Analysis of Temperature Variation with different Parameters using Linear Regression

Submitted To:
Prof. Gaurav Garg, IIM Lucknow

Prepared By:

Naveen Kumar (IPMX13024)
Anubha Mishra (IPMX13057)
Pushpendra Patel (IPMX13083)
Riddhi Kalaria (IPMX13086)
Sagar Wahal (IPMX13088)

Table of Contents

Motivation

From means of transport to food sources, from our holiday plans to the clothes we wear, from our life choices to various sources of livelihood, the climate that we live in has an immense impact on nearly every aspect of our lives. Climate can be defined as the pattern of weather conditions of specific geography.

Since the onset of the 21st century, climate change has become a significant factor impacting various government policies all around the world. The activities of humans have been creating a huge impact on climatic conditions since the 19th century.

We believe multivarious analysis can be used to help us understand the significant factors that are impacting climatic changes across the world. We believe that this analysis may help us understand which factors we can concentrate on to reduce the impact on climate change.

Dataset Description

The file ClimateChange.csv contains climate data from 1984 May to 2000 December. It has been sourced from MITx: 15.071x The Analytics Edge program offered by MIT OpenCourseWare (OCW) in the Spring Semester 2014.

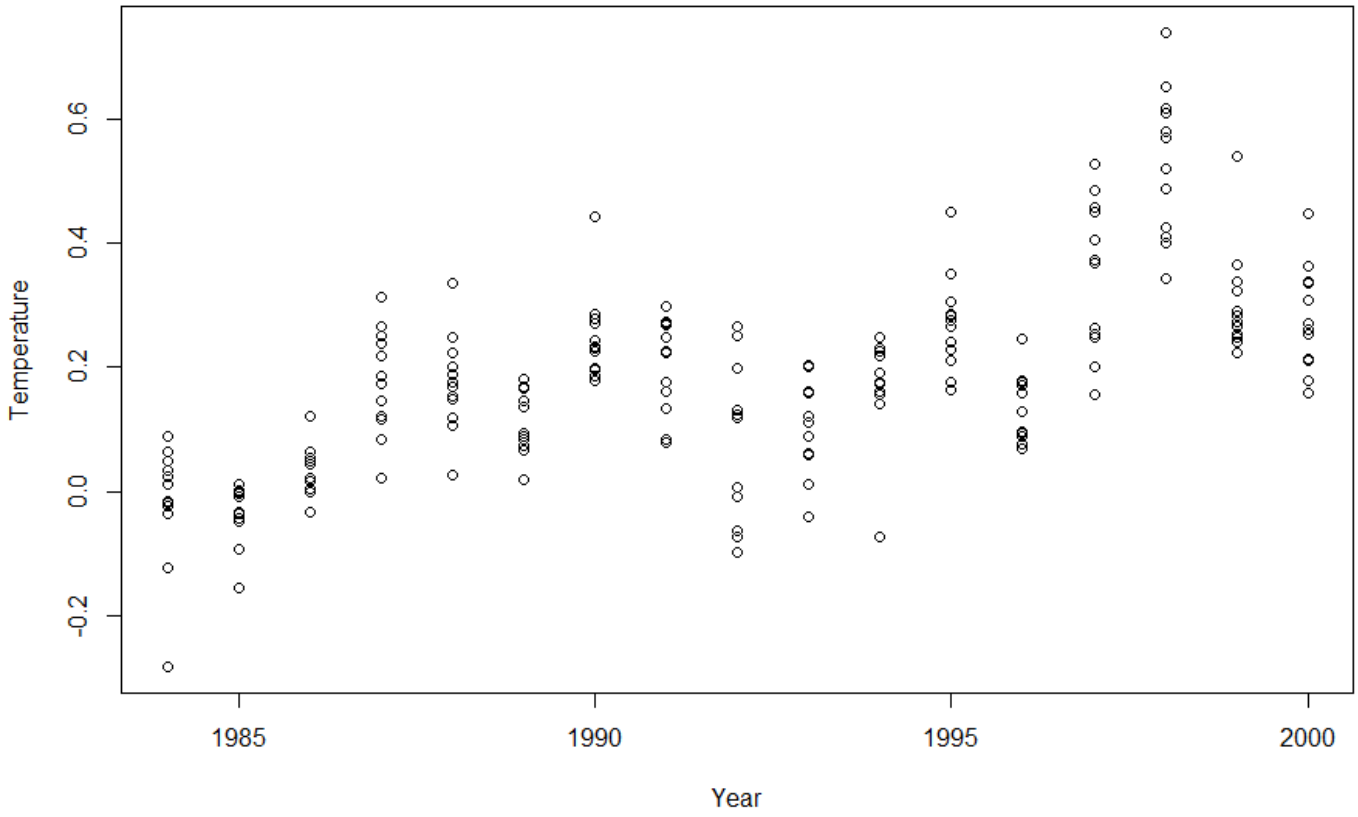
Data:

https://drive.google.com/drive/folders/1DerwfAgIyXxKBBGaU4n7-lQwJCAXGlDj?usp=sharing

