**Predicting Cardiovascular Mortality Rates with “Just” Poverty Rates**

# Background

Around the world, rising healthcare costs are worrying policymakers. In 2015, the United States spend close to 3 times on healthcare as compared to the average of other countries with comparable incomes. They spend more on healthcare than any other country, but with average health outcomes. According to the World Health Organization, the US is ranked 30th in the world in terms of life expectancy.

In regards to Cardiovascular Disease, the American Health Association projects that the costs associated with this disease will exceed $1 Trillion by 2035.

## Project Purpose

The goal of the project is to answer one specific question. Can we predict Cardiovascular Mortality Rates in the United States in a satisfactory way, simply by using the poverty rate?

This purpose raise natural questions. Why such a limited scope? Why are we not including other factors that we inherently know would affect mortality rates?

All those questions bring up valid points. However, we must understand the realities of the political climate in the United States. If we can find a factor that can accurately predict these mortality rates, then there is a greater chance of finding the political will to improve this factor’s outcomes. The 80/20 principle directly applied. Ideally, attacking a problem of this magnitude on multiple fronts is preferred. However, the likelihood is highly unlikely today.

## Dataset

The dataset is actually a combination of two datasets. The US Census Poverty Rate dataset and the University of Washington’s Institute for Health Metrics and Evaluation Cardiovascular Mortality Rate dataset.

The University of Washington’s dataset is quite rich. It provides mortality rates for more than just cardiovascular disease. It certainly provides many avenues for further research.

Both datasets were untidy and required cleaning. The biggest problem was that fact that both datasets had differing levels of detail. One dataset went down to the county level; the other dataset was limited to the state level. One dataset has entries every 5 years; the other dataset had annual entries.

After cleansing both datasets, we combined them into one master dataset using Excel. The cleansing process is documented in the cleanse\_data\_set.py file.

# Analysis

## Exploratory Data Analysis

### Outlier Detection

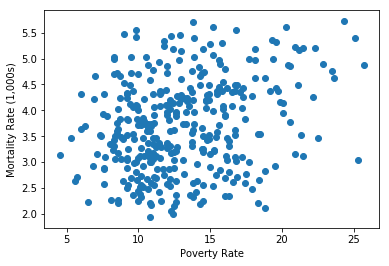
Our initial goal was to determine if there were outliers in the data and the best way of dealing with them.

We decided to use the well-known method of 1.5 \* IQR to determine whether a point was considered an outlier.

Grouping the data by year resulted in five outliers. They were all from the Poverty Rate column. Surprisingly, there were no mortality rate outliers. Another surprising finding was the fact that were exactly one outlier a year, except for the year 2000.

Curious about this anomaly, we decided to slice the data another way. We grouped the data by state instead of by year. Shockingly, we went from five outliers to 17 outliers. If we did group the data in any way, there were four outliers.

Going back to our target question, we graphed a scatterplot of the mortality rate vs Poverty Rate.

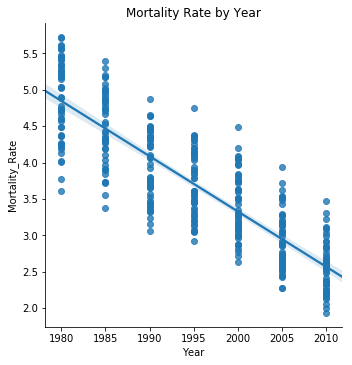


There are no clear outliers and so we decided that the best optimal solution was to ignore any outliers that we had previously flagged. In addition, we would group the data by state.

See the EDA - Outliers Notebook for the code and in-depth analysis.

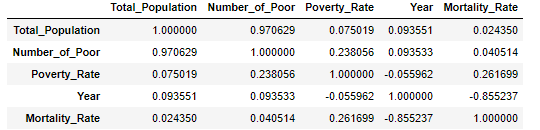
### Correlated Variables

Our first step was to graph the mortality rate over time to see if we could spot a trend.



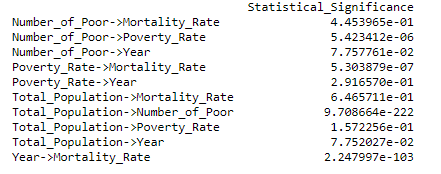
There is a clear downward trend, but the next question was if this was true for every state. After graphing all fifty states, it was clear that the trend held for all states.

Our next step was to determine all the pairwise correlations.



Only the year variable had a strong correlation with the mortality rate. Unfortunately, the poverty rate did not have a strong correlation to the mortality rate.

We also calculated the statistical significance of the Pearson Correlations.

