Homework 1

Pankti Antani and Yukta Butala

a) After running the linear regression model on python to predict the dependent variable Temp, using CO2, CH4, N2O, CFC-11, CFC-12, Aerosols, TSI and MEI as features, we found the following in-sample and out-of-sample R2, MSE, and MAE.
in-sample R2: 0.6920595959984741
in-sample MSE: 0.008731426409911177
in-sample MAE: 0.07260918612938931
out-of-sample R2: -0.541325583402297
out-of-sample MSE: 0.012206974835137179
out-of-sample MAE: 0.09312747891276273

(b) After running the linear regression model on python to predict the dependent variable Temp, using N2O, Aerosols, TSI, and MEI as features, we found the following in-sample and out-of-sample R2, MSE, and MAE.
in-sample R2: 0.6490120806760372
in-sample MSE: 0.009952007429105784
in-sample MAE: 0.07666650280233205
out-of-sample R2: 0.20031861104556226
out-of-sample MSE: 0.006333308611894036
out-of-sample MAE: 0.06154027269393422

(c) Between the two models built in parts (a) and (b), the model built in part (a) performs better in-sample, as its R-squared is higher and error metrics are smaller. However, the model in part (a) has a negative R-squared for out-of-sample, indicating poor performance compared to using averages as predicted values. In contrast, the model in part (b) explains roughly 20% of the out-of-sample variability in predicted values. Therefore, the model in part (b) performs better out-of-sample.

(d) Model 1 has N2O coefficient = -3.48478075e-02
Interpretation: If we keep all the parameters and features constant and increase the N2O content by 1 unit, the temperature will decrease by -3.48478075e-02 units.
Model 2 has N2O coefficient = 0.02427612
Interpretation: If we keep all the parameters and features constant and increase the N2O content by 1 unit, the temperature will increase by 0.02427612 units.

(e) We feel that we should prefer the Model 2, ie. use only N2O, Aerosols, TSI, and MEI as features to predict the temperature. The first model suggests that as the quantity of N2O increases, the temperature drops which is against the scientific opinion, given N2O is a greenhouse gas that increases temperature and not decreases.

Additionally, out-of-sample R^2 of the first model is negative, making the model unreliable to understand the true impact of independent variables, including N20, on the dependent variable.

The second model still includes N20's impact on temperature and has a better out-of-sample R^2, meaning it is more effective in making accurate predictions on new data.