The given variables include:

- ***Year, Month***: The observation year & month.
- ***Temp***: The average global temperature in that period in degrees Celsius. This data is sourced from the Climatic Research Unit at the University of East Anglia.
- ***CO2***, ***N2O***, ***CH4***, ***CFC.11***, ***CFC.12***: The data for the given parameters comes from the ESRL/NOAA Global Monitoring Division.
    - The atmospheric concentration of Carbon Dioxide (CO2), Nitrous Oxide (N2O), and Methane (CH4) is expressed in parts per million by volume (ppmv).
    - Atmospheric concentrations of Trichlorofluoromethane (CCl3F; CFC-11) and Dichlorodifluoromethane (CCl2F2; CFC-12) are expressed in parts per billion by volume (ppbv).
- ***Aerosols***: The variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, affecting how much of the sun's energy is reflected into space (Mean stratospheric aerosol optical depth at 550 nm). This data is derived from the Godard Institute for Space Studies at NASA.
- ***TSI***: Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. The rate of deposition of the sun's energy per unit area, known as total solar irradiance (TSI), is in W/m2. This data is acquired from the SOLARIS-HEPPA project website.
- ***MEI***: MEI (Multivariate El Nino Southern Oscillation Index) measures the Pacific Ocean's weather effect that affects global temperatures. This data is sourced from the ESRL/NOAA Physical Sciences Division.

Data Analysis



Temperature vs Year

Here we can see that there is overall rise in temperature with years.

```
> summary(climate)
      Year          Month             MEI                CO2              CH4
 Min.   :1984   Min.   : 1.00   Min.   :-1.5860   Min.   :341.4   Min.   :1630
 1st Qu.:1988   1st Qu.: 3.75   1st Qu.:-0.4657   1st Qu.:350.6   1st Qu.:1701
 Median :1992   Median : 6.50   Median : 0.2375   Median :356.3   Median :1746
 Mean   :1992   Mean   : 6.50   Mean   : 0.3260   Mean   :356.8   Mean   :1735
 3rd Qu.:1996   3rd Qu.: 9.25   3rd Qu.: 1.0000   3rd Qu.:363.2   3rd Qu.:1768
 Max.   :2000   Max.   :12.00   Max.   : 3.0010   Max.   :371.8   Max.   :1803
      N2O            CFC.11          CFC.12           TSI          Aerosols
 Min.   :304.1   Min.   :197.2   Min.   :363.4   Min.   :1365   Min.   :0.00210
 1st Qu.:306.7   1st Qu.:244.6   1st Qu.:443.6   1st Qu.:1366   1st Qu.:0.00540
 Median :310.0   Median :264.6   Median :505.4   Median :1366   Median :0.00865
 Mean   :309.7   Mean   :253.8   Mean   :482.7   Mean   :1366   Mean   :0.02111
 3rd Qu.:311.8   3rd Qu.:268.9   3rd Qu.:530.4   3rd Qu.:1366   3rd Qu.:0.01678
 Max.   :316.2   Max.   :271.5   Max.   :542.4   Max.   :1367   Max.   :0.14940
  Temperature
 Min.   :-0.2820
 1st Qu.: 0.0760
 Median : 0.1760
 Mean   : 0.1825
 3rd Qu.: 0.2660
 Max.   : 0.7390
>
```

