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), 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” 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.

I was the sole contributor to this project. I preprocessed the data, created three different models, trained and tested the models, and analyzed the results. Because there was so little data, I also decided to create a second dataset based on numbers I looked up online. I then tested this data against the model as well to see how it performed.

Approach

I used the Adult Mortality Rate (2019-2021) dataset from Kaggle. This dataset contains three ground truth values: Adult Female Mortality Rate, Adult Male Mortality Rate, and Crude Death Rate. I used three linear regression models, one for each of the ground truth values. I could have simply used Crude Death Rate as a singular ground truth value, but I believe that it is so strongly correlated to Adult Female Mortality Rate and Adult Male Mortality Rate as to make the rest of the data meaningless.

Each of the three linear regression models contained the same seven features: country, continent, average population, average GDP, average GDP per capita, average healthcare expenditure per capita, and development level. I was interested to see whether the different models would have different weight for these features. Do certain economic statistics impact male mortality more than female mortality?

The dataset is very small, containing only 156 rows. Since there are 195 countries in the world, I wondered whether I could find data for additional countries on my own. Many of the countries which were missing from the dataset do not have accurate reports of data from one or more columns. With that in mind, I understood that even among those countries I could find additional data for, the data might be less accurate and therefore not fit the model well.

I was able to find data for 19 additional countries which were not in the original dataset. The following countries were never included in either set of data: North Korea, Laos, Liechtenstein, Micronesia, Monaco, Palau, Saint Lucia, Saint Vincent, San Marino, Slovakia, Somalia, Syria, Taiwan, Tanzania, Turkey, Ukraine, Vatican City. I also excluded Russia because it is located in both Europe and Asia, and therefore did not fit into the data format.

The original dataset and the dataset I created were each preprocessed. The data did not contain any missing values. There were three categorical features: countries, continent, and development level. I chose to convert these into numeric features using label encoding. I could have used one-hot encoding instead, but it would have significantly increased the number of columns. Since there were very few attributes to begin with, I thought that might place too much importance on a single attribute in the model.

After preprocessing, the original dataset was split into different training and testing data for each of the three models. Since there was limited data I used a 80:20 split. The training data was fit to a linear regression model for each of the three ground truth values. Then the testing data was used to determine the error for each model. Upon finding the error, each model was tested with both a ridge and lasso hyper parameter. I used Mean Average Error, Mean Squared Error and Root Mean Squared Error to determine model performance before and after adding the hyper parameters. Surprisingly, the models for Adult Male Mortality Rate and Crude Death Rate performed better without a hyper parameter. I suspect that with such a small dataset, the algorithm was already very simple and adding a hyper parameter only flattened the results. For Adult Female Mortality Rate, a ridge hyper parameter slightly reduced the error.

After training and testing each of the three models and adding the one hyper parameter, I decided to test the models on the second dataset which I had created. I did not retrain the model with this dataset at all. I checked the Mean Average Error, Mean Squared Error and Root Mean Squared Error for each model to determine whether it worked on the new data.

Results

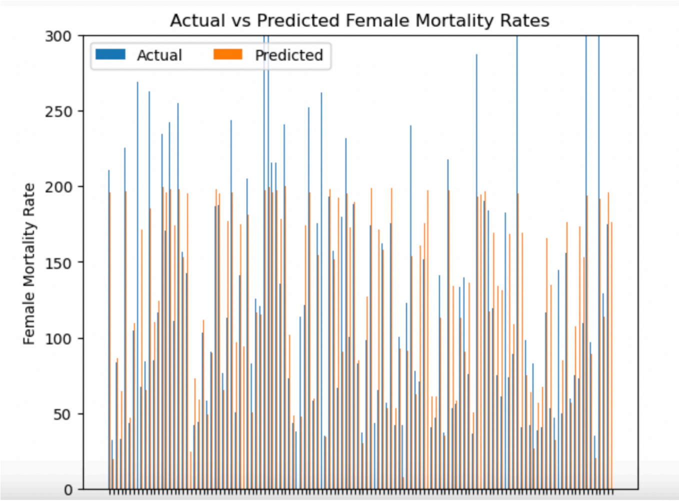
Since the datasets I worked with were very small, the results were predictably not very accurate. However, the models were certainly able to detect a general trend for each of the three ground truth values.

