DATA 11900 Midterm Project Report

Group 14: Gianina Flegueras, Jenny Lee, Neil Pandya, Jayson Tung

Overview and Description of Dataset

Our project was broadly inspired by the UNDP's Global Multidimensional Poverty Index (MPI), which aims to measure the incidence and severity of poverty experienced in different countries and is currently used to assess acute poverty in developing countries. [1] The indicators included in the MPI are categorized into 3 dimensions: health, education, and standard of living. One example of existing work on this topic is "Factors Determining Differences in the Poverty Degree among Countries", a paper by Eduardo Cuenca García, Margarita Navarro Pabsdorf, and Juan Carlos Moran Alvarez [2], where the authors selected predictors within the 3 MPI dimensions and by applying the Rasch Model, they were able to "quantify" poverty. We aim to follow a similar process through identifying variables in each dimension that are strong predictors of poverty. Our research question therefore is: **Using the 3 MPI dimensions as a guideline for identifying predictors, what are some of the key variables in predicting poverty?**

The persistence of poverty remains a pressing global issue today, and we believe that these predictors of poverty could be indicative of areas that countries need to focus on for development. It is thus crucial to ascertain the factors that tend to be the most correlated with poverty and predict the degree to which they affect the poverty level across different countries. In order to do this, we decided to first use a Lasso Regression model since the regularization process allows us to determine which variables have substantial coefficients (i.e., are strong predictors) and which variables have coefficients that are very small or equal zero (i.e., are weak predictors). Second, we trained a kNN classification model on the selected predictors to classify countries into Lower Income and Upper Income, as categorized by the World Bank. The focus of this report however, will be on the Lasso Regression model.

The data we are using is the World Development Indicators (WDI) dataset from the World Bank [3]. WDI is generally considered a good dataset to use when analyzing global socio-economic development because it provides time series data for 266 economies (countries). For most variables, there is data for each country for the years 2000, 2005, and years 2010 to 2023. The World Bank classifies each country into High, Upper-Middle, Lower-Middle and Low income classes. We created a new column, *Binarized Income*, which classifies a country as **Low Income (0)** for Lower-Middle and Low income classes, or **High Income (1)** for High and Upper-Middle income classes. Our final cleaned and filtered dataset includes 3472 observations of 43 variables, where each observation includes data for one year of one country. Our project will focus on determinants of poverty by predicting poverty headcount ratios (Model 1) and using predictors to classify upper versus lower income economies (Model 2). More specifically, the goal for Model 1 is to predict the percentage of people below the poverty line in a country using the *"Poverty headcount ratio at $3.65 a day (2017 PPP) (% of population)"* variable. For the report, we will shorten this to *Poverty Headcount Ratio*.

Exploratory Data Analysis and Feature Selection

In our initial data exploration, we created scatterplots and correlation heatmaps of various indicators against *Poverty Headcount Ratio* to select features for our initial model. However, after creating our first regression model, we found that our residual diagnostic plot (Figure 3) indicated high heteroscedasticity. We thus decided to transform our response variable using a square-root transformation, and we explain in more detail why we chose to do this in the "Modeling: Lasso Regression" section below. This is why the scatter plots that we use below have *sqrt(Poverty Headcount Ratio)* as our response variable.

Educational Indicators: There were 5 educational indicators in our dataset, including government expenditure on education, percent of people who completed different levels of education, and net enrollment rates for

primary education. Following initial data exploration, we unfortunately found that many of the educational indicators had a significant number of missing values. To accommodate for this and still include a relevant educational indicator, we selected the indicator that had a strong correlation with *sqrt(Poverty Headcount Ratio)* as well as a lower amount of missing values. Ultimately, the most appropriate indicator was *Adjusted net enrollment rate, primary (% of primary school age children)* with a correlation coefficient of -0.56 against *sqrt(Poverty Headcount Ratio)*. As primary students who cannot attend school at such a young age are unable to have the foundational knowledge needed to have a well-paid job in the future, this likely contributes to potential increased poverty headcount.

