Kyle, Josh, Ariyan
Professor Fontenot
CS 5394
February 7, 2022

# World Population in 2122 Projection

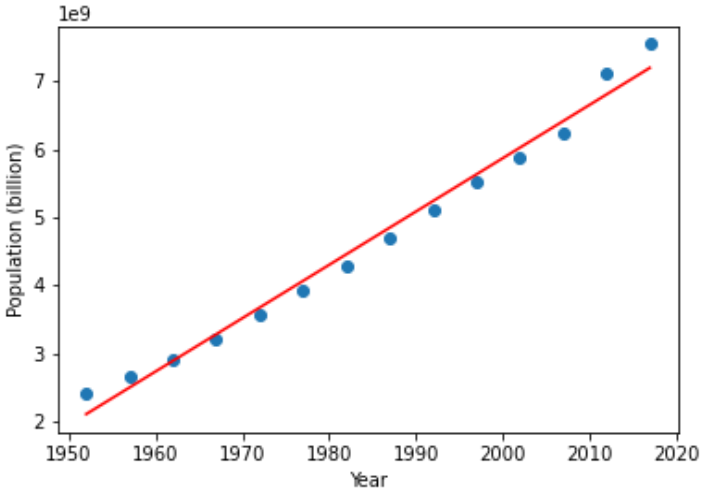
To begin the prediction project, we decided to research existing population projections and the methods used to achieve them. We encountered several machine learning models and algorithms that seemed adequate. However, most seemed rather complex and daunting compared to our current understanding of the field. Through our research, we found predictions that used logistic regression, logarithmic regression, multiple regression, and many others. We ultimately chose to start with a linear regression model as we were more familiar with this type of analysis, and it seemed appropriate for projecting the future population.

After this consensus was made, we focused on what kind of datasets we would use for the prediction. This would prove to be one of the more challenging parts of the assignment. We considered examining not only the world population but also growth rate over time, mortality rate, fertility rate, average age, and age over 65. We decided to try using the world population first because it made the most sense to us. Our first dataset was taken from Our World in Data, a scientific publication that focuses on compiling data regarding global problems such as hunger, climate change, and poverty. This dataset was appealing because it offered extensive population information dating back to 10,000 B.C. The only drawback was the way the data was formatted. Our initial goal was to find a set that offered the total world population and the progress of that information throughout the years. However, we could not find a website that offered this. Almost all the data sets we found listed population by country and not for the world as a whole. To solve this problem, we decided to add each country's population by year; this way, we could have a world population that reflected its respected year instead of having multiple countries with their individual populations.

We imported the data into JupyterLab and began organizing it using a pandas data frame. The first problem we noticed with the dataset was that the population numbers remained the same for many of the earlier years. We also realized that having population data from these earlier years was not relevant or valuable to our analysis. Due to this, we decided to remove all the years from the dataset before the year 1800. We then began

summing each country's population based on the year to get a world population for each year. After getting our world population calculations, we noticed our numbers were way too high. We were positive the method we used to calculate the world population for each year was correct, so we thought the problem was with the dataset. Due to this, we discarded the dataset and searched for something different. We found another dataset with similar formatting and recalculated our world population numbers. However, similar to the last dataset, our numbers were still too high. We began investigating the data and realized that many countries were duplicated. For example, many countries were listed as individual entries and members of their continents or encompassing regions (i.e., Australia being a part of Oceania). The lack of structure to the countries and the groups they belonged to made it difficult to extract the information we needed. The dataset did not list individual countries first, followed by groups. Instead, groups and countries were listed throughout the dataset in what seemed like a random order. Because of this, we decided to search for an easier-to-use dataset.

