We developed a multiple linear regression model to infer properties about how a handful of socioeconomic and cultural indicators impact suicide rates. An important distinction is that the model is intended to be used for inferential, rather than predictive purposes. Our objective is to discover relationships between variables to inform relevant public policy and future research in the area.

The first step in developing the model was to transform the outcome variable. This transformation (Box-Cox) was used to make the outcome variable ‘more normal’, and it helped to characterize relationships between variables in our data.

The next step was to remove outliers. We used regression diagnostics and visual data exploration to identify unusual data points. Brief qualitative research was then conducted on the country represented by each point to confirm whether or not the point should be removed.

After the removal of the outlier points, we utilized a stepwise algorithm to identify which variables should be included in our model. This ‘automatic’ procedure yielded the set of variables that we would analyze more closely.

The following variables were selected and included in our model: labor force participation rate (female-male ratio), GDP per capita (PPP), liters of alcohol consumption per capita, and the prevalence of a national suicide prevention strategy. The algorithm excluded the following variables from the model: current health expenditure as a percentage of GDP, the number of psychiatrists working in the mental health sector (per 100k pop.), and the number of mental hospitals (per 100k pop.).

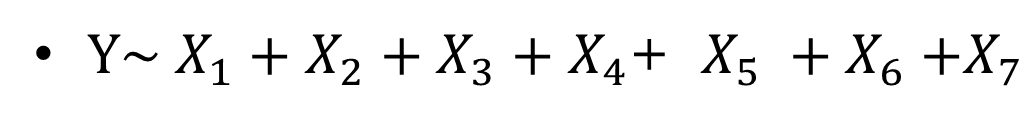
The final step in the development of our model was to implement an iterative algorithm that adjusted the weights for each of our data points (Iteratively Reweighted Least Squares). This allowed us to further limit the influence of outliers on our data.

In the development of our model, we relied on a few assumptions about the quality of our data. The first is regarding GDP per capita, which is assumed to be an appropriate indicator to reflect the wealth of a country. The second relates to the prevalence of a national suicide prevention strategy. It is assumed that the presence of such a strategy is indicative that the country has taken the time to develop a comprehensive and data driven approach to suicide, based on solid evidence. We also assume that the liters of alcohol consumed per capita reflects the tendency for individuals in the given country to consume excessive amounts of alcohol.

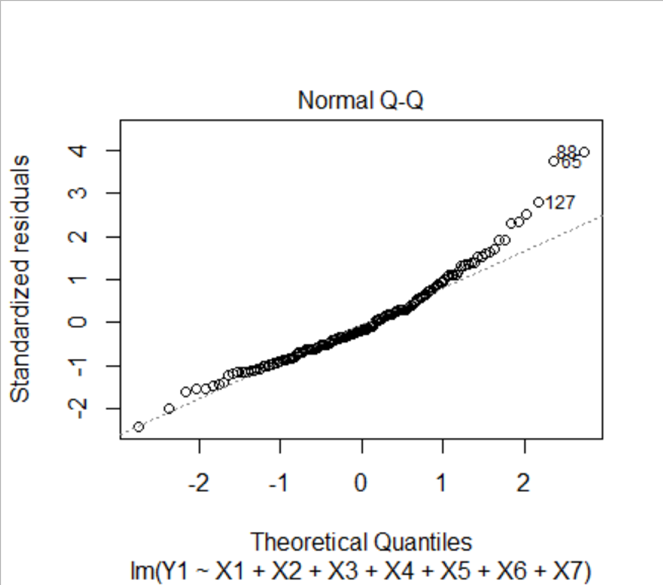
BELOW SHOULD BE IN THE APPENDIX

We developed a multiple linear regression model to infer properties about how a handful of socioeconomic and cultural indicators impact suicide rates.

