Contents


```
clc;
clear variables;
close all;

% The below contains the location of the data file on my laptop. Please
% change it with the location of yours of ghg-concentrations_1984-2014.xlsx

data = readmatrix("C:\Users\ojasp\OneDrive\Desktop\IITM\SEM4\CH5440\assign2\ghg-concentrations_1984-2014.xlsx");

year = data(:, 1);
CO2_conc = 1000*data(:,2); % I multiplied by 1000 so as to bring the entire system in ppb units
CH4_conc = data(:,3);
N2O_conc = data(:,4);
O3_conc =  2.69*data(:,5); % Similar reason as above
```

Part A

```
X = [CO2_conc, CH4_conc, N2O_conc, O3_conc];
Y = data(:,7); % Target value

X_s = [];

for i = 1:4
    X_s = cat(2,X_s,X(:,i)-mean(X(:,i)));
end

Y_s = Y - mean(Y);
m = (X_s'*X_s)\X_s'*Y_s;
c = mean(Y) - [mean(X(:,1)),mean(X(:,2)),mean(X(:,3)),mean(X(:,4))]*m;

Y_estimate = X*m + c;
SSE = Y - Y_estimate;

error_variance = sum(SSE.*SSE)/29;
fprintf("The slope obtained by MISO type regression model of CO2, CH4, N2O, O3 is %0.4f, %0.4f, %0.4f, %0.4f\n", m(1), m(2), m(3), m(4));
```

```
The slope obtained by MISO type regression model of CO2, CH4, N2O, O3 is 0.0001, 0.0059, -0.1465, 0.0030
```

A) Calculating Correlations:

```
COV1 = cov(CO2_conc, Y);
rho_CO2_Y = COV1(1,2)/std(CO2_conc)/std(Y);

COV2 = cov(CH4_conc, Y);
rho_CH4_Y = COV2(1,2)/std(CH4_conc)/std(Y);

COV3 = cov(N2O_conc, Y);
rho_N2O_Y = COV3(1,2)/std(N2O_conc)/std(Y);

COV4 = cov(O3_conc, Y);
rho_O3_Y = COV4(1,2)/std(O3_conc)/std(Y);

fprintf("\nThe values of correlation coefficients of CO2, CH4, N2O, O3 are %0.4f, %0.4f, %0.4f, %0.4f. \n", rho_CO2_Y, rho_CH4_Y, rho_N2O_Y, rho_O3_Y)
disp("As the value of correlation coefficient of O3 is -ve, hence it shows negative correlation coefficient");

fprintf("\nHowever, it is interesting to note that despite the presence of a negative slope for N2O, the correlation coefficient of N2O is +ve\n");
fprintf("At the same time, the slope of Ozone despite being +ve shows -ve correlation coefficient\n")

fprintf("\n**********************************************************************************************************************************************\n")
```

```
The values of correlation coefficients of CO2, CH4, N2O, O3 are 0.8863, 0.8822, 0.8837, -0.0591.
As the value of correlation coefficient of O3 is -ve, hence it shows negative correlation coefficient
```

```
However, it is interesting to note that despite the presence of a negative slope for N2O, the correlation coefficient of N2O is +ve
At the same time, the slope of Ozone despite being +ve shows -ve correlation coefficient

**********************************************************************************************************************************
```