Health Indicators: The overall health of a country's population can directly impact poverty headcount ratios because for each individual, the prevalence of chronic illnesses or disabilities can limit job opportunities, decrease productivity, and increase medical expenses, all of which contribute to poverty. We explored 8 health indicators on population health and healthcare spending, and healthcare capacity. Unfortunately, we found that most of the health indicators had many missing values. As a result, we chose 4 indicators total: *log(Health Expenditure)*, *Annual Population Growth as a Percentage, Total Life Expectancy at Birth,* and *Probability of dying among youth ages 20-24 years (per 1000)*.

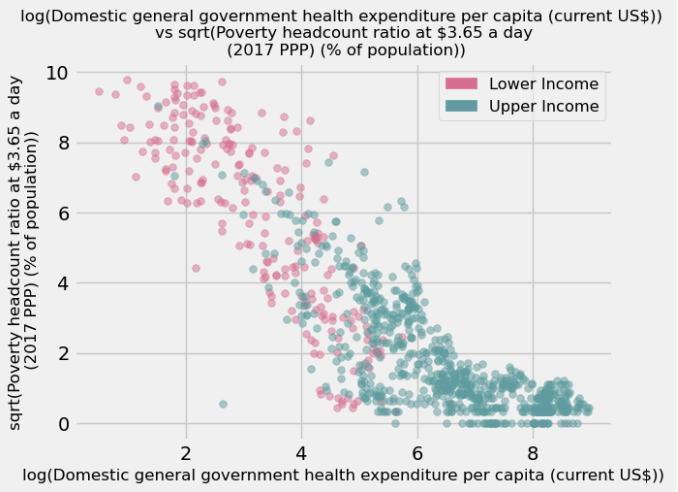

Figure 1: Scatterplot of log(Government Health Expenditure per Capita) against sqrt(Poverty Headcount Ratio)

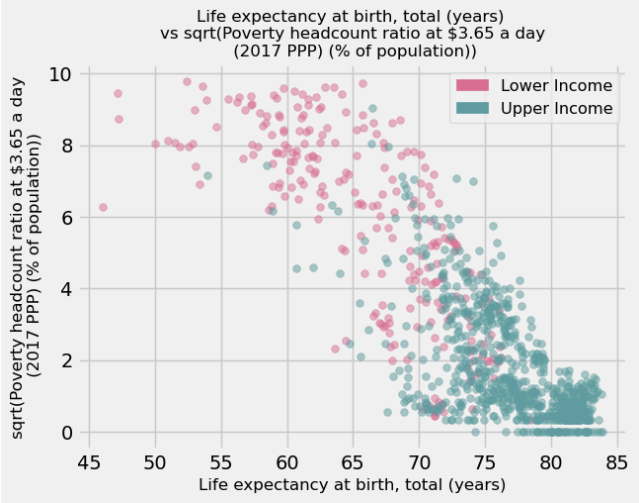

Figure 2: Scatterplot of Total Life Expectancy at Birth against sqrt(Poverty Headcount Ratio)

Two health indicators that we chose are log-transformed *Government Health Expenditure per Capita* and *Life Expectancy at Birth*. First, health expenses can be significant, and especially for people in poverty, these costs can be prohibitive, leading to a lack of access to essential healthcare services and exacerbating poverty. As seen in Figure 1, we observe this expected strong negative linear correlation (corr = –0.87) between *log(Govt Health Expenditure per Capita)* and *sqrt(Poverty Headcount Ratio)*. Second, *Life Expectancy at Birth* is an overall population mortality indicator, so it represents the health status of the country. An increase in *Life Expectancy* means that overall living standards have improved, so we expect *Poverty Headcount Ratio* to decrease, and this negative relationship (corr = -0.84) is indeed observed in Figure 2. From both scatterplots above, we see that these health predictors appear to be linearly correlated with *sqrt(Poverty Headcount Ratio)* and there are little to no significant outliers, making these valid predictors for our regression model.

