DSC180B - Capstone Project Report

"Causal Effects of Socioeconomic and Political Factors on Life Expectancy in 166 Different Countries"

Emily Ramond, Max Levitt, Adam Kreitzman

ABSTRACT

This project examines causal relationships between various socioeconomic variables and life expectancy outcomes in 166 different countries, with the ability to account for new, unseen data and variables with an intuitive data pipeline process with detailed instructions and the PC algorithm with updated code to account for missingness in data. With access to this model and pipeline, we hope that questions such as "do authoritarian countries have a direct relation to life expectancy?" or "how does women in government affect perceived notion of social support?" will now be able to be answered and understood. Through our own analysis, we were able to find intriguing results, such as a higher Perception of Corruption is distinctly related to a lower Life Ladder score. We also found that higher quality of life perceptions is related to lower economic inequality. These results aim to educate not only the general public, but government officials as well.

1. Introduction

In this project, we aim to establish causality between various socioeconomic variables and life expectancy outcomes in roughly 166 different countries, noting the strongest connections between economic and political factors with the length of life expectancy. This report accompanies a user interface which displays the main causes of life expectancy for each selected country in order to educate the general public. [to be finished after conclusion]

- The point/ reasoning for our website / its function / repeated goal/outcome.

2. Background

The World Happiness Report (WHR) is published by the United Nations Sustainable Development Solutions Network and contains rankings of national happiness based on survey respondents' ratings of their own lives. Most of the variables are answered on a scale, such as from 1 to 10, or taken as a national average from binary {0,1}, {yes, no} responses and are used to track progress of various countries. The survey and summaries are released on a yearly basis. Included in the WHR is the World Health Organization's (WHO) life expectancy at birth value. This prediction is concluded based on a vast variety of indicators, as shown on the website in the References section. The WHO states that life expectancies are actually getting longer, with an eight percent increase in the past twenty years. The data, along with their indicators, are available below, but are not directly used in our project.

3. Data

<u>Background</u>

We combined several different datasets in order to gather enough data to properly examine our question of choice. To begin, we utilized the World Happiness Report (WHR)

(https://worldhappiness.report/faq/) which details, based on surveys, various factors that contribute to a country's overall happiness and wellbeing. Below are some examples of important variables we examine:

Life Ladder: asks the surveyor to rank their overall life on a scale of 1 to 10

Log GDP per Capita: the value of goods and services produced by the nation's economy

Healthy Life Expectancy: World Health Organization's life expectancy at birth.

Social Support: National Average of Binary Response {0, 1} to being asked if there are friends or relatives you can rely on if you needed help.

Freedom to Make Life Choices: Are you satisfied or dissatisfied with your freedom to choose?

Generosity: National Average of have you donated money to a charity in the past month?

Perceptions of Corruption: National average of response to is corruption widespread throughout the government or businesses?

If you would like to read in more detail about these and other variables, as well as some summary statistics, please see this PDF: https://happiness-report.s3.amazonaws.com/2021/Appendix1WHR2021C2.pdf.

It is important to be transparent about the validity of data and the impact of our results. Because the World Happiness Report is gathered using surveys, and does not survey the entire population, there is room for bias to make a negative impact on the validity of the data. The WHR states that, with confidence, they survey enough of the population in each country in order to have a reasonable estimate of the national average. Thus, we believe that the data is still valid enough to make assumptions, but should be done so with recognition of the possible bias.

Our second dataset comes from the Comparative Political Data Set (CPDS) which includes a collection of political and institutional data from 1960 to 2019. Because the dataset is fairly large, we will not go into detail about each variable and instead encourage you to visit https://www.cpds-data.org/images/Update2021/Codebook_CPDS_1960-2019_Update_2021.pdf to read more about the specific variables used. In summary, the dataset includes detailed information on party composition, reshuffles, duration, types of government, and additional economic, socioeconomic, and demographic variables to add to the World Happiness Report and our examination of the connections between these variables and Life Expectancy.

