Analysis of fragility index through male population distributions

0.1 Abstract:

PURPOSE: Investigate & evaluate the effectiveness of linear regression across proportions of different age groups of male populations to predict a country's fragility index.

METHODS: The linear regression was computed across each age group. The smallest RMS for both positive and negative correlations was found. K-fold cross validation was then computed across these regressions.

RESULTS: Linear regression across the proportion of male populations within certain age groups is an effective method for predicting the fragility index. It can provide results with the RMSE of predicted fragility values within $\pm 10 - 20\%$ of the actual range.

CONCLUSIONS: The algorithm implemented over the given dataset met the metrics evaluation criteria specified in the scientific question, with all RMSE values falling within $\pm 10 - 20\%$ of the range. This method is highly effective for this use-case, and further sociological exploration may yield more insightful results for underlying causes of this relationship.

Word Count: 140

0.2 Introduction:

The objective of this study was to investigate the effectiveness of the usage of linear regression across male population data of different age groups for predicting the fragility index of a country.

The data is not standardized for two reasons. The first is due to a loss of precision when dealing with some datasets. In this case, the rank of the data matrix decreases when going from raw to standardized, posing potential problems to our calculations. Additionally, this data is based of percentages and proportionality, and as a result, does not necessarily benefit significantly from having zero-mean and unit-variance.

The method for selecting observations in each fold that was used in this study was randomized permutations. This was chosen because across many iterations, it shows a generally even relation between the training and testing RMS values. One benefit is that certain countries may be linked in terms of fragility index and alphabetical order due to some unknown variable or also coincidentally. This method minimizes this issue.

Linear regression with an intercept is a method of modelling the relationship between a dependent and independent variables. It predicts the dependent variable using a linear equation of the form $y = mx + b$. For our purposes, we have the dependent variable has vector c, the weight vector w, the data matrix A, and the intercept vector b, such that $c = Aw + b$ [8]. Regression with the addition of an intercept typically increases performance, especially for negative correlations in our specific use case where all data is in quadrant 1.

The RMSE (root mean squared error) is a measure of the magnitude of a varying quantity. It helps define the variability of the data points with linear regression. The RMSE equation is defined as RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ [8]

Data standardization is a method of making all variables statistically equally valuable, and it involves two steps. The first is to make the data zero mean, which is to subtract the mean of each variable from each observation. The second is the make the observations unit variance, which is to say that for each variable, the sum of the observations is 1.

K-fold cross validation is a method of measuring and verifying the quality of a linear regression over K Folds of a dataset. K is typically 5, which means the dataset is sliced into 5 equal folds. Linear regression is performed on a majority of each fold, known as the training set [9]. This regression is then tested on the rest of each fold, known as the test set. Comparing the test and train values across each fold helps determine if the regressions are consistent throughout the dataset [9].

The scientific question to be looked at is, given a dataset of male population distributions,

can the Linear regression provide us with predictions where the RMSE value is within $\pm 10 - 20\%$ of the range of the fragility index.

0.3 Methods:

There are two main algorithms used in this report. The first finds the two variables in the data that best explain the dependent variable. There is one variable for a positive correlation and one for a negative correlation, and this is found through Linear Regression. The second performs K-Fold cross validation of the regression using random permutations.

0.3.1 Variable of Best Fit:

This algorithm requires that the data be processed through an external resource provided by the instructor. The algorithm is divided into two components, the first of which is an iterative linear regression calculation. The initial step is to isolate the observations for a specific age group, and then concatenate it with the ones vector. This allows for a unique line of best fit for each age group that also has an intercept. The w vector is then approximated using the independent variable, in this case the observations for a specific age group, represented by matrix A, and the c vector, which is representative of the dependent variable, which are the fragility indices. The second component of this algorithm is a computation that isolates an age group with the lowest RMSE (Root Mean Squared Error) value for both positive and negative correlations.

The final part of this section are the plots of the aforementioned age groups. The figures show both the line of best fit and the scatter plot of the specific age group, one for positive correlation and one for negative.

0.3.2 Cross Validation of Regression:

The second significant algorithm is a K-fold Cross Validation of the Regression for the Variable of Best Fit. This analysis specifies 5 folds, where $4/5$ of the data observations are used for training and the last $1/5$ are used for testing. The code first performs a random permutation of the rows of the dependent and independent variables, after which it calculates the regression for each fold. The most complicated computation is the definition of the start and end indices for each fold. This is found by specifying a size of each fold, such that the first 4 are always the same size, and the final set goes up until the end of the observation table. The results of the training and testing are stored for each fold and compared at the end. The folds are divided using the equation $\text{fold\_size} = \lceil \frac{n}{k} \rceil$, ensuring an equitable distribution of data for each fold.