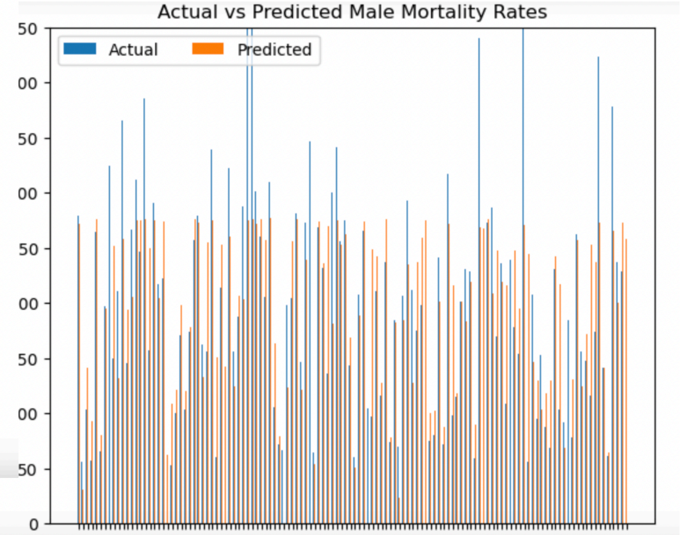
Adult Female Mortality Rate had a mean of 130.89 in the original dataset. The Mean Average Error was 40.07. The Mean Squared Error was 3112.27. The Root Mean Squared Error was 55.79. In the new dataset the mean was 128.59. The Mean Average Error was 92.00. The Mean Squared Error was 13021.13. The Root Mean Squared Error was 114.11.

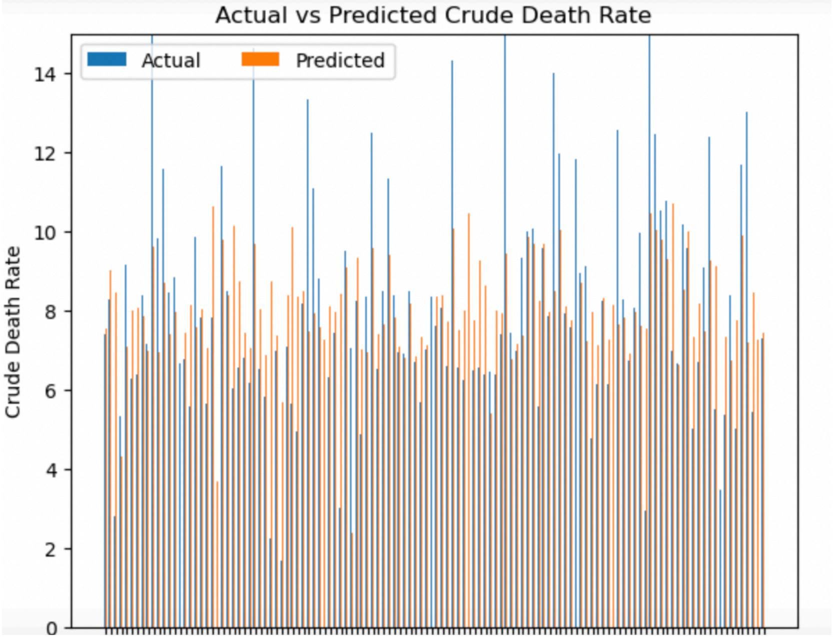
Adult Male Mortality Rate had a mean of 202.38 in the original dataset. The Mean Average Error was 51.57. The Mean Squared Error was 4997.13. The Root Mean Squared Error was 70.69. In the new dataset the mean was 217.31. The Mean Average Error was 145.96. The Mean Squared Error was 27407.83. The Root Mean Squared Error was 165.55.

Crude Death Rate had a mean of 8.14 in the original dataset. The Mean Average Error was 1.94. The Mean Squared Error was 6.90. The Root Mean Squared Error was 2.63. In the new dataset the mean was 8.71. The Mean Average Error was 7.39. The Mean Squared Error was 65.60. The Root Mean Squared Error was 8.01.

