Project 1: Draft

Econ 1680: Machine Learning, Text Analysis, and Economics

Anushka Kataruka

March 8, 2024

1 Introduction

A state's apparatus and government play a significant role in the socioeconomic well-being, prosperity and happiness of its people. A government's ability to accomplish its policy goals (also known as state capacity), level of freedom, corruption, and transparency, can all determine whether a state's institutions and governments are ”strong” enough to guide its development. Recent global crises, including the COVID-19 pandemic, climate change, financial crises, etc. evidence the weakening of state capacities around the world, and thus could have grave implications for the well-being and economic prosperity of current and future generations (Lindsey (2021)). Multiple other scholars of disciplines across political science, economics, and sociology, have attempted to draw relationships between state capacities and people's well-being, including Dincecco (2017) and Asadullah and Savoia (2018). This is an attempt to do the same, and possibly contribute to previous research, using machine learning and econometrics methods to quantify the relationships between state capacity and structure with development. This project has also partially been inspired by Our World in Data's work on state capacity (Herre et al. (2023)).

I used data from various sources including the Varieties of Democracy project, the UNDP, and various quantitative studies on state capacity to evaluate the impact of different state capacity indicators on human development. I did an initial analysis using Lasso regression model to obtain initial coefficient values for my independent variables. I will be using a multilayer neural network, after aggregating the data using PCA, for future analyses to obtain more accurate results, and compare the final predictions to those produced by the linear Lasso model. I expect most of my coefficients to have a significant impact on changes in the human development index.

2 Data Sources and Descriptions

The data for each variable comes from different sources. I used the democracy and state capacity indicators collated and developed by V-Dem (Varieties of Democracy Institute)Michael et al. (2023) alongside the state capacity index developed by Hanson and Sigman (2021). All the data produced by the V-Dems project is available for the years 1789-2022. Table 1

Table 1: Summary Statistics

	HDI	rigor_admin	rule_of_law	regime	civil_liberties	corruption
Count	5116.0	5116.0	5116.0	5116.0	5116.0	5116.0
Mean	0.66	0.50	0.55	4.81	0.69	0.50
Stdv	0.17	1.49	0.31	2.92	0.25	0.30
Min	0.216	-3.617	0.009	0.0	0.026	0.002
Median	0.685	0.35	0.551	5.0	0.768	0.546
Max	0.962	4.046	0.999	9.0	0.977	0.969

	inf_capacity	years_colonized	state_capacity	taxation	territory_control
Count	1618.0	5066.0	3841.0	4349.0	5055.0
Mean	0.778	100.25	0.49	0.20	91.93
Stdv	0.135	125	0.95	0.12	9.44
Min	0.183	0.0	-2.029	0.0009	39.0
Median	0.73	62.0	0.35	0.178	95.2
Max	1.0	514.0	2.964	0.609	100.0

Note: HDI is our dependent variable, while the rest are the regressors for our model. There are a total of 10 indicators which will be used for measuring different aspects of state capacity and civil liberties.

shows the summary statistics used for each variable.

Following is the description of each variable, including its source:

1. *HDI* (Human Development Index): Our dependent variable comes from the United Nations Development Program's Human Development 2021-22 Report (UNDP (2022)). The index measures ”human development” combining indicators of health, economic well-being, and education. Health is measured by life expectancy at birth, economic well-being is measured by Gross National Income per capita, and education is measured by average years of schooling (or expected years of schooling for school-aged children). Data is recorded for the years 1990-2021.
2. *rigor_admin* (Rigorous and impartial administration index): This has been sourced from the data compiled by the V-Dems project (Michael et al. (2023)), and ranges from 0 to 1. It estimates the extent to which public officials respect the law, and administer it without arbitrariness and bias (like through nepotism, cronyism, discrimination, etc.)
3. *rule_of_law* (Rule of Law index): This has also been sourced from the data compiled by the V-Dems project (Michael et al. (2023)), and ranges from 0 to 1. It estimates the enforcement of laws ”transparently, independently, predictably, impartially, and equally” along with government officials' compliance with the law.
4. *regime* (Type of regime): The source is again the V-Dems project (Michael et al. (2023)). It classifies each state into its political regime based on the competitiveness of access to power and liberal principles. It ranges from 0 to 9, with its data type being integers, where 0 represents a closed autocracy and 9 represents a liberal democracy. For the purpose of this

project, I one-hot encoded the data to produce more accurate results.

