Access to Drinking Water and Mortality rate

Ryan Díaz-Pérez

January 16, 2020

1. Introduction

1.1 Background

Clean drinking water is one of the most important resources for civilization. Increasing clean water access across the world is essential to reduce negative health outcomes and disease burden. Some regions have greater difficulty accessing clean drinking water, and yet these same regions tend to have the highest rates of population growth. With a growing population, there is a greater need for accessible water and better water management practices. One of the regions with a fast-growing population and a lack of drinking water is Sub-Saharan African. Sub-Saharan African also shows the highest number of deaths due to the lack of clean drinking water. If we can understand the relationship between population growth and the available water resources we can have a better understanding of the level of water scarcity that future generations will face.

1.2 Problem

The that we face is that we have an understanding of population growth, but we don't understand how this population growth may affect water resources. This project aims to find the correlation, if any between deaths due to unsafe water and population growth by analyzing different regions of the world. This will include the use of total population, population growth, access to improved drinking water, and deaths due to unsafe water data. The UN predicts that by the year 2050 Africa's population will grow to 2.5 billion, or 25.6% of the global population. This will mean that even if the continent is able to supply 80% of its residents with clean water, there will still be 0.5 billion people without clean water.

1.3 Interest

World governments would be interested in the impact of population growth and clean water resources. As well as the people in the regions that will be most heavily affected since clean water is an essential right and resource.

2. Data Acquisition and Cleaning

2.1 Data Sources

The data source for this project is found in https://ourworldindata.org, an online publishing site that focuses on major world problems with a research group based at the University of Oxford. Our World in Data provides data on the percentage of the population with access to improved drinking water over time, and share of deaths from unsafe water sources. I obtained the world population by region, and the population growth rate by region data from https://data.worldbank.org, an international financial institution that provides loans and grants to the governments of poorer countries for the purpose of pursuing capital projects. The share of the population with access to improved drinking water data is from the year 1990 to 2015 giving us 26 years of data. This is the smallest data set and

due to this constraint, 1990-2015 was defined as the time period of focus for the data exploration and analysis.

2.2 Data Cleaning

Data was downloaded from multiple sources and were individually cleaned. The first step to extract the data from the regions of interest; Sub-Saharan Africa, East Asia & Pacific, South Asia, Latin America & Caribbean, Middle East & North Africa, Europe & Central Asia, and North America. This allows us to separate the world into regions with similar economic status, cultures, and geographies. The next step was to select data from the year 1990 to 2015, due to limited data on the percentage of the population with access to improved drinking water. I repeated this process for the total population by region, and population growth by region data sets.

The share of population with access to drinking water and share deaths due to population data sets were setup differently, and required manipulating the data sets to have the same format as the data sets from https://ourworldindata.org. This involved selecting the data from each region individually and assigning a variable to it. A new database was created from these variables with the same format as the other data sets. There were no missing values for the 1990-2015 time period, which made data manipulation easier.

2.3 Feature Selection

After cleaning there were four data sets of the share of deaths due to unsafe water, share of the population with access to improved drinking water, population growth rate, and total population by region. An improved drinking water source includes piped water on premises (piped household water connection located inside the user's dwelling, plot or yard), and other improved drinking water sources (public taps or standpipes, tube wells or boreholes, protected dug wells, protected springs, and rainwater collection). Each data set contains 26 years of data for 7 regions of the world.

The next step is plotting the data to understand the features of each region. We can see from Figure 1 that Sub-Saharan Africa has the lowest access to improved drinking water. In the year 2015 67.54% of the population in Sub-Saharan Africa had access to improved drinking water, that is 24.83% less than the second lowest region in the world South Asia, and 3.66% less than the second lowest region in the year 1990, East Asia & Pacific. This large gap in water resources means that there is a large number of people that have to obtain there drinking water from unimproved water sources that do not have protection against contamination. Having access to improved water sources does not guaranteed that the water is safe to drink, but these sources are more likely to provide safe drinking water and to prevent contamination.