B) Bootstrapping and obtaining confidence intervals

```
k = 29;
CO2_conc_comb = nchoosek(X(:,1),k);
CH4_conc_comb = nchoosek(X(:,2),k);
N2O_conc_comb = nchoosek(X(:,3),k);
O3_conc_comb = nchoosek(X(:,4),k);
Y_comb = nchoosek(Y,k);
m_est = [];
c_n = zeros(465, 1);

for i=1:465
    X_new = [CO2_conc_comb(i,:)',CH4_conc_comb(i,:)',N2O_conc_comb(i,:)',O3_conc_comb(i,:)'];
    Y_new = Y_comb(i,:)';
    X_s_new = [];

    for j = 1:4
        X_s_new = cat(2,X_s_new,X_new(:,j)-mean(X_new(:,j)));
    end

    Y_s_new = Y_new - mean(Y_new);

    m_new = inv(X_s_new'*X_s_new)*X_s_new'*Y_s_new;
    m_est = [m_est m_new];

    c_new = mean(Y_new) - [mean(X_new(:,1)),mean(X_new(:,2)),mean(X_new(:,3)),mean(X_new(:,4))]*m_new;
    c_n(i) = c_new;
end

m1_avg = mean(m_est(1,:));
m2_avg = mean(m_est(2,:));
m3_avg = mean(m_est(3,:));
m4_avg = mean(m_est(4,:));
c_final = mean(c_n);

m1_sst = m_est(1,:) - m1_avg; m1_est = sqrt(sum(m1_sst.*m1_sst)/464);
fprintf("\nThe estimates of slope of CO2 Concentrations lies from [%0.4f , %0.4f]\n", m(1) - 2*m1_est, m(1) + 2*m1_est);

m2_sst = m_est(2,:) - m2_avg; m2_est = sqrt(sum(m2_sst.*m2_sst)/464);
fprintf("The estimates of slope of CH4 Concentrations lies from [%0.4f , %0.4f]\n", m(2) - 2*m2_est, m(2) + 2*m2_est);

m3_sst = m_est(3,:) - m3_avg; m3_est = sqrt(sum(m3_sst.*m3_sst)/464);
fprintf("The estimates of slope of N2O Concentrations lies from [%0.4f , %0.4f]\n", m(3) - 2*m3_est, m(3) + 2*m3_est);

m4_sst = m_est(4,:) - m4_avg; m4_est = sqrt(sum(m4_sst.*m4_sst)/464);
fprintf("The estimates of slope of O3 Concentrations lies from [%0.4f , %0.4f]\n\n", m(4) - 2*m4_est, m(4) + 2*m4_est);

fprintf("**********************************************************************************************************************************\n")
```

```
The estimates of slope of CO2 Concentrations lies from [0.0000 , 0.0001]
The estimates of slope of CH4 Concentrations lies from [0.0047 , 0.0071]
The estimates of slope of N2O Concentrations lies from [-0.1900 , -0.1031]
The estimates of slope of O3 Concentrations lies from [0.0008 , 0.0052]

**********************************************************************************************************************************
```

Plots

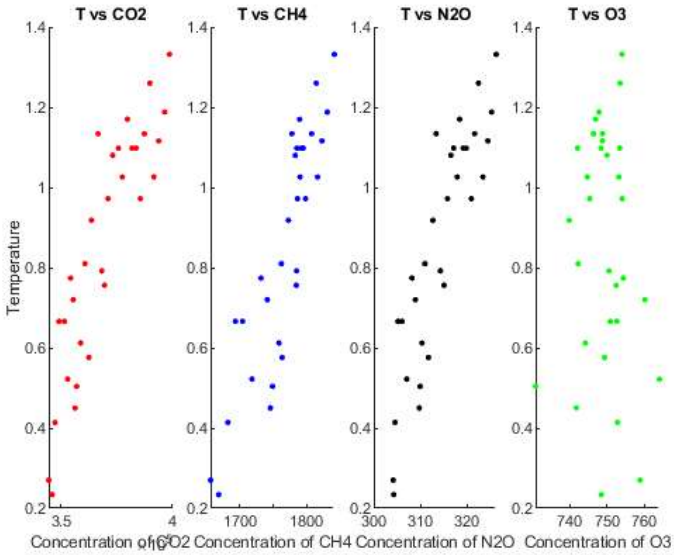
```
subplot(1, 4, 1);
scatter(CO2_conc, Y, 10, 'r', "filled"); xlabel("Concentration of CO2"); ylabel("Temperature"); title("T vs CO2");

subplot(1, 4, 2);
scatter(CH4_conc, Y, 10, 'blue', "filled"); xlabel("Concentration of CH4"); title("T vs CH4");

subplot(1, 4, 3);
scatter(N2O_conc, Y, 10, 'black', "filled"); xlabel("Concentration of N2O"); title("T vs N2O");

subplot(1, 4, 4);
scatter(O3_conc, Y, 10, 'green', "filled"); xlabel("Concentration of O3"); title("T vs O3");

% As per the above plots, it looks like there is a positive correlation of CO2, CH4, N20 for
% temperature, while there does not seem to be a proper relation for Ozone
% by eyeballing
```