5. *civil_liberties* (Civil liberties index): Also sourced from the V-Dems project (Michael et al. (2023)), this measures the extent to which governments/states respect the civil liberties of the people.

6. *corruption* (Political corruption index): Also sourced from the V-Dems project (Michael et al. (2023)), this measures the pervasiveness of political corruption in a state. It incorporates judicial, executive, legislative corruption, and public sector corruption, and incorporating both ’petty’ and ’grand’ corruption. Corruption is primarily measured through instances and magnitude of public bribery and embezzlement, with the overall index ranging from 0 to 1 (0 representing no corruption).

7. *inf_capacity* (Information capacity): The data for measuring the information capacity of a country comes from the study conducted by Brambor and Teorell (2019). It aggregates multiple indicators of information capacity (or a state’s ability to gather and organize information on itself and its people for the purpose of governance). The indicators include the measure of when a country first established a statistical agency, whether the country has in place a civil register and a population register, the ability of the country to carry out censuses consistently and the ability of the country to publish a statistical yearbook over a ten-year window. The data however is based only on a sample of 86 states, and is only uptil 2015, which makes its utility for the purpose of our project questionable, which is why it has been excluded from the analysis for the purpose of this draft. However, information capacity is a vital aspect of state capacity, and excluding it would make our model inaccurate as well. I hope to replace this with some other measure of information capacity when compiling the revised draft.

8. *years_colonized* (Years a country has been colonized): This combines data from the Colonial Dates Dataset (COLDAT) developed by Becker (2019), which itself sources the data from various historical datasets, with the processing on the data conducted by Our World in Data (Herre et al. (2023)). It measures the number of years a country has been colonized by overseas European colonial powers, specifically by Belgium, United Kingdom, France, Germany, Netherlands, Portugal, Spain, and Italy. This variable functions as one of the controls for our model, as colonization is known to negatively impact state capacity and stability, especially for recently decolonized states. This takes into account both extractive and settler colonialism.

9. *state_capacity* (State capacity index): As mentioned earlier, this is the index developed by Hanson and Sigman (2021). It measures state capacity by combining 21 different indicators related to three key dimensions: extractive capacity, coercive capacity, and administrative capacity. The data is available only uptil 2015.

10. *taxation* (Tax Revenues as share of GDP): This data comes from the UNU-WIDER Government Dataset (UNU-WIDER (2023)), with some processing done by Our World in Data. The ability to collect taxes effectively reflects their administrative/bureaucratic efficiency, as tax collection is often complex, especially for larger countries. Some variation in the data also comes from tax policy differences between countries. The data is available from 1980-2022.

11. *territory_control* (Percentage of territory controlled by the government): This has been sourced from the V-Dems dataset, with some processing done by Our World in Data (Michael et al. (2023), Herre et al. (2023)). This is also a control indicator, as governments which are

not the main authority over certain territories of a state or are rejected by the population in those territories, will not be able to implement policies effectively, thus negatively impacting state capacity.

I combined all the country-level data, left merging on the HDI dataset. Thus the combined dataset is available from the years 1990-2021, with some gaps for certain variables, as seen in Table 1.

