Examining Worldwide Income Inequality

Abstract

Are the rich becoming richer while the poor become poorer? Income inequality is a hot topic in the media and politics lately. The goal of our statistical exploration was to examine various development indicators from the World Bank and discover how they might be influencing, or influenced by, income inequality as measured by the GINI index. We want to find out the determinants and know which variables are highly correlated with a GINI index score. We hypothesized that the more developed a country, the lower the income inequality would be. First we investigated which year had the most data on the GINI index. Then, we selected variables that were available that year amongst all of the countries containing a GINI index entry. We were left with 93 possible variables from the year 2010, but narrowed it down using variable selection techniques and removing variables that were similar in their measurements or were highly correlated. Then, we split our data set randomly into two, a training set with 75% of the countries for constructing a model, and a testing set with the remaining 25% of the countries for testing the validity of our model. Finally, we analyze our model and discuss how its different factors might lead to inequality worldwide.

Introduction

The Gini index is the measurement of income distribution of residents within a country ranging from perfect equality, represented by 0, to perfect inequality, represented by 1. In our experiment the GINI index is viewed as a percentage rather than in decimal form. The line of equality (show below) is an idealized version of society, which shows the income distributed equally amongst the population. However, in the real world, there is no such thing as perfect equality. The Gini index is a way to measure how much inequality exists. The commonly used measures of Gini index is calculated by comparing the Lorenz curve which shows the total income (percentage) earned by the cumulative population (percentage) with the line of equality which is a 45 degree line (De Maio, 2007), because it represents the ideal case of equality in income distribution. (Appendix 7) For example, the lowest or the poorest 20% of the population is supposed to earn 20% of the total income to have equality. Basically, the Gini index is the area between the equality line and the curve divided by the total area under the equality line (De Maio, 2007).

The Gini index is important because it lets one know how equal the income is in any given country. It is a simple and direct way to see if the income inequality changes over time since the Gini index is a statistical value. Also, it allows us to see the income inequality of the different income classes. This can be seen by looking at how much difference there is between the Lorenz curve and the equality line for certain percentage of the population. For instance, the lower 40% of the population can just earn 15% of the total income depending on how the curve looks up to that certain point of the population. In order to determine the Gini index which is expressed as a percentage for our dataset, many factors are involved and we would like to know which variables would help to determine the Gini index and lead to income inequality.

In this research, we will focus on the questions: What factors lead to income inequality? As a country's economy improves, is there any effect on its income distribution?

Research Design and Hypotheses

We retrieved our data from the World Bank database, specifically the Development Indicator data set. The raw data set contained over 500 variables for each year between 1960 and 2015. We used regression methods to conduct our analysis, and did not intend to include time as a factor. That meant restricting our analysis to a single year.

The World Bank Development Indicator data set is not perfect. Some years do not contain data for every country, and most countries do not contain entries for a given variable. To conduct our analysis, we wanted the largest sample we could find. Since our response variable was the GINI index, we wanted to find the year with the most countries to have entries for it. Using SQL queries (Appendix 4), we found that 2010 had the most entries for the GINI index, with 78 countries present. Then, we found which indicators that all 78 countries had entries for, and came up with a list of 93 possible indicators to analyze. From the list of 93 possible indicators, we manually narrowed down the list to 20 indicators based on logical reasoning to avoid the common problem in statistics -- “Correlation does not imply causation”.

We decided that measures of the economy and quality of life would be good indicators of developed countries. The reasoning is that countries with large economies, as measured by GDP, would have more wealth per person and thus more to share. Countries with higher quality of life would have more of the population with access to education and health care, leading to less children born per person. This leads to our initial intuition based hypothesis:

The more developed a country is (good economy, higher quality of life), the less income inequality there will be (GINI index closer to 0).

From the 20 indicators that were related to our null hypothesis, we trimmed them down to a few. The correlation matrix (Appendix 6) was used to remove correlated variables as to avoid multicollinearity in our final model. The darker and larger a circle is in a box, the more correlated the two variables that index that box. We threw out variables that had low or no correlation with the GINI index.

