LINEAR REGRESSION AND FIVE-FOLD CROSS-VALIDATION

1. ABSTRACT

PURPOSE: To analyze the relationship between age cohorts of the male population in a country and the fragility index of that country by performing linear regression, and conducting a five-fold cross-validation.

METHODS: In order to find the index of the variable that best explained the fragility index, linear regression was performed, by calculating the Root Mean Square (RMS) error, which is a measure of the average magnitude of the differences between predicted and observed values. To assess the performance of the chosen variable, a five-fold cross validation technique is used, wherein the model is trained and evaluated five times, each time using a different fold as the test set and the remaining four folds as the training set. The relationship is visualized between the proportion of the male population in a specific age group and the fragility index through two linear fits, one each for the negative and positive slope, which are straight lines that best approximates the trend or relationship between two variables.

RESULTS: The Root Mean Square (RMS) errors for linear regression on each variable is found in an array using ‘rmsvars’, and the errors for each of these 21 columns range between 14.50 – 20.50 units of fragility. The index of the variable with the smallest RMS error for a positive slope, and negative slope, is 2 and 18 respectively. An array with the RMS errors for the training phase of a 5-fold cross-validation is created using the variable identified in ‘lowIndexPositive’, and an array with the RMS errors is created for the testing phase of the same 5-fold cross-validation, with the values for both of these arrays ranging between 15 – 19 units of fragility.

CONCLUSIONS: Higher RMS values within ‘rmsvars’ indicate a poorer fit of that variable to the fragility index. Since the index of the variable with the smallest RMS error for a positive and negative slope is 2 and 18 respectively, the variables at index 2 and 18 are the ones with the most significant positive and negative influence on the fragility index respectively. Lastly, the ‘rmstrain’ and ‘rmstest’ arrays provide RMS errors for 5-fold cross-validation, which indicate how well the linear regression model generalizes to new data.

2. INTRODUCTION

The scientific objective that was posed here was to investigate the association between age cohorts in a country's male population and the fragility index through linear regression, while assessing the model's performance using a five-fold cross-validation.

To understand this report, one must have knowledge with regards to some of the background topics addressed in this report. This report is based upon the relationship between the fragility index of countries and the proportion of male populations across various age groups. One of the main concepts used in the report is linear regression, which is an algorithm that models the relationship between dependent and independent variables using a linear equation. The linear regression is completed through the Root Mean Square (RMS) Error, which is a measure of the average deviation between predicted and actual values. By using the Root Mean Square Error, a relationship is created between the fragility index and the proportion of males in a specific age group. An index of the variable that best explains the dependent variable (fragility index) and distinguishes between positive and negative relationships is also found using this process. The lowest positive index is then used to split the dataset into five subsets for training and testing to assess the model's generalization performance. This is known as five-fold cross validation, and by calculating the RMS errors for both training and testing sets in each fold, the five-fold cross validation also provides insights into the model's generalization capability and robustness. To illustrate the relationship, two plots are created, one for the index of the negative slope and one for the positive slope, between the proportion of male populations in a specific age group and the fragility index. This visualization helps interpret the linear fits and provides a graphical representation of the observed trends. Overall, the report is based around understanding demographic factors influencing fragility and evaluating the predictive performance of the linear regression model.

To test the scientific question, multiple steps were taken. Linear regression was performed on each variable by firstly, creating a design matrix that includes the predictor variable and a ones vector. This allowed for an intercept term to be created, and the linear regression equation was then solved. The RMS errors were then calculated, which allowed for the identification of the variables that minimally contributed to the RMS errors, considering both positive and negative slopes separately. The five-fold cross validation took place after this, by partitioning the dataset into approximately five equal folds. The linear regression models were trained using the training set, and the trained model was then used to predict the fragility index for both the training and testing sets, by calculating the Root Mean Square (RMS) errors for each set. The values for each iteration were stored in the training and testing sets, which thus resulted in two different arrays, one for each set.