3 Method

I will be using multiple machine learning and econometrics methods for this project. I will first be using some dimension reduction (principal component analysis and clustering (KMeans/Hierarchical) on my inputs, namely the democracy and state capacity indicators, to make the final analysis and regression simpler. I will aggregate all the dependent variables I'm using except for $state\_capacity$ and $inf\_capacity$, as both are already aggregated indices. For the purpose of this draft, I skipped this step, and proceeded with the regression analysis directly.

I split the data into training and test sets, after which I applied OLS, Lasso, and Ridge regression on the training data, and fit them accordingly to the outputs, to observe the magnitude of impact of different aspects of state government have on HDI, along with evaluating the mean-squared error for each of these models. From the analysis, it is observed that Lasso produces the least error, which is why I shall be using Lasso as my linear model. Further, I will be comparing the results of this model with a multi-layer perceptron regression model as well, and will apply a similar penalty on the coefficients, such that both models will be of the form:

$$\begin{gathered} HDI_i = \beta_0 + \beta_1 state\_capacity_i + \beta_2 rigor\_admin_i + \beta_3 rule\_of\_law_i + \beta_4 regime_i \\ + \beta_5 civil\_liberties_i + \beta_6 corruption_i + \beta_7 years\_colonized_i + + \beta_8 taxation_i \\ + \beta_9 territory\_control_i + \varepsilon_i \end{gathered}$$

where, $HDI_i \in \{0, 1\}$. I will be using Tensorflow's Keras Dense model for the deep learning analysis, and mean-squared error as the loss metric for both models. The following will be the loss function, taking the Lasso penalty into account ($\hat{HDI}$ being the predictions of the model):

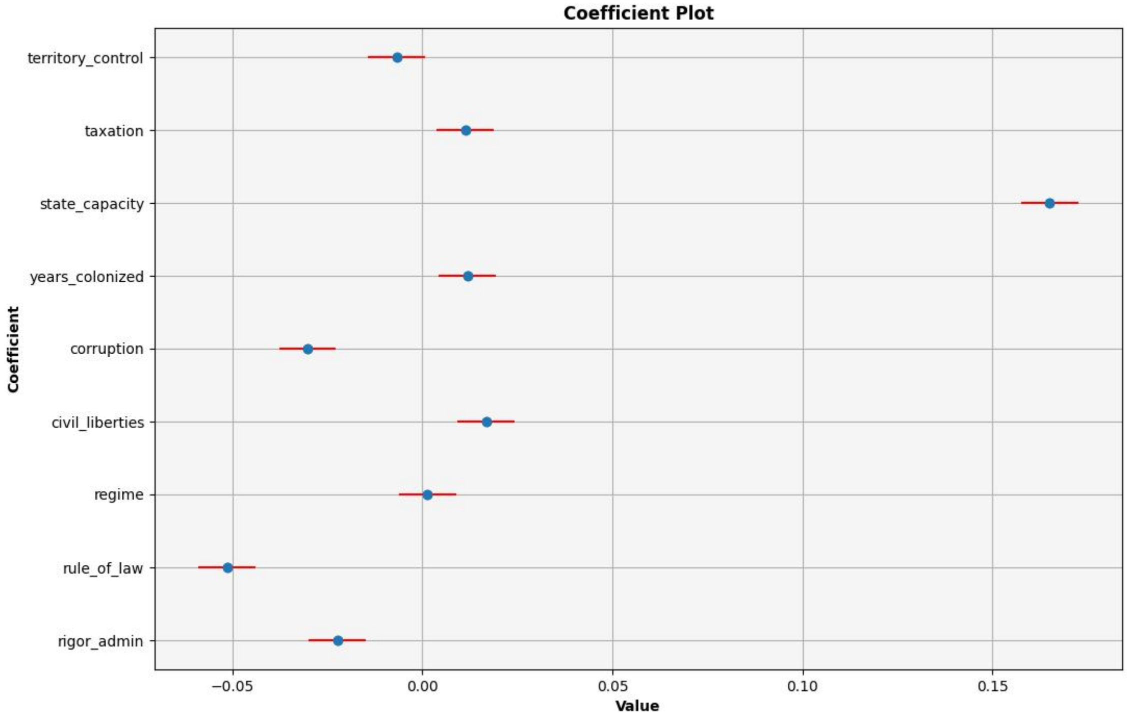
$$\hat{\beta} = argmin\{\Sigma_{i=1}^{N}(HDI_i - \hat{HDI}_i)^2 + \alpha * \Sigma_{j=1}^{p}|\beta_j|\}$$

. If this does not yield statistically significant results, I will also factor in country fixed effects, to control for variations in HDI between countries due to inherent characteristics.

