   iqr = Q3-Q1
+
+   upper_limit = Q3 + (iqr*1.5)
+   lower_limit = Q1 - (iqr*1.5)
+
+   x > upper_limit | x < lower_limit
+ }
> remove_outliers <- function(df, cols = names(df)) {
+   for (col in cols) {
+     df <- df[!outliers(df[[col]]),]
+   }
+   df
+ }
> vehicles=vehicles[,c(4,5,8,9,10,11,12)]
> data=remove_outliers(vehicles,cols = names(vehicles))
> names(data)=c("X1","X2","X3","X4","X5","X6","Y")
> head(data)
   X1 X2   X3  X4   X5 X6   Y
1 2.0  4  9.9 6.7  8.5 33 196
2 2.4  4 11.2 7.7  9.6 29 221
4 3.5  6 12.7 9.1 11.1 25 255
5 3.5  6 12.1 8.7 10.6 27 244
6 3.5  6 11.9 7.7 10.0 28 230
7 3.5  6 11.8 8.1 10.1 28 232
```

After removing the outliers and the categorical variables, we have 6697 observations and 7 variables. Also, the column names of explanatory variables and dependant variable are renamed as

X1= Engine Size(L)

X2= Cylinders

X3= Fuel consumption city(L/100km)

X4= Fuel consumption Hwy(L/100km)

X5= Fuel Consumption Comb (L/100 km)

X6= Fuel Consumption Comb (mpg)

Y = CO2 Emissions(g/km)

Econometric Analysis:

1. Initial fitting of Model:

Taking Y as dependent variable, a multiple linear regression model was fitted and the Xi's (i=1 to 6) as explanatory variables.

$$Y=B0+B1X1+B2X2+B3X3+B4X4+B5X5+B6X6+U$$

where U is the disturbance term

Bi's are the ith parameter associated with explanatory variable Xi

The fitted model RM1 is:

```
> #REGRESSION
> X=cbind(data$X1,data$X2,data$X3,data$X4,data$X5,data$X6)
> Y=data$Y
> RM=lm(Y~X)
> RM

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5           X6
    193.540        4.333        4.961       -3.614        3.328       10.478       -3.031
```

- **ANOVA of the model RM is:**

Hypothesis testing:

H0 : B0=B1=B2=B3=B4=B5=B6=0

H1 : Atleast one of Bi is not equal to zero. (i=0 to 6)

```
> anova(RM)
Analysis of Variance Table

Response: Y
            Df   Sum Sq Mean Sq F value    Pr(>F)
X            6 14497496 2416249   12645 < 2.2e-16 ***
Residuals 6690  1278391     191
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistics obtained from ANOVA is 12645 with it's p value being less than 0.05. Thus, taking the level of significance at 5%,we are able to reject the null hypothesis and conclude that atleast one of Bi's is not equal to zero.

- **Significance of the parameters obtained:**

Hypothesis testing:
H0 : The model is not of significant fit. ($R2=0$)
H1 : The model is of significant fit. ($R2\neq0$)

```
> summary(RM)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-131.175   -3.871    0.163    5.125   56.199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 193.5396     6.8907  28.087  < 2e-16 ***
X1            4.3328     0.4248  10.200  < 2e-16 ***
X2            4.9605     0.3264  15.199  < 2e-16 ***
X3           -3.6137     2.1508  -1.680  0.09297 .  
X4            3.3284     1.7719   1.878  0.06036 .  
X5           10.4776     3.9092   2.680  0.00738 ** 
X6           -3.0309     0.1231 -24.630  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.82 on 6690 degrees of freedom
Multiple R-squared:  0.919,	Adjusted R-squared:  0.9189 
F-statistic: 1.264e+04 on 6 and 6690 DF,  p-value: < 2.2e-16
```

Adjusted R2 value obtained is 0.9189 which indicates a very good fit. The corrosponding p value is less than 0.05. thus taking significance level at 5%, we are able to reject H0 and conclude the model is of significant fit.

On further examining of parameters obtained, p-value of X3 and X4 is more than 0.05, so maybe they are not significant at 5% level of significance. So, we need to further examine the model for multicollinearity, heteroscedasticity and autocorrelation.