Summary of the data

```
> cor(climate)
                    Year        Month          MEI          CO2          CH4          N2O
Year         1.00000000  0.000000000 -0.01116715  0.96592367  0.94173347  0.98687362
Month        0.00000000  1.000000000 -0.03485837 -0.12019015  0.07739003  0.06178369
MEI         -0.01116715 -0.034858368  1.00000000 -0.02721314  0.03470189 -0.03741654
CO2          0.96592367 -0.120190152 -0.02721314  1.00000000  0.89908764  0.94669991
CH4          0.94173347  0.077390035  0.03470189  0.89908764  1.00000000  0.93571728
N2O          0.98687362  0.061783689 -0.03741654  0.94669991  0.93571728  1.00000000
CFC.11       0.79619069  0.053052463  0.15958411  0.75219805  0.87761861  0.76364978
CFC.12       0.94748299  0.055976594  0.07559866  0.90191681  0.96038679  0.92426094
TSI          0.22765309  0.010018866 -0.17247976  0.27594125  0.28066510  0.29515405
Aerosols    -0.13242242 -0.006595161  0.34102838 -0.16132257 -0.04157781 -0.11502035
Temperature  0.65336874 -0.135838491  0.25817254  0.67369576  0.61031826  0.63215430
                CFC.11       CFC.12          TSI      Aerosols Temperature
Year        0.79619069 9.474830e-01  0.22765309 -1.324224e-01   0.6533687
Month       0.05305246 5.597659e-02  0.01001887 -6.595161e-03  -0.1358385
MEI         0.15958411 7.559866e-02 -0.17247976  3.410284e-01   0.2581725
CO2         0.75219805 9.019168e-01  0.27594125 -1.613226e-01   0.6736958
CH4         0.87761861 9.603868e-01  0.28066510 -4.157781e-02   0.6103183
N2O         0.76364978 9.242609e-01  0.29515405 -1.150203e-01   0.6321543
CFC.11      1.00000000 9.455873e-01  0.32224885  1.181986e-01   0.5242837
CFC.12      0.94558728 1.000000e+00  0.28583367  5.706456e-05   0.6171170
TSI         0.32224885 2.858337e-01  1.00000000  1.114075e-01   0.2383832
Aerosols    0.11819858 5.706456e-05  0.11140753  1.000000e+00  -0.2755793
Temperature 0.52428367 6.171170e-01  0.23838318 -2.755793e-01   1.0000000
>
```

Correlation between different variables

Model A

We run Model A with all the variables.

```
> modelA = lm(Temperature ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 + TSI + Aerosols, data = climate)
> summary(modelA)

Call:
lm(formula = Temperature ~ MEI + CO2 + CH4 + N2O + CFC.11 + CFC.12 +
    TSI + Aerosols, data = climate)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26197 -0.06426 -0.00629  0.05802  0.33974

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.357e+02  2.803e+01  -4.841 2.62e-06 ***
MEI          6.824e-02  7.466e-03   9.140  < 2e-16 ***
CO2          7.967e-03  3.051e-03   2.611  0.00973 **
CH4          3.020e-04  6.729e-04   0.449  0.65401
N2O         -2.753e-02  1.504e-02  -1.831  0.06868 .
CFC.11      -6.250e-03  2.734e-03  -2.285  0.02336 *
CFC.12       3.969e-03  1.734e-03   2.290  0.02312 *
TSI          1.030e-01  2.205e-02   4.672 5.55e-06 ***
Aerosols    -1.722e+00  2.385e-01  -7.223 1.11e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09628 on 195 degrees of freedom
Multiple R-squared:  0.6502,    Adjusted R-squared:  0.6358
F-statistic:  45.3 on 8 and 195 DF,  p-value: < 2.2e-16
```

CH4 with p value 65% is removed.

Model B

From the base model CH4 variable is removed and the model is re-run.

```
> modelB = lm(Temperature ~ MEI + CO2 + N2O + CFC.11 + CFC.12 + TSI + Aerosols, data = climate)
> summary(modelB)

Call:
lm(formula = Temperature ~ MEI + CO2 + N2O + CFC.11 + CFC.12 +
    TSI + Aerosols, data = climate)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26198 -0.06599 -0.00448  0.05818  0.33386

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.343e+02  2.781e+01  -4.831 2.73e-06 ***
MEI          6.811e-02  7.445e-03   9.148  < 2e-16 ***
CO2          7.982e-03  3.045e-03   2.621  0.00945 **
N2O         -2.508e-02  1.398e-02  -1.794  0.07440 .
CFC.11      -6.009e-03  2.676e-03  -2.246  0.02585 *
CFC.12       3.965e-03  1.730e-03   2.292  0.02295 *
TSI          1.018e-01  2.184e-02   4.662 5.78e-06 ***
Aerosols    -1.726e+00  2.379e-01  -7.255 9.13e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09608 on 196 degrees of freedom
Multiple R-squared:  0.6498,    Adjusted R-squared:  0.6373
F-statistic: 51.95 on 7 and 196 DF,  p-value: < 2.2e-16
```

N2O with p value 7.4% is removed from the model.