3. METHODS

The goal, as part of the task assigned, was to assess the relationship between the fragility index and various demographic variables, by specifically focusing on age groups in the male population. In order to reach this goal, it was important to first prepare the data. This was done by loading the fragility data, which includes the fragility index and proportions of males in different age groups for assessed countries. Once this was done, the focus leaned towards linear regression and the RMS calculation. For each demographic variable, a linear regression model was constructed using the least-squares method. The design matrix was formed, by including the selected variable and a constant term, and the linear regression coefficients were computed. Then, the root mean square (RMS) error was calculated as a measure of how well the linear regression model fit the fragility index. Through this process, the selection of the variable took place. In order to do so, the variable that best explained the fragility index was determined by identifying the variable with the lowest RMS error. Both the positive and negative variables were tracked separately, based upon the consideration of whether the slope of the linear fit was positive or negative. This gave the basis or possibility of visualization, as a scatter plot was generated for the index with a negative and positive slope against the fragility index. The linear regression line was overlaid on the plot, providing a visual representation of how well the variable explains the fragility index.

Once the variable had been selected, the process of five-fold cross validation was used by originally splitting the data. The dataset was randomly shuffled, and the data was divided into five approximately equal-sized folds. The model was trained and tested, by using a linear regression model on each fold and using the training data, which focused on the previously identified optimal variable. The model's performance was then evaluated on both the training and testing sets within each fold. The evaluation of the performance came through only by specifically looking at the RMS error calculation. The RMS errors for both the training and testing sets were computed for each fold. This meant that the process of finding RMS errors was repeated five times, therefore providing a comprehensive assessment of the model's predictive performance across different subsets of the data. One of the key ideas that was also needed in order to provide the same five-fold cross validation each time was through reproducibility. To ensure scientific reproducibility, the code set a consistent seed for the random number generator, by using “(rng('default'))”. This step was crucial to obtain consistent and comparable results in each run.

Overall, the onus was upon providing a rigorous evaluation of how well demographic variables, particularly age groups in the male population, predict the fragility index. The use of linear regression models, RMS errors, and 5-fold cross-validation allowed for a systematic and statistically robust analysis of the relationships between variables. The transparency in data preparation, model training, and reproducibility measures ensures the reliability of the findings in addressing the task's objectives.

4. RESULTS

The results of the analysis are centered around the root mean square (RMS) errors, revealing the accuracy of linear regression models in predicting the fragility index based on demographic variables. The RMS errors for individual variables exhibit varying degrees of influence, with specific variables standing out as particularly impactful, whether positively or negatively. The associated plot of the given data against linear regression highlights the model's fit to the dependent data, providing a visual representation of the relationship. Furthermore, the RMS errors of the folds in the 5-fold cross-validation illuminate the generalization performance of the models, offering insights into their robustness and reliability across different subsets of the dataset. The interplay of these metrics provides a comprehensive overview of the regression models' effectiveness and their ability to capture the underlying patterns in the data.

Table 1: Value of RMS Error for each column number and age group within original CSV file. Index of the age group that best explains the fragility index is Column 2 and the age range is between 5-9 years old.

Column Number	Age Group	Value of RMS Error (in units of fragility)
1	0 – 4	14.8983
2	5 – 9	14.7841
3	10 – 14	14.8839
4	15 – 19	15.6132
5	20 – 24	19.6997
6	25 – 29	23.3833
7	30 – 34	23.1422
8	35 – 39	20.8596
9	40 – 44	17.5837
10	45 – 49	15.5265
11	50 – 54	15.7893
12	55 – 59	15.8134
13	60 – 64	15.6797
14	65 – 69	14.9863
15	70 -74	15.7994
16	75 – 79	16.0824
17	80 – 84	15.2327
18	85 – 89	14.7603
19	90 – 94	15.2526
20	95 – 99	17.7945
21	100+	20.0580