Figure 2 depicts the vast difference in deaths due to unsafe water. Two regions in particular stand out, Sub-Saharan Africa and South Asia. These two regions experience the largest rates of deaths by unsafe water sources. Although around 90% of South Asia's population has access to improved water sources their mortality rate remains high compare to the rest of the world. This could be cause by the water being contaminated and the improved water systems not being properly tested. Sub-Saharan Africa has the fastest growing population, as shown on Figure 3. This can make improving the water sources for its people much more difficult.

With this insight, we can determine that the region that is most negatively affected by the lack of drinking water sources is Sub-Saharan Africa. I will use this region as the main focus of my analysis, because if there is a correlation between mortality rate due to unsafe water and population growth rate than this region will be mostly affected. This analysis will require creating a database for the Sub-

Saharan Africa region. This can be done by using SQL but I decided to use Python since the analysis and model development will be done through the use of Python.

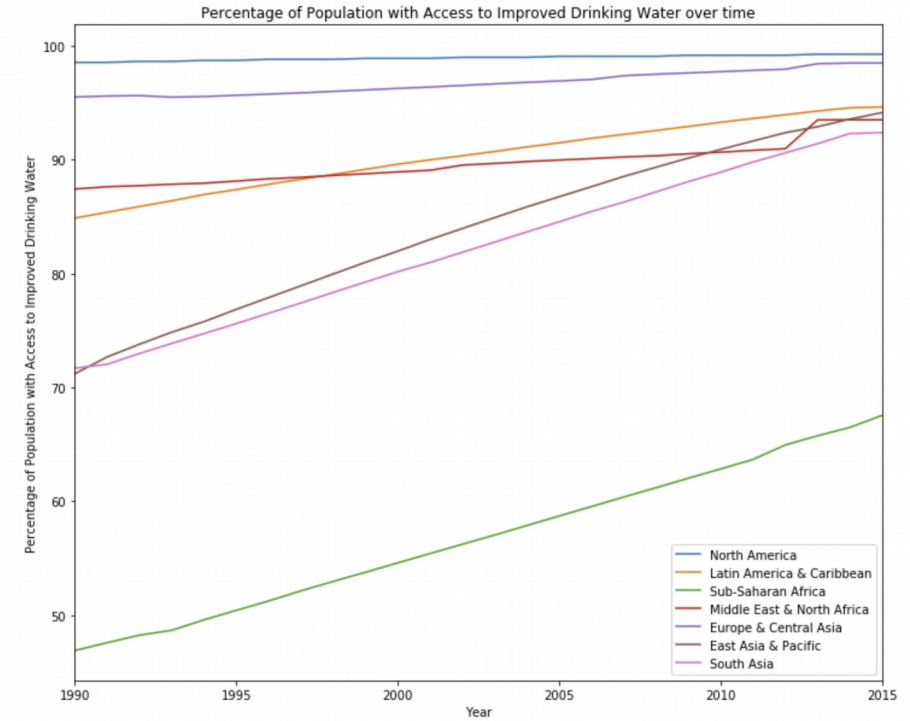

Figure 1. Sub-Saharan Africa has the largest number of people in need of drinking water, with only 67.54% of the population having access to improved drinking water.