Governance Indicators: Given that the government plays a crucial role in ensuring a decent standard of living for its population, we decided to explore 6 indicators of governance quality (*Control of Corruption, Government Effectiveness, Political Stability and Absence of Violence/Terrorism, Regulatory Quality, Voice and Accountability,* and *Rule of Law*) called "Worldwide Governance Indicators (WGI)" as a proxy for the

MPI's standard of living dimension. These indicators are normalized to the units of a standard normal random variable, ranging from approximately -2.5 (worst) to 2.5 (best) since the aim of the WGI project was to create a cross-country measurement for governance. [4] From the scatterplots that we created, there seemed to be a moderately strong negative correlation between *sqrt(Poverty Headcount Ratio)* and most of the governance indicators, with the correlation coefficients for these indicators ranging from -0.57 to -0.69. Though all 6 features were correlated with each other, there was no perfect collinearity since they capture different aspects of governance. Thus, we included all of them within the Lasso regression, as the regularization process will help determine the best predictors of the poverty ratio.

Modeling: Lasso Regression

Rationale: The variables we have chosen are likely to be correlated with each other due to interdependence. For example, *log(Health Expenditure)* is likely to be associated with *Government Effectiveness* since *Government Effectiveness* captures the perception of the quality of public services, which includes public healthcare services. Thus, in order to improve our feature selection, we decided to use Lasso regularization for our model in order to identify insignificant and redundant variables. This would shrink the irrelevant coefficients towards zero, and help us account for potential multicollinearity problems.

Preparing the Data – Transformations and Standardization: After selecting our indicators, we had 643 observations without any missing or "NaN" values. We ran the regression over all the available data points for all countries across all years. Our initial model aimed to predict *Poverty Headcount Ratio* as our response variable, which quantifies the percent of the population that fall below this poverty line. However, after running our regression against this response variable, we found that our residuals were clustering on the left, as seen in Figure 3, which suggests that the homoscedasticity assumption may be being violated. Despite attempting various transformations to our predictor variables, our residuals plot did not improve much. Thus, we ultimately decided to use a square root transformation on our response variable, and used *sqrt(Poverty Headcount Ratio)* as our response variable instead. This slightly changes our initial research question, since we would be predicting *sqrt(Poverty Headcount Ratio)* instead of *Poverty Headcount Ratio*. Although some of our indicators displayed a weaker linear relationship against the transformed response variable as a result, others displayed a stronger linear relationship. Thus, there is a trade-off between the variables, but ultimately, we believe that this transformation was necessary to reduce heteroscedasticity. As seen in Figure 4, transforming our response variable significantly improved our residuals plot, and the residuals are spaced out with closer to equal variance over the scaled *sqrt(Poverty Headcount Ratio)* values.

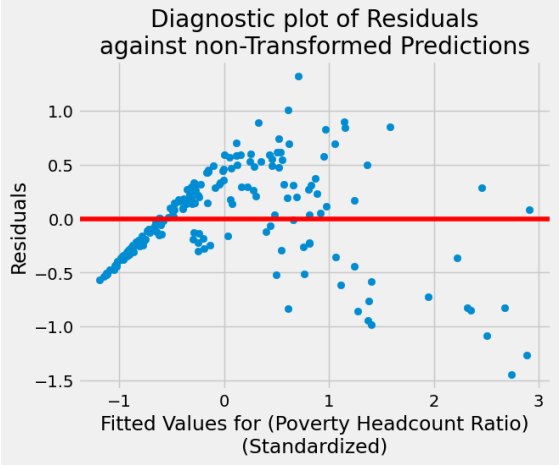

Figure 3: Diagnostic Plot of Residuals against standardized non-Transformed Fitted Values for Poverty Headcount Ratio, $r^2 = 0.7893$

