Introduction

Income inequality, measured by the GINI coefficient, is the unevenness of the distribution of income throughout a population as a result of several factors (i.e., education, globalization, labor markets, wealth concentration, etc.). Due to its growing relevance in today's society, various studies have been conducted to find the relationship between income inequality and other economic variables, such as economic growth defined as the increase in the output per capita per annum. This research has offered contrasting results. Suwoto and Thai on their paper, "Income inequality as a Determinant of Economic Growth", concluded that there is a positive relationship between income inequality and economic growth. Caballo et al. paper tittle, "Income Inequality and Economic Growth" stated that low-income countries are negatively affected by inequality, while for high-income countries, income inequality fosters economic growth. Forbes' paper, "A reassessment of the relationship between Economic Growth and Inequality", defends the view that overall income inequality has a negative effect on growth. All of these studies share a characteristic in common: either they focus on develop or very underdeveloped nations. More importantly, few studies study Central America in isolation, and if they do, they tend to focus more on the political instability and other non-quantifiable factors than statistics and data. This, along with the interest of finding a final answer to the relationship between income inequality and economic growth, has motivated the research question of this study: What is the impact of income inequality on the economic growth of Central America?

Methods

Load Penn World Tables dataset into R. From the main data set, create a sub-dataset containing only countries in Central America excluding Belize. Then proceed to add GINI coefficient data and GNI data from the World Bank Data to your sub-dataset. Now add an extra column to your sub-dataset; in this extra column calculate the average annual growth for country i in time t $\{i,t = 1, 2, 3, \ldots\}$. This is to minimize the effect of short-run recessions when studying the behavior of economic growth over a long period of time. Also make sure all of the other variables correspond to a for country i in time t-1 (i.e., potential predictors in the row of year t correspond to year t-1) as we attempt to uncover what effect the previous state of economic variables has in the future state of economic growth. Subdivide the sub-dataset into test-dataset (1990-1994) and model-dataset (1995-2019).

Proceed with EDA by plotting scatterplots and histograms of each numeric variable in the model-dataset. Observe and acknowledge the existence of influential points with the help of EDA, but don't delete them as that would be unethical. Delete potential predictors that exhibit a strong pairwise linear relationship to avoid multicollinearity, if and only if, they are analogous to each other or one is a factor of another in economic theory (e.g., expenditure side rGDP and output side rGDP). When choosing what potential predictor to delete, delete that that has more variation and/or economic theory states it is less influential to economic growth. Furthermore, observe your scatterplots and histogram to uncover future potential problems.

Fit a linear regression model with the remaining variables as predictors and the average annual growth for country i in time t $\{i,t = 1, 2, 3, \ldots\}$ as the response variable using

model-dataset. Remember to include GINI index as one of your predictors. Run a summary in R of the model coefficients and delete coefficients with large p-values for the t-test. Repeat this until the remaining variables are all influential.

Now, create the residuals versus predictors and versus fitted values plots. If our plots indicate that we have model violation check condition 1 and 2 by creating a response against fitted values plot and a pairwise plot of all predictors respectively. If condition 1 and 2 are met, we apply Box-Cox on predictors and response together. Depending on hypothesis testing, apply transformation to both, neither or response/predictors. After applying transformations, check again for model violations by doing residual plots. Assuming residual plots show no evidence of model violation, create a QQ plot to check normality assumption. If normality doesn't hold, apply Box-Cox as previously stated in this paragraph and later proceed to check all assumptions again.

Assuming all assumptions in the model hold, test different other possible model variations using AIC-based stepwise selection. After finalizing your model, proceed to test it using your test-dataset. Fit the model using test-dataset and observe the coefficients obtained. Furthermore, test assumptions by creating residual plots, QQ plot and plots for condition (1) and (2) as outlined in the previous paragraph. If conditions hold, and coefficients are fairly similar to that of the original model, the model is validated.

Results

Visualizing the data in the dataset gives us clues regarding the problems we will encounter while building our model. Before alteration, our model-dataset contains 151 observations of 20 variables. However, after performing EDA and deciding upon potential predictors using guidelines mentioned in methods part of this investigation, we end up with 7 variables, two of which are used for reference and one which is the response variable. The numerical summary of our variables of interest can be seen in Table 1 below, while the histogram of these variables, Figure 5, is at the Appendix:

Table 1. Numeric Summary of Predictors and Response Variable using Model-data.