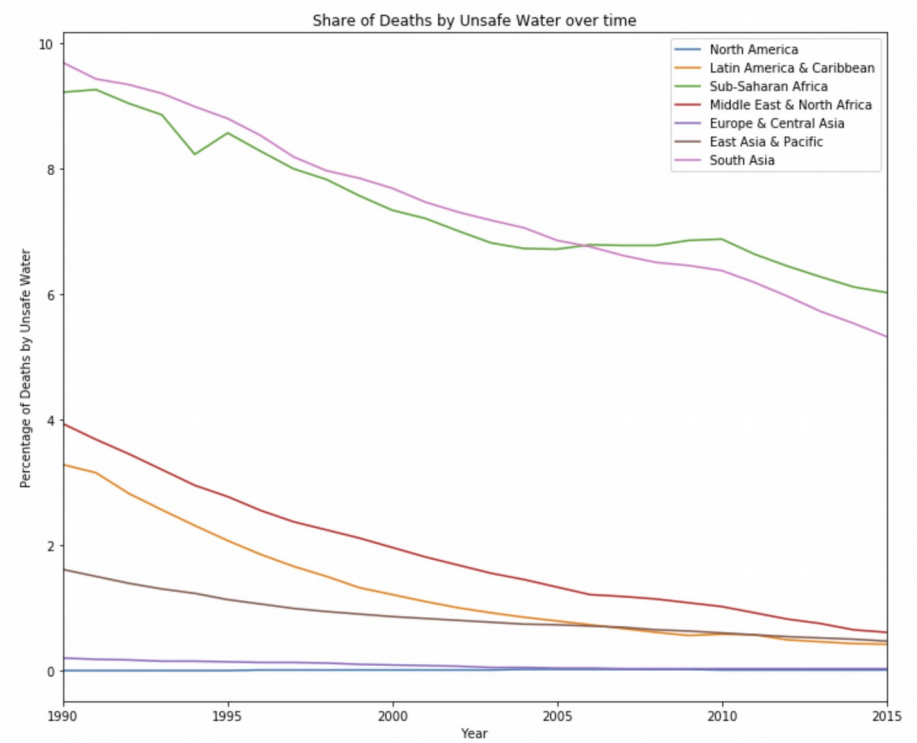

Figure 2. Percentage of deaths due to unsafe water from 1990 to 2015. Sub-Saharan Africa and South Asia have the largest rates of deaths by a factor of 10.

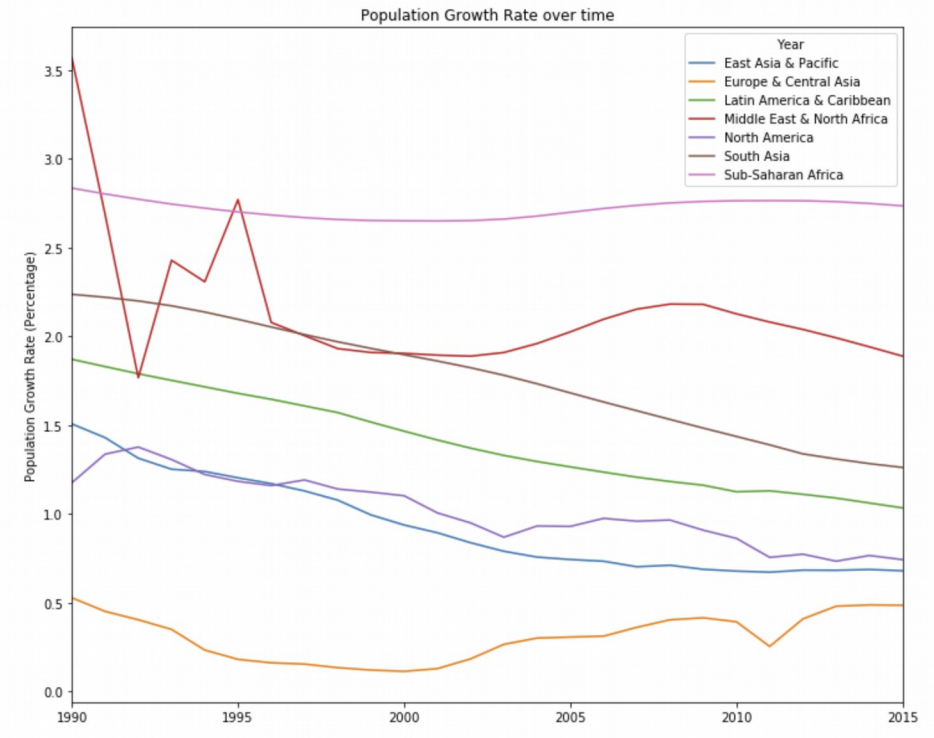

Figure 3. Population growth rate by region from 1990 to 2015. Sub-Saharan Africa has the largest population growth rate of 2.74% in 2015.

