Using Linear Regression to Examine the Correlation Between Access to Drinking Water and Life Expectancy

Julia Andronowitz
B.S., Mathematics and Statistics

December 9, 2022

Abstract

The aim of this paper is to assess the relationship between a population's access to potable drinking water and their average expected longevity. Data from the World Health Organization is filtered in Python before linear regression is performed. The paper then goes on to assess the correlation that exists as well as the assumptions for the model. Finally, improvements for the model and possible implications for the data are discussed along with overarching conclusions for the public health sector.

University of Connecticut
Department of Statistics
Professor Jun Yan, Ph.D.

1 Introduction

Many of us in the United States often take the necessity of safe drinking water for granted, but there are parts of the country that still struggle with availability of clean water. In fact, "in 2015, nearly 21 million people relied on community water systems that violated health-based quality standards" (Allaire et al., 2018). Furthermore, it is still a pressing issue in many parts of the world. Current research shows that unsafe water is responsible for millions of deaths (Ritchie and Roser, 2021), and from the *World Health Organization*, "Contaminated water and poor sanitation are linked to transmission of diseases such as cholera, diarrhoea, dysentery, hepatitis A, typhoid and polio" (WHO, 2017). These diseases are largely preventable, and an increase in both access to and quality of safe drinking water can dramatically increase life expectancy (Angelakis et al., 2021).

Although there are numerous studies on how drinking contaminated water can have tremendous negative effects on one's health, there are very few studies that examine how the percentage of people with access to clean water in a certain area affects statistics regarding the community's overall health. In previous studies, the quality of drinking water and sanitation facilities lie on a continuum. In Wolf et al. (2014), the quality of drinking water was measured by water and sanitation facilities on a scale from poor to good. Conclusions were drawn based on the improvements to water systems, rather than the amount of people with access to potable water, stating that "...there are large potential reductions in diarrhoeal disease risk through improvements to both water and sanitation in low- and middle-income settings" (Wolf et al., 2014). Moreover, this study was based on the specific health implications regarding diarrhoeal disease risk.

In fact, studies regarding drinking water have been done to draw conclusions about the presence of certain diseases, like the study above, when clean drinking water is limited. Yet, there are many other external effects of maintaining access to clean drinking water. From of Disease WaSH Collaborators (2020), "Low access to safe water and sanitation has also been linked to broader social outcomes such as reductions in school attendance (particularly for girls who are menstruating), losses to economic productivity, and undue burden on women of time spent collecting water". Additionally, the distance the source of the water is from one's home is extremely important. From the work done in Cassivi et al. (2018), "Across all countries in this study, results show that up to 40% of the national population needs more than 30 minutes to fetch water irrespective of the type of water source used by the household...as reducing the collection burden required to fetch water is essential to enhancing water quantity available for households and improving general health and quality of life".

Besides the poor health effects, there are various other negative effects that are present alongside low access to potable water. Chronic stress in one's life whether related to nutrition, health, financial well-being, or other factors is linked to a reduction in life expectancy. This study aims to quantify how access to drinking water (and the negative mental and physical health effects associated with it) affects a population's overall life expectancy at birth. Such a study could help emphasize the need for accessible clean water worldwide, as it has numerous health benefits beyond longevity (Popkin et al., 2010).

In the following sections, two data sets will be used to analyze how a population's access to clean drinking water can affect the average life expectancy. First, we will outline the background of the data. Then, we will go on to describe how to filter the data to what we are looking for and how linear regression is applied once the data is cleaned. Following this, we will discuss how the model was coded in Python and how various graphs were created. A summary of the model including the correlation we found and associated generated figures will be described. Finally, limitations of the study as well as future directions will be discussed.