2. Checking for the presence of multicollinearity in the model

- **Method of partial correlation:**

Method of partial correlation is used to find the partical correlation between the explanatory variables to check with significance that which explanatory variables are a cause of multicollinearity. Using the “ppcor” package in R, partial correlation between the Xi’s and its p value is obtained.

```
> #CHECK FOR MULTICOLLINEARITY
> library(ppcor)
> ppcor::pcor(X)
$estimate
              [,1]          [,2]          [,3]         [,4]         [,5]          [,6]
[1,]  1.0000000000  8.069393e-01  0.0004936812 -0.003732183  0.016619711  4.941470e-02
[2,]  0.8069392939  1.000000e+00  0.0182245545 -0.018988509  0.000718234 -5.783607e-06
[3,]  0.0004936812  1.822455e-02  1.0000000000 -0.971780929  0.991705448  1.020754e-04
[4,] -0.0037321827 -1.898851e-02 -0.9717809292  1.000000000  0.985387700  1.046441e-02
[5,]  0.0166197112  7.182340e-04  0.9917054477  0.985387700  1.000000000 -8.720439e-02
[6,]  0.0494146978 -5.783607e-06  0.0001020754  0.010464414 -0.087204389  1.000000e+00
```

Here we can observe that variables (X3,X4),(X3,X5),(X4,X5) have significant correlations with each other as their partial correlation is greater than 0.90 and thus they are a cause of multicollinearity in the model.

So now we will check for Variance inflation factor (VIF) to identify which of them is the major cause of multicollinearity.

- **<u>Variance Inflation Factor:</u>**

Using the "mctest" package in R, Variance Inflation factor (VIF) was calculated for each explanatory variable. Variables having a VIF of 10 and above were subjucted to suspicion for cause of multicollinearity.

```
> library(mctest)
> imcdiag(RM)

Call:
imcdiag(mod = RM)


All Individual Multicollinearity Diagnostics Result

         VIF    TOL          Wi          Fi Leamer     CVIF Klein   IND1   IND2
X1    8.9284 0.1120   10609.722   13264.14 0.3347  -0.1876     0 1e-04 0.9336
X2    8.2946 0.1206    9761.672   12203.91 0.3472  -0.1742     0 1e-04 0.9246
X3 1131.7093 0.0009 1513115.215 1891676.70 0.0297 -23.7734     1 0e+00 1.0505
X4  315.3249 0.0032  420629.646  525865.64 0.0563  -6.6239     1 0e+00 1.0481
X5 2543.7215 0.0004 3402669.969 4253973.14 0.0198 -53.4350     1 0e+00 1.0510
X6   17.7787 0.0562   22453.243   28070.75 0.2372  -0.3735     1 0e+00 0.9923

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

X3 , X4 , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.919
```

We can observe that the VIF of X3,X4,X5 and X6 is greater than 10. Firstly, we will remove the variable with maximum VIF, i.e., X5 having VIF=2543.7215.

3. Multicollinearity removal and revised model:

- **Removing X5 variable:**

Variable X5 is dropped from the model, now the revised model RM1 is:

Y=B0+B1X1+B2X2+B3X3+B4X4+B6X6+U

where U is the disturbance term

Bi's are the ith parameter associated with explanatory variable Xi

The fitted model RM1 is:

```
> #X5 has max VIF
> #removing x5
> data1=data[,-5]
> X=cbind(data1$X1,data1$X2,data1$X3,data1$X4,data1$X6)
> Y=data1$Y
> RM1=lm(Y~X)
> RM1

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5  
    195.144        4.352        4.961        2.103        8.008       -3.060  
```

- **ANOVA of the model RM1 is:**

Hypothesis testing:

H0 : B0=B1=B2=B3=B4=B6=0

H1 : Atleast one of Bi is not equal to zero. (i=0,1,2,3,4,6)

```
> anova(RM1)
Analysis of Variance Table

Response: Y
            Df   Sum Sq Mean Sq F value    Pr(>F)    
X            5 14496123 2899225   15158 < 2.2e-16 ***
Residuals 6691  1279764     191                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F statistics obtained from ANOVA is 15158 with it's p value being less than 0.05. thus, taking the level of significance at 5%,we are able to reject the null hypothesis and conclude that atleast one of Bi's is not equal to zero.

- **Significance of the parameters obtained:**

  Hypothesis testing:
  H0 : The model is not of significant fit. ($R2=0$)
  H1 : The model is of significant fit. ($R2\neq0$)

```
> summary(RM1)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-130.944   -3.904    0.206    5.009   56.342 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 195.1442     6.8678  28.414  < 2e-16 ***
X1            4.3517     0.4249  10.241  < 2e-16 ***
X2            4.9611     0.3265  15.193  < 2e-16 ***
X3            2.1031     0.2766   7.604 3.26e-14 ***
X4            8.0080     0.3019  26.522  < 2e-16 ***
X5           -3.0596     0.1226 -24.947  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.83 on 6691 degrees of freedom
Multiple R-squared:  0.9189,    Adjusted R-squared:  0.9188 
F-statistic: 1.516e+04 on 5 and 6691 DF,  p-value: < 2.2e-16
```

Adjusted R2 value obtained is 0.9188 which indicates a very good fit. The corrosponding p value is less than 0.05. thus taking significance level at 5%, we are able to reject H0 and conclude the model is of significant fit.

On further examining of parameters obtained, p-value of all the explanatory variables are less than 0.05, so they are significant at 5% level of significance.

- **VIFs are again calculatef for the model RM1 :**