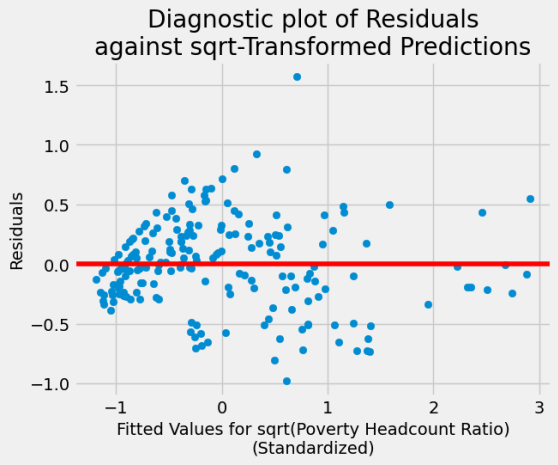

Figure 4: Diagnostic Plot of Residuals against standardized sqrt-Transformed Fitted Values for Poverty Headcount Ratio, $r^2 = 0.8514$

To prepare our data for the Lasso Regression, we used StandardScaler to standardize all our predictor variables and our response variable after transforming all the data accordingly.

Assumptions: In making our model, we did not control for time using time-fixed effects. We thus had to assume that there are no unobserved variables that vary over time that could systematically affect *sqrt(Poverty Headcount Ratio)* in a way that is related to our predictor variables. We had to assume that there aren't any underlying time trends that affect *sqrt(Poverty Headcount Ratio)* that are not sufficiently captured by our predictor variables, which may be a limitation to our model since this is unlikely to be the case. We also had to assume that all our observations were independent of each other. Further, we had to assume linearity – since we used a square root transformation on our response variable, we had to assume that the relationships between *sqrt(Poverty Headcount Ratio)* and each term in our regression are linear.

Training and Testing: We split the non-NaN data into training and testing data (with 30% test size), ultimately resulting in a training and testing size of 450 and 193 observations respectively. To ensure that our training model is a good basis to use when making predictions on testing data, we performed a k-Fold on the training data to see whether the same predictions can be made within the training dataset itself. We observed an average adjusted r^2 value of 0.7545 using 10 splits, indicating that our model is a strong predictor of *sqrt(Poverty Headcount Ratio)*. To provide a final validation of the Lasso regression, we fit a model using the entire training data and validated it with the testing data, obtaining a best **$\alpha = 0.00923$**. As we see in Figure 5, the Lasso regularization shrunk some of our coefficients to zero at a very low α level. With this validation, we calculated an **adjusted r^2 score of 0.8334**, indicating a very good fit of our Lasso Regression model against the testing data.