2 Data

The two data sets we will be using are from the World Health Organization and provided by Kaggle under the name "World Health Statistics 2020—Complete—Geo-Analysis" (Kaggle, 2022). The first is called Basic Drinking Water Services and contains information regarding basic drinking water services across various countries and in different years. From WHO, basic drinking water services is defined as "drinking water from an improved source, provided collection time is not more than 30 minutes for a round trip. Improved water sources include piped water, boreholes or tubewells, protected dug wells, protected springs, and packaged or delivered water" (Bank, 2022). There are 3455 rows with 4 columns: Location, Period, Indicator, and Tooltip. The indicator column is a text cell that says "Population using at least basic drinking-water services", and the Tooltip column gives the percentage of the population where the indicator is true. In this data set, we will be looking at the country, year, and percentage of population using at least basic drinking water services.

The second data set is called Life Expectancy at Birth and contains information about life expectancy in years for various countries at the time of birth. It has 2197 rows and 5 columns: Location, Period, Indicator, Dim1, and Tooltip. In this data set, the indicator is the life expectancy at birth in years. The Dim1 column indicates which sex the life expectancy is

referring to - male, female, or both sexes. For this data set, we will have to filter the rows to just contain those with a life expectancy for both sexes, as the other data set is not filtered by gender. We will be using the country and year columns to look at life expectancy.

3 Methods

For the regression model, we start by loading the libraries that we will need. These include the Numpy, Pandas, and Bokeh libraries.

```
import numpy as np
import pandas as pd
import seaborn as sns
from bokeh.plotting import figure
from bokeh.io import show, output_notebook
from bokeh.models import ColumnDataSource, HoverTool
from bokeh.layouts import gridplot
output_notebook()
```

Next, we import the data:

```
rawData1=pd.read_csv('archive/basicDrinkingWaterServices.csv')
rawData2=pd.read_csv('archive/lifeExpectancyAtBirth.csv')
```

If we would like to see the data, use the command

```
rawData1
```

An output of the data is in Figures 1 and 2. For the drinking water data, we take only the columns with the country, year, and the percentage of the population using at least basic drinking water services.

```
rawData1=rawData1[["Location", "Period" , "First Tooltip"]]
```

We do the same for the life expectancy data, only taking the columns with the country, year, sex, and average life expectancy at birth.

```
rawData2=rawData2[["Location", "Period", "Dim1", "First Tooltip"]]
```

```
In [3]: 1 rawData1
```

Out[3]:

	Location	Period	Indicator	First Tooltip
0	Afghanistan	2017	Population using at least basic drinking-water...	57.32
1	Afghanistan	2016	Population using at least basic drinking-water...	54.84
2	Afghanistan	2015	Population using at least basic drinking-water...	52.39
3	Afghanistan	2014	Population using at least basic drinking-water...	49.96
4	Afghanistan	2013	Population using at least basic drinking-water...	47.56
...	...	...	...	...
3450	Zimbabwe	2004	Population using at least basic drinking-water...	57.94
3451	Zimbabwe	2003	Population using at least basic drinking-water...	58.59
3452	Zimbabwe	2002	Population using at least basic drinking-water...	59.23
3453	Zimbabwe	2001	Population using at least basic drinking-water...	59.88
3454	Zimbabwe	2000	Population using at least basic drinking-water...	59.88

3455 rows × 4 columns

Figure 1: Access to Basic Drinking Water Services Data

```
In [4]: 1 rawData2
```

Out[4]:

	Location	Period	Indicator	Dim1	First Tooltip
0	Afghanistan	2019	Life expectancy at birth (years)	Both sexes	63.21
1	Afghanistan	2019	Life expectancy at birth (years)	Male	63.29
2	Afghanistan	2019	Life expectancy at birth (years)	Female	63.16
3	Afghanistan	2015	Life expectancy at birth (years)	Both sexes	61.65
4	Afghanistan	2015	Life expectancy at birth (years)	Male	61.04
...	...	...	...	...	...
2192	Zimbabwe	2010	Life expectancy at birth (years)	Male	49.58
2193	Zimbabwe	2010	Life expectancy at birth (years)	Female	53.21
2194	Zimbabwe	2000	Life expectancy at birth (years)	Both sexes	46.57
2195	Zimbabwe	2000	Life expectancy at birth (years)	Male	45.15
2196	Zimbabwe	2000	Life expectancy at birth (years)	Female	48.12