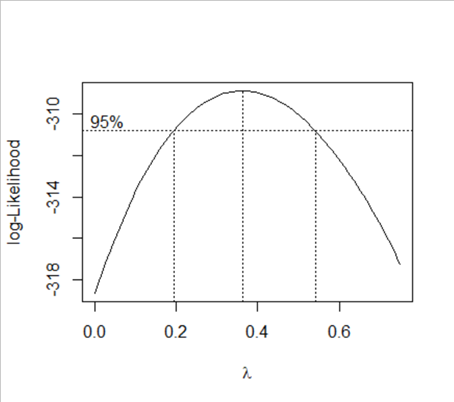
Initial Model



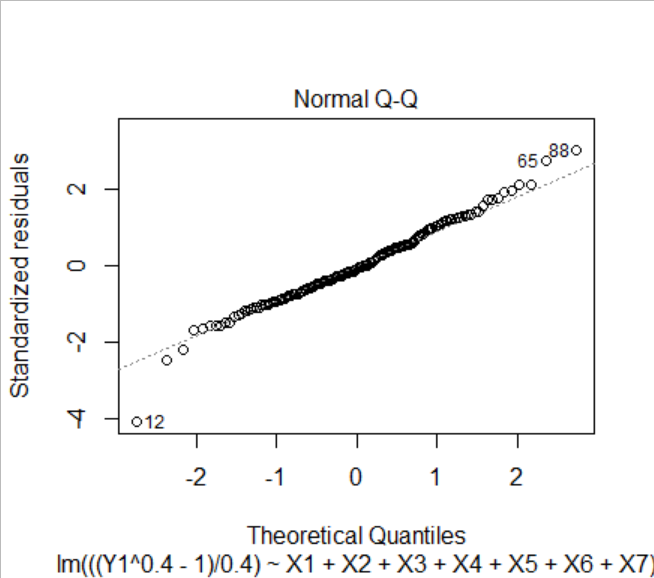
First, we analyzed the diagnostic plots to understand if our model appropriately fits the data that we have. A major red flag here was the Normal Q-Q plot, which shows if residuals are normally distributed.



The residuals deviate from the reference line at the higher quintiles. In order to correct for this, our next step was to try a Box Cox transformation on Y. Below is the log-likelihood plot to determine the lambda value for the Box Cox transformation.



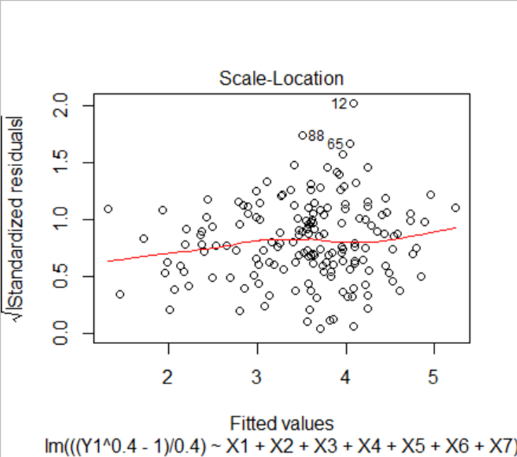
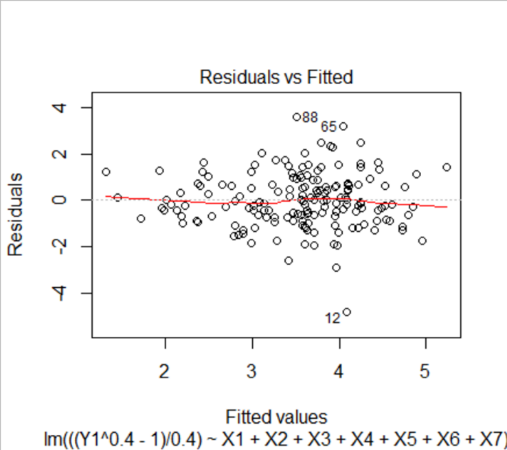
A lambda value of 0.4 was chosen.



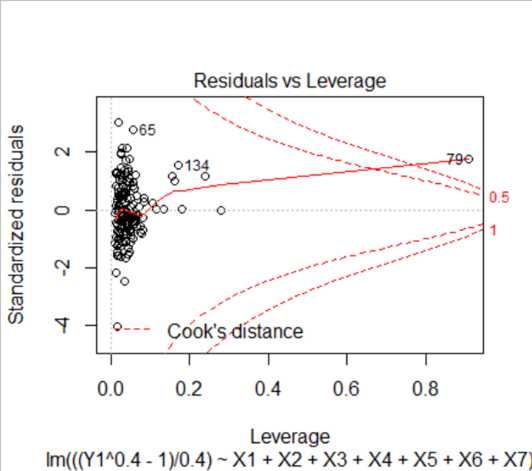
As a result of the Box Cox transformation, the points in the Normal Q-Q plot are much closer to the reference line.

The next step in model development was to remove outliers.

Points 12, 65, and 88 were identified as outliers on the Residuals vs Fitted, Scale-Location, and Normal Q-Q plots.



In addition, point 79 was identified as an outlier that should be removed, as it had very high leverage in the model.