3. Methodology

3.1 Calculating Correlation Between Variables

It is unknown which parameter has the greatest correlation with the share deaths due to unsafe water. I determined the correlation between all of my parameters by using the Pandas method corr() to find the feature that is most correlated with the share deaths due to unsafe water. As shown in Figure 4 the parameter with the highest correlation with share deaths due to unsafe water is access to improved drinking water, with a correlation of -0.95. Population growth rate has the lowest correlation with the share deaths due to unsafe water. This means that the model that will be developed will be highly influenced by the access to improved drinking water in the region.

	Improved Drinking Water	Deaths by Unsafe Water	Total Population	Population Growth Rate
Improved Drinking Water	1.000000	-0.946741	0.997519	0.091015
Deaths by Unsafe Water	-0.946741	1.000000	-0.926161	0.182624
Total Population	0.997519	-0.926161	1.000000	0.139385
Population Growth Rate	0.091015	0.182624	0.139385	1.000000

Figure 4. Correlation between all of the parameters. The parameter with the highest correlation with share deaths due to unsafe water is access to improved drinking water, with a correlation of -0.95.

3.2 Residuals

I used residual plots to better understand what type of model will be best suited for this set of data. A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate. The residual shown on Figure 5 is not random, so a linear model will not appropriate to fit the share deaths due to unsafe water. Possibly a multi-regression model or a polynomial model would be best for this data set.