Lastly, our project is built as a pipeline, ready for any additional data to be added and utilized by the algorithm. This means that as more reliable and detailed data is made publicly available, we can easily update our algorithm and output to accommodate this new knowledge, and change the conclusions of the report accordingly.

Data Overview

For our analysis, we had three main sources of data including World health report, Comparative Political Data Set or CPDS, and Income inequality data set. The World health report was our starting data set which had variables such as GDP, Life expectancy, Life Ladder, a measure of happiness, and more. This data set was a good starting point because it had good measures of quality of life, which would end up being our main variables of interest. These variables are life expectancy and life ladder score, which is a measure of happiness. The next dataset we added in was the Comparative Political Data Set, or CPDS. This data set includes features such as the

amount of right and left wing members of government, government type, women percentage in government, and more government related features. This data set was useful to us to be able to use our analysis to focus on how governmental factors impact quality of life. Our last data set that we included was an income inequality data set. This data had features such as Gini coefficient, median income, and poverty rate which are all measures of inequality and wealth. This dataset was useful for us to be able to identify how financial factors affect quality of life.

Cleaning Data

The first step of cleaning the data was converting all values to be numerical. Next I got rid of all columns that did not have quantitative data. I then added in code to change some common country name spellings to the ones utilized by our data sets. To see the naming conventions for all countries in our dataset, you can refer to the list here. My code uses regular expressions to detect the index columns for country and year if it is included. My code will then remove all entries missing an index value since their data is not usable. These cleaning steps will be repeated for all data sets that are loaded in and the datasets will be assigned into either data with year or data with no year. All datasets with year our then combined together. This data will then be reformatted to be indexed by country name with values for every feature for every year. To remove all missing values, if a country has no data for any feature, then that country will be removed from the reformatted data. If a country has missing values, then they will be predicted with linear regression based on the other feature values for that country. This process results in the reformatted data having no missing values. Once all the data with years is reformatted, the data with no years will then be added in. Whenever data sets are merged together, to retain all features with no missing values, only countries present in both data sets will be kept in merged

data. After all the inputted data is merged and reformatted, it will then be written to a csv file name of the user's choosing with the final_data folder in the src folder. The outputted data of this cleaning function will be in the correct format to have the PC algorithm run properly on it.

When adding in your own datasets there are a few prior cleaning steps that may have to be done. There must be a column with country names that has "country" in the column name. The country names also must follow the same naming conventions as other data that you are merging your data with. If you wish to have an included column with years, then "year" must be in the column's name. When combining datasets with a year column, the pipeline will only include years that are in both datasets, so make sure the years are overlapping in their span. For any other unexpected issues that you encounter with using our pipeline, feel free to reach out to Max Levitt for assistance at mglevitt@ucsd.edu.

4. Algorithm Methods

We will be using Independence-based causal discovery, leading us to use the popular PC algorithm, created and named after Spirtes and Glymour, that utilizes the idea that two statistically independent variables are not causally linked. We start with nodes all connected by undirected edges, ensuring there are no set distributions in place. We also use a significance level of .05. The lower the significance value, the smaller number of edges there will be in the graph. There are three steps the PC algorithm cycles through.

Step 1:

In the first step, the PC algorithm identifies the skeleton by starting with a complete undirected graph. It removes the edge X-Y were X |_| Y | Z for some Z. In other words, edge X - Y is deleted if the corresponding variables X and Y are independent.

Step 2:

After step 1, all left connected edges go through conditional independence testing. If there is no dependence, the variable is a separation set. This continues until there are no tests left to run.

Step 3:

Now for any paths X-Z-Y in our working graph from the previous step, where

1. There is no edge between X and Y as determined by Step 1
2. Z was not in the conditioning set that makes the variables X and Y conditionally independent.
3. If these are true, then we know X-Z-Y forms an immortality

Any edge Z-Y part of a partially directed path of the form X→ Z - Y where there is no edge connecting X and Y can be oriented as Z -> Y.