2197 rows × 5 columns

Figure 2: Life Expectancy at Birth Data

Next, we want to convert our pandas dataframes into numpy arrays so that we can use the appropriate mathematical functions.

```
data1=np.array(rawData1)
data2=np.array(rawData2)
```

We want to have as many constants as possible in our model. Since the data was collected starting in 2000, we must use the same years in both data sets. We will use the year 2015 as the benchmark. The data can then be filtered using a loop and appending the 2015 data to a new list.

```
dw2015=[]
for x in data1:
    if x[1]==2015:
        dw2015.append(x)
```

This must be done for both data sets. We now do it for the life expectancy data set. As previously mentioned in the "Data" section, we also want to use only the rows with both sexes.

```
le2015=[]
for x in data2:
    if (x[1]==2015) and (x[2]=='Both sexes'):
        le2015.append(x)
```

These lists must also be converted into numpy arrays.

```
DW2015=np.array(dw2015)
lifeExpectancy2015=np.array(le2015)
```

Now, we see the shapes of both these arrays are different. The drinking water array contains 193 rows and 3 columns while the life expectancy array contains 183 rows and 4 columns. This means that there are some countries that are not in both lists. So, we need to delete the data for countries that are only in one or the other. We do this by checking if the countries in the drinking water array are also in the life expectancy data. If both countries are in both data sets, we append them to our new drinking water list, which is what we will use going forward.

```
drinkingWater2015=[]
for x in DW2015:
```

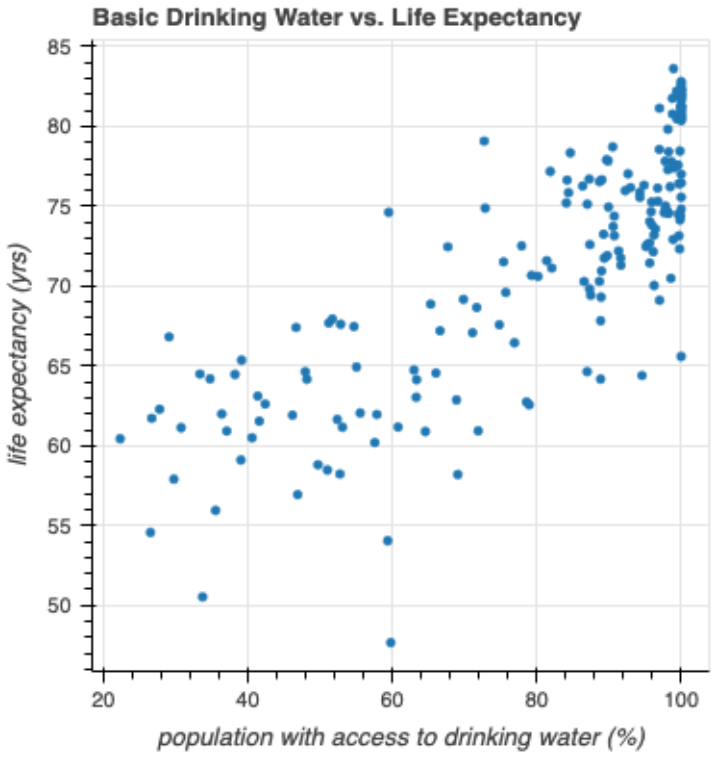

Figure 3: Life Expectancy vs. Drinking Water

```
    if x[0] in lifeExpectancy2015:
        drinkingWater2015.append(x)
drinkingWater2015=np.array(drinkingWater2015)
```

Now, we can see the lists are the same length. If the lengths were still different, we could do the same process with the life expectancy list, i.e. making sure every country in the life expectancy list is also in the drinking water list and removing elements that are not. Next, we want to assign x to the percentages of people with access to basic drinking water services and y to the life expectancies in 2015.

```
x2015=drinkingWater2015[:,2]
x2015=x2015.astype('float64')
y2015=lifeExpectancy2015[:,3]
y2015=y2015.astype('float64')
```

Now that we have standardized data across the board, we can plot the life expectancy in years and the percentage of the population with access to basic drinking water services. The output is shown in Figure 3.

