Does Economic Development Predict Democratization?

Chiara Bondi and John Madigan

Abstract

The relationship between a country's level of democracy and its economics develop is a much debated topic within economics and political science. It is believe that a country's economic development causes liberalization. However, with some notable exceptions like China and the Gulf countries, this relationship needs to be further analyzed. In this paper, we attempt to discover which economic variables can predict a country democracy index. To do this, we analyze 2019 World Bank Data and the EIU Democracy Index by conducting a LASSO regression. We found 16 variables that predict a country's level of democracy including positive coefficients with mortality from cancer and diabetes and military expenditure.

1 Background and Significance

The relationship between economic variables and Democracy is a highly debated topic within economics and political science. For example, many authors such as Francis Fukuyama predicted that the liberalization of China would occur once it reached certain economic factors (Fukuyama (1992)). As economic rights expand (women in the work force, ease of doing business, property rights, etc), Fukuyama's theory states that political liberalization follows. This theory dates back to Montesquieu's *Doux commerce*, "stating that commerce tends to civilize people, making them less likely to resort to violent or irrational behaviors." ("Doux Commerce" (2021)) More recently, the Arab Emirate's have been of interest due to their Islamic regimes while having vast wealth of oil. Due to the examples above, we are interested in creating a model to determine whether these states are outliers, or if there are any economic variables that can predict whether a state is a Democratic regime. If this were true, defenders of Democracy would prioritize more of their economy and its success and well being. Because of this, we expect to find high predictive power of the prevalence and utilization of economic rights.

2 Methods

a. *Data collection.* To create our model, we analyze 806 variables selected from the World Development Indicators data set ("World Development Indicators" (n.d.)) using R (R Core Team (2019)) and RStudio (RStudio Team (2020)). To assess a country's level of democracy, we use the EIU Democracy index ("World Bank" (n.d.)). All data points are from 2019. An important thing to note is our data set contains 174 observations, each one representing a country. Due to the high dimensionality of the data, we will use regularization to properly create a regression model to predict each country's democracy index in that year. The World Bank sources its data by collecting reports and other statistical work collected by its member countries. Because of this, the data quality depends on how well each country is able to collect and report its data. The Economist Intelligence Unit (EIU) index of democracy comprises of 5 indicators, each having a scale from 1 to 10. In this paper, we choose to focus on the Political Stability and Absence of Violence indicator. This category is composed of 12 dichotomous or 3-point scales. These categories are then summed and converted into a 0 to 1 scale. Our dataset merges both the World Bank and EIU reports, to create a single data frame with the Economic indicators of 2019 as our predictors, and the Democracy index for that same year as our response variable. In total, our data set contains 174 observations, each corresponding to a country, and 806 variables including our response variable.

b. *Variable creation.* Our current working data set is a merger of the EIU and World Bank data sets. To begin our variable creation process, we started by properly formatting the World Bank data using tidyverse in R. Specifically, we pivoted the data table from a long format to a wide format. This allowed each row to represent a country and our variables be the different economic indicators. Furthermore, we eliminated columns where more than 50% of the values were null. Then, we split our data into 70% training and 30% test data. To impute the missing data in our testing set, we calculated the median in our training data and filled all missing values in our testing set with the median value from the training set. All of our economic indicators are continuous numeric variables, representing different aspects of economic development such as GDP, labor force participation, etc. The data set contains 2 identification variables, identifying the country of our observations: Country Name and Country Code. Because both variables represent the same identification form, we dropped the Country Code variable from our set.

c. *Analytic Methods.* The data set used is a high-dimensional set. Because of this, we cannot apply traditional regression techniques such as OLS Regression to predict the Democracy index. Once all of the data was imputed and no more values were missing, we ran a LASSO Regression on our data set to select which variables could account for most of the variation. We then went on to create a set of possible lambda values ranging from $\lambda = 10^{10}$ to $\lambda = 10^{-2}$. Only using our training set, we performed a Lasso Regression using the glmnet (Friedman, Hastie, and Tibshirani (2010)) and Amelia (Honaker, King, and Blackwell (2011)) packages on our training set using a grid of lambda values. We continued by performing a cross-validation

to identify the best value of lambda as well as see which variables would be selected to model Democracy index.

We choose LASSO over ridge regression because we wanted the less important coefficients to be zero which does not occur when performing ridge regression. We also decided not to use Elastic Net regression due to its computational complexity. Also, while Elastic Net could theoretically be the best of both LASSO and ridge regression, our best Elastic Net models included variables with incredibly small coefficients. This marginally improved our predictions by adding 24 variables, but because we are first and foremost interested in variable selection, we will focus on our LASSO results in the rest of the paper. For more information on the other models we tested, see appendices.