Data Analysis

Before finding our model, we partitioned our dataset into two for training and testing. The training partition contained 75% of the entries for 2010, and is what we would use to build our model. The testing partition contained 25% of the entries for 2010, and is what we would test our model on for validity and whether it could accurately predict a country's GINI index, and therefore the amount of inequality in a country.

Note: In our analysis, the GINI index is treated as a percentage with values between 0% and 100%. This a 100x scaled version of the actual GINI index, which is between 0 and 1.

Using multiple variable selection techniques (stepwise, backward, and forward), we converged on the following model:

GINI = 24.803 + (.742* GDP GROWTH) + (.207 * AGE DEPENDENCY RATIO TO WORKING AGE) - (.147 * TRADE SERVICES)

The residual assumptions were satisfied with this model, but when we applied the model to our test data to check whether it would correctly predict the GINI index, we found some outliers in our residuals. The standard errors for the prediction values were relatively small (1 - 2%), except for a few countries with large errors (3 - 4%). One country, Luxembourg, stood out as an extreme outlier with an error of 13.3%. We found that the Trade Services variable was causing the regression to be less accurate. In particular, Luxembourg has a ridiculously high GDP, and a major part of its economic success can be attributed to its Banking services (Appendix 1b). We decided to try our model without the Trade Services variable because of its magnitude of error and lower correlation (compared with the other predictors) to the GINI index.

With GDP Growth and Age Dependency as our predictor variables, SPSS gave us the following model from the input ‘training’ dataset:

GINI = 21.345 + (.925 * GDP GROWTH) + (.207 * AGE DEPENDENCY RATIO TO WORKING AGE)

The predictors we have chosen, GDP Growth(%) and Age Dependency (Ratio of older dependents to the working age population) are significant with P-values of 0.000 and 0.001, respectively. From this model, we can tell that GDP and Age Dependency are positively correlated with the GINI index and that as these variables increase, so too does the GINI index (more inequality.) Holding the other variable constant, a 1% increase in GDP Growth corresponds with a .00925 increase in the GINI Index (unscaled, 0 - 1); a 1% increase in Age Dependency corresponds with a .00207 increase in the GINI Index (unscaled, 0 - 1).

We found the model to be satisfactory in predicting the outliers and still satisfied the model assumptions (Appendix 5). We had eliminated correlation amongst predictors at the beginning of our analysis. We found that most of the points on a Normal QQ-plot fell approximately on a straight line which means our residuals are still assumed to be normal. Our fitted vs residual plot has no clear pattern that indicates deviation from constant variance. Our R^2 value was not very high, but given the inconsistencies in the input data, we are happy with the final model.

Conclusion

Based on our final model, we are left with the following conclusions:

Firstly, the GDP growth has a significant impact on the income distribution of a country. In general, a country that experiences an increase in its economy experiences an increase in income inequality. This appears to go against our initial hypothesis that the more developed a country, the better its income equality. As a result, further research which takes into account more factors and time frame may be required. It could be that in the long run, a continual increase in GDP would eventually lead to better income equality.

Secondly, the age dependency ratio to working age of a country's population has a significant impact on its income distribution. This on the other hand appears to support our hypothesis. The higher this ratio is in a country, the greater we would expect income inequality to be.

Though the outcome of our project are different what we have expected at the beginning, oone crucial point that we would like to point out is that GINI index shouldn't be overemphasized as it is one of the index to measure inequality, but not the only one. Other alternatives include Atkinson's index and Hoover index. In the future, if you happen to read an article about GINI and the article claims that some percentage of wealth of the society is held by some percentage of population or the income inequality of a country, think about these questions: What is the difference between wealth and income? What is the relationship between the GINI index of wealth and that of income? Do we really need to concern about the alert level of inequality at 0.4 of the index, which is popular in the media? Is this a good indicator of the economics overall? Remember that nowadays these questions are still debatable among the economists.

