It is obvious that the emission of CO2 is caused by many aspects of our society. The population of human beings, for example, could directly contribute to the emission of CO2 into the atmosphere since the action of exhaling produces CO2. Meanwhile, human actions such as powering transports, producing goods, and destructing forests all cause changes in the environment, releasing large amounts of CO2 while removing the plants that are responsible for the uptake of CO2.

Our group has considered various aspects, but it seems impossible to include all variables for our prediction. With that said, we strive to produce models that mimic reality the most with limited amounts of variables.

To begin with, we set time as the independent variable and the level of CO2 in the atmosphere as the dependent variable. The first model that could be set would be the model of linear regression.

Model 1: Linear Regression

* 1. Introduction to the model:

This model is used to predict the relationship between two variables by applying a linear equation to the given data. It is commonly used for predictive analysis that includes an explanatory (independent) variable and a dependent variable. But before using this model, it is essential to prove that there is a relationship between the two variables.

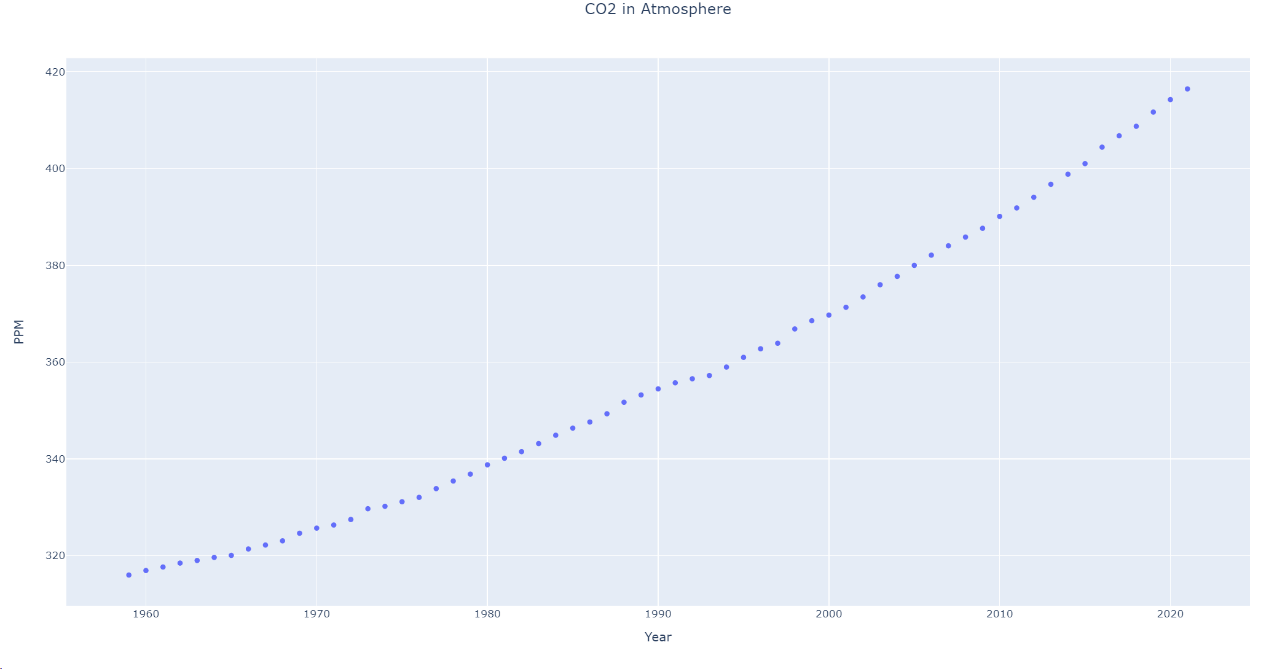
* 1. Variables of the model:

x: the time (year number)

y: the level of CO2 in the atmosphere (in ppm)

1.3 Justification of the use of this model:

To justify the use of linear regression, all we have to do is to justify that there is a close relationship between the year number and the level of CO2 in the atmosphere. As a visualized graph, a scattered plot may be used to indicate such correlation:



As shown above, the trend of the line is sloping upwards, indicating that there is a positive relationship between the year number and the level of CO2: The level of CO2 increases as the year increases. And since the line is relatively flat, it seems reasonable for us to use linear regression as a model to predict the future level of CO2 with respect to the year.

Now we have the idea of a correlation, we may use a correlation coefficient to prove numerically that our conclusion is not only a conjecture. Correlation coefficient formulas are used to measure the relationship between two variables. It will return a value between -1 and 1. The magnitude of the value means how strong the relationship is (with 1 indicating the strongest relationship). The sign of the value means how the two values are related. If it results is negative, then the two variables have a negative relationship; if the result is positive, then the two variables have a positive relationship. If the result is 0, then there is no relationship at all.

The most commonly used type of formula is Pearson’s correlation coefficient formula, which is the one we will use here:

is the sample covariance.

is the standard deviation of values of x.

is the standard deviation of values of y.

Expanding the formula will give us:

Where n is the sample size.

The result is 0.9912, which means that there is an extremely strong and positive correlation between the two variables. Our use of the model is valid.

To insert our regression line into the XY plot, we use the method of least squares. This process determines the best-fitting line for the given data by reducing the sum of the squares of the vertical deviations from each data point to the line.

[photo of the code]

[explanatory of the code]

[photo of the resulted line + equation of the line]

The linear line is the line of best fit for the data given. Yet it is impossible for the line to be absolutely straight in the reality. This is why we include our random effect model

Model 2: Random Effect Model

But, if we view the entire graph in a larger overview, it seems like both models previously doesn’t fit exactly into reality. Starting from around 2010, the level of CO2 in the atmosphere seems to be increasing in a larger speed than before. It appears to be reasonable that the speed will continue to increase as the technology advances even more. However, the speed of increase in both of our models seems to be relatively constant. For the linear model, the speed of increase is simply the slope of the line. For the random effect model, the speed of increase in based on the basic line and the randomly generated numbers. This is why we comes to the idea of making a quadratic regression model.

Model 3: quadratic regression model