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 Python software.

Data Cleaning:

```
import numpy as np
import pandas as pd
```

```
vehicles = pd.read_csv('CO2 Emissions_Canada.csv',index_col='Make', parse_dates=True)
vehicles.head()
```

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

```python
import pandas as pd

def remove_outliers(df, cols=None):
    if cols is None:
        cols = df.columns

    for col in cols:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1

        upper_limit = Q3 + (1.5 * IQR)
        lower_limit = Q1 - (1.5 * IQR)

        df = df[(df[col] >= lower_limit) & (df[col] <= upper_limit)]

    return df

# Assuming you have a DataFrame named 'vehicles'
# Selecting specific columns as in your R code
vehicles = vehicles[['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'Y']]

# Remove outliers from selected columns
data = remove_outliers(vehicles)

# Display the head of the resulting DataFrame
print(data.head())
```

```
        X1  X2    X3   X4    X5  X6    Y
Make
ACURA  2.0   4   9.9  6.7   8.5  33  196
ACURA  2.4   4  11.2  7.7   9.6  29  221
ACURA  3.5   6  12.7  9.1  11.1  25  255
ACURA  3.5   6  12.1  8.7  10.6  27  244
ACURA  3.5   6  11.9  7.7  10.0  28  230
```

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:

	coef
const	193.5396
X1	4.3328
X2	4.9605
X3	-3.6137
X4	3.3284
X5	10.4776
X6	-3.0309

- **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)

```
# Regression model
X = data[['X1', 'X2', 'X3', 'X4', 'X5', 'X6']]
Y = data['Y']
X = sm.add_constant(X)  # Add a constant term to the independent variables
RM = sm.OLS(Y, X).fit()

# Display the regression results
print(RM.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.919
Model:                            OLS   Adj. R-squared:                  0.919
Method:                 Least Squares   F-statistic:                 1.264e+04
Date:                Fri, 12 Jan 2024   Prob (F-statistic):               0.00
Time:                        00:14:45   Log-Likelihood:                -27088.
No. Observations:                6697   AIC:                         5.419e+04
Df Residuals:                    6690   BIC:                         5.424e+04
Df Model:                           6
Covariance Type:            nonrobust
```

The F statistics obtained from ANOVA is 1.264e+04 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)

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.919
Model:	OLS	Adj. R-squared:	0.919
Method:	Least Squares	F-statistic:	1.264e+04
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	0.00
Time:	00:14:45	Log-Likelihood:	-27088.
No. Observations:	6697	AIC:	5.419e+04
Df Residuals:	6690	BIC:	5.424e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	193.5396	6.891	28.087	0.000	180.032	207.048
X1	4.3328	0.425	10.200	0.000	3.500	5.166
X2	4.9605	0.326	15.199	0.000	4.321	5.600
X3	-3.6137	2.151	-1.680	0.093	-7.830	0.603
X4	3.3284	1.772	1.878	0.060	-0.145	6.802
X5	10.4776	3.909	2.680	0.007	2.814	18.141
X6	-3.0309	0.123	-24.630	0.000	-3.272	-2.790

Adjusted R2 value obtained is 0.919 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.

Correlation Matrix:

	X1	X2	X3	X4	X5	X6
X1	1.000000	0.934640	0.821684	0.748813	0.809999	-0.776221
X2	0.934640	1.000000	0.804618	0.719563	0.788362	-0.757244
X3	0.821684	0.804618	1.000000	0.927991	0.991300	-0.963537
X4	0.748813	0.719563	0.927991	1.000000	0.968397	-0.939083
X5	0.809999	0.788362	0.991300	0.968397	1.000000	-0.971247
X6	-0.776221	-0.757244	-0.963537	-0.939083	-0.971247	1.000000

Here we can observe that variables (X1,X2),(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.

- **Variance Inflation Factor:**

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.

VIF:

	Variable	VIF
0	const	1651.500434
1	X1	8.925887
2	X2	8.294624
3	X3	18.696183
4	X4	9.147917
5	X6	17.643490

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:

	coef
const	195.1442
X1	4.3517
X2	4.9611
X3	2.1031
X4	8.0080
X6	-3.0596

- **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)

Regression after removing X5:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.919
Model:	OLS	Adj. R-squared:	0.919
Method:	Least Squares	F-statistic:	1.516e+04
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	0.00
Time:	16:54:18	Log-Likelihood:	-27092.
No. Observations:	6697	AIC:	5.420e+04
Df Residuals:	6691	BIC:	5.424e+04
Df Model:	5		
Covariance Type:	nonrobust		

The F statistics obtained from ANOVA is 1.516e+04 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)

Regression after removing X5:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.919
Model:	OLS	Adj. R-squared:	0.919
Method:	Least Squares	F-statistic:	1.516e+04
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	0.00
Time:	16:54:18	Log-Likelihood:	-27092.
No. Observations:	6697	AIC:	5.420e+04
Df Residuals:	6691	BIC:	5.424e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	195.1442	6.868	28.414	0.000	181.681	208.607
X1	4.3517	0.425	10.241	0.000	3.519	5.185
X2	4.9611	0.327	15.193	0.000	4.321	5.601
X3	2.1031	0.277	7.604	0.000	1.561	2.645
X4	8.0080	0.302	26.522	0.000	7.416	8.600
X6	-3.0596	0.123	-24.947	0.000	-3.300	-2.819

