Justin Le
Professor: Michael Leung
TA: Yilin Li
ECON 124

Final Paper

Introduction

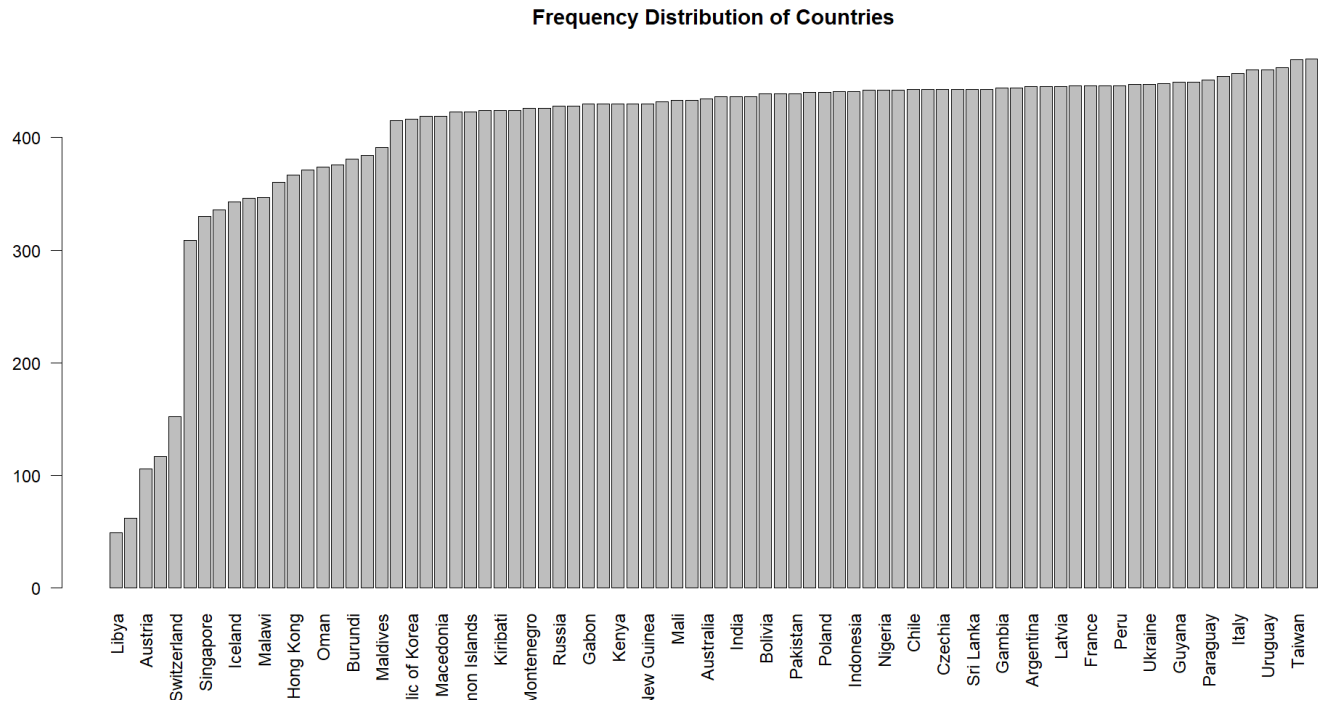
The question for this research paper is to analyze and predict rigidity of the average of the 13 coded Public Health Measures of the dataset, which are: *school, domestic, travel, travel_dom, curf, mass, elect, sport, rest, testing, surveillance, masks, and state*. Where it is valued at 0.5 if they are localized or partial, 1 if they are national or strict, and 0 indicates no measures. This is interesting because of how some countries act and perform differently compared to other countries even with similar variables and we can measure how the public's health is generally doing and how the government is responding to these Public Health Measures. It is only measured if there are at least 10 out of 13 measures coded.