4 Results or Expected Results

Figure 1 shows the initial results of our linear Lasso model. Due to the data restrictions of $inf\_capacity$, that variable has been excluded from my draft analysis.

Figure 1: Lasso Coefficients



regime was one-hot encoded for the purpose of the Lasso analysis. We see an initial high coefficient magnitude assigned to *state_capacity* and *rule_of_law*. *corruption* and *rigor_admin* are also significant.

We observe a high magnitude for coefficients of *state_capacity*, *rule_of_law*, *corruption*, *rigor_admin* on HDI, implying that on initial analysis, these variables have a significant impact on human development. As expected, higher state capacity is associated with higher HDI, and higher corruption is associated with lower HDI. Surprisingly more rule of law, and a more rigorous admin is associated with a lower HDI.

For future analysis, we expect our results from our linear model to be in similar directions. Territory control, taxation, state capacity, corruption, rule of law, and administration rigor/impartiality should be positively associated with the human development index. Corruption and years colonized should be negatively associated with the HDI. I'm not sure exactly what to expect for regime and civil liberties, as specifically economic development can still be high regardless of democracy indicators.

5 Conclusion

I am starting to be able to answer my research question. My analysis currently relies entirely on coefficients produced by a linear Lasso regression model. It is not entirely robust. The data needs more processing, and possible clustering to achieve more significant results. The multi-layer neural network should be a robust analysis, fitting the data better to accurately predict human development based on state capacity indicators.

A significant limitation of my analysis for now is limitations of the data itself, with certain variables not having enough data to cover the entire time period and all the countries I'm analysing. To overcome this, my plan is to do more research into other available data, and maybe come up with my own indices if appropriate data is non-existent.

References

Asadullah, Niaz and Antonio Savoia (2018) "Poverty reduction during 1990–2013: Did millennium development goals adoption and state capacity matter?" *World Development*, 105, 70–82.

Becker, Bastian (2019) "Introducing COLDAT: The Colonial Dates Dataset."

Brambor, Agustín Goenaga Johannes Lindvall, Thomas and Jan Teorell (2019) "The Lay of the Land: Information Capacity and the State.," *Forthcoming in Comparative Political Studies.*

Dincecco, Mark (2017) "State Capacity and Economic Development," *Cambridge University Press.*

Hanson, Jonathan K. and Rachel Sigman (2021) "Leviathan's Latent Dimensions: Measuring State Capacity for Comparative Political Research," *The Journal of Politics*, 83 (4).

Herre, Bastian, Pablo Arriagada, and Max Roser (2023) "State Capacity," *Our World in Data*, https://ourworldindata.org/state-capacity.

Lindsey, Brink (2021) "State capacity: what is it, how we lost it, and how to get it back," *Niskanen Center.*

Michael, Coppedge, Carl Henrik Knutsen John Gerring, Staffan I. Lindberg et al. (2023) "V-Dem Dataset [Country-Year/Country-Date] v13,"Technical report, Varieties of Democracy (V-Dem) Project.

UNDP (2022) "Human Development Report 2021-22: Uncertain Times, Unsettled Lives: Shaping our Future in a Transforming World,"Technical report, UNDP (United Nations Development Programme).

UNU-WIDER (2023) "UNU-WIDER Government Revenue Dataset,"Technical report, UNU-WIDER.