0.4 Results:

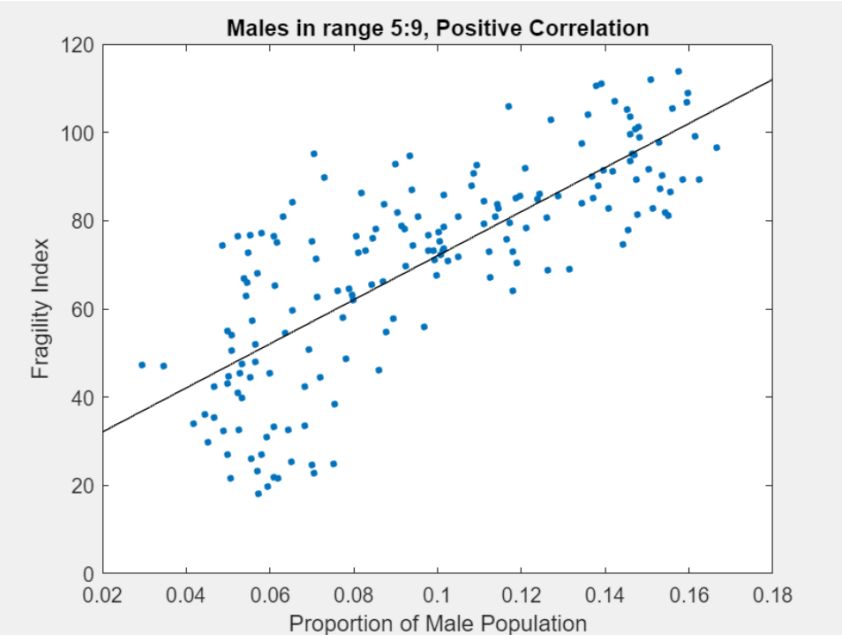

Figure 1: The above figure is a scatter plot with a line of best fit, demonstrating a positive correlation between the age group of males 5-9 and the fragility index. Each data point of the scatter plot, in blue, show the relation between the fragility index and the proportion of males of age 5-9, and the line of best fit uses the slope and intercept from the linear regression.

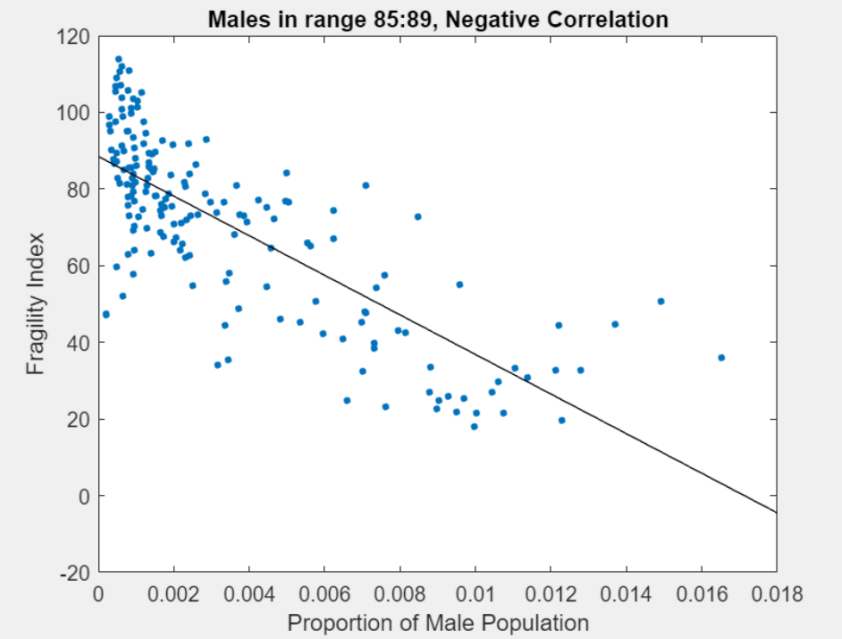

Figure 2: The above figure is a scatter plot with a line of best fit, demonstrating a negative correlation between the age group of males 85-89 and the fragility index. Each data point of the

scatter plot, in blue, show the relation between the fragility index and the proportion of males of age 85-89, and the line of best fit uses the slope and intercept from the linear regression.

Table 1: The RMS values of the linear regression of each age group, split into 3 rows of values, all with their respective index labelled above. The age group with the lowest RMS is age group 18, which is males of age 85-89.

Index	1	2	3	4	5	6	7
Value	14.8983	14.7841	14.8839	15.6132	19.6997	23.3833	23.1422
Index	8	9	10	11	12	13	14
Value	20.8596	17.5837	15.5265	15.7893	15.8134	15.6797	14.9863
Index	15	16	17	18	19	20	21
Value	15.7994	16.0824	15.2327	14.7603	15.2526	17.7945	20.0580

0.5 Discussion:

The findings of this experiment are mostly contained within the figures and table in the results section. The rows of the table describe the RMSE of the linear regression for each age group from 1 to 21. The values are relatively similar, ranging from approximately 14.8 to 23.4. The first graph shows the age group with the lowest RMSE that has a positive correlation between the Nation's fragility index and the proportion of males of age 5 to 9. The points are evenly distributed across the graph, all of which are fairly close to the line of best fit. The second graph shows essentially the same for a negative correlation, and the age group is specified as males of age 85 to 89.