Model C

From the model B, N2O variable is removed and the model is re-run.

```
> modelC = lm(Temperature ~ MEI + CO2 + CFC.11 + CFC.12 + TSI + Aerosols, data = climate)
> summary(modelC)

Call:
lm(formula = Temperature ~ MEI + CO2 + CFC.11 + CFC.12 + TSI +
    Aerosols, data = climate)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25365 -0.06565 -0.00412  0.05569  0.33571

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.158e+02  2.597e+01  -4.461 1.37e-05 ***
MEI          6.753e-02  7.480e-03   9.028  < 2e-16 ***
CO2          7.756e-03  3.060e-03   2.535    0.012 *
CFC.11      -2.002e-03  1.482e-03  -1.351    0.178
CFC.12       1.252e-03  8.436e-04   1.484    0.139
TSI          8.284e-02  1.921e-02   4.311 2.56e-05 ***
Aerosols    -1.732e+00  2.392e-01  -7.242 9.74e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09662 on 197 degrees of freedom
Multiple R-squared:  0.644,     Adjusted R-squared:  0.6332
F-statistic: 59.41 on 6 and 197 DF,  p-value: < 2.2e-16
```

CFC.11 & CFC.12 with p values 17.8% & 13.9% respectively, are removed from the model.

Model D

From the model C, CFC.11 & CFC.12 variables are removed and the model is re-run.

```
> modelD = lm(Temperature ~ MEI + CO2 + TSI + Aerosols, data = climate)
> summary(modelD)

Call:
lm(formula = Temperature ~ MEI + CO2 + TSI + Aerosols, data = climate)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25678 -0.06521 -0.00543  0.05831  0.34173

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.071e+02  2.476e+01  -4.326 2.40e-05 ***
MEI          6.603e-02  7.237e-03   9.125  < 2e-16 ***
CO2          1.186e-02  9.354e-04  12.681  < 2e-16 ***
TSI          7.547e-02  1.820e-02   4.146 5.01e-05 ***
Aerosols    -1.705e+00  2.306e-01  -7.396 3.84e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09667 on 199 degrees of freedom
Multiple R-squared:   0.64,     Adjusted R-squared:  0.6328
F-statistic: 88.46 on 4 and 199 DF,  p-value: < 2.2e-16
```

Here all the variables are significant and there is multicollinearity.

VIF Test

Further vif test is conducted to check for multicollinearity on base model A and consider values over 10 as insignificant.

```
> vif(modelA)
       MEI        CO2        CH4        N2O     CFC.11     CFC.12        TSI   Aerosols
  1.293618  12.236339  16.901699  51.089893  76.770598 208.193992   1.751334   1.330008
> modelE = lm(Temperature ~ MEI + TSI + Aerosols, data = climate)
```

After this we recheck the model for adjusted R squared as in model E below.

Model E

```
> summary(modelE)

Call:
lm(formula = Temperature ~ MEI + TSI + Aerosols, data = climate)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35456 -0.09197  0.00402  0.07013  0.44145

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.021e+02  3.166e+01  -6.384 1.18e-09 ***
MEI          7.566e-02  9.653e-03   7.838 2.67e-13 ***
TSI          1.481e-01  2.318e-02   6.390 1.14e-09 ***
Aerosols    -2.363e+00  3.014e-01  -7.840 2.63e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1297 on 200 degrees of freedom
Multiple R-squared:  0.3492,    Adjusted R-squared:  0.3394
F-statistic: 35.77 on 3 and 200 DF,  p-value: < 2.2e-16
```

Here after conducting the vif test and removing variables as per vif test, we see that adjusted R squared has decreased significantly in model E, so we will go with model D.

Conclusion

The summary statistics tell us that there are no nulls and no outliers in the data.

We began by looking at the following variables, which are *CO2*, *N2O*, *CH4*, *CFC.11*, *CFC.12, Aerosols, TSI, and MEI,* which might display a strong correlation temperature result.

Via Model A, we proved the alternative hypothesis to be true, meaning that at least one variable has a linear relation with the result.

Post that, we began eliminating the variables to create a robust model with a better adjusted R^2 value. Finally, we zeroed in on Model D, which gave us the best R^2 value. The following variables showed the strongest relation to the temperature: MEI, CO2, TSI, and Aerosols.