Project Proposal

Thuong Pham
Maoting Zeng
Andy Aliakseichuk

December 9, 2021

Abstract

For our final project we chose to do linear regression of life expectancy by country. In this project, we used a life expectancy data set with data from countries from around the world. In this project, we use linear regression to predict the age of death world-wide and by country in the years after the dataset. In this experiment, We get the dataset from the website https://ourworldindata.org/life-expectancy by CSV file and we pushed it into GitHub to convert it to the raw file and use the link from GitHub to generate the dataset."https://raw.githubusercontent.com/timphamvn33/ML_Project/Tim/life-expectancy.csv" The data source has different countries' name and their ages. The countries' human ages were collected from years 1500 to 2019. However, we only manipulate the data from 1850 to 2019 because in this time frame, in those decades, people's age is not much affected by wars and epidemics, so the stability and accuracy will be improved. We split the dataset in half. Half of the dataset will be used for the training set and the rest for the testing set. This experiment will use a linear regression algorithm that we have learned in Big Data and Machine Learning class and Python Sklearn library to predict the testing set. Then We will compare the advantages and disadvantages of using those differences between them, which is one of the most significant ways for our solution. In the first section, we apply the linear regression algorithm to find the linear equation Y = B0 + B1X. We will implement the model method with two parameters,B_0 and B_1. and return the linear prediction equation. And another process is called loss. That method will assist me in finding the linear prediction equation. In the second section, we will apply Sklearn to find the prediction linear.

0.1 Review Linear Regression

Our goal is to estimate how long humans will live. We choose linear regression because it is a valuable tool for determining how some factors influence other variables and because its answer is simple to explain. Linear regression is a simple and straightforward statistical regression method for demonstrating the relationship between continuous variables and predicting outcomes. Linear regression illustrates a linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis) as the name suggests (Y-axis). The dependent variable is the one from which we will learn, and the regression model must solve in order to provide a prediction. The term "simple linear regression" refers to regression using only one input variable (x). Many linear regression is a kind of linear regression that uses multiple input variables. A slanted straight line describing the relationship between the variables is produced by the linear regression model. To calculate the best fit line linear regression uses a traditional slope-intercept form

$$Y = B_0X + B_1$$

Y=$DependentVariable$.

X=$IndependentVariable$.

B_0 = Intercept of the line.

B_1 = Linear regression Coefficient.

. The independent variables are the variables that do not get affected by anything. When we use linear regression, there will be dependent variables and independent variables. In this project, the lifetime is the dependent variable, and the year is the independent variable.

0.2 Preparing the data

First and foremost, we generate data on human ages from 1850 to 2019 based on the original dataset. The independent factors are year data variables, while the dependent variables are human ages data variables. Our dataset will now have two columns. We divide the years (1850 to 2019) into distinct groups (0, 169). We made computing and anticipating the linear equation easier by observing human ages in a more transparent and more accessible range. We identify the lowest value in the list of years and subtract it from an array list of years to accomplish so. The ultimate result will be a different array of independent variables but with the same number of years and distances between each variable. Next, we split the dataset that we collected and sorted from above into half. Half of the dataset will be training-set, and the rest will be for prediction and testing.

0.3 Applying Linear-Regression Algorithm to Find The Linear Line

We will get a prediction line at the end of linear regression, and we can use a math equation to solve it. There will be regression coefficients and random errors in the regression equation. An estimated regression function will be used to display the relationship between variables. We want the prediction to be as near to our real data as possible, thus we'll try to keep the sum of squared residuals as low as possible. The residuals are the sum of all the observations' differences. To begin, we'll develop a model method. It has 2 parameters are B_1 and B_0. Our goal is to discover those two factors. The model method returns the value of the prediction y, which is calculated using a linear equation with an independent value of x. The other way, loss, was then implemented. This approach will calculate the distance between the predicted linear regression line and the testing set by modifying the parameters B_0 and B_1 that are our Intercept of the line and Linear Regression Coefficient. Finally, we find the initial B_0, B_1, and delta. In this experiment, $B_0 = 45$ and $B_1 = 0$. We picked those numbers for our intercept and coefficient because our picked dataset has a minimum age is 45 years old. We assumed that the linear starts at those initial values. The value of delta is the number of B_0 will change until the loss is minimized. After the loop finish, we have the prediction of intercept and coefficient. We apply those prediction numbers into the linear regression function and predict the dependent value y, human ages based on the independent value x, years.

