IDS 702 Team Project

Part II: Statistical Analysis Plan

Yellow Team - Suzy Anil, Sukhpreet Sahota, Xianchi Zhang, Yuanjing Zhu

2022-11-04

Overview

Our dataset was selected from the World Inequality Database, which is one of the most extensive databases on the evolution of world distribution of income and wealth within and between countries. The database is open-access and has compiled valid data from national databases, surveys, fiscal data, and wealth rankings. With its vast array of features, there are many key economic and social inequality questions that could be answered with access to this data. Our group has decided to focus our statistical analysis on the impact of certain economical features on a nation's carbon footprint (Total National CO_2 Footprint). For clarification, a nation's total carbon footprint is equal to the combination of CO_2 footprint and footprint of other greenhouse gases.

This leads to two distinct research questions:

1. **How do income brackets (top 10%, middle 40%, bottom 50%) affect a nation's carbon footprint?** From this inference question, we hope to understand how the difference economic status groups affect a nation's carbon footprint. Our assumption would be that emissions are comparable across the three income brackets. However, there has been sufficient evidence that the wealthiest bracket emits notably more tons of carbon compared to the bottom bracket *(https://ourworldindata.org/co2-by-income-region)*.

    a. H_O: Carbon emissions are the same across all income brackets.

    b. H_A: Carbon emissions vary across income brackets.

2. **How have changes in a nation's average wealth affected their carbon footprint over time?** For this analysis, we are looking to predict whether a country is a high or low carbon emitter. Our hopeful end goal is to use the model created around this question to predict whether other countries (unseen data) are high or low carbon emitters.

Models

To better quantify the impact to a nation's carbon footprint by each income bracket and understand the relationship between CO_2 levels and income bracket, we will be using a multi-linear regression model. As mentioned above, the regression output/outcome is the amount of total carbon emission, which is a continuous variable. For this question, we also have multiple predictor variables since the income brackets are separated into three categories: top 10%, middle 40% and bottom 50%. During our EDA, we found a linear relationship between GDP and CO_2 and found that the most of wealthiest countries tend to have the higher CO_2. While this was the general conclusion, we also observed a few wealthy countries did not show positive trends. Hence, we want to discover the impact of the relationship based on the separate income brackets.

We will implement an ordinal logistic regression model as we look at the different factors/predictors that help categorize/classify a country as a high vs low carbon emitter, resulting in a binary outcome. We will classify high or low carbon emitter countries based on average carbon emissions across all 10 observed countries over the 20-year period (2000 – 2020). We will then compare our mean/average carbon emission to each country and categorize them depending on whether they fall above or below this threshold. For this reason, we will also use country as an interaction term to ensure it is included within our model as well as to generate one overall model.

Variable Selection

From the original vast dataset, we distinguished several key predictors for our analysis based on prior research. These predictors are the 10 selected countries, year (from 2000 to 2020), total population of respective nation, national income for each nation, the respective nation's gross domestic product, the income inequality within a nation (which is denoted as income brackets: top 10%, middle 40%, bottom 50%), the respective nation's market-value national wealth (which is defined as the total value of assets owned by the national economy minus its debts), and the respective nation's wealth-to-income ratio (which assesses the net national wealth to the net national income). To distinguish among a set of possible models using the above mentioned predictors, we plan to detect any outliers and repeatedly perform backward selection using AIC (Akaike Information Criterion) as our model selection criterion to find the most optimal multiple linear regression model. Backward elimination has the advantage of allowing to evaluate the combined power of the predictors since all predictors are included in the model from the beginning of the process. It can also eliminate the least significant factors early on, leaving the model with only the most significant predictors. While our sample size is small, we will generate the most representative logistic model by understanding the effectiveness and the accuracy of our model via a confusion matrix as well as using a ROC curve/AUC before applying the model on test data (or data point).

Challenges

There were some challenges we addressed in our exploratory data analysis that presented issues when working with the World Inequality Database. For this plan, we dove a little deeper and proposed some solutions for these challenges.

1) To obtain data from the database, we needed to select which predictors to retrieve and from those predictors, which countries and years we wanted. We specified the ten distinct countries based on continent and national income based on the obtained data from 2000-2020. However, to make our model better, we wanted to include additional years starting from 1980 (enabling us to observe trends over the last 50 years compared to 20 years). Unfortunately, a lot of possible predictors did not contain data from that far back. It can be assumed that for smaller countries, some key data points may not have been/are not available due to limited resources. In addition, we found that as the World Inequality Database has become more widely used for research purposes, data collection of interesting data points varies and is only available and can only go as far back based on the sources/when the variable was created. While we may have preferred to garner a more significant understanding/trend, we are unable to due on this limitation.
2) The standardization of the data across countries due to multiple factors: population size, currency, etc. While we have tried to control these factors by including population as a variable and normalizing all financial data to US dollars, this is an important limitation to outline now and remember for future model creation, etc.
3) The resistance to including more variables was due to the fact that many variables are dependent on the primary independent variables, resulting in multi-collinearity. For example, wealth income is made up of assets and debt, so we were unable to include those in our analysis. Therefore, our solution is to include the predictors that are aggregated/composed of the dependent variables for our model to be more well-rounded.