Data

The data is a county's government response to COVID19 and puts together all the measures implemented by governments worldwide in response to COVID19. The data mainly consists of binary variables, with the exception of economic measures and public health measures. The limitations of this dataset is that it is the government making these claims and assumptions about their county in response to the pandemic, maybe not what the people of the country actually want to happen and their general response of how the country's public health is doing. In addition, the amount of missing values might play an issue to how the prediction models perform. An observation in my dataset represents a country, the date that the observation was taken, the government's responses to the pandemic, as well as the rigidity of public health measures and economic measures. The sample of the dataset is not that representative if you are trying to predict the overall general public health measure of all the observations, however it is representative if you are looking at a specific country/region since there are many observations that are near other observations such as South American countries. Representativeness is important because it helps validate the prediction and dataset. If a sample is not representative, it skews the prediction and does not make an accurate prediction and data of what you are trying to represent.

There are additional variables that are in the dataset, however they are mainly pertaining to the economic measures of how the country is doing.

The number of missing values from the dataset is 158,532. This eliminates many countries in our observations. And as said previously, this may play an issue to how the prediction models perform, especially on the countries that have very few observations compared to the other countries.



What is graphed is the Frequency distribution of the countries in the dataset after removing missing values. For countries with low frequencies such as Libya and Austria, there might be a less accurate prediction than other countries with higher frequencies. This might be due to the type of government they have and how much their government actually cares about the pandemic.

Some variables that I think are most important pertaining to public health are: *school, domestic, travel, travel_dom, curf, mass, sport, rest, testing, masks, and state*. These variables that I thought to be important make sense to me to the rating that the country's public health measure is.

The *school* variable measures if schools were closed, *domestic* measures if there was a domestic lockdown, *travel* measures if travel restrictions were implemented, *travel_dom* measures if travel restrictions within the country were implemented, *curf* measures if a curfew was implemented, *mass* measures if bans on mass gatherings were implemented, *sport* measures if bans on sporting and large events were implemented, *rest* measures if restaurants were closed, *testing* measures if there was a public testing policy, *masks* measures if the obligations to wear masks in public spaces was implemented, and *state* measures if state of emergency is declared.

In this research, I want to analyze more emphasis on the variables *domestic, testing* and *state,* which, to me, are the most important factors to a country's public health measures.

Below is a table showcasing key variable's mean and standard deviation:

	Mean	Standard Deviation
school	0.6384	0.4804838
domestic	0.24	0.4270903
travel	0.8302	0.3754681
travel_dom	0.3671	0.4820213
curf	0.4123	0.4922576
mass	0.8446	0.3623192
sport	0.7254	0.4463047
rest	0.642	0.4794097
testing	0.2062	0.4045488
masks	0.7082	0.4545889
state	0.452	0.4976955

Methodology

The methods that I will be using are linear regressions and lasso regression in addition to using interactions between the regressor and their binary flag to see how it will affect the over regression. Also providing in sample and out of sample analysis of the models to see how good of a fit the models are of predicting in and out of sample samples.

Using linear regressions, we are able to see the linear relationship between the regressors and the outcome and how it affects it. For example, the *elect* regressor, a binary variable equal to 1 if some elections were postponed and 0 otherwise, might not have a linear or significant effect on the rigidity of public health measures. In addition, using a lasso regression model to find the best beta estimates for the regressors, like said earlier, *elect* may not be significant to determining the rigidity of public health measures. These methods are very useful in determining the "fit" of a prediction model, how well it performs to the sample outcome and as well as the out of sample sample predictions.

Since the data was collected during the COVID19 pandemic, it is mainly applicable to that time, however if there is another pandemic, there is a good chance that these observations can be used to predict future pandemics to see how public health will perform and the overall economic measures as well. There are other controls that the dataset does not include such as the number of positive cases and deaths that might contribute to the causal inference of the public health measures.