```
> imcdiag(RM1)

Call:
imcdiag(mod = RM1)


All Individual Multicollinearity Diagnostics Result

       VIF    TOL       Wi       Fi Leamer    CVIF Klein  IND1   IND2
X1  8.9259 0.1120 13260.01 17682.65 0.3347 -0.2442     0 1e-04 0.9762
X2  8.2946 0.1206 12203.91 16274.31 0.3472 -0.2270     0 1e-04 0.9669
X3 18.6962 0.0535 29605.71 39480.18 0.2313 -0.5116     1 0e+00 1.0406
X4  9.1479 0.1093 13631.47 18178.00 0.3306 -0.2503     0 1e-04 0.9792
X5 17.6435 0.0567 27844.56 37131.63 0.2381 -0.4828     1 0e+00 1.0371

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

* all coefficients have significant t-ratios

R-square of y on all x: 0.9189
```

We can observe that the VIF of X3 and X5 is greater than 10. So now we will remove the variable X3 which has maximum VIF, i.e., VIF=18.6962.

- **Removing X3 variable:**

Variable X3 is dropped from the model, now the revised model RM2 is:

Y=B0+B1X1+B2X2+B4X4+B6X6+U

where U is the disturbance term

Bi's are the ith parameter associated with explanatory variable Xi

The fitted model RM2 is:

```
> #REMOVING X3
> data2=data[,-3]
> X=cbind(data2$X1,data2$X2,data2$X4,data2$X6)
> Y=data1$Y
> RM2=lm(Y~X)
> RM2

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)           X1           X2           X3           X4
    229.866        4.779        5.327        8.578       -3.689
```

- **<u>ANOVA of the model RM2 is:</u>**

  Hypothesis testing:
  H0 : B0=B1=B2=B4=B6=0
  H1 : Atleast one of Bi is not equal to zero. (i=0,1,2,4,6)

```
> anova(RM2)
Analysis of Variance Table

Response: Y
            Df   Sum Sq Mean Sq F value    Pr(>F)
X            4 14485063 3621266   18774 < 2.2e-16 ***
Residuals 6692  1290824     193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F statistics obtained from ANOVA is 18774 with it's p value being less than 0.05. thus, taking the level of significance at 5%,we are able to reject the null hypothesis and conclude that atleast one of Bi's is not equal to zero.

- **Significance of the parameters obtained:**

  Hypothesis testing:
  H0 : The model is not of significant fit. (R2=0)
  H1 : The model is of significant fit. (R2≠0)

```
> summary(RM2)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-129.978   -4.340    0.198    5.119   55.364 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 229.86551    5.15184   44.62   <2e-16 ***
X1            4.77856    0.42299   11.30   <2e-16 ***
X2            5.32696    0.32434   16.42   <2e-16 ***
X3            8.57784    0.29373   29.20   <2e-16 ***
X4           -3.68879    0.09092  -40.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.89 on 6692 degrees of freedom
Multiple R-squared:  0.9182,    Adjusted R-squared:  0.9181 
F-statistic: 1.877e+04 on 4 and 6692 DF,  p-value: < 2.2e-16
```

Adjusted R2 value obtained is 0.9181 which indicates a very good fit. The corrosponding p value is less than 0.05. thus taking significance level at 5%, we are able to reject H0 and conclude the model is of significant fit.

On further examining of parameters obtained, p-value of all the explanatory variables are less than 0.05, so they are significant at 5% level of significance.

- **<u>VIFs are again calculated for the model RM2:</u>**

```
> imcdiag(RM2)

Call:
imcdiag(mod = RM2)


All Individual Multicollinearity Diagnostics Result

      VIF    TOL       Wi       Fi Leamer    CVIF Klein  IND1   IND2
X1 8.7701 0.1140 17335.20 26006.68 0.3377 -0.3434     0 1e-04 1.0005
X2 8.1146 0.1232 15872.62 23812.48 0.3510 -0.3177     0 1e-04 0.9901
X3 8.5844 0.1165 16920.86 25385.07 0.3413 -0.3361     0 1e-04 0.9977
X4 9.6138 0.1040 19217.42 28830.44 0.3225 -0.3764     0 0e+00 1.0118

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

* all coefficients have significant t-ratios

R-square of y on all x: 0.9182
```

We can observe that VIF of all the explanatory variables is less than 10. So therefore no significant multicollinearity is present in the model RM2.

