Covid 19 Analysis – from the Perspective of Macro Metrics

Introduction

This project aims to analyze linear relationship between the effectiveness of different countries' responses to Covid-19 pandemic and macroeconomic metrics. The macroeconomic metrics consist of the feature vector. The effectiveness, which would be defined later, is the label. After analyzing the relationship with all features, reduce the dimensions of feature vectors by Principle Component Analysis (PCA) to analyze linear relationship with subset of features. In terms of the beneficiaries, as different macro measurements would demonstrate national power from different aspects, if there is strongly or approximately linear relationship between the input and output, this would possibly imply the existence of macro policies that reduce the cost caused by current plague and lower the probability of such disaster. Therefore, the primary stakeholders would be the policymakers, people working in medical and educational sectors, and the public. One potential ethical problem could be data privacy when macro metrics is gradually refined. Possibly, the access to more detailed data should be restricted.

Methods

1. Data Representation

For the linear regression task, each (x^i, y^i) pair is the data from a country. In the preliminary work, five macro metrics are used as a feature vector $x^i = (x_1^i, x_2^i, x_3^i, x_4^i, x_5^i)$ for instance i. For the dataset, X stands for matrix with each instance's feature vector being one row. Specific features are in the following table.

x_1	Human Development Index (HDI)[1] of a country
x_2	Average years in school of the country
x_3	Share of the population with access to electricity of the country
x_4	Share of the population with access to clean fuels and technologies for cooking of a country
x_5	Literacy rate

The way of representing effectiveness (labels) is similar to what has been done in P7 (calculate x, y). However, in this project, there are 4 values aggregated to represent the label rather than features. These four values are represented as $(y_1^i, y_2^i, y_3^i, y_4^i)^T$ for instance i. $y_1^i$and y_2^i correspond to x and y computed from calculate_x_y(time_series) in P7. Similarly, if n is number of Covid 19 cases confirmed, the days of cases growing from n/1000 to n/100 is y_3^i, and y_4^i is the days of cases growing from n/10000 to n/1000. Label vector y would be $\sum_{i=1}^{4} y^i$. Then, the larger y is, the more effective a country's responses are since it took more time for number of cases to achieve exponential growth. For PCA, input is X.

2. Data Collection

Data were gathered from the reports[2] of United Nation Development Programme x_1, the git repository[3] of Search Johns Hopkins University y on May 4th, and websites[4] of the organization of Our World in Data (x_2, x_3, x_4, x_5). All data is stored in csv files.

3. Data Preprocessing

Since not all the countries have all the features, it is necessary to preprocess the dataset. After preprocessing, there are 119 countries/regions with all 5 features mentioned above. Convert data into matrix and vector. Then X is a 119

[1] https://en.wikipedia.org/wiki/Human_Development_Index - Explanation of HDI

[2] http://hdr.undp.org/en/data - HDI source

[3] https://github.com/CSSEGISandData/COVID-19 - case source

[4] https://ourworldindata.org/ x2 – x5 come from here

by 5 matrix and y is a 119 by 1 vector.

4. **Model Training and Dimension Reduction**

   Two tasks are mainly done via Python Machine Learning tool kit Scikit-learn[5] and linear-algebra library numpy[6].

   a) **Model Training**
      1) Normalize each feature in X.
      2) X and y are shuffled randomly. Take first 100 instances as training set. The rest of 19 would be test set.
      3) Feed training set data into learning regression model. Then use test set data to evaluate coefficient of determination (R2 score).
      4) Since R2 score is a random variable due to shuffling X and y, repeat from the first step for 10000 times.
      5) Calculate and record the mean R2 score.

   b) **Dimension Reduction**
      1) Try use single feature as the only feature to train the linear regression model.
      2) Let reduced dimension $r = 1$.
      3) Reduce dimensions of feature matrix by r with PCA.
      4) Do model training specified in a).
      5) r = r + 1.
      6) Repeat from 3) until r is 5.

5. **Analysis**

   Based on R2 score for a linear model, analyze the performance or accuracy of the model. Compare different models trained differently. Find out potential problems for the linear model.

Preliminary Result and Analysis

1. **Linear Regression Results**

For the linear regression model trained with all 5 features, the R2 score is around 0.15 – 0.17.

2. **Dimension Reduction & Linear Regression Results**
   1) For the linear regression models using only one feature, different models have different R2 scores. They are shown on the right table.
      (R2 average for 100 times of training shown in "()")
   2) If using PCA to reduce the dimension of the feature matrix X, then train the linear regression model, R2 scores are shown on the right.
      (Only average for 100 times of fitting is shown.)