0.4 Applying Python Sklearn Library

We used the Sklearn library to forecast the test set in this section. Now that we've separated our data into training and testing sets, it's time to train our algorithm. Implementing linear regression models with Scikit-Learn is relatively simple. And it gives us a faster result than the Linear Regression Algorithm we discussed in the previous section.

Here is the output of the human lifespan prediction

[51, 52, 52, 52, 52, 53, 53, 53, 54, 54, 54, 54, 55, 55, 55, 56, 56, 56, 56, 57, 57, 57, 58, 58, 58, 58, 59, 59, 59, 60, 60, 60, 60, 61, 61, 61, 62, 62, 62, 62, 63, 63, 63, 64, 64, 64, 64, 65, 65, 65, 66, 66, 66, 67, 67, 67, 67, 68, 68, 68, 69, 69, 69, 69, 70, 70, 70, 71, 71, 71, 71, 72, 72, 72, 73, 73, 73, 73]

And this is the actual human lifespan prediction

[52, 54, 54, 51, 58, 56, 61, 61, 50, 50, 51, 51, 52, 52, 53, 53, 54, 54, 55, 55, 56, 56, 57, 57, 57, 58, 58, 59, 59, 59, 60, 60, 60, 61, 61, 62, 62, 62, 63, 63, 63, 64, 64, 64, 65, 65, 65, 65, 66, 66, 66, 66, 66, 67, 67, 67, 67, 67, 68, 68, 68, 69, 69, 69, 69, 70, 70, 71, 71, 71, 72, 72, 72, 72, 73, 73, 73, 73]

The final stage is to assess the algorithm's performance. This phase is especially important when comparing the performance of multiple algorithms on a given dataset. Three evaluation metrics are typically employed for regression algorithms.

Mean absolute error: 1.076

Root mean squared error: 1.954

Those figures revealed that while our prediction was not perfect, it was temporary acceptable and fairly near to the true value. This indicates that Linear Regression performed admirably.

0.5 What We Learned

In this project, we learned how to apply linear regression to a dataset with multiple columns of information. We learned how to sort the data by the columns and produce graphs that interpret the results from different data types within the dataset(data from other countries). We used Sklearn's linear regression algorithm, and we learned to predict age data from the dataset. After we studied linear regression, we found that linear regression can be split into different assumptions. Here are some basic assumptions: linearity, normality, homoscedasticity, and independence. The dependent variable's linearity will be linearly related to the independent variables. The normalcy is the dependent variable, and the independent variable should be normally distributed, which means a probability distribution is symmetric from the mean. It shows the data near the mean has more frequency of occurrence than the data far away from the standard. The homoscedasticity means the variance of the error terms should be constant, which means the separation of residuals should have a regular for the independent variable. The independence means the variable should be independent, which means there is no connection between those variables. In Sklean's we used a library called r2-score. This is a regression score function. The best prediction score will be one, which can be negative. If we use a constant model to make the prediction and don't consider input, then the score will be zero. We also used the metrics from Sklean. It helps to implement further losses, scores, and utility functions to help us measure the performance. We used the mean absolute error regression loss and the mean squared error regression loss. The mean absolute error is the difference between the prediction value and the actual value. It is essential because it can tell us how big of a mistake we can expect from the prediction on average.We know the residuals are a measurement of how far the data points are from the regression line. The mean squared error regression loss measures how those residuals are split in the graph. As a result, it will tell us how concentrated the data around the best fitting line is. The mean absolute error and the mean squared error are similar but different. The mean squared error finds the difference between the prediction and the predicted value. The mean fundamental error is a quantity used to measure how close predictions are compare with the actual.