3 Results

In Table 1, we list the selected coefficients from our LASSO model.

Table 1: LASSO Variables and Coefficients

	Variables	Coefficients
1	Intercept	0.16
2	Labor force participation rate, female (% of female population ages 15-64) (modeled ILO estimate)	1.00e-3
3	GNI per capita, Atlas method (current US$)	6.33e-7
4	Military expenditure (% of general government expenditure)	-5.83e-4
5	Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70 (%)	-2.23e-5
6	Women Business and the Law Index Score (scale 1-100)	5.25e-4
7	Grants, excluding technical cooperation (BoP, current US$)	-2.76e-11
8	GDP per capita, PPP (current international $)	7.39e-10
9	Adjusted net national income per capita (current US$)	5.98e-7
10	Immunization, DPT (% of children ages 12-23 months)	8.94e-5
11	Price level ratio of PPP conversion factor (GDP) to market exchange rate	6.28e-2
12	Ease of doing business score (0 = lowest performance to 100 = best performance)	2.69e-3
13	Employment in services, female (% of female employment) (modeled ILO estimate)	3.59e-4
14	Source data assessment of statistical capacity (scale 0 - 100)	1.90e-4
15	Statistical performance indicators (SPI): Pillar 2 data services score (scale 0-100)	1.72e-4
16	Statistical performance indicators (SPI): Pillar 4 data sources score (scale 0-100)	3.02e-4
17	Computer, communications and other services (% of commercial service imports)	3.78e-4

This table shows the coefficients of each variable that was selected in our LASSO regression. To see how well these variables can predict a country's Democracy index, we used Multiple Linear Regression to generate a MLR model with the Democracy index as our predictor. One important thing to note is that the coefficients generated through MLR are not the same as those generated by LASSO regression, and this is because with LASSO we are doing variable selection as well as shrinkage of coefficients. To see more information about the coefficients generated through MLR, please refer to Appendix (C). Below are the diagnostic plots for our linear model using the predictors selected through LASSO.

Below in Figure 2, we see that cross validation plot from multiple LASSO regressions. We choose the lambda value corresponding with the minimum RSS which is $\lambda = 0.024$. As we can see from our diagnostic plots in Figure 1, the selected variables seem to model very well the Democracy index. Specifically, the QQ-plot shows that our residuals are predominately normally distributed, where at the end they skew from the linear trend, but the majority follow such linearity. Moreover, in the residuals vs fitted values plot, there is no pattern in our residuals, meaning they are randomly distributed and are not dependent on our fitted values. When creating predictions using our testing set using the LASSO model, we find an $R^2 = 0.4871$ approximately. This means that the LASSO regression model using the selected variables listed above explains almost 50% of the variation in the response variables (democracy index) in our testing data.

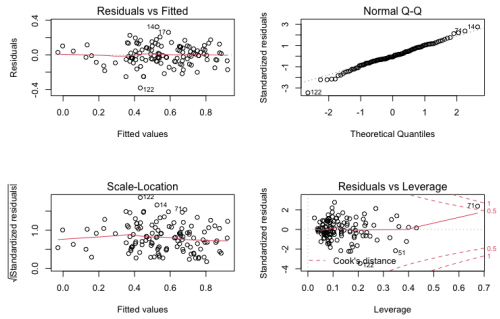

Figure 1: Diagnostic plots for our Multiple Linear Regression Model

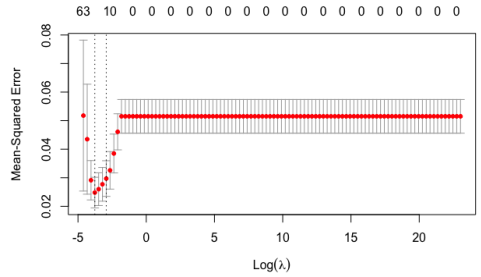

Figure 2: Cross Validation Plot for LASSO Regression

4 Discussion/Conclusions

As seen in the table above, we can use LASSO to generate a list of variables that will predict how democratic a country is. While female labor participation and the Ease of Doing Business index represent rights found in liberal democracies, variables such as military expenditure and the necessity to import computer and communication services pose interesting questions. For example, if a countries computer services are imported, they are not homegrown like in countries such as China. While the dominant American technology companies mostly abide by American rules when it comes to free speech, Chinese and North Korean platforms operate under their own government mandates when it comes to speech.