Adjusted R2 value obtained is 0.919 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 calculated for the model RM1 :**

```
# Calculate VIF for each variable
X = data[['X1', 'X2', 'X3', 'X4', 'X6']]
Y = data['Y']
X = sm.add_constant(X)  # Add a constant term to the independent variables
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print("\nVIF:")
print(vif_data)
```

```
VIF:
  Variable          VIF
0    const  1651.500434
1       X1     8.925887
2       X2     8.294624
3       X3    18.696183
4       X4     9.147917
5       X6    17.643490
```

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+B3X3+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:

	coef
const	229.8655
X1	4.7786
X2	5.3270
X4	8.5778
X6	-3.6888

- **ANOVA of the model RM2 is:**

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)

```
Regression after removing X3:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.918
Method:                 Least Squares   F-statistic:                 1.877e+04
Date:                Fri, 12 Jan 2024   Prob (F-statistic):               0.00
Time:                        00:25:58   Log-Likelihood:                -27120.
No. Observations:                6697   AIC:                         5.425e+04
Df Residuals:                    6692   BIC:                         5.428e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
```

The F statistics obtained from ANOVA is 1.877e+04 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)

```
Regression after removing X3:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.918
Method:                 Least Squares   F-statistic:                 1.877e+04
Date:                Fri, 12 Jan 2024   Prob (F-statistic):               0.00
Time:                        00:25:58   Log-Likelihood:                -27120.
No. Observations:                6697   AIC:                         5.425e+04
Df Residuals:                    6692   BIC:                         5.428e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        229.8655      5.152     44.618      0.000     219.766     239.965
X1             4.7786      0.423     11.297      0.000       3.949       5.608
X2             5.3270      0.324     16.424      0.000       4.691       5.963
X4             8.5778      0.294     29.203      0.000       8.002       9.154
X6            -3.6888      0.091    -40.573      0.000      -3.867      -3.511
==============================================================================
```

Adjusted R2 value obtained is 0.918 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 calculated for the model RM2:**

```
VIF:
  Variable         VIF
0    const  921.498389
1       X1    8.770147
2       X2    8.114574
3       X4    8.584426
4       X6    9.613815
```

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. Checking for the presence of heteroscedasticity:

Goldfield Quandt test is used for the purpose.

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 using Goldfeld-Quandt test
# Perform Goldfeld-Quandt test for heteroscedasticity
test_result = het_goldfeldquandt(RM2.resid, X)

# Extract F-statistic and p-value
F_statistic = test_result[0]
p_value = test_result[1]

# Display the F-statistic and p-value
print("Goldfeld-Quandt test F-statistic:", F_statistic)
print("Goldfeld-Quandt test p-value:", p_value)
```

```
Goldfeld-Quandt test F-statistic: 0.6189670270368012
Goldfeld-Quandt test p-value: 0.9999999999999999
```

We see that the GQ value is 0.61897 and its p value is 0.99. 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.

Hypothesis testing:

H0 : There is no presence of autocorrelation.

H1 : There is presence of autocorrelation.

```
# Check for autocorrelation using Durbin-Watson test
dw_statistic = durbin_watson(RM2.resid)
print("\nDurbin-Watson statistic:", dw_statistic)
```

```
Durbin-Watson statistic: 1.6454950308622205
```

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.

```
# Apply Cochrane-Orcutt correction if needed
if dw_statistic < 1.5 or dw_statistic > 2.5:  # Rule of thumb for Durbin-Watson statistic
    RM3 = sm.OLS(Y, X).fit(cov_type='HAC', cov_kwds={'maxlags': 1})
    print("\nRegression after Cochrane-Orcutt correction:")
    print(RM3.summary())
else:
    print("\nNo Cochrane-Orcutt correction needed. Durbin-Watson statistic within acceptable range.")
    print("Original regression:")
    print(RM2.summary())
```

```
No Cochrane-Orcutt correction needed. Durbin-Watson statistic within acceptable range.
Original regression:
```

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.

No Cochrane-Orcutt correction needed. Durbin-Watson statistic within acceptable range.
Original regression:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.918
Model:	OLS	Adj. R-squared:	0.918
Method:	Least Squares	F-statistic:	1.877e+04
Date:	Fri, 12 Jan 2024	Prob (F-statistic):	0.00
Time:	17:20:12	Log-Likelihood:	-27120.
No. Observations:	6697	AIC:	5.425e+04
Df Residuals:	6692	BIC:	5.428e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	229.8655	5.152	44.618	0.000	219.766	239.965
X1	4.7786	0.423	11.297	0.000	3.949	5.608
X2	5.3270	0.324	16.424	0.000	4.691	5.963
X4	8.5778	0.294	29.203	0.000	8.002	9.154
X6	-3.6888	0.091	-40.573	0.000	-3.867	-3.511

Omnibus:	3171.510	Durbin-Watson:	1.645
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35786.552
Skew:	-1.980	Prob(JB):	0.00
Kurtosis:	13.610	Cond. No.	919.

We have adjusted R2 for the model RM3 as 0.918, and the F value is 1.877e+04. 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:

	coef
const	229.8655
X1	4.7786
X2	5.3270
X4	8.5778
X6	-3.6888

$\hat{Y}$=229.86+(4.78*X1)+(5.32*X2)+(8.57*X4)-(3.69*X6)+U

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

Adjusted R^2= 0.918 which means that the model explains 91.8% 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