0.6 Conclusion

We are forecasting the lifespan of humans in this study. We think this is an interesting subject to research because medical science and technology have advanced considerably in the last few decades. People are concerned about what they consume and are eager to undergo health screenings. To put it another way, people desire to live longer. So, if we can anticipate a person's lifespan, it may assist those who are unconcerned about their health and encourage them to consider it. When people find statistics to back up their claims, they may alter their behavior.

0.7 The Graphs of Human Lifespan from 1550 to 2020, Using Linear Regression Algorithm and Sklearn Library to Predict The Testset

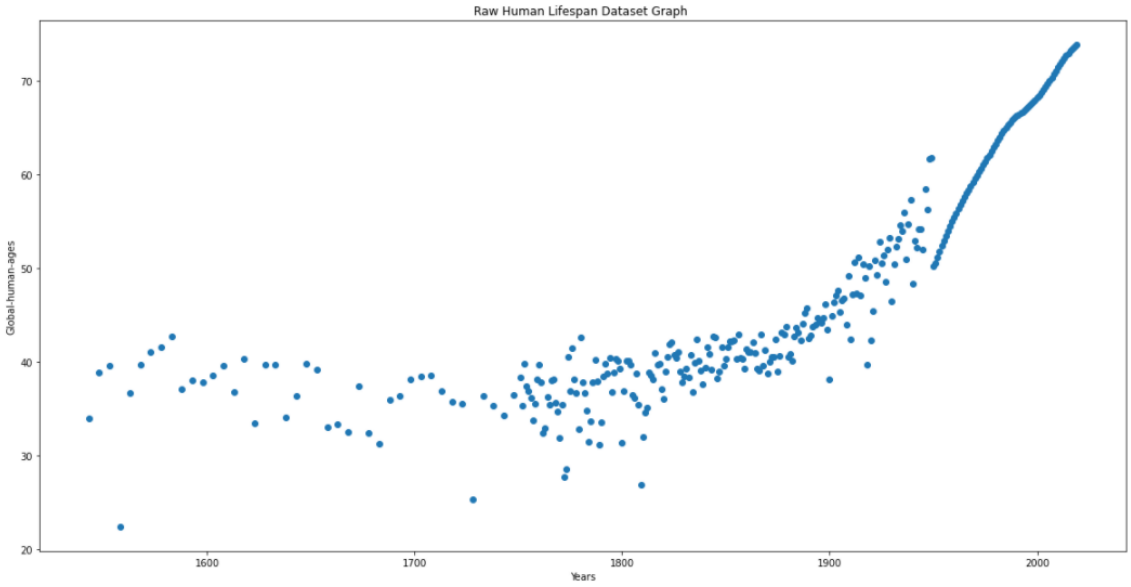

Figure 1: The Lifespan of Humans from 1550 to 2020

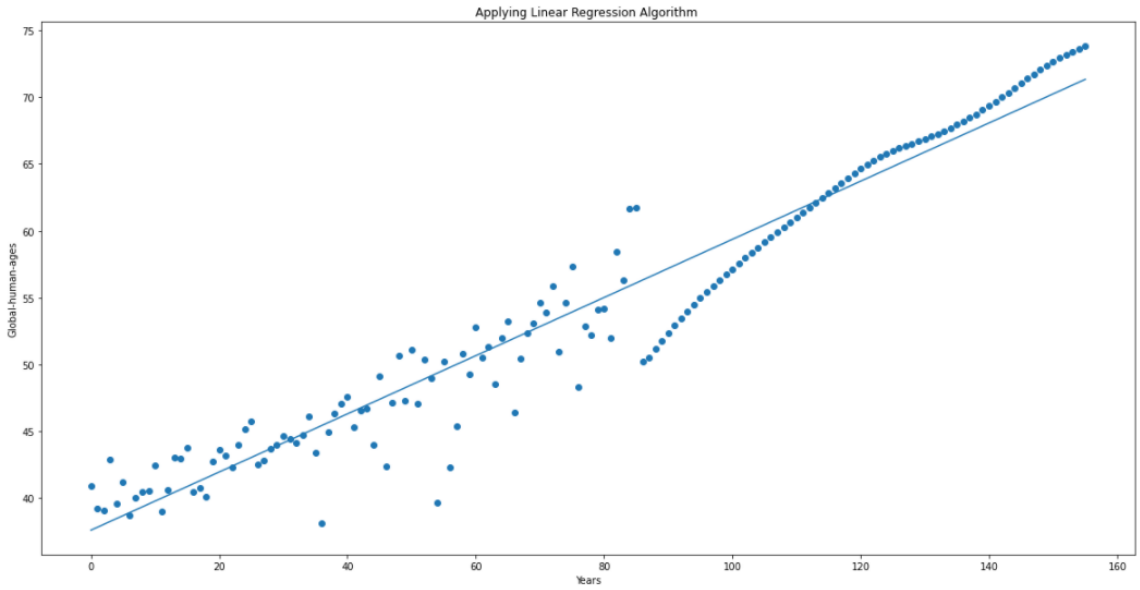

Figure 2: Applying The Linear Regression Algorithm to Predict the human ages from 1950 to 2020

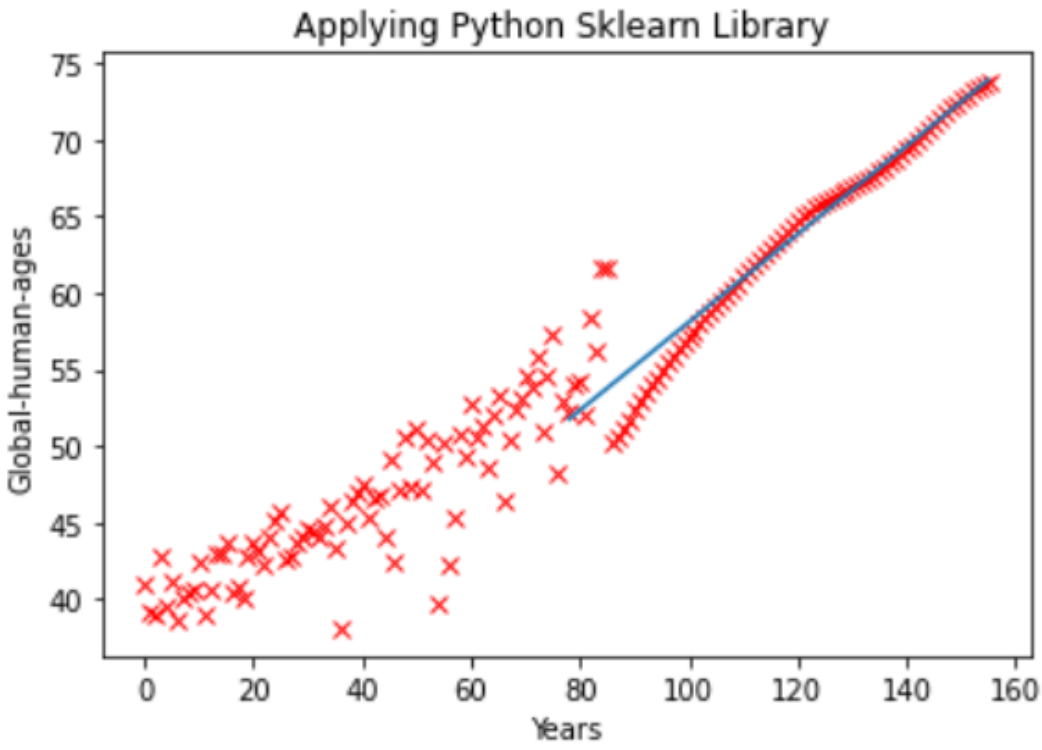

Figure 3: Applying Python Sklearn Library to Predict the human ages from 1950 to 2020