Variable Name	*avg_t	*GINI_T_1	*Income_t_1	*hc_t_1	*workers_t_1
Min.	208.90	0.3800	7.946	0.3911	0.8667
1st Quartile	948.10	0.4730	8.482	0.4671	1.7179
Median	1273.2	0.4950	8.828	0.5171	2.2196
Mean	1486.2	0.5023	8.961	0.5390	2.5708
3rd Quartile	2104.6	0.5395	9.449	0.6258	2.9427
Max.	3439.5	0.5910	10.27	0.7182	7.0934

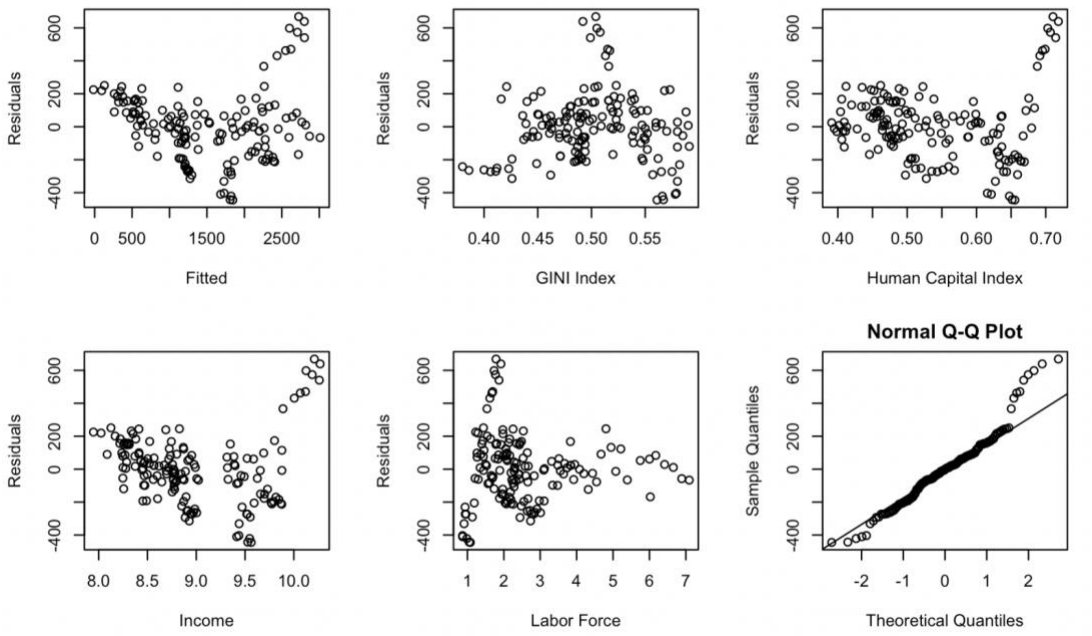
*avg_t is the response variable and stands for average annual growth for country i in time t {i,t = 1, 2, 3, …}.
*GINI_T_1, *Income_t_1, *hc_t_1 and *workers_t_1 are the predictors and stand for Gini Index, Log of Gross National Income, Human Capital Index and Total Workforce for country i in time t-1.

Figure 5 shows that no variable is normally distributed. Similarly, all variables have outliers at different years, more predominantly between 1990-1994.

Running a summary of such model in R, we find out that all coefficients share small p-values for the t-test, with hc_t_1 having the highest of 1.15×10^{-6}. As all of our

variables are influential, we decide to continue with this model and create residual plots and a Q-Q plot for the model. The following is what we obtained:

Figure 1. Residual Plots and Q-Q Plot of Untransformed Model



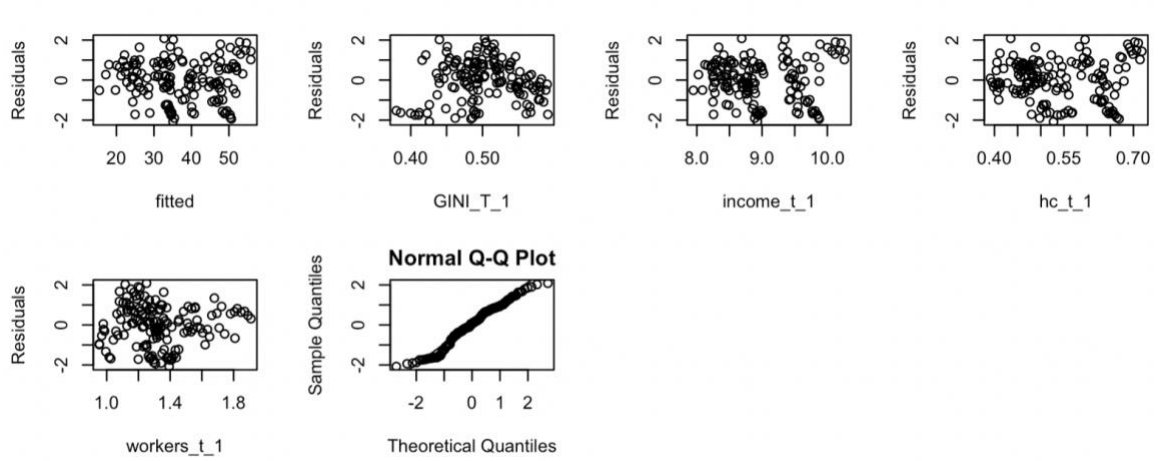
Residual plots show uncorrelated errors assumption doesn't hold for any variable, non-constant variance doesn't hold for Gini Index and Labor force and linearity holds for every variable. Observing the QQ plot also tells us we will have a potential issue with our normality assumption too (as our histograms predicted). Therefore, we continue to plot response against fitted values plot and a pairwise plot of all predictors to check condition 1 and 2. Condition 1 holds and condition 2 holds, henceforth we advance to apply Box-Cox to our model to satisfy our broken assumptions.

The hypothesis testing tells us to apply transformations to both predictors and response, making our this our equation with rounded powers:

$\sqrt{avg_{growth_{t,i}}} = B_0 + B_2 GINI_{t-1,i} + B_3 Income_{t-1,i} + B_4 Human\ Capital_{t-1,i} + B_5 \sqrt[3]{Workers}_{t-1,i} + \in, for\ country\ i, time\ t$. Relevantly, the IQR of the residuals of our model goes from 216.79 to 2.71.

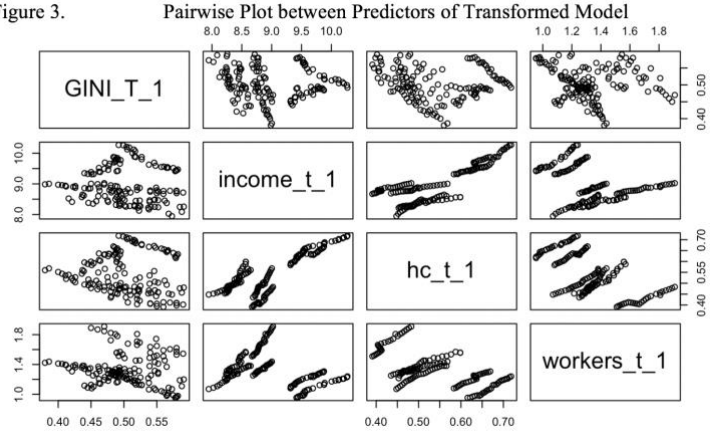
Checking model assumptions for our transform model, we get the following:

Figure 2. Residual Plots and Q-Q Plot of Transformed Model



Linearity holds, uncorrelated errors holds better than untransformed model, non-constant variance is satisfied, and normality holds better than previously.

Figure 3. Pairwise Plot between Predictors of Transformed Model



Now, assessing condition (1&2) from figure 3 we can see that condition 2 holds. We don't worry about multicollinearity as all our VIF are less than 5.

Figure 4. Average Annual Economic Growth versus Fitted Values

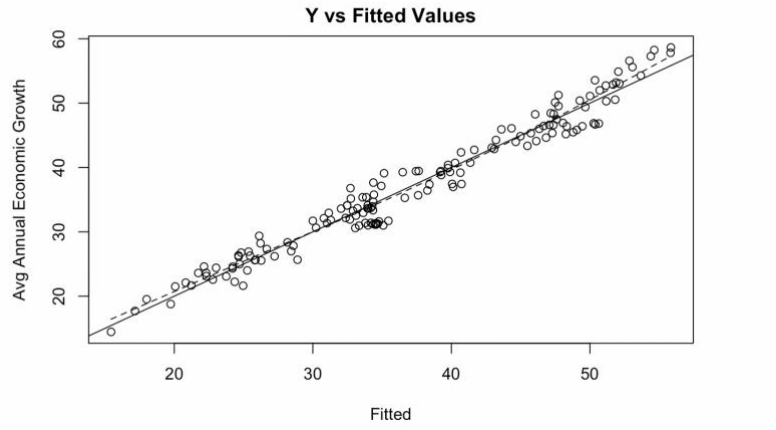


Figure 4 to the left tell us that condition 1 is satisfied. As our dataset doesn't has any more potential predictors after we cleaned it, we skip the step of using AIC-based stepwise selection and go directly into testing our model with the test-dataset.