Once these steps have been cycled through and completed, and there are no more checks to run, the graph with directed edges is complete and we can start inferring information from that graph and associated variables, and in this case, causal dependencies have been oscillation patterns. It is important to note that a downside of conditional independence testing is that it is only efficient with infinite data.

[update conversation from adam - talking to developers of algorithm about missing PC; our results since trying another algorithm and/or accounting for missingness]

5. Conclusions

To facilitate better understanding of our results, we decided to designate certain variables that we wanted to analyze as "variables of interest," these were the following: Life Ladder, Social Support and Freedom to Make Life Choices.

Our logic was that these three variables would allow us to explore how certain governmental policy and leadership could influence how long people live, as well as how people's individual choices could impact other variables as well.

We found the following relationships:

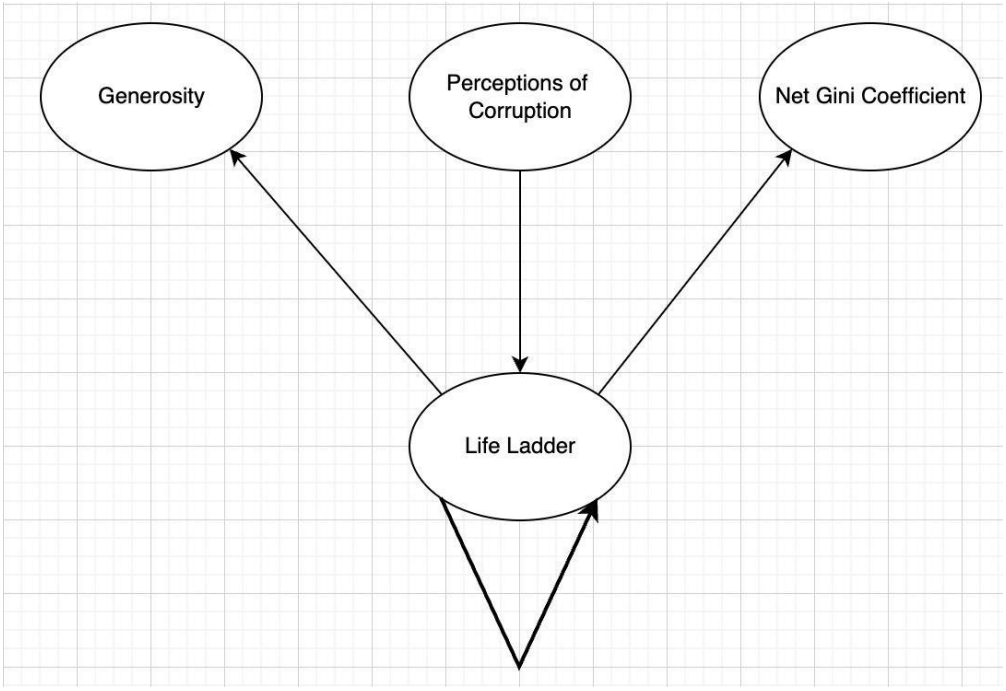

Figure 1: Life Ladder

For the first variable, Life Ladder (see Figure 1), the biggest connection that we witnessed was with itself, meaning there was a causal relationship between countries with a high life ladder and maintaining a high life ladder in the future. The next variable that influenced Life Ladder was

Perceptions of Corruption. Since the algorithm itself doesn't specify whether the causality is positive or negative, we plotted Life Ladder against Perceptions of Corruption to see the nature of the relationship and found that a higher Perception of Corruption was distinctly related to a lower Life Ladder. In other words, the more corruption in a country, the lower quality of life and life expectancy in that country.

Life Ladder was also causally related towards two other variables. The first of these variables was Generosity. This meant that a higher quality of life caused people to be more generous towards others. Life Ladder also was causally related with the Gini Coefficient, which meant having a higher quality of life was causally related with Gini Coefficient. We also plotted Gini Coefficient against Life Ladder to see the nature of the relationship and saw that the higher the Life Ladder score, the lower the Gini Coefficient. This meant that a higher quality of life caused lower economic inequality.