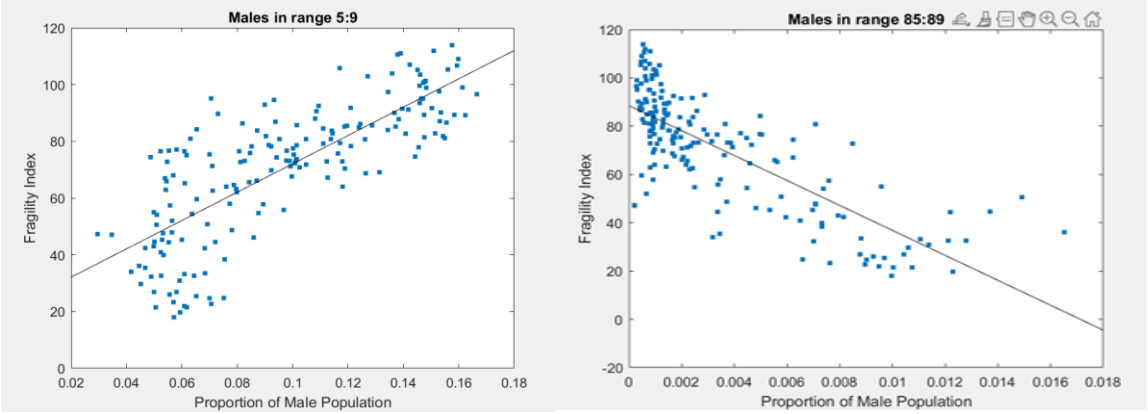

Figure 1: Going from left to right, plots that show relationship between given data and a linear regression to the dependent data for the indices with the smallest RMS error for a positive and negative slope respectively.

Table 2: Value of RMS Error for Training and Testing Data based upon value number within both arrays. Index of the age group that best explains the fragility index is Column 2 and the age range is between 5-9 years old.

Value Number	RMS Error for Training Data (in units of fragility)	RMS Error for Testing Data (in units of fragility)
1	17.0355	15.4396
2	16.6601	16.9917
3	16.2530	18.5121
4	16.6038	17.2239
5	17.0351	15.3443

5. DISCUSSION

To understand the relationship between age cohorts of the male population and a country's fragility, linear regression analysis was applied, with the usage of the RMS errors as a pivotal metric. The primary aim was to identify the age group that best explains the fragility index, both in terms of positive and negative correlations. This objective is fulfilled through a detailed examination of the RMS errors, encapsulated in Table 1. The scientific process involved systematically evaluating each age group as a potential explanatory variable in the linear regression model. We computed the RMS errors for every age group, providing a quantitative measure of how well the fragility index could be predicted based on the proportion of the male population within each group. This analytical approach adheres to the fundamental principles of linear regression, aiming to minimize the discrepancy between predicted and observed fragility values. Table 1 presents the RMS errors for each age group, and from observing the table, the specific age cohorts that yield the most accurate predictions

of fragility can be found. The lowest RMS errors signify a stronger correlation between the demographic composition of a given age range and the fragility index. Consequently, the identification of the age group with the minimum error serves as a direct fulfillment of our scientific objective. The differentiation between positive and negative correlations allows for a nuanced understanding of how varying proportions of the male population in distinct age cohorts influence fragility, and this will be discussed in the following paragraphs.

