**Estimating a Country’s Mortality Rates Using Linear Regression**

Heron Ziegel

**Introduction**

Understanding mortality in different countries is an important metric when making decisions in nonprofit work and foreign aid. Some countries are missing in the dataset used ([Adult Mortality Rate](https://www.kaggle.com/datasets/mikhail1681/adult-mortality-rate-2019-2021)), possibly due to a lack of substantial or reliable data. Creating a model based on the countries which have provided data could help make predictions for those countries which have not provided mortality data. The model weights could also offer insight into which factors are most important when studying mortality rates.

The article “[Machine Learning Can Unlock Insights Into Mortality](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8495631/)” explains the necessity for this research, the limitations of mortality data, and how machine learning can help bridge the gap. Essentially, many countries base mortality data on deaths in hospitals when 72% of deaths occur outside of hospitals. It is important to note that the data I am using is susceptible to this bias as well. However, this machine learning method should be able to produce an algorithm that will give a larger weight to more strongly correlated attributes, making it more reliable than traditional statistical methods.

**Proposed Work**

There are seven features in this data: country, continent, average population, average GDP, average GDP per capita, average healthcare expenditure per capita, and development level. There are three ground truth columns: adult female mortality rate, adult male mortality rate, and average crude death rate. Using linear regression is ideal for this data. My goal is to create an estimate of the ground truth values for mortality rate based on the first seven columns of data and known ground truth values. I will remove all three of these ground truth columns from the testing data and create three separate linear regression models, one for each ground truth.

With this in mind, the data only needs to be preprocessed once. I will check for missing values and categorical features and normalize the data. Then for each ground truth I will split the preprocessed data into a training and testing set in order to learn model parameters and evaluate the learned model. I will train the three different linear regression models individually. Finally, I will evaluate each model using Mean Average Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

In order to improve the accuracy of the model, I will add a regularization term to avoid overfitting the data. Because there are only 156 records and 7 feature columns, I don’t anticipate needing to use dimensionality reduction.

After creating a successful model, I will apply the model to several countries which were left out of the dataset, assuming I can find most of the other data columns available as public record. I will compare the results to existing ground truth data when possible. If there is a significant discrepancy, it could indicate either that my model needs to be improved or that the reported data is inaccurate.

**Timeline**

* **03/30/24 Project Proposal due**
* **04/01/24 Preprocess the data**
  + Find missing values and either remove the record or replace with median
  + Transform categorical features into numeric values
  + Normalize data
* **04/07/24 Progress Report 1 due**
* **04/08/24 Split, train and test data**
  + Split into training and testing sets for 3 different ground truth values
  + Test each of the 3 models
* **04/14/24 Progress Report 2 due**
* **04/15/24 Evaluate the model**
  + Use MAE, MSE, RMSE to evaluate accuracy
  + Make improvements if necessary
  + Gather outside data and run it through the models
* **04/23/24 & 4/25/24 Lightning Talk**
* **04/26/24 Start writing Final Report**
* **04/28/24 Final Report due**

**References**

Dataset: Adult Mortality Rate. <https://www.kaggle.com/datasets/mikhail1681/adult-mortality-rate-2019-2021>

“Machine Learning Can Unlock Insights Into Mortality.” <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8495631/>