Residual Analysis

As 95% confidence interval included all the points, I'm considering a 90% confidence interval for the errors in Y which corresponds to 1.645*sigma

```
fprintf("\nThe size of X matrix (data matrix of concentrations) is %0.0d, %0.0d\n", size(X, 1), size(X, 2));
high = 0.5;
y_rem = [];

while(high>1.645*sqrt(error_variance))
arr = [];
for i=1:size(Y,1)
    arr = cat(1, arr, (abs(Y(i) - (m'*X(i, :)' + c))));
end

j=0;
for i=1:size(Y,1)
    if (arr(i) == max(arr))
        j=i; break;
    end
end
high = max(arr);
y_rem(end+1) = Y(j);
X(j, :) = []; Y(j) = []; arr(j) = [];
end

fprintf("The values which got removed were %0.4f, %0.4f, %0.4f, %0.4f\n", y_rem(1), y_rem(2), y_rem(3), y_rem(4));
fprintf("The size of the new X matrix (data matrix of concentrations) after removing outliers is %0.0d, %0.0d \n", size(X, 1), size(X, 2));
```

```
The size of X matrix (data matrix of concentrations) is 31, 4
The values which got removed were 1.1340, 0.5760, 1.0800, 0.5220
The size of the new X matrix (data matrix of concentrations) after removing outliers is 27, 4
```

Now there are no outliers

```
X_s = [];
for i = 1:4
    X_s = cat(2,X_s,X(:,i)-mean(X(:,i)));
end

Y_s = Y - mean(Y);
Y_estimate = X*m + c;
SSE = Y - Y_estimate;

error_variance1 = sum(SSE.*SSE)/(size(Y, 1) - 2);
m = (X_s'*X_s)\X_s'*Y_s;

c = mean(Y) - [mean(X(:,1)),mean(X(:,2)),mean(X(:,3)),mean(X(:,4))]*m;

Y_estimate = X*m + c;
SSE = Y - Y_estimate;

fprintf("\nThe newly calculated slope is %0.4f, %0.4f, %0.4f, %0.4f\n", m(1), m(2), m(3), m(4));
```

```
The newly calculated slope is 0.0001, 0.0071, -0.2095, 0.0050
```

```
X = X(:, 1:3);

X_s = [];
for i = 1:3
    X_s = cat(2,X_s,X(:,i)-mean(X(:,i)));
end

error_variance2 = sum(SSE.*SSE)/(size(Y, 1) - 2);
m = (X_s'*X_s)\X_s'*Y_s;

Y_s = Y - mean(Y);
Y_estimate = X*m + c;
SSE = Y - Y_estimate;

c = mean(Y) - [mean(X(:,1)),mean(X(:,2)),mean(X(:,3))]*m;

Y_estimate = X*m + c;
SSE = Y - Y_estimate;

fprintf("The calculated slope after dropping the unimportant(insignificant) variable is %0.4f, %0.4f, %0.4f\n\n", m(1), m(2), m(3));

fprintf("**********************************************************************************************************************************************\n\n")

fprintf("The error variance originally was %0.4f\n", error_variance);
fprintf("The error variance after removing the outliers became %0.4f\n", error_variance1);
fprintf("The error variance after removing the unimportant column of ozone concentration was %0.4f\n", error_variance2);
```

```
The calculated slope after dropping the unimportant(insignificant) variable is 0.0001, 0.0066, -0.2160

**********************************************************************************************************************************************

The error variance originally was 0.0149
The error variance after removing the outliers became 0.0100
The error variance after removing the unimportant column of ozone concentration was 0.0093
```

d) GWP Calculations

```
fprintf("\nThe ratio of slope of CH4 and CO2 is %0.1f which is close to the actual value of 86\n", m(2)/m(1));
fprintf("The ratio of slope of N2O and CO2 is %0.1f which is very far from the value of 289\n", abs(m(3)/m(1)));
disp("GWP is a complex metric, and regression coefficients alone cannot fully capture its nuances ")
disp("For accurate assessments, consulting established GWP values and considering the specific GHGds behavior over the desired time horizon is necessary.")
```

```
The ratio of slope of CH4 and CO2 is 76.2 which is close to the actual value of 86
The ratio of slope of N2O and CO2 is 2490.8 which is very far from the value of 289
GWP is a complex metric, and regression coefficients alone cannot fully capture its nuances
For accurate assessments, consulting established GWP values and considering the specific GHGds behavior over the desired time horizon is necessary.
```

Published with MATLAB® R2023a