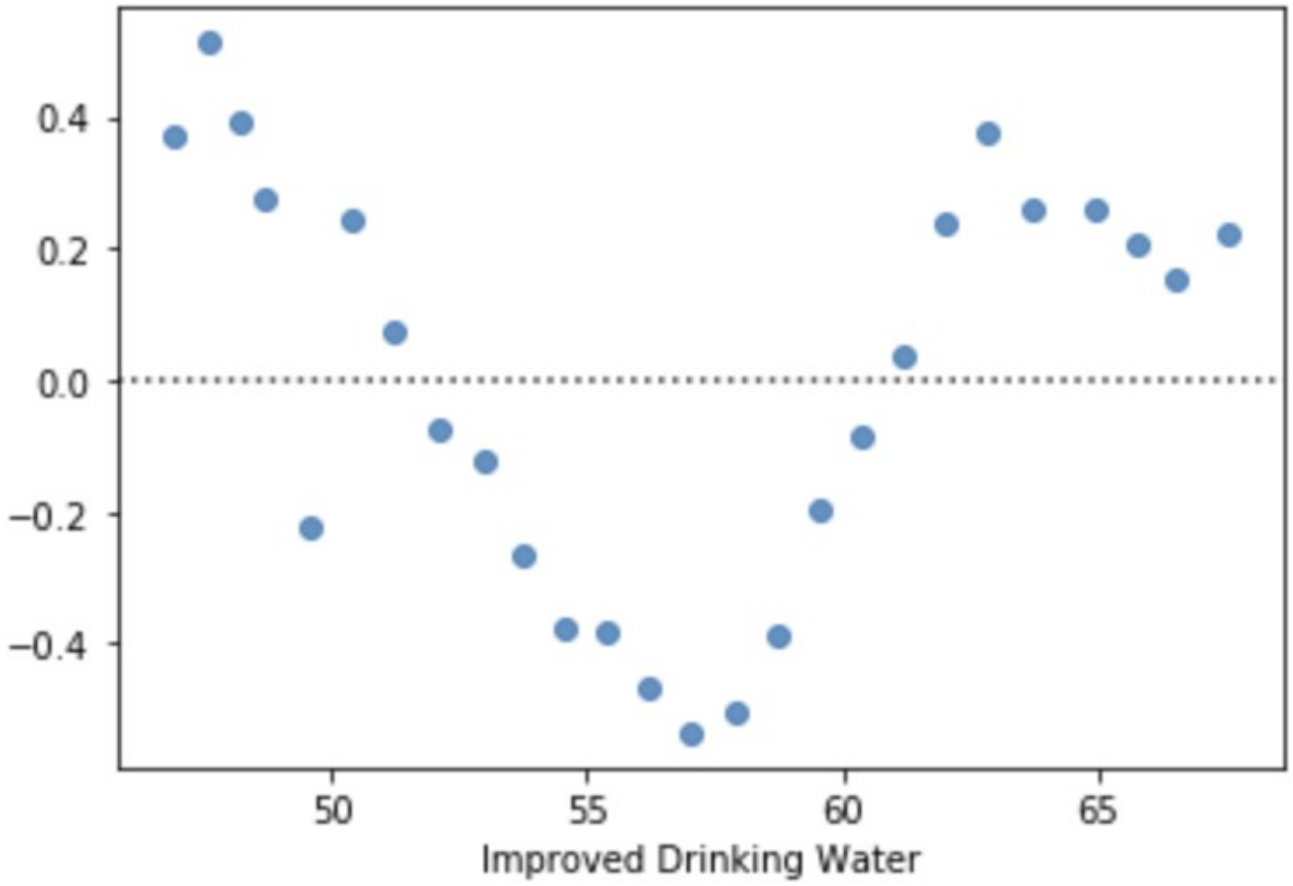

Figure 5. Residual between access to improved drinking water and share deaths due to unsafe water.

3.3 Model Development

There were four models that I used to fit the data, Linear Regression, Multi-Regression, Polynomial Fitting, and Non-linear Regression. Because the residual is not random one would expect that the polynomial model will find the best fit. I determine the performance of the models by calculating there R-squared value, and there mean squared error (MSE). The R-squared value also known as the coefficient of determination, is the proportion of the variance for a dependent variable that is explained by the variance of one or multiple independent variables. Put simply R-squared is the measure of how close the model and the data are to each other. The mean squared error is the average squared difference between the estimated values and the actual values. The best fit is indicated by a R-squared close to one and a mean squared error close to zero.

3.4 Linear Regression Model

Linear models are used mostly when the residual between two parameters are random. Although in work the residuals were not random, for the purpose of properly testing all regression models. In this model I used the access to improved water sources as the independent parameter, and the share deaths as our dependent parameter. This choice was made because this two parameters are highly correlated with each other. This model is expected to perform the worst as the residuals have a pattern. The calculated R-squared value is 0.89 and the mean squared error is 0.09.

3.4 Multi-linear Regression Model

Multi-linear regression models use multiple parameters as independent parameters to solve for our dependent parameter. By using multiple parameters you can find a more accurate fit because every independent parameter shapes the model depending in it's contribution. The calculated R-squared value is 0.97 and the mean squared error is 0.02.

3.5 Polynomial Regression Model

Because the residual plot had a pattern to it, a polynomial fit is a appropriate model for fitting. I used a 5^{th} order polynomial to fit the share deaths due to unsafe water with access to improved water because these two parameters had the greatest correlation. The calculated R-squared value is 0.98 and the mean squared error is 0.01.

3.6 Non-linear Regression Model

It is standards to use a non-linear model, for example a sigmoid function, to fit data with structured residuals. This type of model is best suited for data sets that don't have a polynomial or linear behavior like exponential functions. The calculated R-squared value is 0.23 and the mean squared error is 0.01.