HDI	0.150 – 0.161 (0.15)
Average years in school	0.027 – 0.035 (0.030)
Access to electricity	0.148 – 0168 (0.157)
Access to energy for cooking	0.065 – 0.080 (0.073)
Literacy rate	-0.003 – -0.010(-0.007)

1 component left	0.103
2 components left	0.088
3 components left	-0.124
4 components left	-0.160

3. **Analysis**
   1) About the model. The linear relationship seems to be weak for all the features. However, with more components left, the linear relationship becomes stronger and stronger (The absolute value of R2 is becoming larger). This possibly suggests that with proper selected features, linear model could potentially be used to measure the effectiveness of responses in different countries. The features used here are only a small fraction of macro statistics. Nonetheless, it is unclear about the potential of linear model for this task.
   2) HDI policy: As HDI is a comprehensive index that measures the living standard of the public in a country, it gives a more comprehensive description of the performance of the overall economy. It has a stronger linear relationship with the effectiveness of responses to Covid 19. However, raising HDI is hard to be done in a few

[5] https://scikit-learn.org/stable/ - Machine learning library

[6] https://numpy.org/ - linear-algebra library

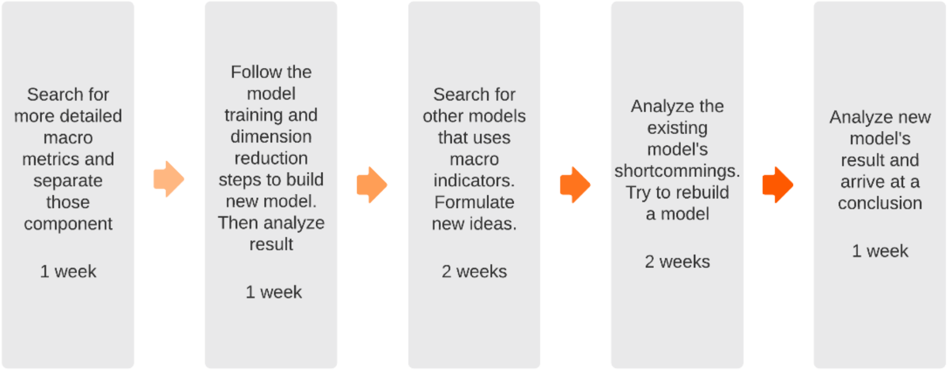
weeks or months even when the economy booms. So, keeping HDI as high as possible should be a long term goal for the policymakers.

3) Energy policy. Notice that although linear relationship seems not so applicable, R2 score could possibly tell us the linear representativeness of different variables. For both features representing people's access to energy show that probably modern energy ensures the basic nutrition supply and medical instruments that require electricity. Keeping the power supply uninterrupted would be necessary to cope with this pandemic.
4) The "linear relationship" assumption of this model is naive. However, the goal of this preliminary work is not to build a precise predictor. The analysis here only reflects some possible measures that reduce the damage of Covid 19.
5) Choice of Label. The label a model is regressing to is crucial for the success of the linear model. The way of choosing labels are not thoroughly investigated here. Death cases and recovered cases can be incorporated.
6) Use of macro data. Macro data could hardly describe details in a system. Macro data could give us intuition about the tendency. However, for sure, different countries will be influenced by different factors differently. That is what macro model cannot incorporate.

Related Work

There is analysis with macro data focusing on impacts[7] of Covid 19 on the macro economy and the potential applicable polices[8]. However, as far as the author know, there is no research that tries to capture covid 19 cases with such macro data. Probably epidemic models are more useful when predicting the number of cases. However, this model links some of the economic implications to the pandemic directly.

Future Milestones


Search for more detailed macro metrics and separate those component

1 week

Follow the model training and dimension reduction steps to build new model. Then analyze result

1 week

Search for other models that uses macro indicators. Formulate new ideas.

2 weeks

Analyze the existing model's shortcommings. Try to rebuild a model

2 weeks

Analyze new model's result and arrive at a conclusion

1 week

Evaluation:

1. HDI is a complex index that deserves decomposition. Besides, expanding the area of indices might be helpful. Currently metrics considered are among the domain of national development (potentially health since HDI includes life expectancy), education, energy. There could be more variables like National Happiness Index, clinical and medical researches, political stability, people's expectations etc. Next goal is to investigate data in those areas.
2. With more detailed features, train the model then do the same analysis that is introduced above.
3. Search for 2-3 macro-metric analysis literatures and summarize their strengths and weaknesses.
4. Rebuild a model that takes advantages of the experience of literature then analyze what it tells. As more data would be published at that time, such model could not be limited to predict things related with number of cases. It could be used as a predictor of the impacts of Covid 19 on the economies and societies.
5. Arrive at a conclusion based on modeling. Analyze the model's advantages and disadvantages. For use of such model, give proper assumptions and contexts.

[7] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547729

[8] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3560337