Our model fails its test. Coefficients are vastly different from that of the model built with model-data. However, the assumptions hold (figure 6,7 at Appendix). This means we overfitted the model and the transformations we did to our model using model-data only worked for such data. A further analysis will be done in the next section.

Final Results

$$\sqrt{avg_{growth_{t,i}}} = B_0 + B_1 GINI_{t-1,i} + B_2 Income_{t-1,i} + B_3 Human\ Capital_{t-1,i} + B_4 \sqrt[3]{Workers}_{t-1,i} + \in$$

	Data Model	Test Model
Coefficients	$B_0 = -187.7854, B_1 = 44.127, B_2 = 20.7252, B_3 = -31.669\ B_4 = 25.643$	$B_0 = -171.769, B_1 = 52.143, B_2 = 12.894, B_3 = 44.444, B_4 = 35.360$
Assumptions	All Assumptions Hold	All Assumption Hold
Residuals	Min = -4.065, IQR = 2.710, Max = 4.005	Min = -12.533, IQR = 4.532, Max = 6.512

Discussion

The results of this paper show that there is a positive relationship between income inequality and economic growth of Central America. Thus, this paper is able to fulfill its goal as it adds knowledge to the existing knowledge of the affinity income inequality and economic in Central America, consequently making these results relevant.

The biggest limitation of the results is that the model cannot accurately predict data outside its train-dataset. This is mainly because the test-dataset been very different from our train-dataset (mainly because between the years 1990-1994 economic growth in Central America boomed and then slowed down significantly while the other variables increased at their normal rates). This means there are many influential points in our test-dataset, which also lead to that failure. Furthermore, the test-dataset is small in econometrics terms, but there isn't much to do about this as there is limited information concerning economic variables of Central American countries prior to 1990. Another limitation is the multicollinearity present in the model. When creating a model that predicts economic growth, multicollinearity is somewhat expected as most economic variables are intercorrelated. Lastly, a big limitation of the results is some of the data used for these results is extrapolated as some Central American governments didn't release official economic data before the past decade, and some don't do it now because of political reasons. All of these limitations are not corrected as they are out of my control and changing the data-points to better fit my model would be unethical.

Citations

Barro, R. J. (1991). Economic Growth in a Cross Section of Countries. *The Quarterly Journal of Economics*, *106*(2), 407–443. https://doi.org/10.2307/2937943

Kwon, J. K., & Paik, H. (1995). Factor Price Distortions, Resource Allocation, and Growth: A Computable General Equilibrium Analysis. *The Review of Economics and Statistics*, *77*(4), 664–676. https://doi.org/10.2307/2109814

Caraballo Pou, M. Á. (2017). Income inequality and economic growth revisited: A note. *Journal of international development : the journal of the Development Studies Association, 29*(7), .

Forbes, K. J. (2000). A Reassessment of the Relationship between Inequality and Growth. *The American Economic Review*, *90*(4), 869–887. http://www.jstor.org/stable/117312_

Suwoto, & Zhai. (n.d.). *Income Inequality as a Determinant of Economic Growth: A Cross-Country Analysis*. Retrieved October 24, 2021, from https://smartech.gatech.edu/bitstream/handle/1853/56034/tsuwoto_yzhai_econometric_analysis.pdf.