Analysis of the rmstrain and rmstest lists show that the randomly permuted K-Fold cross-validation is mediocre at best for the testing, but it performs much better for training. For testing, the RMSE values typically range from 5-50, while they can also reach up to around 100. This can be attributed to the small size of the dataset, implying that a monte carlo analysis may prove more fruitful [8]. The training RMSE values were much better, on average between 4 to 7, which implies that there may not be too many statistical outliers [8].

The table on its own can be interpreted as proof that applying linear regression on the proportion of males in certain groups to find a nation's fragility index is in fact plausible and can be quite accurate. Because the fragility index is between 0:120, RMSE values from 14.8:23.4 can be thought of as $\pm 10 - 20\%$, which statistically is significant justification for the use of this method. However, from just the table alone, it is difficult to see exactly what effect the size of each age group has on the fragility index. This can be extrapolated from the two figures, where the younger age group has a strong positive correlation, while the older age group has a negative correlation. This shows that countries with large elderly male populations are typically more politically and economically stable than countries with high percentages of younger male populations. Further analysis of the linear regression for each age group shows that, as the age of each group increases, the correlation gradually transforms from positive, to nearly 0, then deep into negative values. This implies that there must be underlying political or socioeconomic variables that are driving these results. There are a myriad of reasons why this could be, and a preliminary analysis of external studies helps offer some potential candidates.

The first possibility has to do with the average number of children that a household has. Countries with very low fragility indices often have a very low number of children per household. For example, Sweden has about 1.5 [5], while Finland has around 1.84 [4]. Meanwhile on the other side of the spectrum, countries like Afghanistan, which both have fragility indices above 100, typically have more than 4 children per household [6]. Numerous studies have

shown that in many countries with significant societal difficulties, there is a strong culture around having many children. A driving factor for the development of this culture is the very fact that many children don't make it to adulthood in these nations, due to a lack of resources or war and civil unrest. By having more children, parents have a higher chance of at least one child reaching an elderly age. On the contrary, countries that do not have this kind of culture have far larger elderly populations. This can be linked back to our findings. Having a large percentage of males in younger ages, especially children 5 to 9, is likely due to economic and political unrest in each country, as children do not reach adult age, and in the places that the culture of having many children is present, the fragility index is on average much higher. On the flip-side, in countries that have a smaller average number of children, there is much less of a worry for them to survive. More children growing up and reaching 80+ falls in line with figure 2, where a higher percentage of elderly males is negatively correlated to the fragility index.

Another possible cause for this correlation could be the amount of natural resources in the country that are in the control of that country's government. Nations with abundant natural resources that are in the control of the native government include Canada and the United States, both of which are fairly low on the fragility index, and also have very large proportions of elderly male populations. These countries are able to utilize their resources for economic purposes, which results in better government infrastructure and a higher quality of life for the population. Meanwhile, countries like South Sudan and Niger also have abundant natural resources, but most of them are in the control of foreign powers [7]. This means that the citizens of these countries are not able to get access to the scarce resources available to them, and their governments have minimal economic status, which in-turn results in significantly worse infrastructure and quality of life. What this means is children in these countries are once again not able to reach elderly status, either due to a lack of necessary resources or becoming victims of violence and war due to the disagreements over the distribution of resources.

In conclusion, the linear regression analysis met the evaluation criteria proposed in the scientific question, where on average the RMSE is within $\pm 10 - 20\%$ of the Fragility index, and countries with larger proportions of male children are typically more politically fragile. These relations can potentially be attributed to children not being able to grow up in countries with poor fragility indices, but further sociological exploration may yield more accurate results.

0.6 References:

1. Study on demographic changes and national stability. Available from: https://www.nature.com/articles/s42949-021-00023-z
2. Analysis of aging populations and societal impact. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513069/
3. Cultural influences on family dynamics. Available from: https://cascw.umn.edu/wp-content/uploads/2014/04/guides_somali.WEB_a.pdf
4. Statistics Finland - Family statistics in Finland. Available from: https://www.stat.fi/til/perh/2020/perh_2020_2021-05-28_tie_001_en.html
5. Life in Sweden: The average family. Available from: https://sweden.se/life/people/the-average-anderssons
6. Demographics of Afghanistan. Available from: https://en.wikipedia.org/wiki/Demographics_of_Afghanistan
7. Countries that developed without natural resources. Available from: https://www.afterschoolafrica.com/49749/top-7-countries-that-developed-without-natural-resources/
8. Cross-Validating Linear Regression. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class13.pdf
9. Patterns- Linear Regression. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class12.pdf