Points 12, 65, 79, and 88 were removed from the dataset. It is important to note that each of these points also had country specific reasons for being excluded:

* 12 – Barbados:
  + Caribbean's leading tourism island, transitioned from agricultural to service based economy very successfully
  + “very high human development‟ status in terms of the UNDP’s human development index
  + Extremely low suicide rate
* 65 – Guyana:
  + Extremely poor island country largely made up of agricultural villages.
  + Very high alcohol and suicide statistics
  + Ministry of health identified poverty, pervasive stigma about mental illness, access to lethal chemicals, alcohol misuse, interpersonal violence, family dysfunction and insufficient mental health resources as key factors causing one of the highest suicide rates in the world.
* 79 – Japan:
  + Notoriously overworked and over stressed population, although the country is very wealthy
  + Long cultural history of considering certain types of suicides honorable, relatively high cultural tolerance for suicide
  + Very high suicide rate when compared to other rich nations
* 88 – Lesotho:
  + Small, landlocked, mountainous country in Africa
  + Highest suicide rate in Africa
  + High levels of child labor
  + Very poor general health outcomes, ex. second highest instances of tuberculosis and HIV/AIDS in the world

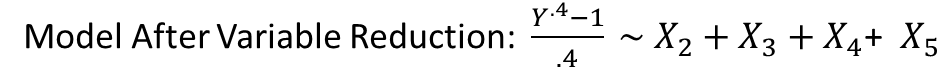
After the removal of outliers, we implemented a stepwise algorithm based on AIC to remove variables that were redundant or unnecessary in the model.

Removed Variables:

* + X1 – Health Expenditure as a percentage of GDP
  + X6 – Psychiatrists working in mental health sector (per 100 000 population)
  + X7 – Mental hospitals (per 100 000 population)

Included Variables:

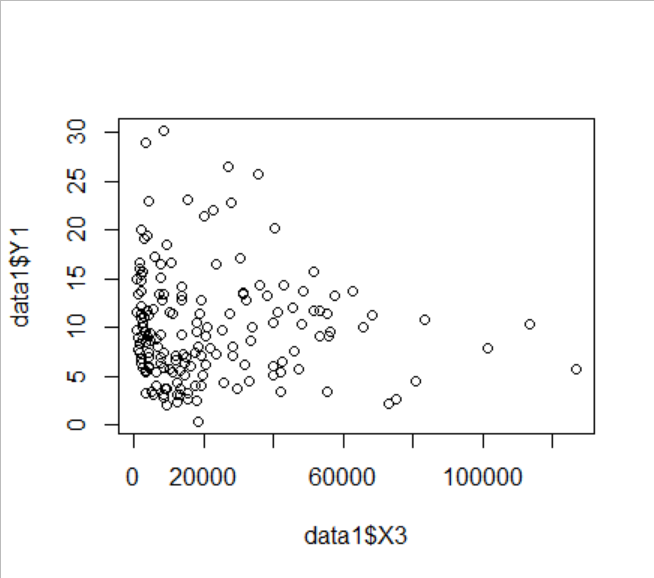
* + X2 – Male to Female ratio of the labor participation rate
  + X3 – GDP per capita, PPP
  + X4 – Liters of alcohol consumption per capita
  + X5 – Prevalence of a national suicide prevention strategy



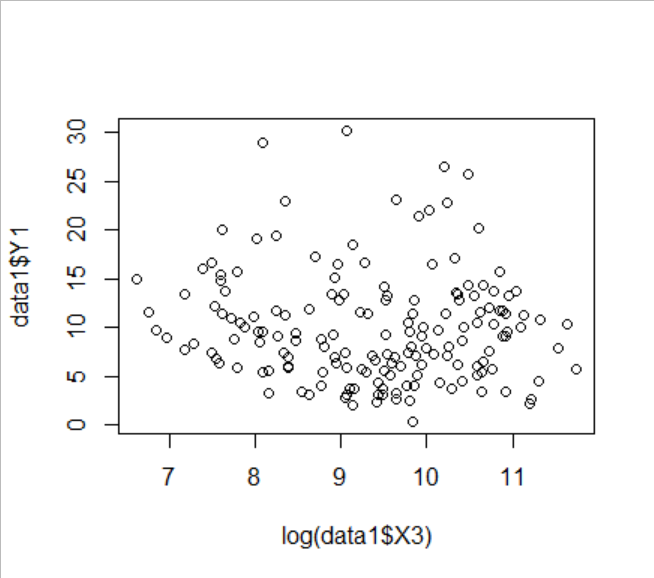
After the removal of variables, further analysis was performed to decide whether a transformation on any of the independent variables may be appropriate.

We discovered that a log transformation of GDP per capita (X3) would better represent the relationship between this variable and the outcome variable. Below are the plots of X3 against the outcome variable before and after the transformation:

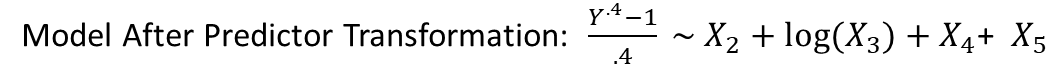
Before Transformation:



After Transformation:



Clearly, the log transformation of the GDP per capita variable yields a better relationship.



The final step in preparing the model was to implement the iteratively weighted least squares algorithm to properly weight each instance in our data. This was more effort to mitigate the effect of outliers on our model. We performed 10 iterations of this algorithm.

Final Model Summary