Now we create our data matrix x, which is the x array transformed into a column vector and appended to a column of ones.

```
x2015=x2015.reshape(-1,1)
X2015=np.concatenate([x2015,np.ones(shape=
    (x2015.shape[0],1))],axis=1)
```

Since this is now a column vector, we also want to transform y into a column vector.

```
Y2015=y2015.reshape(-1,1)
```

Now, we can compute the least squares line and the matrix M that will give us the slope and the y-intercept for the line of best fit.

```
D2015=X2015.transpose()@X2015
np.linalg.inv(D2015)
M2015=(np.linalg.inv(D2015)@X2015.transpose())@Y2015
```

The output is the array $[[0.27859093], [49.15736534]]$, which tells us the slope of the line is 0.279 and the intercept is 49.157. If we want to compare the actual values to predicted values, we use the following code:

```
Yhat2015 = np.dot(np.dot(np.dot(X2015,np.linalg.inv(D2015)),
                                X2015.transpose()),Y2015)
```

A visual representation of the least squares line can be seen once executing

```
f2015.line(x=x2015.ravel(),y=Yhat2015[:,0],color='black')
show(f2015)
```

The output is shown in Figure 4.

Assumptions

As seen in Figure 3, we can assume a linear regression is appropriate for the model based on the data. There looks to be a linear correlation between the variables rather than a quadratic or other non-linear relationship. To see the R-Squared value of our linear regression, we can use the following code (Python, 2022):

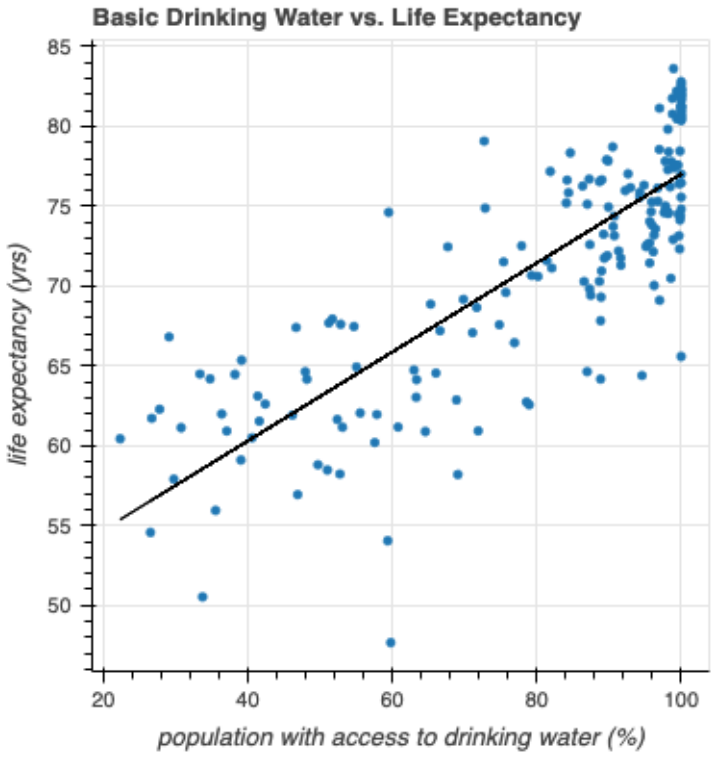

Figure 4: Plot with Least Squares Line

```
corr_matrix = np.corrcoef(Y2015[:,0], Yhat2015[:,0])
corr = corr_matrix[0,1]
R_sq = corr**2
```