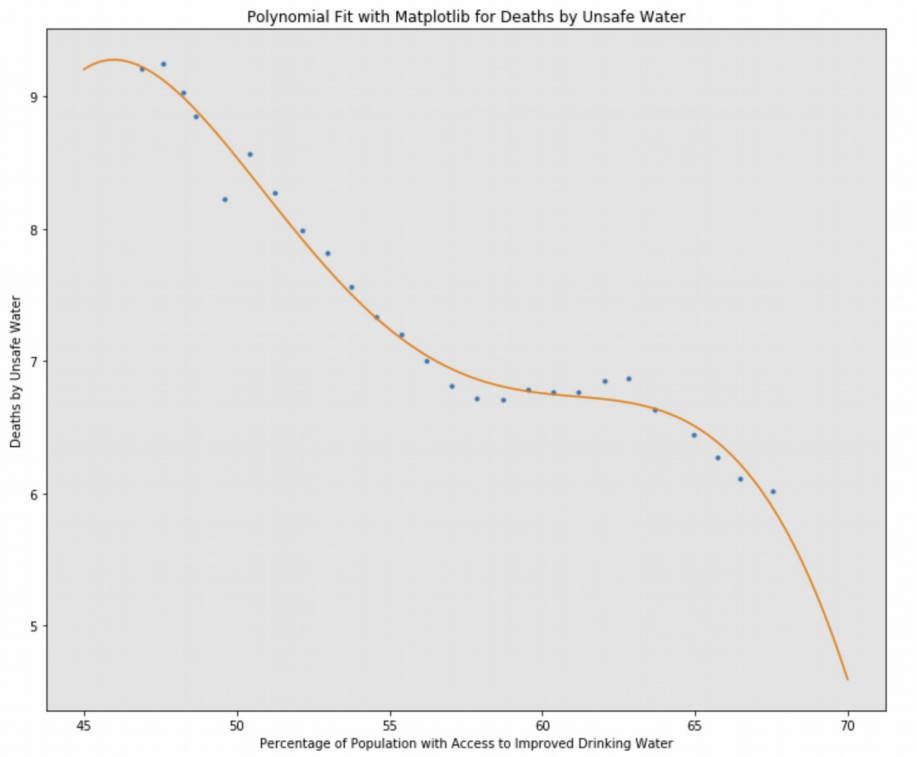

Figure 6. Polynomial fit of the percentage of deaths due to unsafe water sources. The resulting polynomial is a 5^{th} order polynomial.

