Sardar Patel Institute of Technology, Mumbai
Department of Electronics and Telecommunication Engineering
B.E. Sem-VII (2022-2023)
OEIT6 - Data Analytics

Experiment: Linear Regression

Name: Pushkar Sutar **Roll No. 2019110060**

Objective : Building a linear regression model for climate change dataset.

Dataset Description:
The file climate_change (CSV) contains climate data from May 1983 to December 2008. The available variables include:

- Year: the observation year.
- Month: the observation month.
- Temp: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.
- CO2, N2O, CH4, CFC.11, CFC.12: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division.
- CO2, N2O and CH4 are expressed in ppmv (parts per million by volume -- i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere)
- CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).
- Aerosols: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Goddard Institute for Space Studies at NASA.
- TSI: the total solar irradiance (TSI) in W/m2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.
- MEI: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

Code and Output:

Problem 1.1 - Creating the model.

We first import all the necessary modules and read the csv file using Pandas.

```
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn import datasets, linear_model, metrics
import matplotlib.pyplot as plt
```

```
df = pd.read_csv("/content/climate_change.csv")
```

As per the instructions we will split the dataset into train and test data.

```
df_train = df[df.Year <= 2006]
df_test = df[df.Year > 2006]
```

```
df_train.tail()
```

	Year	Month	MEI	CO2	CH4	N2O	CFC-11	CFC-12	TSI	Aerosols	Temp
279	2006	8	0.759	380.45	1762.66	319.930	248.981	539.682	1365.7067	0.0041	0.482
280	2006	9	0.793	378.92	1776.04	320.010	248.775	539.566	1365.8419	0.0043	0.425
281	2006	10	0.892	379.16	1789.02	320.125	248.666	539.488	1365.8270	0.0044	0.472
282	2006	11	1.292	380.18	1791.91	320.321	248.605	539.500	1365.7039	0.0049	0.440
283	2006	12	0.951	381.79	1795.04	320.451	248.480	539.377	1365.7087	0.0054	0.518

```
df_test.head()
```

	Year	Month	MEI	CO2	CH4	N2O	CFC-11	CFC-12	TSI	Aerosols	Temp
284	2007	1	0.974	382.93	1799.66	320.561	248.372	539.206	1365.7173	0.0054	0.601
285	2007	2	0.510	383.81	1803.08	320.571	248.264	538.973	1365.7145	0.0051	0.498
286	2007	3	0.074	384.56	1803.10	320.548	247.997	538.811	1365.7544	0.0045	0.435
287	2007	4	-0.049	386.40	1802.11	320.518	247.574	538.586	1365.7228	0.0045	0.466
288	2007	5	0.183	386.58	1795.65	320.445	247.224	538.130	1365.6932	0.0041	0.372

Obtaining statistical description of the data using describe().

```
df_train.describe()
```

	Year	Month	MEI	CO2	CH4	N2O	CFC-11	CFC-12	TSI	Aerosols	Temp
count	284.000000	284.000000	284.000000	284.000000	284.000000	284.000000	284.000000	284.000000	284.000000	284.000000	284.000000
mean	1994.661972	6.556338	0.341923	361.414261	1745.841479	311.657225	252.487092	494.217546	1366.101437	0.017721	0.247799
std	6.845996	3.446768	0.929639	11.439691	45.669846	4.758513	20.987671	59.046642	0.401283	0.030014	0.181136
min	1983.000000	1.000000	-1.586000	340.170000	1629.890000	303.677000	191.324000	350.113000	1365.426100	0.001600	-0.282000
25%	1989.000000	4.000000	-0.323000	352.315000	1716.347500	307.657000	249.557750	462.543000	1365.754550	0.002700	0.118000
50%	1995.000000	7.000000	0.308500	359.890000	1758.605000	310.849500	260.373500	522.089000	1366.054500	0.006200	0.232500
75%	2001.000000	10.000000	0.898000	370.585000	1781.637500	316.129250	267.448000	540.972750	1366.399275	0.014000	0.406500
max	2006.000000	12.000000	3.001000	384.980000	1808.150000	320.451000	271.494000	543.813000	1367.316200	0.149400	0.739000

Splitting the features and the target, X as independent variable and y as the dependent variable for both test and train data.

```
X_train = df_train.iloc[:,:-1]
y_train = df_train['Temp']
X_test = df_test.iloc[:,:-1]
y_test = df_test['Temp']
```