Appendix:

1. **References:**

   a. De Maio, F. G. (2007, October). Income inequality measures. Retrieved May 15, 2016, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652960/

   b. Economy_of_Luxembourg. (n.d.). In Wikipedia. Retrieved May 13, 2016, from http://en.wikipedia.org/wiki/Economy_of_Luxembourg

2. **Data Sources:**

   a. The World Bank, World Development Indicators (2010). [Data file]. Retrieved from https://www.kaggle.com/worldbank/world-development-indicators

3. **Description of Variables:**

   a. GINI Index: Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution.

   b. GDP Growth(%): Annual percentage growth rate of GDP at market prices based on constant local currency

   c. Age Dependency (Ratio of older dependents to the working age population): Age dependency ratio, old, is the ratio of older dependents--people older than 64--to the working-age population--those ages 15-64

   d. Population Growth (We converted this variable to categorical data): Negative Growth or Positive Growth.

   e. Trade In Services: Trade in services is the sum of service exports and imports divided by the value of GDP, all in current U.S. dollars

4. **SQL Queries**

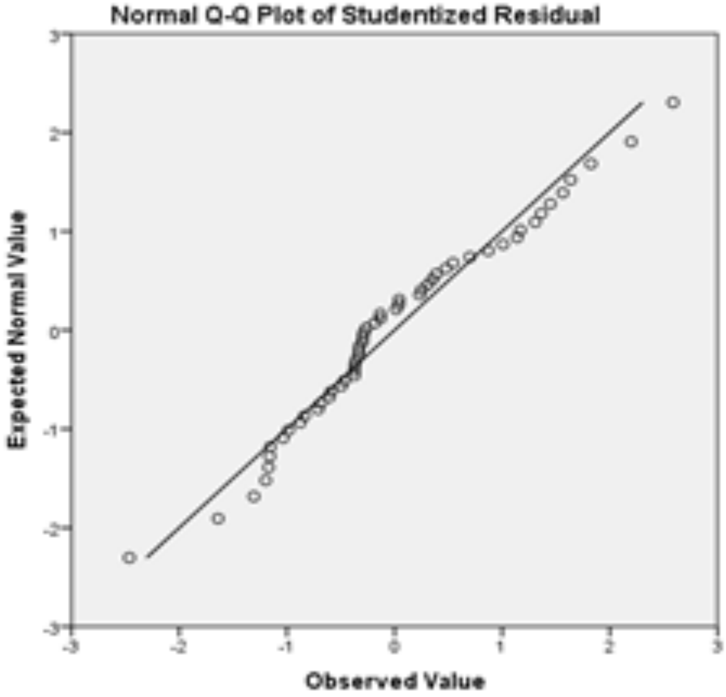
   a. Using the SQLite database retrieved from Kaggle (which was retrieved from World Bank), we found Countries that contained entries for the GINI Index from 2010 using SQL.

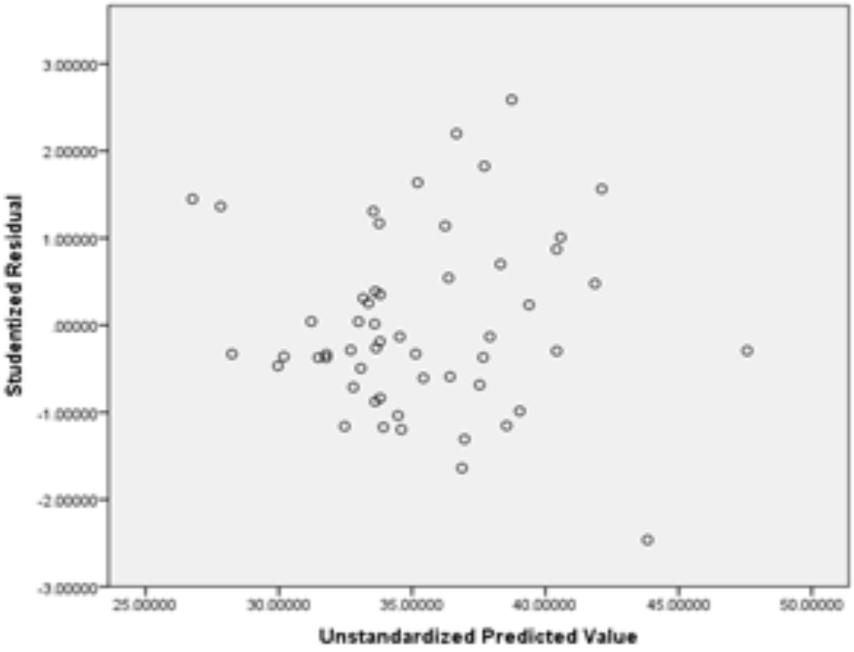
```
SELECT DISTINCT(CountryName) from Indicators where Year = 2010 and
CountryName in (Select CountryName from Indicators where Year = 2010 and
IndicatorName = "GINI index (World Bank estimate)");
```