We see the R^2 value is 0.66348. So, 66.348% of the variability in the data is explained by the model. While this is on the lower end, our model is still supported. We would also like to check that the mean of the residuals is close to zero. This can be calculated with the following code.

```
residuals=Yhat2015[:,0]-Y2015[:,0]
np.mean(residuals)
```

We see the mean is -1.2813065726821479e-14 which is very close to zero. Next, we need to check for homosteadasticity among the residuals. To look at the residuals, the difference between the measured and predicted values, we can use the following function as detailed in Teitlebaum (2021).

```
def comparison_plot(x,Y, Yhat):
    '''Plots Predicted vs True values for analysis of regression'''
```

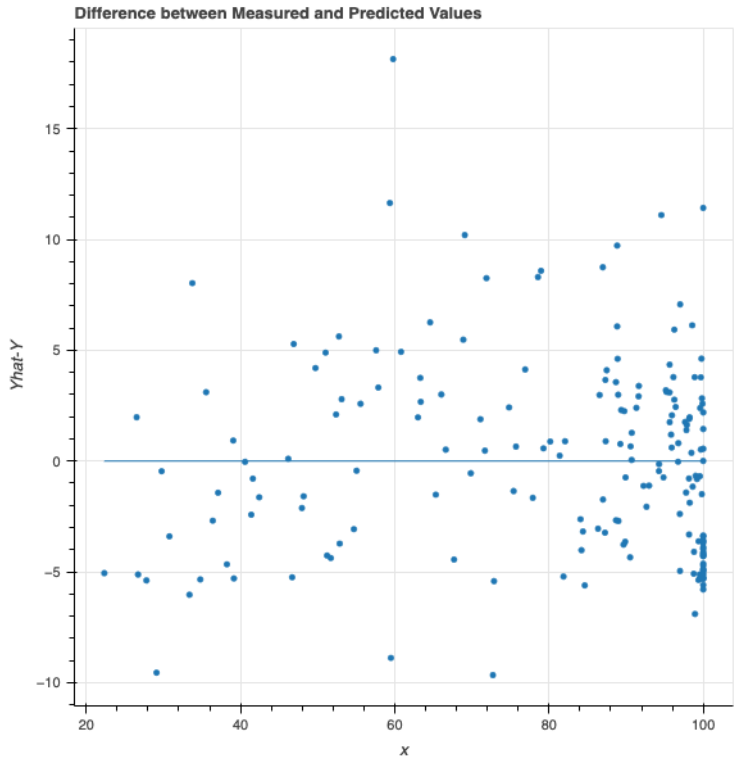

Figure 5: Residuals Plot

```
    comparison_plot =
        figure(title='Difference between Measured and Predicted Values')
    comparison_plot.xaxis.axis_label='x'
    comparison_plot.yaxis.axis_label='Yhat-Y'
    comparison_plot.scatter(x=x,y=Yhat-Y)
    comparison_plot.line(x=[x.min(),x.max()],y=[0,0])
    return comparison_plot
show(comparison_plot(x2015.ravel(),Y2015[:,0], Yhat2015[:,0]))
```

In Figure 5, we see there are no patterns among the residuals, so this assumption is validated.

Next, we would like to check for normality among the residuals. This can be achieved by plotting the distribution (Iyyer, 2020).

```
p = sns.displot(residuals,kde=True)
```

We see the graph in Figure 6 is slightly bimodal but almost normally distributed as it is very hard to get perfectly normally distributed curves with real life.