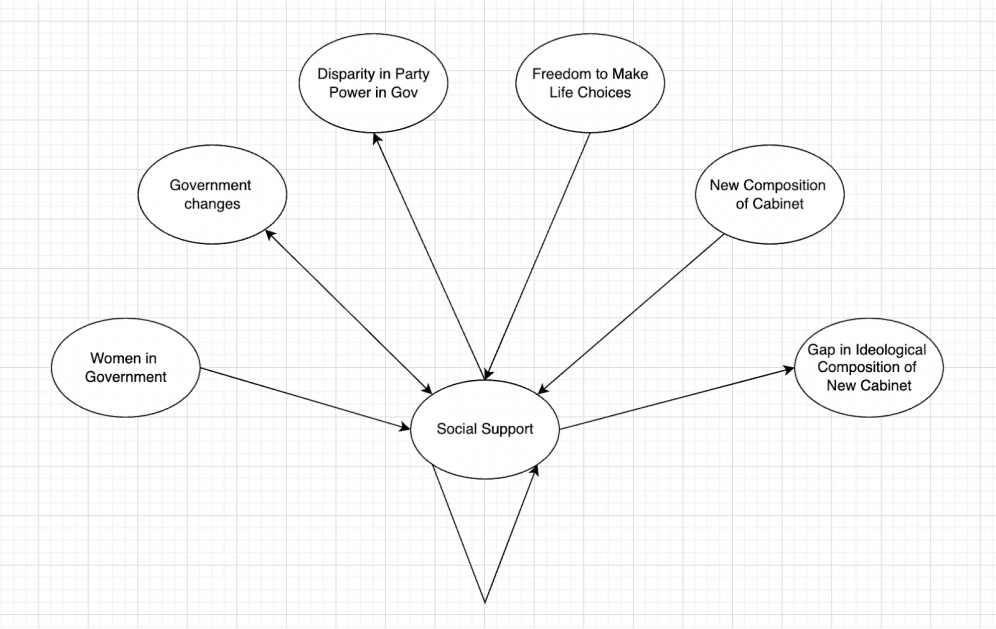

Figure 2: Social Support

Our second variable of interest was Social Support (see Figure 2). We found that it was causally influenced by four variables other than itself (Countries with higher Social Support typically maintained a high Social Support year after year). The first variable was Women in Government. This was not too surprising to see as women are typically more liberal than men and thus more likely to support governmental policies that lead to more Social Support for people. The second and third variables were changes in the government and a new majority power in government. This essentially signaled that changing government and having new people in power leads to more Social Support. This makes sense, because people typically campaign on promises to help people and engage voters to vote for them for these reasons. Finally, Freedom to Make Life Choices was causally related with Social Support as well, meaning that if people have more freedom in a country, it causes them to enact more systems of social support.

Additionally, Social Support causally influenced several variables. The first of these was a disparity in government power. This means that higher Social Support caused the power dynamic in government to be more in favor of one party. This is likely due to the fact that countries with a higher level of social support are more likely to be happy with those who are in power in government. The second variable was a gap in the ideological composition of the new cabinet, which is similar to what was discussed in the paragraph prior.

