Econ 434

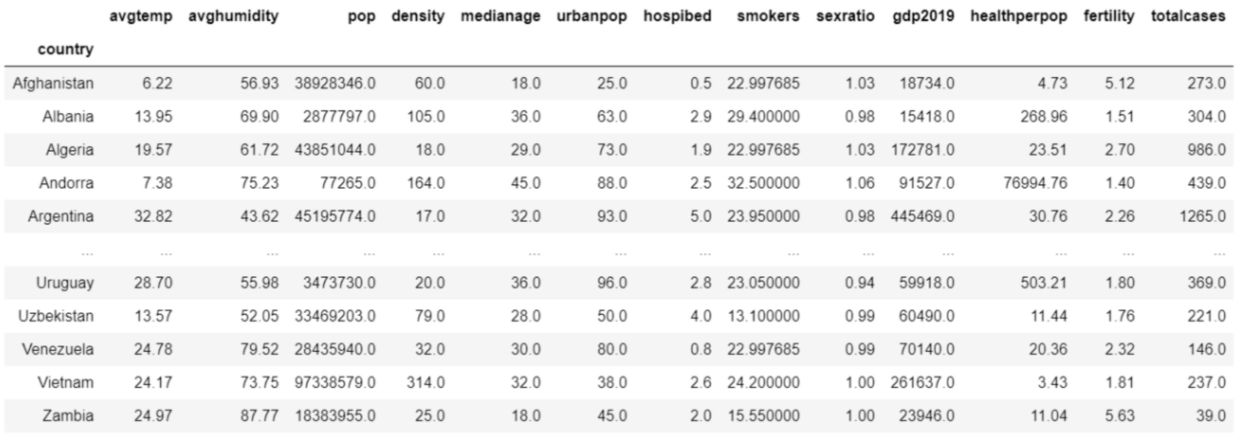
May 23, 2020

Coronavirus and Temperature

The purpose of this project is to investigate whether temperature has any effect on the transmission of coronavirus, so that hopefully the spread of coronavirus may slow down in the summer because of the higher temperature.

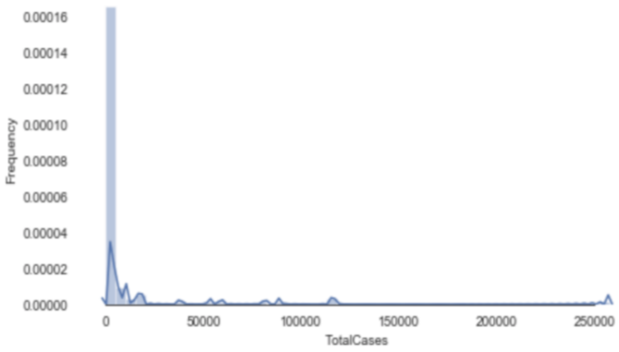
Data Description

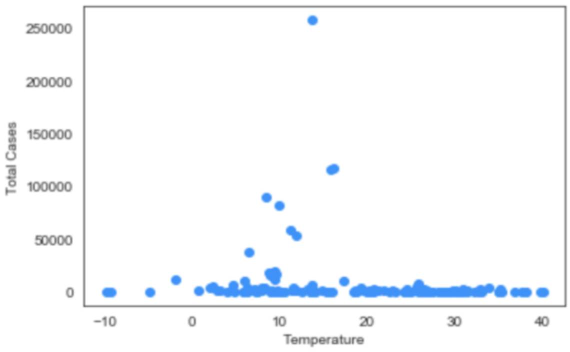
Our original dataset, covid19countryinfo.csv, contains essential demographic information for 196 countries related to COVID-19. After cleaning the data, we have 143 countries. The dependent variable is the *totalcases* of each country or the *totalcases per capita*. The independent variable is *avgtemp*, which is the average temperature of March. We selected 11 control variables, which are *avghumidity*, *pop*, *density*, *medianage*, *urbanpop*, *hospibed*, *smokers*, *sexratio*, *gdp2019*, *healthperpop*, *fertility*.



	avgtemp	avghumidity	pop	density	medianage	urbanpop	hospibed	smokers	sexratio	gdp2019	healthperpop	fertility	totalcases
country													
Afghanistan	6.22	56.93	38928346.0	60.0	18.0	25.0	0.5	22.997685	1.03	18734.0	4.73	5.12	273.0
Albania	13.95	69.90	2877797.0	105.0	36.0	63.0	2.9	29.400000	0.98	15418.0	268.96	1.51	304.0
Algeria	19.57	61.72	43851044.0	18.0	29.0	73.0	1.9	22.997685	1.03	172781.0	23.51	2.70	986.0
Andorra	7.38	75.23	77265.0	164.0	45.0	88.0	2.5	32.500000	1.06	91527.0	76994.76	1.40	439.0
Argentina	32.82	43.62	45195774.0	17.0	32.0	93.0	5.0	23.950000	0.98	445469.0	30.76	2.26	1265.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Uruguay	28.70	55.98	3473730.0	20.0	36.0	96.0	2.8	23.050000	0.94	59918.0	503.21	1.80	369.0
Uzbekistan	13.57	52.05	33469203.0	79.0	28.0	50.0	4.0	13.100000	0.99	60490.0	11.44	1.76	221.0
Venezuela	24.78	79.52	28435940.0	32.0	30.0	80.0	0.8	22.997685	0.99	70140.0	20.36	2.32	146.0
Vietnam	24.17	73.75	97338579.0	314.0	32.0	38.0	2.6	24.200000	1.00	261637.0	3.43	1.81	237.0
Zambia	24.97	87.77	18383955.0	25.0	18.0	45.0	2.0	15.550000	1.00	23946.0	11.04	5.63	39.0

From the distribution of total cases, it could be found that outbreak of COVID-19 didn't happen in most of the countries until March 2020. There are several spikes, implying the U.S., Spain, Italy, Germany, China, France and so on. And from the plot of total cases and temperature, there's no straightforward relationship so we would try the double Lasso for the estimation.