One of our findings that moves the field forward is the the realization that people in democracies die from certain types of deaths more than others. While childhood mortality may be higher in countries that are not democracies, deaths from cancer and diabetes predict whether or not they a country is a democracy. This may be because of the association between democracies and increase medical technology or the association between countries that are not democratic and the prevalence of civil wars or political violence.

One limitation of this research is that the coefficients that LASSO produces are do not imply causation. If a country wanted to increase democracy in a region, increasing computer and communication imports will not necessarily cause an increase in democracy. Another possible limitation is our rudimentary method of imputing data. While we could have used KNN or another method of imputing data, we leave it up to future researchers to implement more rigorous method. Finally, another limitation of our analysis is the age of our data. While we wanted to use more up-to-date data, it contained too many null values for proper research consider our imputation methods.

5 References

"Doux Commerce." 2021. *Wikipedia*. Wikimedia Foundation. https://en.wikipedia.org/wiki/Doux_commerce.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://www.jstatsoft.org/v33/i01/.

Fukuyama, Francis. 1992. *The End of History and the Last Man*. HarperCollins.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7): 1–47. https://www.jstatsoft.org/v45/i07/.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

"World Bank." n.d. *Economist Intelligence Unit - World Bank*. https://info.worldbank.org/governance/wgi/Home/downLoadFile?fileName=EIU.xlsx.

"World Development Indicators." n.d. *DataBank*. https://databank.worldbank.org/reports.aspx?source=world-development-indicators.

6 Appendix

6.1 Appendix (A): Ridge Regression

Throughout our analysis process, other than LASSO Regression we decided to apply Ridge Regression as well to see how our minimal tuning parameter compared to the one determined through LASSO Regression. Moreover, we wanted to see how well our predictions using a Ridge Regression model compared to the observed values from our testing set as well as how accurate they are compared to our LASSO Regression model. To compute the Ridge Regression model, we used the same grid of possible tuning values as we did for the LASSO Regression. To choose the best tuning parameter λ, we also used cross-validation to select the parameter yielding lowest MSE. Here is the plot resulting from our cross-validation, showing all λ values as well as the MSE. From the plot below, the best tuning parameter came to be $\lambda = 2.656$, significantly larger than our tuning parameter for LASSO, which was only 0.024.

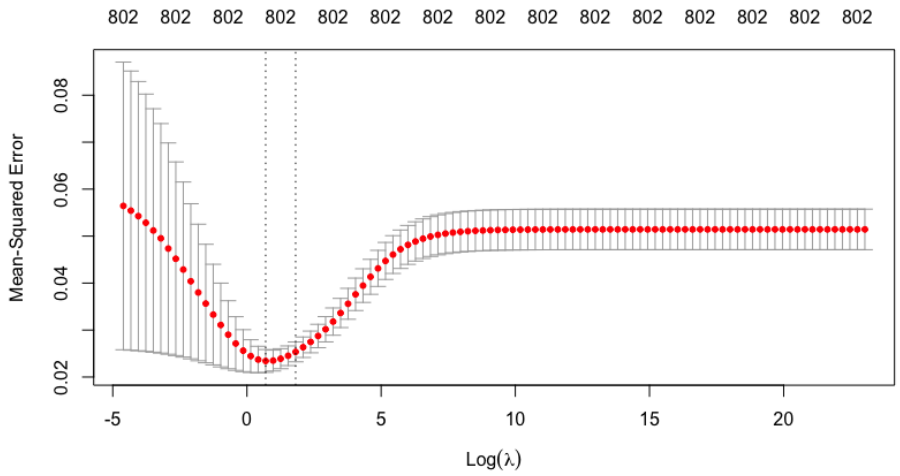

Figure 3: Cross-Validation plot for Ride Regression

As we can see from the plot, the tuning parameter with lowers MSE is approximately 2.009. We the used this to make predictions for our testing set, which contains 50 observations, none of which were used in the training set. Once we had our predictions, we computed their R^2 and RMSE. These came out to be $R^2 = 0.4405$ and RMSE $= 0.1717$. Recall that for our LASSO Regression, our predicted values had a $R^2 = 0.4871$ and RMSE $= 0.1644$. This means that our LASSO model explains almost 4% more of the variation of the democracy index. Despite the difference of R^2 values and RMSE between our Ridge and LASSO Regression models is marginal, what we wanted was to understand what variables best model the Democratization index. Ridge Regression is a good model when you know all of your variables are important. So, because we want to do feature selection as well as regression, Elastic Net and LASSO Regression are better methods to apply. To see how we applied Elastic Net regression and its results, refer to Appendix (B).