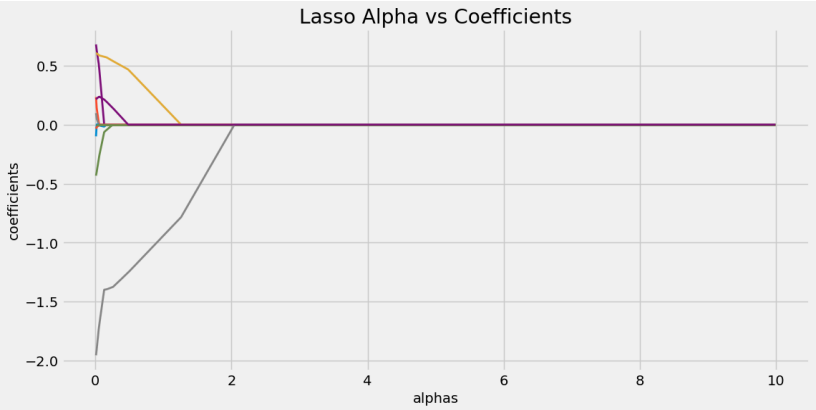

Figure 5: Graph of Lasso Alpha against Coefficients

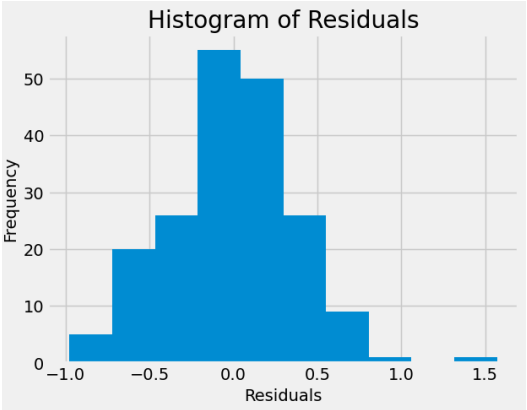

Figure 6: Histogram of Residuals

Model Evaluation and Interpretation

As seen in our diagnostic residuals plot in Figure 4, the residuals seem to be approximately evenly distributed around the y = 0 line, especially for the residuals where *sqrt(Poverty Headcount Ratio)* is positive. However, there is some clustering on the left side of the plot where *sqrt(Poverty Headcount Ratio)* is negative, and the residuals seem to show an increasing pattern. This suggests that the constant variability assumption may not have been fully satisfied in our model. However, from Figure 6 above, our histogram of residuals looks approximately normal and symmetric apart from a few outliers, which suggests that a multilinear regression model is generally appropriate for this dataset.

Our Lasso regression model yielded the following regression equation and coefficients:

Regression Equation:

$$\text{sqrt(Poverty headcount ratio at \$3.65 a day (2017 PPP) (\% of population))}$$
$$= -2.2385 \times 10^{-16} + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{12} X_{12}$$

Table 1: Regression Coefficients

Variable	Coefficient
X_1 : Adjusted net enrollment rate, primary (% of primary school age children)	−0.0
X_2 : Control of Corruption: Estimate	0.0566
X_3 : Government Effectiveness: Estimate	−0.0
X_4 : Political Stability and Absence of Violence/Terrorism: Estimate	−0.1659
X_5 : Regulatory Quality: Estimate	0.0196
X_6 : Voice and Accountability: Estimate	0.263
X_7 : Rule of Law: Estimate	−0.0
X_8 : Probability of dying among youth ages 20-24 years (per 1,000)	0.254
X_9 : Life expectancy at birth, total (years)	–0.0
X_{10} : log(Domestic general government health expenditure per capita (current US))	−0.7979
X_{11} : Population growth (annual %)	0.0935
X_{12} : Binarized Income	−0.0086

Figure 6: Regression Equation and Table of Regression Coefficients

The Lasso regularization helped to shrink the β_1, β_3, β_7 and β_9 coefficients to 0, which are highlighted in green in Table 1 above. This effectively drops these variables from our model. *log(Government health expenditure per capita)* had the highest coefficient of –0.7979, which suggests that it had the largest effect on *sqrt(Poverty Headcount Ratio)* compared to the other variables. Given that the *Government health expenditure per capita* indicator and the *sqrt(Poverty Headcount Ratio)* response variable are standardized, and transformed in different ways, the interpretation of this coefficient is complex and not very meaningful. The negative coefficient means that as *log(Government health expenditure per capita)* increases, *sqrt(Poverty Headcount Ratio)* decreases, which matches our expectations since higher health expenditure by the government indicates more investment into public healthcare services, which could translate into better health outcomes for its citizens. This tends to be positively associated with higher productivity and income levels and ultimately lower poverty levels.

Three coefficients (β_2, β_5, and β_6) do not match our expectations because we expected an inverse relationship between their respective predictors and *Poverty Headcount Ratio*, but the model instead predicts a positive coefficient. For example, consider β_6, the coefficient for X_6: *Voice and Accountability*. Our estimate for β_6 is 0.263, which implies a weak positive correlation between X_6 and *Poverty Headcount Ratio*. In other words, our model predicts that if we increase freedom of expression and strengthen state capacity to respond to citizen's demands, then we expect the prevalence of poverty to increase. This prediction is counterintuitive, and we can use the same logic to conclude that the predictions for β_2 and β_5 are also counterintuitive. We offer two explanations as to why these coefficients are positive.