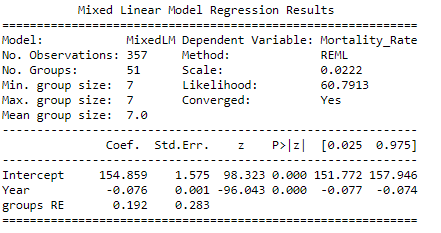


### Statistically Significant Variables

We saw above that the year variable had a strong correlation with the mortality rate. However, we wanted to see if this connection was statistically significant no matter how we grouped the data.

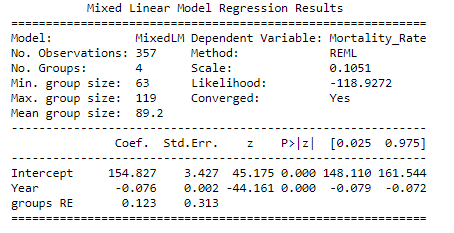
We wanted to answer the question, “Is there a statistical difference in average mortality rates between the different years?”

Due to the longitudinal nature of the data, we could not rely on the usual anova tests. Our data violated the IID assumptions. Luckily, we could have used a repeated measures anova test, but it was not readily implemented in Statsmodels. So we used a similar method called a Linear Mixed Effect Model to answer our question.



Yes, the year variable was statistically significant, when grouping the data by state.

But what if we grouped the data in a different way? Using the US Census specifications, we were able to group each state into their respective regions. Then we reran the analysis.



Again, the year variable was important.

See the Statistical Inference Jupyter Notebook for the correlated variables and statistically significant variables code and in-depth analysis.

## Machine Learning

Now that we have established the year variable is important, it is time to see if we can answer our target question. Can we use poverty rates to accurately predict cardiovascular mortality rates?

The quick answer? No. Please note that we decided that we would focus primarily on using linear regression to model our data.

**Linreg1:** Our first attempt using only the poverty rate resulted in a poorly fitting model with an R^2 value of .066.

**Linreg2:** We knew that the year variable was statistically significant, so we added that to the model. It resulted in a much better result, however, the regression threw a possible multicollinearity error. We were curious in the best possible model, regardless of possible multicollinearity.

**Linreg3:** So we proceed to throw the kitchen sink at the next model. This resulted in an adjusted r-squared of .973. Again, we encountered the multicollinearity error. But, this was to be expected. The number of poor variable was highly correlated with the total population variable.

**Linreg4:** Thus, we got rid of the number of poor variable. We choose this variable because it had a higher correlation with the poverty rate than the total population variable. This resulted in an adjusted r-squared of .973 and another multicollinearity error.

**Linreg5:** Our suspicion was that the multicollinearity error was coming from the year variable. So we removed it and the result was a model with an adjusted r-squared of .369. Again, another multicollinearity error.

**Linreg6:** This time we removed the total population variable because in the previous result, the coefficient was essentially zero. This time, we get an adjusted r-squared of .123 but no multicollinearity error.

**Linreg7:** Surely, we would not receive a multicollinearity error if we just used the year to model the mortality rate. Unfortunately, this did result in a multicollinearity error, but it had an adjusted r-squared value of .731.

**Linreg8:** If we were to ignore the multicollinearity error, then we would have chosen linreg4 because it has the highest adjusted r-squared the least amount of variables. However, we noticed that the total population variable in that model was a coefficient that was close to zero. So what would happen if we removed that variable? A model with an adjusted r-square of .973.

**Linreg9:** But what if we just used the state and year as the predictors? What would happen if we did not include the poverty rate as a predictor at all? A model with an adjusted r-square of .972.

See the Machine Learning Jupyter notebook for the code.

# Conclusions

There are two conclusions that we could potentially reach. Poverty rate data is a good predictor of cardiovascular mortality rates or not. Which conclusion we reach depends on the issue of multicollinearity. If we choose to ignore the multicollinearity error thrown by the code, then we can do an excellent job modeling the mortality rates using the poverty data. If we do not ignore the error, then the poverty data does a poor job modeling the mortality rates.

Unfortunately, this specific instance of multicollinearity is hard to decipher. We clearly saw that the system threw the error even with just one predictor, Year. However, multicollinearity is when one predictor variable in a multiple regression model can be linearly predicted from the others. It is not clear how this applies when there is only one predictor. The literature online is not clear on how to deal with such an occurrence.

Staying on the safe side, we will not ignore the multicollinearity error and assume that the best model is linreg6. Therefore, we conclude that using the poverty rate to predict cardiovascular mortality rate is an ineffective endeavor.

Thus, we would recommend to the client that their money is better spent finding another variable or set of variables that could be used to predict cardiovascular mortality rates.