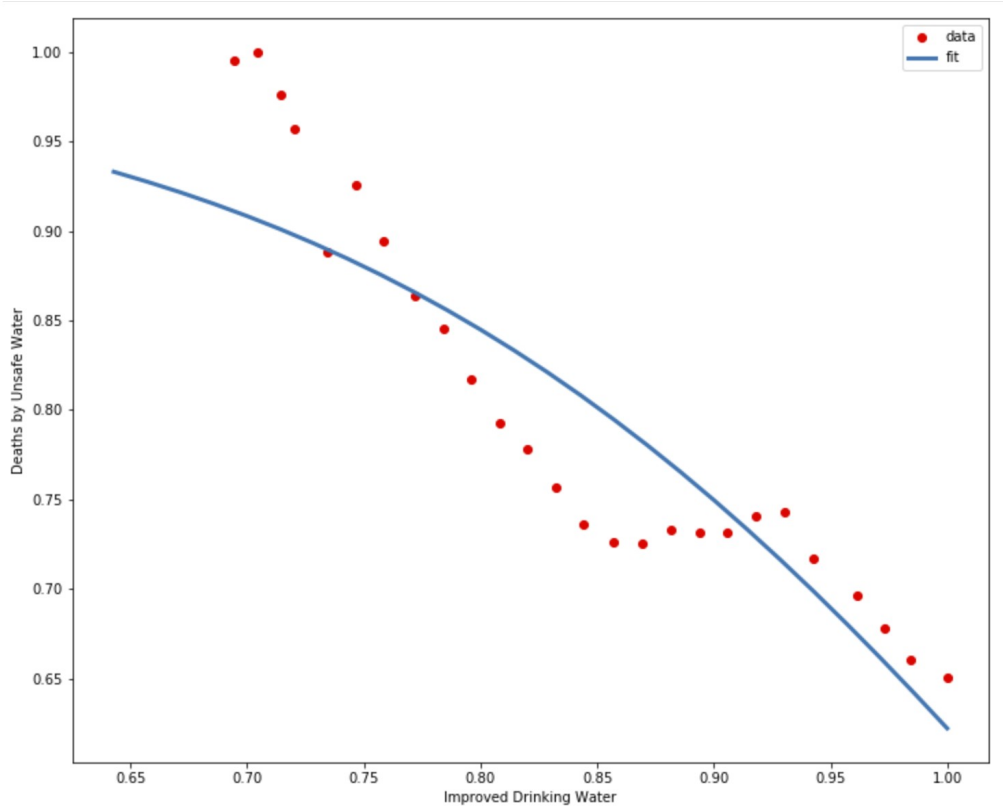

Figure 7. Non-linear regression model, with a R-squared of 0.23 and a MSE of 0.01.

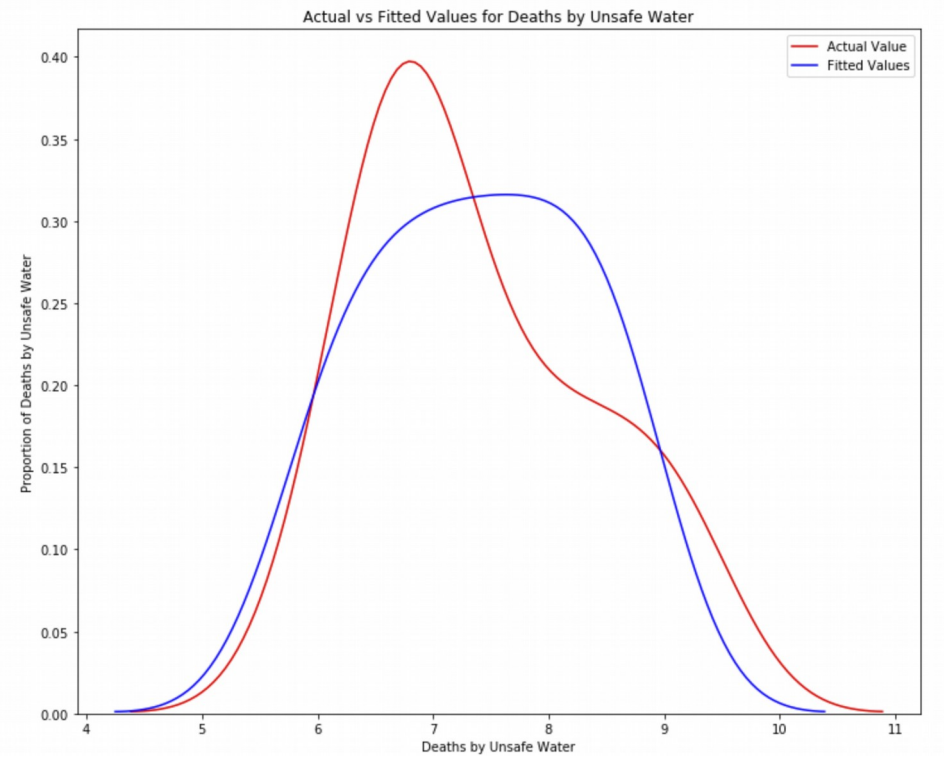

Figure 8. Linear regression model distribution with a R-squared of 0.89 and the mean squared error of 0.09.