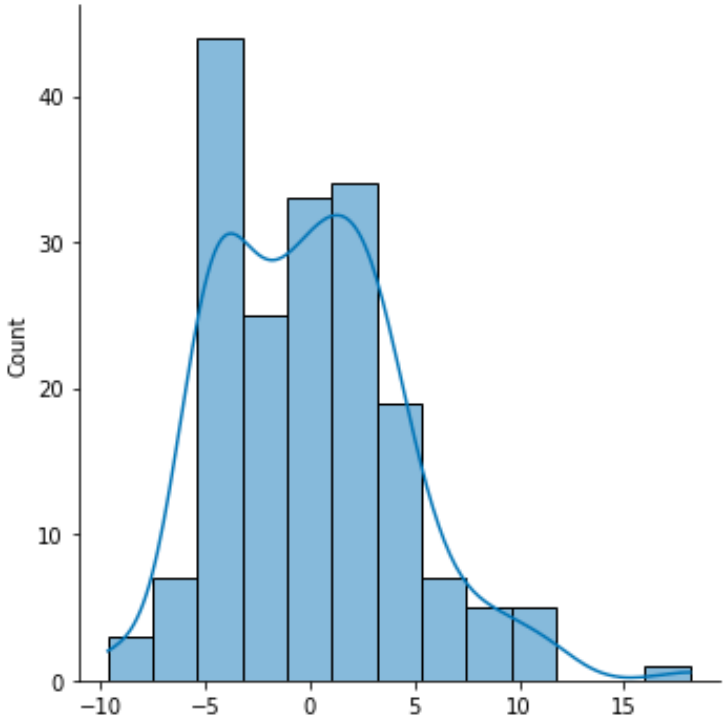

Figure 6: Distribution Plot

Finally, we look at possible outliers. The following code will calculate the studentized residuals and print values that are outside the normal range of (-3,3) (Zach, 2020).

```
df = pd.DataFrame({'x': x2015[:,0],
                   'y': Y2015[:,0]})
model = ols('x ~ y', data=df).fit()
stud_res = model.outlier_test()

for x in stud_res['student_resid']:
    if x > 3 or x < -3:
        print(x)
```

Running the code, we see there is one outlier with a studentized residual of -3.160. This outlier should be removed from the data set.

4 Results

Most of the assumptions for linear regression were validated. The R^2 wasn't extremely high, but it still supported a positive correlation. Additionally, the mean of the residuals was extremely close to zero. Homeosteadasticity among the residuals was validated through a plot of the residuals. Since there were no patterns, this assumption is valid. Next, we checked the normality assumption. We observed a slightly bimodal distribution. Finally, our test for outliers resulted in one observation whose studentized residual was above 3. This outlier should be removed from the data set.

From Figure 3, we see a general positive correlation between the two variables. This relationship illustrates that as the percentage of the population's access to safe drinking water increases, the average life expectancy also increases. Based on prior research, like that in Angelakis et al. (2021) that suggests access to and quality of drinking water can increase an individual's life expectancy, this result was expected. Later on, we discovered the slope of the best fit line to be 0.279, and the R^2 value is 0.6638. This further supports our prediction that there would be a positive relationship between access to drinking water and life expectancy.

From Figure 4 containing the resulting best fit line, we see that some points are very close to the line while others are very far. Checking the plot of the residuals in Figure 5, we see there is a somewhat large variation between the predicted values and the measured values. Some of the values are extremely close to zero, meaning that there was little difference between the predicted value and the actual value. However, some of the values are closer to 10-15 away from zero, indicating a large difference between the predicted and measured values. This confirms the observation made from Figure 4.

5 Discussion

Once the data was adjusted for correlating countries in the same year, the scatterplot suggested a positive linear relationship existed. After the line of best fit was computed, this was confirmed. The slope of the least squares line is 0.279 with an intercept of 49.157. Theoretically, if 0% of the population had access to clean drinking water, we would expect the average life expectancy at birth to be 49.157. Every one-unit increase in the percentage of access to drinking water results in a 0.279 year increase on average in the average life expectancy

of the population. These results highlight the importance for access to basic drinking water services, as we saw a positive correlation between the percentage of the population with access to these basic services and life expectancy at birth.