First, there may be omitted variable bias, meaning that unobserved confounding variables are causing an upwards bias. To illustrate, using X_6: *Voice and Accountability* as an example again, one factor that is not included in our model is *International Aid Dependency*. When a country has greater dependency on support from foreign aid and international NGOs, this means that there are more social workers advocating for the underserved communities, so *Voice and Accountability* increases. However, an increase in international aid

could cause the misallocation of aid towards poorly designed interventions that hinder long-term local development. Thus, we expect *International Aid Dependency* to cause an increase in both *Voice and Accountability* and *Poverty Headcount Ratio*. As a result, our coefficient for *Voice and Accountability* is biased upwards, so $\beta_6 = 0.263$ is an overestimate. Again, we can apply this same explanation for β_2 and β_5, so it is possible that these coefficients are positive because they are an overestimate due to upwards bias. However, we are less inclined to believe that this explanation is true because our model includes multiple indicators from all 3 MPI dimensions, so we believe that our choice of predictors is reasonably comprehensive. A second explanation is that there may have been a nonlinear relationship between each of these predictors and the response variable to begin with, and the true parameters are all close to 0. This is plausible because $\beta_2 = 0.0566$, $\beta_5 = 0.0196$, and $\beta_6 = 0.263$ are all small coefficients and imply a weak correlation anyways, so it is possible that our coefficients are slightly positive due to random sampling variation.

Conclusion

Through our Lasso regression analysis, we were broadly able to answer our research question to identify key predictors of poverty using the 3 MPI dimensions as a guideline. We found that the predictors that had the largest effects on *sqrt(Poverty Headcount Ratio)* were *log(Government health expenditure per capita)*, *Voice and Accountability*, *Probability of dying among youth ages 20-24 years (per 1000)*, and *Political Stability and Absence of Violence/Terrorism*, these indicators stemming mainly from the categories Health and Governance. However, given that the coefficients for *Voice and Accountability* and *Political Stability and Absence of Violence/Terrorism* were contrary to what we expected, our ability to identify these as true key predictors of poverty is limited for the reasons discussed in the previous section. It is also important to note that while we have identified variables that are highly effective in predicting *sqrt(Poverty Headcount Ratio)*, their significance may be understated due to the limited availability of educational data. In addition, another limitation of the model is that several indicators in our model are collinear, especially the governance indicators. While the Lasso regression does adjust for this limitation by reducing the coefficients of correlated features (and in cases of multicollinearity, reducing to 0), the collinearity could still potentially distort the coefficients. In other words, although our regression was extremely effective in predicting *sqrt(Poverty Headcount Ratio)*, due to all the limitations discussed above, our initial research question cannot be answered with complete confidence. This further raises the importance of feature selection, or selecting the most appropriate features for modeling without redundancy or collinearity. If we were to have chosen a wider range of non-collinear, independent indicators from the original World Development Indicators dataset, we would have likely obtained more concrete results and a more complete understanding of the factors influencing poverty.

Works Cited

[1] UNDP (United Nations Development Programme). 2023. 2023 Global Multidimensional Poverty Index (MPI): Unstacking global poverty: Data for high impact action. New York.

[2] Cuenca García, E.; Navarro Pabsdorf, M.; Moran Alvarez, J.C. Factors Determining Differences in the Poverty Degree among Countries. Resources 2019, 8, 122. https://doi.org/10.3390/resources8030122

[3] World development indicators. Washington, D.C. : The World Bank.

[4] Kaufmann, Daniel and Kraay, Aart and Mastruzzi, Massimo, The Worldwide Governance Indicators: Methodology and Analytical Issues (September 2010). World Bank Policy Research Working Paper No. 5430, Available at SSRN: https://ssrn.com/abstract=1682130