The extent the assumptions are plausible in the dataset really depend on when it is being used. If the dataset were to be used during a time of no pandemic, then it would be creating false statistics and inferences that do not make much sense. As said earlier and going back to the representativeness of the data, it should also be noted that the data was collected during a pandemic and should be used to make inferences on a pandemic or similar situations/events. This dataset can be used for past pandemics, however there are some discrepancies such as the use of masks and using surveillance apps.

Main Results

For my first linear regression model, I chose to include the 13 coded Public Health Measures of the dataset. The fit for this model was 85.3388%, which makes sense considering that we are using all the 13 public health measures in the dataset. However, it might be too high since we are using all the variables and maybe considered overfitting.

Below is a table of the largest 4 variable coefficients from the regression:

mass	0.08373350
school	0.08015650
state	0.07088022
curf	0.06689103

Although most of the other variables are fairly similar in value, we can see the most "influencing" variables in this regression model. One of which, *state*, is one of the variables that I was most interested in.

Below is the in and out of sample predictive performance by using 20% of the observations for this linear regression model:

In sample R^2	85.23407%
Out of sample R^2	24.66855%

We can see that there is a high predictive performance within the sample, however there is a low predictive performance for the out of sample samples.

This prediction model, in my opinion, should only be used for when the country has already made these decisions as there is a good predictive probability of correctly measuring a

country's public health measure. And as said in the data portion of this research, it is difficult to forecast a country's public health measure if they are not in a pandemic, which is probably why there is such a low out of sample R^2.

For my second linear regression model, I chose to include the 13 coded Public Health Measures including their binary flag interactions of the dataset. The fit for this model was 97.94073%, which makes sense considering that we are using all the 13 public health measures in the dataset. However, it might be too high since we are using all the variables and maybe considered overfitting just like the previous model.

Below is a table of the largest 4 variable coefficients from the regression:

domestic	0.10187242
sport	0.09541016
travel_dom	0.08765965
mass	0.08598772

Below is the in and out of sample predictive performance by using 20% of the observations for this linear regression model:

In sample R^2	97.9172
Out of sample R^2	3.406103

We can see that we gain similar results to the first linear regression model, there is a high predictive performance within the sample, however there is a low predictive performance for the out of sample samples.

This prediction model, in my opinion, should only be used for when the country has already made these decisions as there is a good predictive probability of correctly measuring a country's public health measure. And as said in the data portion of this research, it is difficult to forecast a country's public health measure if they are not in a pandemic, which is probably why there is such a low out of sample R^2.

For my third linear regression model, I chose to include the 3 coded Public Health Measures, *domestic, testing, and state,* of the dataset that I thought are the most important factors to a country's public health measures. The fit for this model is 29.18098%, which is fairly low, since there are less variables in the regression.

Below is a table of the coefficients from the regression:

domestic	0.136370
testing	0.084712

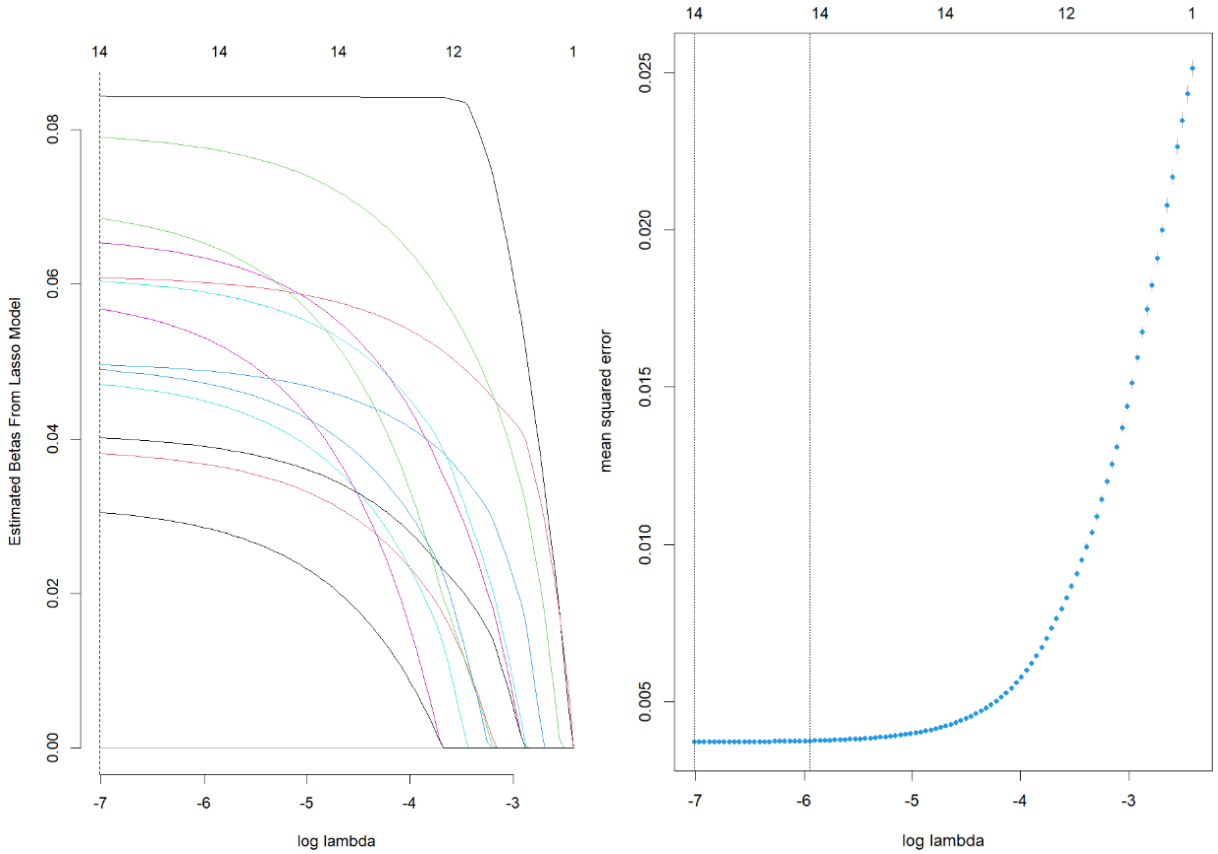
state	0.099752

After creating a few more linear regression models to try to at least achieve a fit of at least 75% and without overfitting, the model to me that makes the most sense, to me, with the regressors that would seem influencing to the model are: *domestic, testing, state, masks, school, travel_dom, sport,* and *rest*. This model produces a fit of 75.22012%. This is the 4th linear model.

Below is a table of the 4 largest coefficients from the regression:

sport	0.0999890
school	0.0975311
state	0.09174517
masks	0.08080142

After performing a lasso regression of the first linear regression model with cross validation with 10 folds, below are the results from the regression model.