   b. Then, we found the indicators that all of these countries had in common (to maximize the number of countries in our analysis). From the list returned by this SQL statement, we selected our variables.

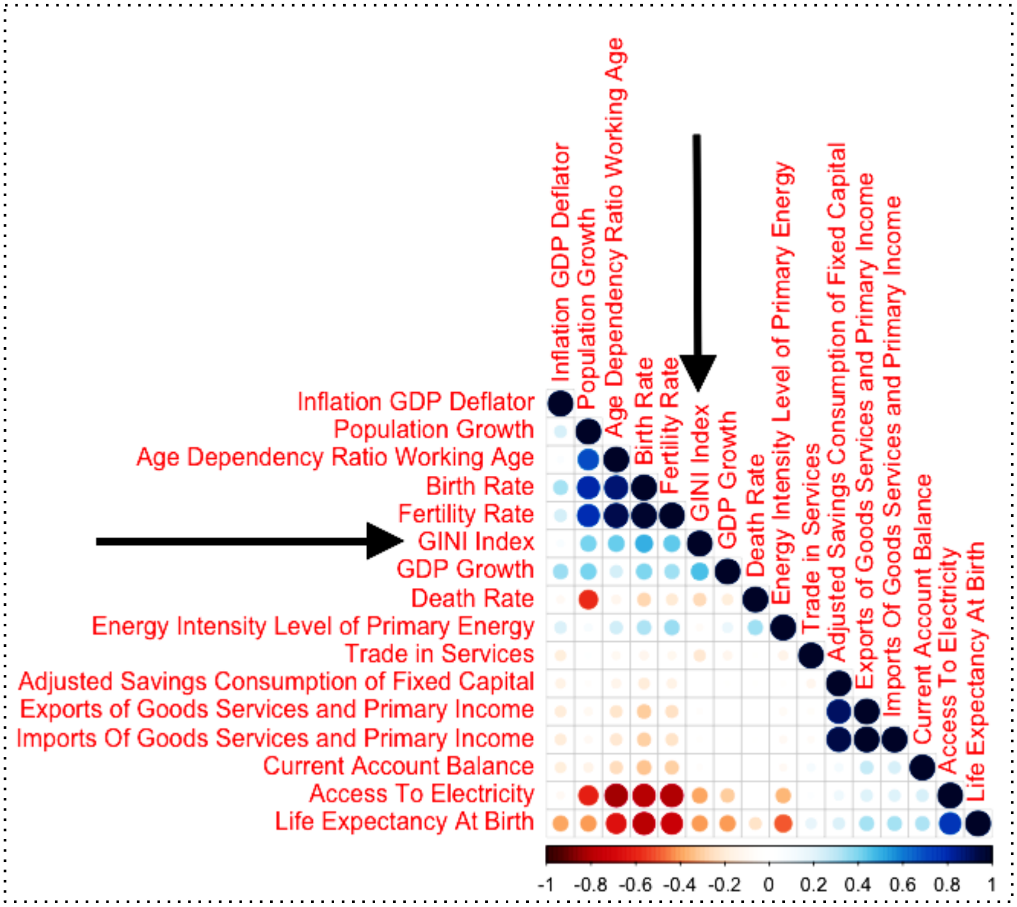
```
SELECT IndicatorName, Count(IndicatorName) as
"NumberOfCountriesWithData"from Indicators where Year = 2010 and
CountryName in (Select CountryName from Indicators where Year = 2010  and
IndicatorName = "GINI index (World Bank estimate)") group by IndicatorName
order by NumberOfCountriesWithData desc;
```