Based on these numbers, it is clear that there is a pattern emerging in the data. There is a large amount of error on the prediction, but it generally follows a trend. This is even more clearly illustrated in the double bar graphs below. Because there was so little original data, it is easy to visually compare the actual ground truth values to the model’s predicted values.


Actual vs Predicted Female Mortality Rates
Actual
Predicted
Female Mortality Rate
300
250
200
150
100
50
0


Actual vs Predicted Male Mortality Rates
Actual
Predicted
50
00
50
00
50
00
50
00
50
0


Actual vs Predicted Crude Death Rate
Actual
Predicted
Crude Death Rate
14
12
10
8
6
4
2
0

The coefficients for the Female Mortality Rate model were as follows: [-1.30919195 -31.91538759   2.24925924 -10.47100809 -30.05653533   8.21642174 21.28995986]

This means that there is a very strong correlation between female mortality rate and continent, which is not surprising. There is also an almost equally strong correlation with the average GDP per capita. The next strongest correlation is development level. The least important column for female mortality rate (aside from country, which is a unique column) is the population size. All of this makes sense. It also makes sense that there is an inverse correlation between female mortality rate and GDP, both overall and per capita.

The coefficients for the Male Mortality Rate model were as follows: [0.05808066 -26.23056549   0.88707583 -13.611748   -44.1267299   5.15672348 24.37309116]

This means that the strongest correlating factor for male mortality rate is average GDP per capita. It was much stronger than the second strongest correlation, continent. This is interesting in that it differs from female mortality rate. It seems that male mortality is more strongly tied to economic factors. After that the trends are similar to female mortality rate, though.

The coefficients for the Crude Death Rate model were as follows: [-3.83513578e-01 3.09071809e-01 1.68552465e-03 8.42206272e-02 -2.42149219e+00 1.94237361e+00 -8.87393072e-01]

This means that the average GDP and healthcare expenditure had by far the strongest impact on crude death rate. It is important to note that unlike male or female mortality rates, crude death rate does not measure how many people out of the general population have died in a given time. It measures how many people died in a given time who were expected to die. So it makes sense that healthcare expenditure would be more relevant to this number than it was to male or female mortality rate. As was the case with male and female mortality rate, crude death rate is the last affected by population size.

Conclusion

The mortality rate data from Kaggle was determined to have three ground truth columns: female mortality rate, male mortality rate, and crude death rate. Three linear regression models were created, one for each ground truth. Each model was tested and trained using the data. Then I searched the web to create a new CSV file with data from some of the missing countries. This new data was tested on each of the three models as well.

The female mortality rate model performed better with a hyper parameter, but the other two models did not. Each model was tested for accuracy using the Mean Average Error, Mean Squared Error, and Root Mean Squared Error. The error for the original data was high, most likely due to the small number of rows in the data. However, there was still a clear pattern which emerged. The error for the newly gathered data was higher than the original, which makes sense because it was gathered separately and from less reliable sources. But the new data still seemed to fit the model somewhat as well.

The most important factors in predicting female mortality rate for a country turned out to be continent and GDP per capita. These factors were also most important for predicting male mortality rate, except that GDP per capita came before continent. This seems to indicate that economic factors are more important in relation to male mortality rate compared to female.

The most important factors in predicting crude death rate were GDP per capita and healthcare expenditure per capita. This makes sense when taking into account that crude death rate is a measurement of how many people die who were expected to die, rather than considering the entire population. For all three ground truths, population size was the least relevant factor.

This was an interesting project in that it displayed just how powerful machine learning models can be even with very limited data. Although some of the errors were high, there were still clear trends which emerged. Predictably, the mortality rate of a country has a lot to do with what part of the world it's in and how much money people in that country have. These findings could be useful when discussing the different risks men and women face when it comes to mortality.

The models could also be used to predict mortality rates in the eighteen countries which were not in either dataset. These countries included: North Korea, Laos, Liechtenstein, Micronesia, Monaco, Palau, Saint Lucia, Saint Vincent, San Marino, Slovakia, Somalia, Syria, Taiwan, Tanzania, Turkey, Ukraine, Vatican City, and Russia. Most of them were excluded from my new dataset for not having data on one or more of the ground truth values. This could provide a way to speculate on mortality rates in countries which may be harder to get data from.

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/