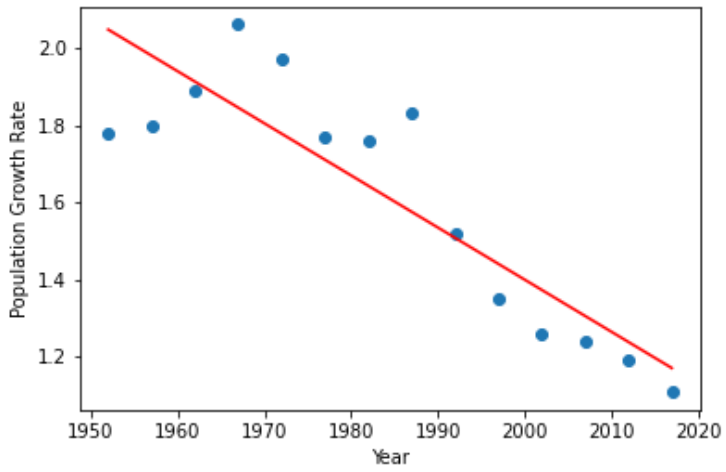
We found another dataset that contained population data between 1952 and 2007. While this data was also sorted by country, it didn't have entries corresponding to collections of countries, so we were able to use it. The dataset had population numbers in five-year intervals, so we thought it might contain insufficient data points. Because of this, we decided to add population data from a different source for 2012 and 2017. Now that we had a functional dataset, we could plot the data and create our first linear regression model.




We felt the regression line in this model accurately depicted the data. We then obtained the slope and intercept of our regression line and used them to calculate our projections. With this, we arrived at our first projection for the population in 2122, which was around 15.4 billion. Based on our research, we knew this number was far too high

but did not know precisely why it was so high. We soon realized that historical data for the world population would only show a steady increase over time and would not account for the declining growth rate. We concluded that it was impossible to predict the inevitable decline of population growth using linear regression with only raw population data. We felt that using another model may yield better results by more clearly reflecting the downward trend in growth rate over time, which led us to attempt to use a logarithmic model next.

This new model would ideally consider the growth rate over time and would therefore result in a prediction lower than a simple linear projection. Unfortunately, when we ran our data through the logarithmic model, we got a final prediction of 15.7 billion, which still seemed too high and contained invalid values. Next, we tried using a multiple linear regression model that took into account both population and growth rate over time. We assumed that since the growth rate has decreased steadily since 1970, we could include it in the regression to get a lower and more accurate prediction. Although this approach gave us a lower prediction like we expected, it was still well over 13 billion, which we still felt was an overestimate. Through this testing, we realized that making use of world population data, which has been trending sharply upwards for the last two centuries, would consistently result in predictions that were too high. As such, we decided to rely solely on population growth rate data for our model. While the growth rate has been decreasing for the past 15 years, the growth rate from 2012 to 2017 was still over 1 percent. This indicates that there will still be an increase in total population for many years until the growth rate dips below 0 percent. Because of this, we felt population growth rate would be our best option to predict the population for 2122. In order to obtain a prediction for the world population in 2122, we needed to estimate the growth rate for the years leading up to 2122. We opted to use a linear regression model to predict this, supplying it with population growth rate data from 1952 to 2017 in 5-year intervals.



From this linear regression line we calculated growth rate values for the years 2023 to 2122, some of which can be seen below.

| Year | Growth Rate |
|---|---|
| 2023.0 | 1.090588 |
| 2024.0 | 1.077116 |
| 2025.0 | 1.063643 |
| 2026.0 | 1.050171 |
| 2027.0 | 1.036698 |
| ... | ... |
| 2118.0 | -0.189302 |
| 2119.0 | -0.202775 |
| 2120.0 | -0.216247 |
| 2121.0 | -0.229720 |
| 2122.0 | -0.243192 |

With these growth rate values we were able to calculate total population predictions for every year from 2023 to 2122. Our prediction for 2122 ended up being around 11.8 billion, which we felt was a reasonable estimate. With this model our predicted populations increased until 2103 reaching a maximum of 12.1 billion. After 2103 our growth rate became negative and the population slowly decreased until 2122. While researching, we noticed that many of the professional projections we looked at predicted population growth would stop around 2100. Most of these projections also predicted the population to be around 11 billion in the year 2122. Although our final projection for 2122 was slightly higher than most others we looked at, we felt it was close enough to be valid and did a good job predicting when population growth would stop. The chart below shows our population predictions from 2023 to 2122.