6 Limitations

This study was only done for the year 2015. While it was necessary to keep the year constant, it would be beneficial to research other years. It would be interesting to see if the correlation stayed somewhat constant or varied among the years. Performing the regression on multiple years would also help validate our results. Additionally, we saw some variation in the residuals, meaning that the model predicted a different outcome than the data points observed. The level of variation is a discrepancy between the predicted life expectancy of the model and the actual observed life expectancy from the data. This most likely has to do with the varying factors among different countries including other health factors such as air quality, sanitation, living conditions, and political stability. More conclusive research would have to be done to verify this. Possible methods such as multivariate regression and an analysis of covariance could be done. It is known that a person's health is affected by many factors, and a complete study to analyze the effect of each factor would be extensive. Though a study of this scale is improbable for this paper, such findings could revolutionize what public health organizations, governments, and citizens prioritize in relation to their health. This type of study has far-reaching effects and could not only improve a person's life expectancy but also their quality of life.

7 Conclusion

Studies regarding drinking water have been done to draw conclusions about the presence of certain diseases when clean drinking water is limited. Research revolves around specific diseases or access to clean drinking water based on a scale. In some cases, the distance the water source is from the community is not measured. This research paper focuses on investigating the effect of the proportion of a population's access to basic drinking water services on their life expectancy. This research filters the kind of access to clean water by requiring availability to be within 30 minutes round trip as well as provides a more widespread conclusion than those regarding specific water-borne illnesses. We have found

a positive correlation between the population's access to basic drinking water servies and life expectancy. This supports existing research that clean water is essential for the body to function, yet expands the relationship to overall life expectancy which is impacted by numerous factors within both physical and mental health. Insights in this field can further promote the basic right to safe water within a reasonable distance from one's home and highlight just how important water is in all aspects of human health.

References

Allaire, M., H. Wu, and U. Lall (2018). National trends in drinking water quality violations. *Proceedings of the National Academy of Sciences of the United States of America 115*(9), 2078–2083.

Angelakis, A., H. S. Vuorinen, C. Nikolaidis, P. S. Juuti, T. S. Katko, R. P. Juuti, J. Zhang, and G. Samonis (2021). Water quality and life expectancy: Parallel courses in time. *Water 13*(752).

Bank, T. W. (2022). Metadata glossary.

Cassivi, A., R. Johnston, E. O. D. Waygood, and C. C. Dorea (2018). Access to drinking water: time matters. *Journal of Water and Health 16*(4), 661–666.

Iyyer, S. (2020). Step by step assumptions - linear regression.

Kaggle (2022). World health statistics 2020—complete—geo-analysis.

of Disease WaSH Collaborators, L. B. (2020). Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000–17. *The Lancet Global Health 8*(9), E1162–E1185.

Popkin, B. M., K. E. D'Anci, and I. H. Rosenberg (2010). Water, hydration, and health. *Nutrition Reviews 68*(8), 439–458.

Python, A. (2022). Coefficient of determination – r squared value in python.

Ritchie, H. and M. Roser (2021). Clean water and sanitation. *Our World in Data*.

Teitlebaum, J. (2021). Lab for linear regression.

WHO (2017). Progress on drinking-water, sanitation and hygiene: 2017 update and sdg baselines. *World Health Organization*.

Wolf, J., A. Prüss-Ustün, O. Cumming, J. Bartram, S. Bonjour, S. Cairncross, T. Clasen, J. M. C. Jr, V. Curtis, J. D. France, L. Fewtrell, M. C. Freeman, B. Gordon, P. R. Hunter, A. Jeandron, R. B. Johnston, D. Mäusezahl, C. Mathers, M. Neira, and J. P. T. Higgins (2014). Systematic review: Assessing the impact of drinking water and sanitation on diarrhoeal disease in low- and middle-income settings: systematic review and meta-regression. *Tropical Medicine and International Health 19*(8), 928–942.

Zach (2020). How to calculate studentized residuals in python.