Appendix

Figure 5. Histogram of Response Variable and Potential Predictors

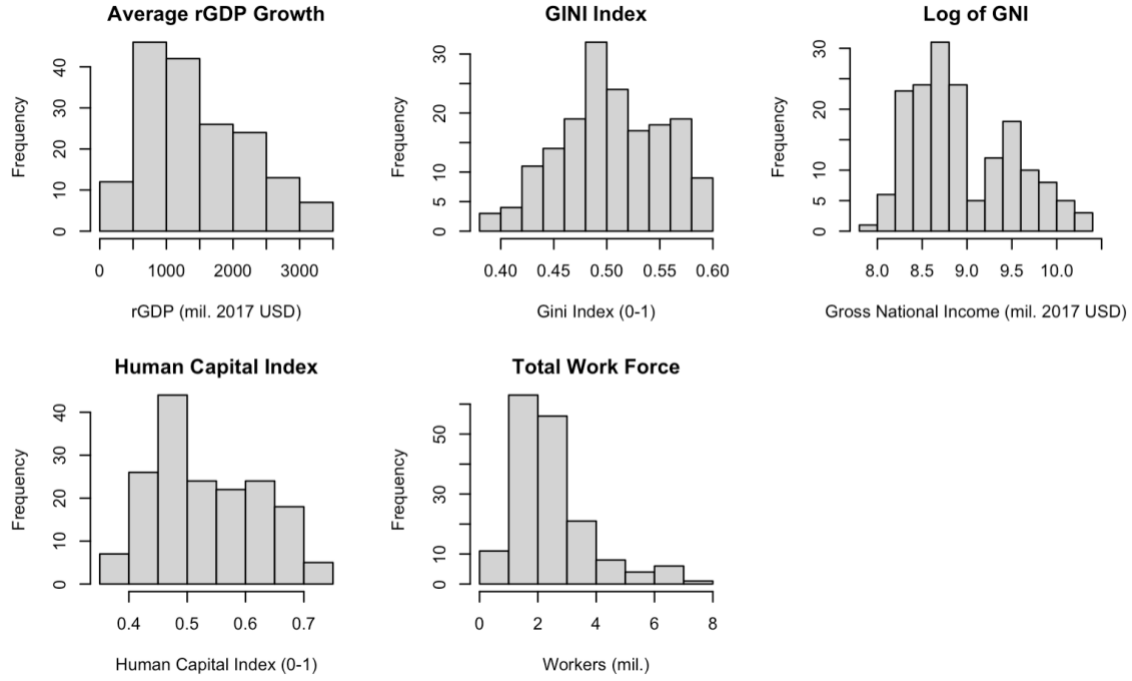


Figure 6. Pairwise Plot between predictors of Transformed Model using Test-Data

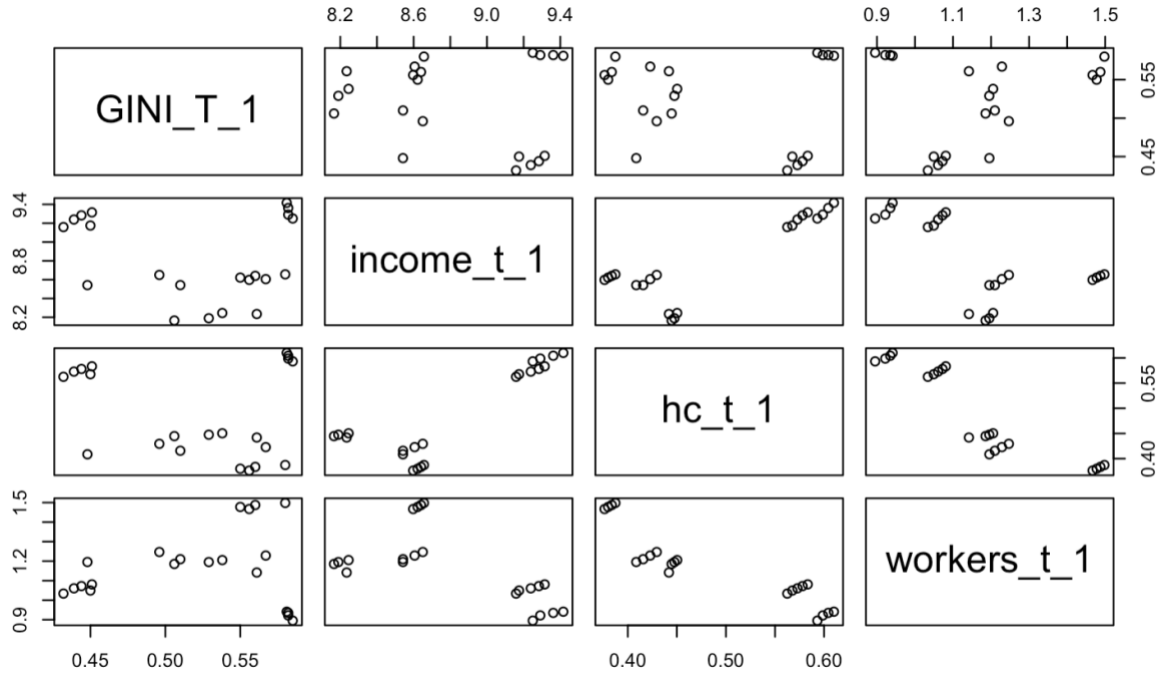


Figure 7. Residual Plots and QQ Plots of Transformed Model using Test-Data