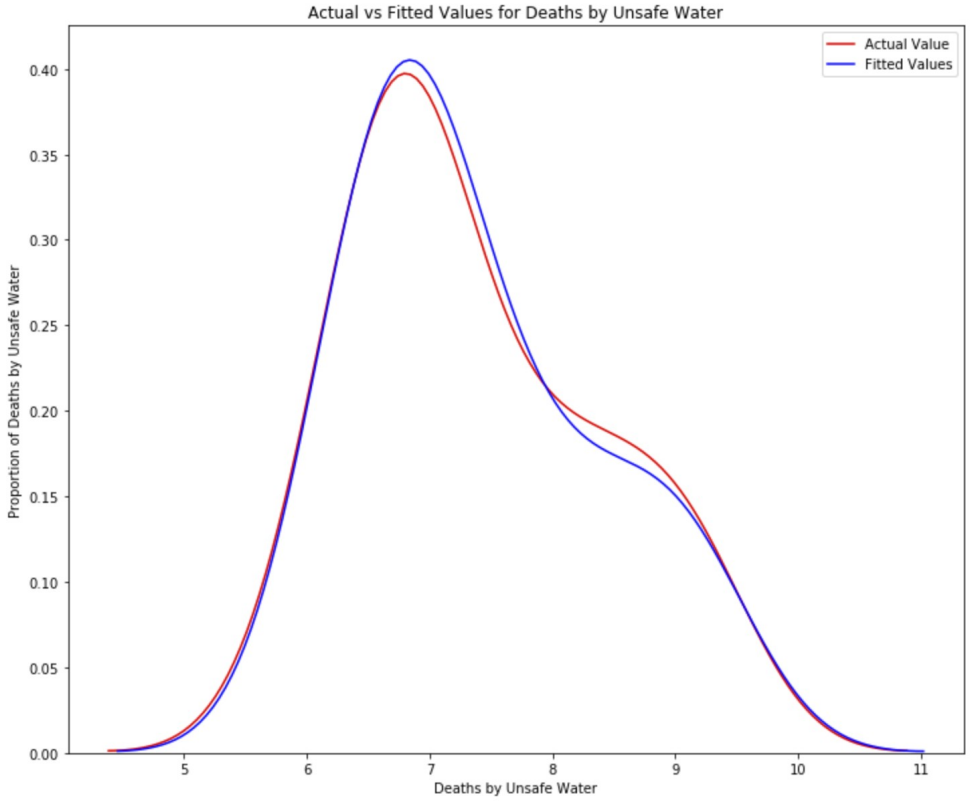

Figure 9. Multi-linear regression model distribution with a R-squared of 0.97 and the mean squared error of 0.02.

4. Results

The best performing model was the polynomial regression model, with a R-squared of 0.98 and a mean squared error of 0.01. As shown in Figure 6 this polynomial model matches the data best. The worst predictive model was the non-linear regression model with a R-squared of 0.23, as shown in Figure 7. The multi-regression model had a very good prediction, much better than the linear regression model. The distribution plots of the multi-regression model and the linear regression model are shown in figures 8 and 9 respectively.

5. Discussion

In this project I analyzed the relationship between the percentage of deaths due to unsafe water, access to improved water sources, population growth, and total population of Sub-Saharan Africa. Sub-Saharan Africa is the region with the greatest need of clean water sources and the region with the fastest growing population. The rate of population growth has a small effect on the share deaths by unsafe water sources, but the biggest contributor is the access to improved water sources. This is a factor that can be greatly improved by improving the financial resources in the region. I would recommend including GDP data to create a better model.

I would also use a larger time period, although the data on the percentage of people with access to improved water sources is limited, this can manage by taking the average rate of change for this parameter, and this should give us a good estimate to the real values. Another solution to this problem would be to create a predictive model of the share access to improved water sources over time, and use it to fill in the missing values.

Sub-Saharan Africa was the region of focus of this project due to the fact that it has the fastest population growth rate and it has the lowest access to improves water sources in the world. Another region of interest is South Asia because its population has similar access to improved water sources as the rest of the world, but the share deaths due to unsafe water sources is similar to Sub-Saharan Africa. It would be very interesting to analyze why this is happening and what contribution led to this.

6. Conclusion

Although there is a small correlation between the share deaths due to unsafe water and the population growth, the biggest impact comes from the percentage of people that have access to improved water sources. This relation can be better expressed and predicted by using a polynomial regression model. The importance of this is that everyone needs clean water, and the regions with the fastest growing population is also the region with the least resources to provide its people with clean water. In order to solve this problem one must find what parameters improve the access to improved water sources and the access to drinking water. Some of these parameters might be the GDP of the region, and the terrain of the population centers.