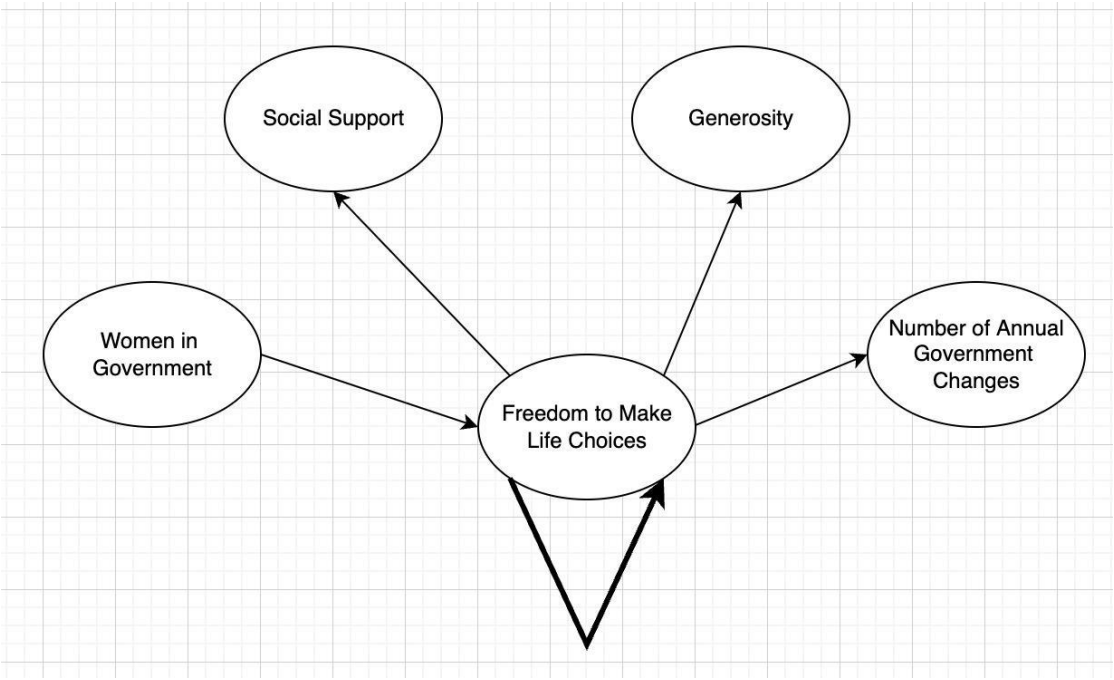

Figure 3: Freedom to Make Life Choices

Our third variable of interest was Freedom to Make Life Choices (see Figure 3). We found that it was highly causally related with itself, meaning that countries that had higher freedoms were likely to maintain freedoms, while those who did not typically stayed in a more authoritarian environment. The one variable we were able to find that causally influenced freedom was having more women in government. On the surface, this definitely makes sense, as typically more authoritarian governments don't have women in power, and most if not all dictators throughout history have been men.

Freedom to Make Life Choices also causally influenced three other variables. Which were Social Support, Generosity and the number of annual changes in the government (i.e. number of elections). This meant that when people were given more freedom over their own life choices, they typically chose to be more generous, and push for more Social Support in their country. It

also led them to want to vote more often so that they could hold elected officials more accountable.

Putting everything together, it is apparent that certain pathways exist that can inform us how different societal measures impact our variables of interest. For instance, we can see that Corrupt governments will lead to a much lower quality of life, whereas giving people more power over their own choices will lead to higher levels of social support and generosity, which indirectly leads to a better Life Ladder (quality of life) score.

Thus, from our results, we can deduce that the best way to set up a society for a high quality of life is to have a transparent government with frequent elections and for the government to include a representative body of people. This will generally lead to people voting for their own self-interest, which will lead to a number of positive outcomes, such as higher quality of life, more generosity among individuals, more social support, and lower economic inequality.

6. References

Center for Causal Discovery. *Tools*. (2020, August 12). Retrieved November 23, 2021, from https://www.ccd.pitt.edu/tools/.

Klaus Armingeon, Sarah Engler and Lucas Leemann. 2021. Comparative Political Data Set 1960-2019. Zurich: Department of Political Science, University of Zurich

WHR Home. (n.d.). Retrieved March 6, 2022, from https://worldhappiness.report/

Wikimedia Foundation. (2022, March 2). *World happiness report*. Wikipedia. Retrieved

March 6, 2022, from https://en.wikipedia.org/wiki/World_Happiness_Report#:~:text=The%20World%20Happiness%20Report%20is,(quality%20of)%20life%20factors.

World Health Organization. (n.d.). *Themes*. World Health Organization. Retrieved March 6, 2022, from https://www.who.int/data/gho/data/