Double Lasso

i. Regression on Total Cases

Firstly, we normalized the independent variable and control variables and run the first Lasso:

$$Y = D\alpha + Z'\gamma + \epsilon,\ E[\epsilon|X] = 0$$

We get $\hat{\alpha}$ and the vector of $\hat{\gamma}$. Then run the second Lasso:

$$D = Z'\phi + \mu,\ E[\mu|Z] = 0$$

And we get $\hat{\phi}$ so as to use analogy principle to estimate $\tilde{\alpha}$ which is about -959.31, implying that temperature would depress the coronavirus infection. Keeping other control variables constant, as temperature increases 1 unit, the infection would decrease nearly 959 cases.

$$\tilde{\alpha} = \frac{\Sigma_{i=1}^{n}(Y_i - Z_i{'}\hat{\gamma})(D_i - Z_i{'}\hat{\phi})}{\Sigma_{i=1}^{n} D_i(D_i - Z_i{'}\hat{\phi})} \approx \text{-959.31}$$

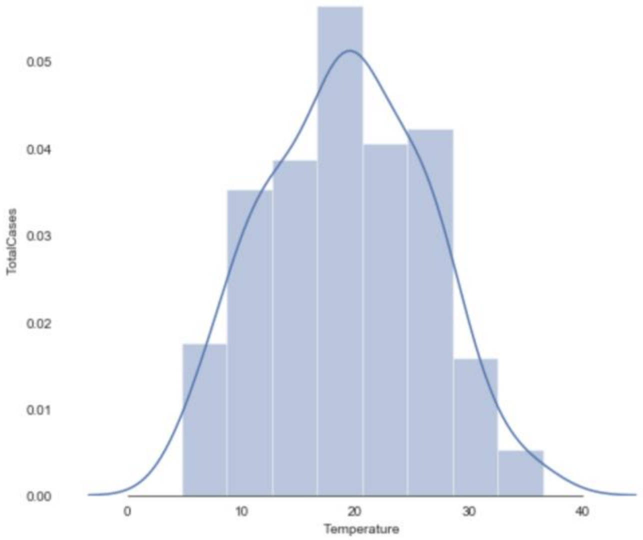
The strength of double Lasso is that we could get more precise outcome and the asymptotic distribution of parameters, also the confidence interval. Then we could calculate the variance of the coefficient using the formula:

$$V = \frac{E[\epsilon^2\mu^2]}{(E[\mu^2])^2}. \approx 2.4692$$

Then we get the asymptotic distribution of $\tilde{\alpha}$: $\sqrt{n}(\tilde{\alpha} - \alpha) \rightarrow\ _D N(0, V)$

The 95% confidence interval of the temperature coefficient is (-959.57, -959.05)

The plot below illustrates the prediction using double Lasso, which is unbiased.



ii. Regression on Total Cases per Capita

We use the same procedure but different dependent variable and get the following results. Run the first stage Lasso on total cases divided by population of each country and get the estimated coefficients. Then run the second stage Lasso on average temperature and calculate the $\tilde{\alpha}$ which is about -0.00011. According to the variance of $\tilde{\alpha}$, we get the asymptotic distribution: $\sqrt{n}(\tilde{\alpha}-\alpha) \rightarrow_D N(0,3.19e^{-10})$, also the 95% confidence interval of $\tilde{\alpha}$: (-0.000111, -0.000105)

From the double Lasso, we get the estimated $\tilde{\alpha}$ which is about -0.00011, implying that temperature would depress the coronavirus infection per capita. Keeping other control variables constant, as temperature increases 1 unit, the total cases per capita of a country would decrease nearly 0.011%.

Reference

https://www.worldometers.info/ - Population, Density, Median Age, Urban Population, Fertility Rate, Patient Zero Detection Date, Confirmed Cases, New Cases, Total Deaths, Total Recovered, Critical Cases.

@benhamner 's link (https://www.kaggle.com/benhamner) - Restrictions Initial dates.

https://worldpopulationreview.com/countries/smoking-rates-by-country/ - % of smokers by country.

https://data.worldbank.org/indicator/SH.MED.BEDS.ZS - Hospital beds per 1000 citizens.

https://en.wikipedia.org/wiki/List_of_countries_by_sex_ratio - Sex ratio by age.

https://www.worldlifeexpectancy.com/cause-of-death/lung-disease/by-country/ - Lung diseases death rate.

https://en.wikipedia.org/wiki/COVID-19_testing - COVID-19 Tests

https://www.worldbank.org/ - GDP 2019, Health Expenses (Whatever was missing was filled with information from Wikipedia)

https://en.climate-data.org/ - Temperature and Humidity raw data.