Fitting the data using linear regression.

```
X_train = df_train.iloc[:,:-1]
y_train = df_train['Temp']
X_test = df_test.iloc[:,:-1]
y_test = df_test['Temp']
```

OLS Regression Results

Dep. Variable:	Temp	**R-squared:**	0.755
Model:	OLS	**Adj. R-squared:**	0.746
Method:	Least Squares	**F-statistic:**	84.10
Date:	Tue, 01 Nov 2022	**Prob (F-statistic):**	2.20e-77
Time:	15:21:22	**Log-Likelihood:**	282.43
No. Observations:	284	**AIC:**	-542.9
Df Residuals:	273	**BIC:**	-502.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-142.4475	55.400	-2.571	0.011	-251.513	-33.382
Year	0.0082	0.021	0.390	0.697	-0.033	0.050
Month	-0.0036	0.002	-1.660	0.098	-0.008	0.001
MEI	0.0645	0.006	9.944	0.000	0.052	0.077
CO2	0.0025	0.003	0.774	0.439	-0.004	0.009
CH4	0.0002	0.001	0.352	0.725	-0.001	0.001
N2O	-0.0163	0.018	-0.883	0.378	-0.053	0.020
CFC-11	-0.0063	0.002	-2.743	0.006	-0.011	-0.002
CFC-12	0.0034	0.002	1.976	0.049	1.27e-05	0.007
TSI	0.0952	0.018	5.151	0.000	0.059	0.132
Aerosols	-1.5430	0.220	-7.021	0.000	-1.976	-1.110

Omnibus:	9.503	**Durbin-Watson:**	0.937
Prob(Omnibus):	0.009	**Jarque-Bera (JB):**	11.419
Skew:	0.305	**Prob(JB):**	0.00331
Kurtosis:	3.769	**Cond. No.**	3.14e+07

The R2 value is 0.755.

Problem 1.2

If we look at the summary of the model, we consider variables as significant only if p value is below 0.05. So MEI, CO2, CFC.11, CFC.12, TSI, and Aerosols are all significant. Only CH4 and N2O are not significant.

Problem 2.1

```
X_train.corr()
```

	Year	Month	MEI	CO2	CH4	N2O	CFC-11	CFC-12	TSI	Aerosols
Year	1.000000	-0.027942	-0.036988	0.982749	0.915659	0.993845	0.569106	0.897012	0.170302	-0.345247
Month	-0.027942	1.000000	0.000885	-0.106732	0.018569	0.013632	-0.013111	0.000675	-0.034606	0.014890
MEI	-0.036988	0.000885	1.000000	-0.041147	-0.033419	-0.050820	0.069000	0.008286	-0.154492	0.340238
CO2	0.982749	-0.106732	-0.041147	1.000000	0.877280	0.976720	0.514060	0.852690	0.177429	-0.356155
CH4	0.915659	0.018569	-0.033419	0.877280	1.000000	0.899839	0.779904	0.963616	0.245528	-0.267809
N2O	0.993845	0.013632	-0.050820	0.976720	0.899839	1.000000	0.522477	0.867931	0.199757	-0.337055
CFC-11	0.569106	-0.013111	0.069000	0.514060	0.779904	0.522477	1.000000	0.868985	0.272046	-0.043921
CFC-12	0.897012	0.000675	0.008286	0.852690	0.963616	0.867931	0.868985	1.000000	0.255303	-0.225131
TSI	0.170302	-0.034606	-0.154492	0.177429	0.245528	0.199757	0.272046	0.255303	1.000000	0.052117
Aerosols	-0.345247	0.014890	0.340238	-0.356155	-0.267809	-0.337055	-0.043921	-0.225131	0.052117	1.000000

The linear correlation of N2O and CFC.11 with other variables in the data set is quite large. The first explanation does not seem correct, as the warming effect of nitrous oxide and CFC-11 are well documented, and our regression analysis is not enough to disprove it. The second explanation is unlikely, as we have estimated eight coefficients and the intercept from 284 observations.

Conclusions :

- We can observe that N2O is highly correlated with CO2, CH4 and CFC-12 which all contribute towards climate change.
- In this particular problem many of the variables (CO2, CH4, N2O, CFC.11 and CFC.12) are highly correlated, since they are all driven by human industrial development.
- From the regression analysis we can conclude that there indeed is global warming due to all these greenhouse gases.
- Linear regression is thus an important tool for statistical analysis. Its broad spectrum of uses includes relationship description, estimation, and prognostication.