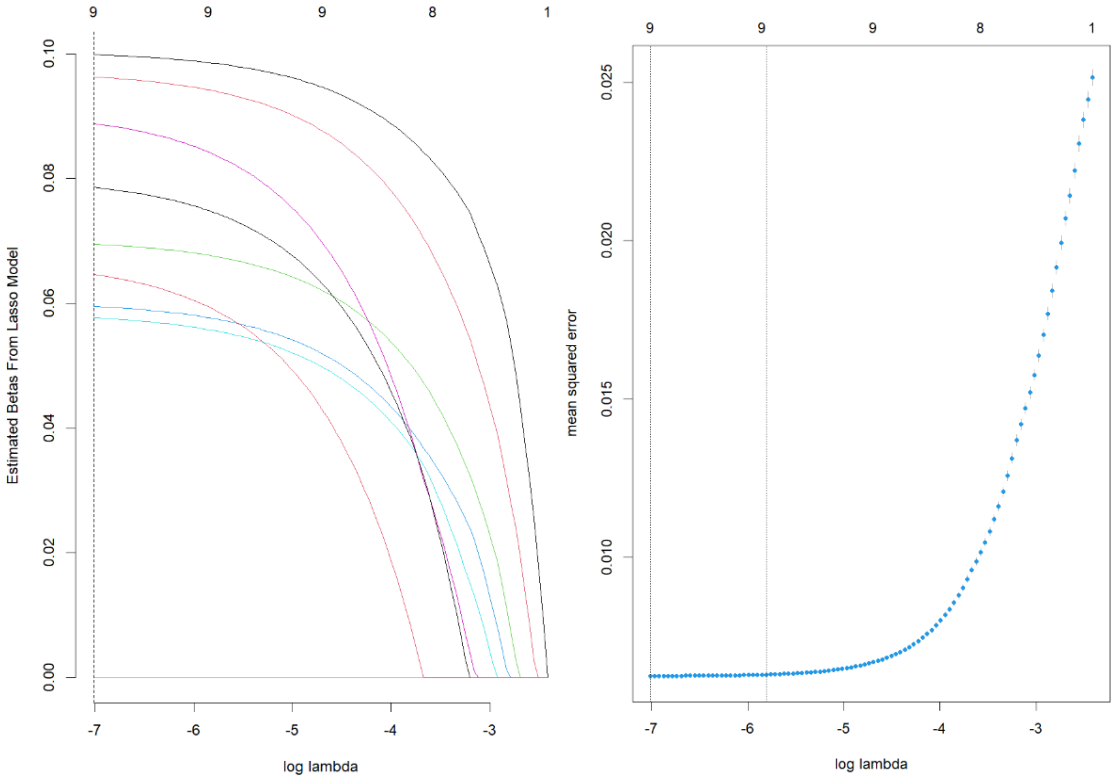


The bias-variance tradeoff graph shows that the optimal lambda for the model is fairly small which gives us the most complex model containing all the regressors.. Since the optimal lambda for this model is 0.000896241.

Below is a table of the 4 largest coefficients of the regression model:

mass	0.08433862
school	0.07903612
state	0.06852059
curf	0.06540124

Creating another lasso regression with the 4th linear model, with the same properties as the first lasso regression model, produces similar results. Below are the results from the regression model.



The bias-variance tradeoff graph shows that the optimal lambda for the model is fairly small which gives us the most complex model containing all the regressors.. Since the optimal lambda for this model is 0.000896241.

Below is a table of the 4 largest coefficients of the regression model:

sport	0.09984789
school	0.09629842
state	0.08884745
masks	0.07860761

After analyzing these results, it is interesting to me that the variables that I thought were most influential to the prediction models: *school, domestic, travel, travel_dom, curf, mass, sport, rest, testing, masks, and state.* We can see that from all the regression models, that I was fairly correct, from them we can see that the variables *domestic, mass,* and *sport* are the most apparent variables considered in the models.

However, looking at the lasso model using the first linear regression with all the variables, we can see that the variable *mass* is the most influential variable in the model. *mass* is a binary variable equal to 1 if bans on mass gatherings were implemented and 0 otherwise. To me, this makes sense because if a country were in the rough stages of a pandemic, it would be wise for the country to implement bans on mass gatherings otherwise more infections would spread through those.

Conclusion

These models, in my opinion, should be considered about a country's public health measures as they are making decisions as to whether they should close down schools, prohibit mass gatherings, implement curfews, and etc. The country should make a "rough draft" of their decisions and use these models to predict their public health measure of how their country is doing and make decisions accordingly. As this dataset is taken from a pandemic, it is only wise to use and make decisions based on whether or not the country or the world might be facing a pandemic such as COVID19. In the past, medicine was not as good as it is today so countries are able to deal with these pandemics in a faster and safer way. However, it is skeptical since the USA, a first world country and a capitalist economy, it is difficult to see how much help it can provide when it can provide.