One of the results that was given as part of this report was that the proportion of males in the 5-9 age range has the strongest positive correlation with the fragility index. This implies that as the proportion of males in the 5-9 age group increases, the fragility index also tends to increase. As an example, some of the countries with a low fragility index such as Canada and Australia can be compared to countries with a higher fragility index such as Afghanistan and Yemen. Australia and Canada's fragility indexes ranges between 25 – 26, and both countries have less than one million people in the age group of 5 – 9. On the other hand, Afghanistan and Yemen's fragility indexes range between 105 – 107, and both countries have nearly 2 million to over 2.5 million people in the age cohort of 5 – 9. Another result that was observed was that of the proportion of males in the 85-89 age range having the strongest negative correlation with the fragility index. This implies that as the proportion of males in the 85-89 age group increases, the fragility index decreases. Once again, the countries with a low fragility index such as Canada and Australia can be compared to those with a higher fragility index such as Afghanistan and Yemen. Australia and Canada's fragility indexes ranges between 25 – 26, and both countries have between 100,000 – 200,000 in the age cohort of 85 - 89. On the other hand, Afghanistan and Yemen's fragility indexes range between 105 – 107, with only 7000 – 8500 people between the age cohort of 85 – 89. The highest RMS error was received in the age cohort of 25 – 29, which shows that regardless of whether the proportion of the people the age cohort of 25 – 29 increases or decreases, its effect on the fragility index is minimal. Canada and Australia, with low fragility indexes, have between 850,000 and 1.25 million people in the age cohort of 25 – 29. Similarly, Afghanistan and Yemen, with high fragility indexes, have between 1.1 million people and slightly over 1.2 million people in the age cohort of 25 – 29. Since all four of the countries have nearly the same people in the age cohort of 25 – 29, it proves that the relationship between the number of people aged between 25 – 29 and the fragility index of a country is very poor. Another important aspect that must be looked at is in regards to the information that was received about the testing data. The RMS error that was received for the testing data differs depending on the specific fold that's being looked at within the dataset. The dataset has been split into five folds, which means that each fold has thirty-five or thirty-six countries. This can provide a way of comparison between different folds, since the RMS error for the testing data varies depending on the fold that's being looked at. For example, in the first fold, for the age cohort of 5 – 9, the RMS value of the testing data is at its lowest. The first fold has countries with low fragility indexes such as Australia and Austria, that have less than 1 million people within the age cohort of 5 – 9-year-olds. On the other hand, countries such as Afghanistan and Chad, with high fragility

indexes, have over a million people in the age cohort of 5 – 9-year-olds, and in Afghanistan's case, have 2.5 million people in the age cohort of 5 – 9-year-olds. Meanwhile, in fold 3, countries such as Maldives and Mauritania, which have high fragility indexes, have less than 275,000 people in the age cohort of 5 – 9-year-olds. On the other hand, a country like Japan, with a low fragility index, still has over 2.75 million people in the age cohort of 5–9-year-olds. This provides considerable evidence as to why the RMS error of the testing data in fold 3 is the highest. The reason why Maldives and Mauritania, have a high fragility index even with less 5-9-year-olds, is potentially because of other factors, one of which might be the smaller population of the country in general in comparison to other larger countries. The scatter plots that have been created for the strongest positive and negative correlation also provide an excellent indication and visualization of how a particular age cohort is either representative of a direct or inverse relationship with the fragility index. The line of best fit that is created is based around the points on the scatter plot, since the line of best fit tries to hit as many points while being created so as to provide a good idea of the relationship between an independent and dependent variable.

In conclusion, the linear regression analysis, supported by RMS errors, highlights distinctive age cohorts with regard to influencing a country's fragility index. The 5-9 age group shows a strong positive correlation, with increased proportions correlating to higher fragility. Conversely, the 85-89 age range exhibits a stronger negative correlation. The 25-29 age cohort demonstrates minimal impact on fragility, evidenced by a high RMS error. Testing data across folds doesn't account for demographic variations, emphasizing the role of country size. The scatter plots visually reinforce these correlations, by portraying the age cohorts' relationships with fragility. Overall, this comprehensive analysis provided through linear regression and five-fold cross validation offers insights into demographic factors that shape fragility indexes across various age groups.

6. REFERENCES

1. Ellis RE. Design Matrix and Standardized Data [Internet]. Queen's University; 2021 [cited 2024 Feb 09]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class09.pdf
2. Ellis RE. Orthogonal Projection [Internet]. Queen's University; 2021 [cited 2024 Feb 09]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class11.pdf
3. Ellis RE. Patterns – Linear Regression [Internet]. Queen's University; 2021 [cited 2024 Feb 09]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class12.pdf
4. 1. Pandian S. K-fold cross validation technique and its essentials [Internet]. 2023 [cited 2024 Feb 9]. Available from:

https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/

5. 2. Brownlee J. Linear regression for machine learning [Internet]. 2023 [cited 2024 Feb 9]. Available from: https://machinelearningmastery.com/linear-regression-for-machine-learning/