4. <u>Checking for the presence of heteroscedasticity:</u>

Goldfield Quandt test is used for the purpose. The “lmtest” package in R has the Goldfield Quandt test.

Hypotheses testing:

H0: There is no presence of heteroscedasticity in the error variance.

H1: There is presence of heteroscedasticity in the error variance.

```
> #CHECK FOR HETEROSCEDASTICITY
> library(lmtest)
> gqtest(RM2)

        Goldfeld-Quandt test

data:  RM2
GQ = 0.61897, df1 = 3344, df2 = 3343, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2
```

We see that the GQ value is 0.61897 and its p value is 1 Thus, taking 5% level of significance, we see p value is greater than 0.05. Hence, we are not able to reject H0 and thus conclude that there is no presence of heteroscedasticity in the model RM2.

5. Checking for the presence of autocorrelation:

Durbin Watson test is used for the same. The "lmtest" package in R contains the Durbin Watson function.

Hypothesis testing:
H0 : There is no presence of autocorrelation.
H1 : There is presence of autocorrelation.

```
> #CHECK FOR AUTOCORRELATION
> dwtest(RM2)

        Durbin-Watson test

data:  RM2
DW = 1.6455, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

We see that the obtained value of DW statistic is 1.6455 which indicative of positive autocorrelation. Furthermore, the p value being less than 0.05. Therefore, taking level of significance at 5 %, we are able to reject H0 and conclude that there is autocorrelation present in the model.

6. Removal of Autocorrelation and revised model:

Cochran Orcutt iterative method is being used for estimating parameters under autocorrelation. "orcutt" package in R has the Cochran Orcutt iterative function.

```
> library(orcutt)
> RM3=cochrane.orcutt(RM2)
> RM3
Cochrane-orcutt estimation for first order autocorrelation

Call:
lm(formula = Y ~ X)

 number of interaction: 14
 rho 0.231386

Durbin-Watson statistic
(original):    1.64550 , p-value: 3.654e-48
(transformed): 2.08758 , p-value: 9.998e-01

 coefficients:
(Intercept)          X1          X2          X3          X4
 255.565484    6.556526    4.733244    6.385986   -4.001039
```

Revised Model RM3 is fitted with Cochran Orcutt iterative procedure.

- **Checking for the significance of fit**

Hypotheses testing:

H0: all the ai's are equal to zero (i=0,1,2,4,6)

H1: at least one of the ai's is not zero.

```
> summary(RM3)
Call:
lm(formula = Y ~ X)

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 255.565484   5.492516  46.530 < 2.2e-16 ***
X1            6.556526   0.468529  13.994 < 2.2e-16 ***
X2            4.733244   0.358778  13.193 < 2.2e-16 ***
X3            6.385986   0.310970  20.536 < 2.2e-16 ***
X4           -4.001039   0.097378 -41.088 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.6032 on 6694 degrees of freedom
Multiple R-squared:  0.8952 ,  Adjusted R-squared:  0.8952
F-statistic: 14289.5 on 1 and 6694 DF,  p-value: < 0e+00

Durbin-Watson statistic
(original):    1.64550 , p-value: 3.654e-48
(transformed): 2.08758 , p-value: 9.998e-01
```

We have adjusted R2 for the model RM3 as 0.8952, and the F value is 14289.5. The corresponding p value is less than 0.05. Thus, taking level of significance at

5%, we are able to reject H0 and conclude that at least one of the ai's is not zero. So, there is no autocorrelation in the model.

On further examining of parameters obtained from the model RM3, p-value of all the explanatory variables are less than 0.05, so they are significant at 5% level of significance. So now our model is free from multicollinearity, heteroscedasticity and autocorrelation.

Conclusion:

Therefore, the final model RM3 is:

```
 coefficients:
(Intercept)          X1          X2          X3          X4
 255.565484    6.556526    4.733244    6.385986   -4.001039
```

$\hat{Y}$=255.56+(6.55*X1)+(4.73*X2)+(6.38*X4)-(4*X6)+U

where U is the disturbance term and Xi (i=1,2,4,6) are the explanatory variables.

Adjusted R^2= 0.8952 which means that the model explains 89.52% of the variation of the dependent variable.

References:

- The data set is taken from Kaggle.
- Techniques used are referred from "Basic Econometrics" by Damodar N Gujrati