6.2 Appendix (B): Elastic Net Regression

Another method we used to analyze our data was through Elastic Net Regression. Using a level of α between 0 and 1 allows us to shrink our coefficients faster and perform some variable selection as well. To determine the optimal value for α we created an array with 100 possible values from 0 to 1, increasing by 0.01 each time. For each value in our array, we ran a regression with that corresponding alpha value, and for each model ran Leave One Out cross-validation to select the optimal λ. To select the optimal value, we stored

all R^2 values, and selected the alpha that resulted in highest R^2. After running the loop, we found that the optimal regression model used $\alpha = 0.23$. Using that alpha value, we ran LOOCV to find the λ with minimal MSE, represented in the plot below.

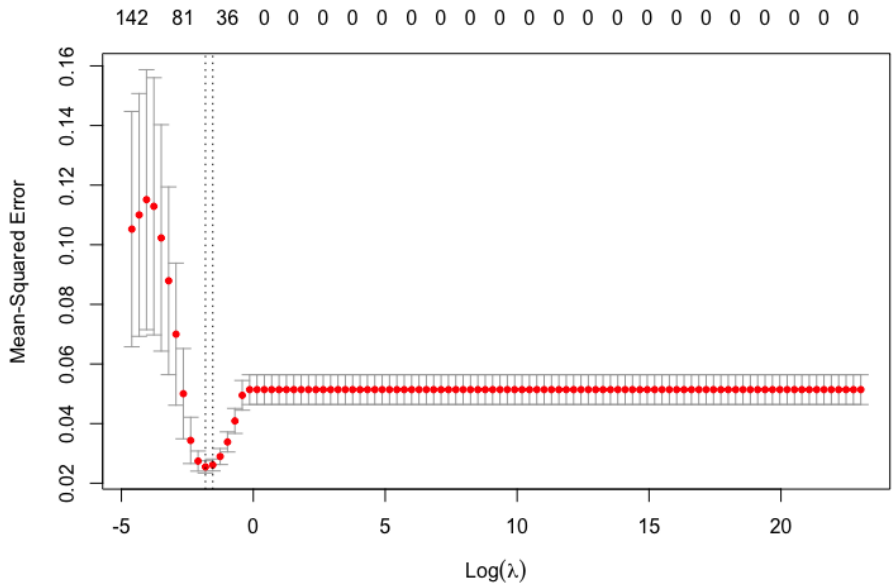

Figure 4: Cross-Validation plot for Elastic Net Regression

To see how all different values of α and λ change the R^2 of our models, here is a 3-dimensional plot with 10000 data points, where x = log(λ) y = α, and $z = R^2$. This 3-dimensional scatter plot shows how our different alpha values and their corresponding optimal tuning parameter value and how these change the R^2 in our predictions. The optimal model can be found by selecting the point with highest R^2, and selecting its alpha value and lambda.

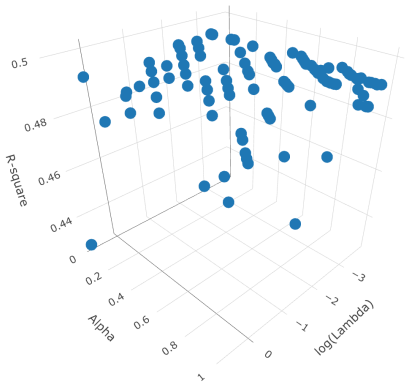

Figure 5: 3D plane with tested α values and their optimal λ

The optimal value for our tuning parameter λ came out to be 0.1629, which is much lower than the Ridge Regression value (2.009), but higher than the optimal λ for LASSO, which is 0.0231. Our Elastic Net Regression model has $R^2 = 0.4929$, the highest out of all 3 models. Moreover, because we did not do

a complete Ridge Regression, many parameters were dropped since their coefficients came out to be 0. However, Elastic Net Regression selected 40 variables, significantly more than LASSO. When we look at the coefficient values, however, we can see that many are on the 10^{-12} level, which is very close to 0. All of the variables selected through LASSO Regression were also selected using Elastic Net Regression, but since our α value was not 1, we are still limited by how many coefficients are estimated to 0 and how many variables we are able to completely drop. Because now we have less parameters than observations, we were able to model the Democracy index against our variables and got the following diagnostic plots:

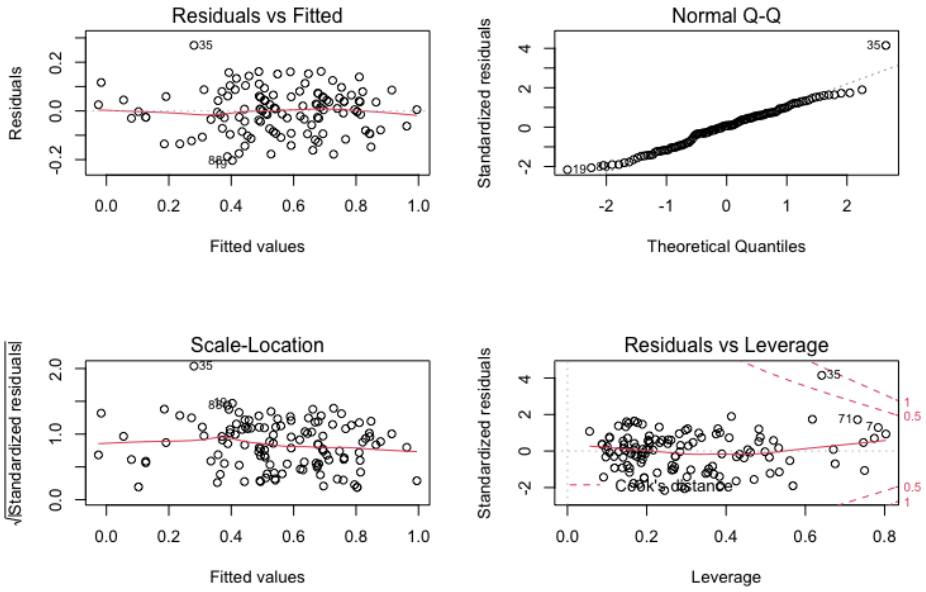

Figure 6: Diagnostic Plot with our Elastic Net Regression Variables

Notice how, in our Residuals vs Fitted values plot, all of our residuals appear to be randomly scattered, with no pattern showing in their dispersion. Moreover, the QQ plot appears to be a relatively straight line, with some outliers appearing at either end of the plot. However, both plots show that there is a strong linear relationship between the 48 selected variables and the Democracy index. Just like with LASSO regression, the linear model generated, although it uses the same variables selected from the Elastic Net regression, it doesn't have the same coefficients as those estimated by the regression. Because of this, the model generate through multiple linear regression is not exactly the same as that generated through Elastic Net. Moreover, we chose to use LASSO over Elastic Net because, although the R^2 value for our predictions from our Elastic Net model is higher, it's not a significant improvement over LASSO. So, since we can get very similar R^2 with less predictors using LASSO, we continued with that regression method instead of Elastic Net.

6.3 Appendix (C): Multiple Linear Regression Model Coefficients

Below is a table with all 16 variables selected through LASSO regression and their corresponding coefficients generated through Multiple Linear Regression.

Table 2: LASSO Variables and MLR Coefficients

	Variables	Coefficients
1	Intercept	-1.53e-1
2	Labor force participation rate, female (% of female population ages 15-64) (modeled ILO estimate)	2.51e-3
3	GNI per capita, Atlas method (current US$)	-3.06e-6
4	Military expenditure (% of general government expenditure)	-1.02e-2
5	Mortality from CVD, cancer, diabetes or CRD between exact ages 30 and 70 (%)	-8.29e-4
6	Women Business and the Law Index Score (scale 1-100)	-6.88e-4
7	Grants, excluding technical cooperation (BoP, current US$)	-5.89e-11
8	GDP per capita, PPP (current international $)	1.79e-6
9	Adjusted net national income per capita (current US$)	1.83e-6
10	Immunization, DPT (% of children ages 12-23 months)	2.24e-3
11	Price level ratio of PPP conversion factor (GDP) to market exchange rate	2.45e-1
12	Ease of doing business score (0 = lowest performance to 100 = best performance)	1.45e-3
13	Employment in services, female (% of female employment) (modeled ILO estimate)	6.78e-4
14	Source data assessment of statistical capacity (scale 0 - 100)	2.18e-3
15	Statistical performance indicators (SPI): Pillar 2 data services score (scale 0-100)	8.68e-4
16	Statistical performance indicators (SPI): Pillar 4 data sources score (scale 0-100)	-5.01e-4
17	Computer, communications and other services (% of commercial service imports)	1.77e-3

As we would expect, most of our coefficients are larger in magnitude than their corresponding value for the LASSO model. This is because LASSO does variable selection as well as shrinkage, making our coefficients as small as possible. However, what is interesting to note is that some of our coefficients predicted through MLR came out negative, but with LASSO are positive. For example, the Women Business and Law Index Score has a negative coefficient for our MLR model. This could be because women in non-democratic countries may be relegated to jobs such as taking care of children or making food, while in democratic countries they can choose to do that or enter the services industry.