5. Checking linear regression model assumptions:

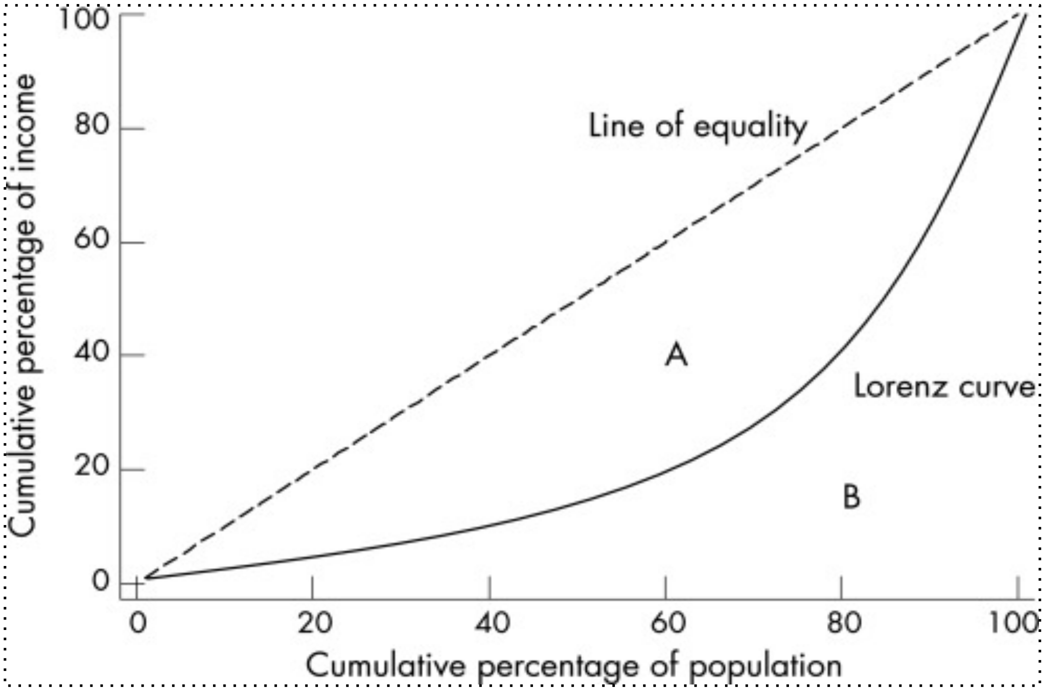




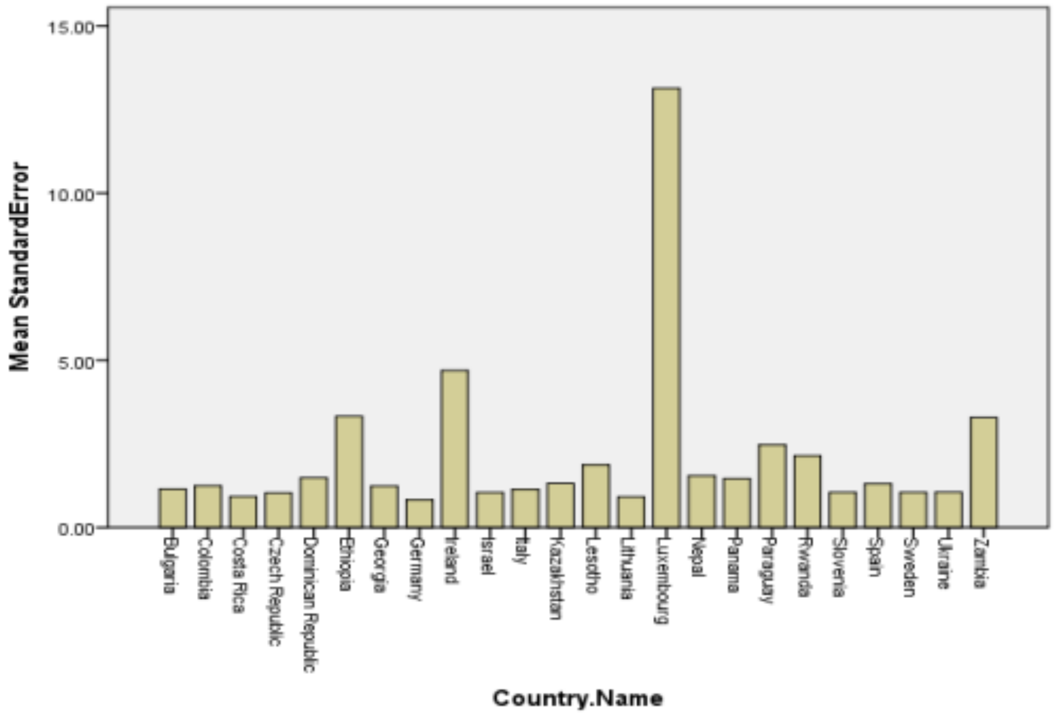
6. Correlation Matrix



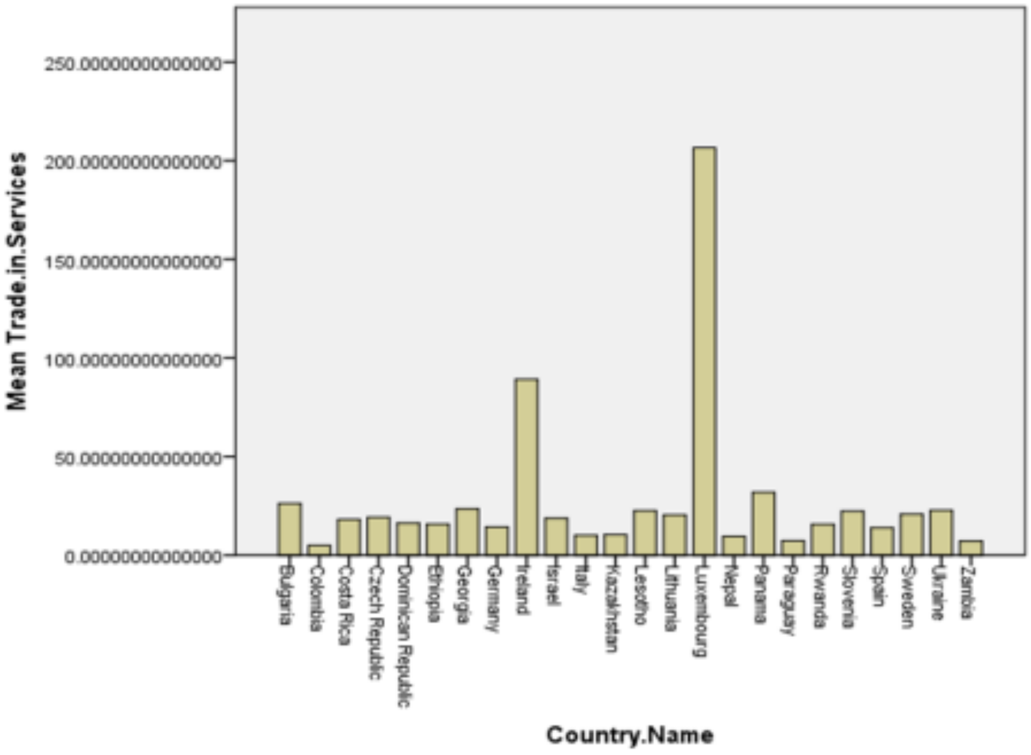
7. Lorenz Curve



8. **Mean Standard Error of the test set**



9. **Mean of the Trade in Services for each country**


Mean Trade.in.Services
Bulgaria
Colombia
Costa Rica
Czech Republic
Dominican Republic
Ethiopia
Georgia
Germany
Ireland
Israel
Italy
Kazakhstan
Lesotho
Lithuania
Luxembourg
Nepal
Panama
Paraguay
Rwanda
Slovenia
Spain
Sweden
Ukraine
Zambia
Country.Name