COVID Data Discussion

Saturday, June 20, 2020

Final Project

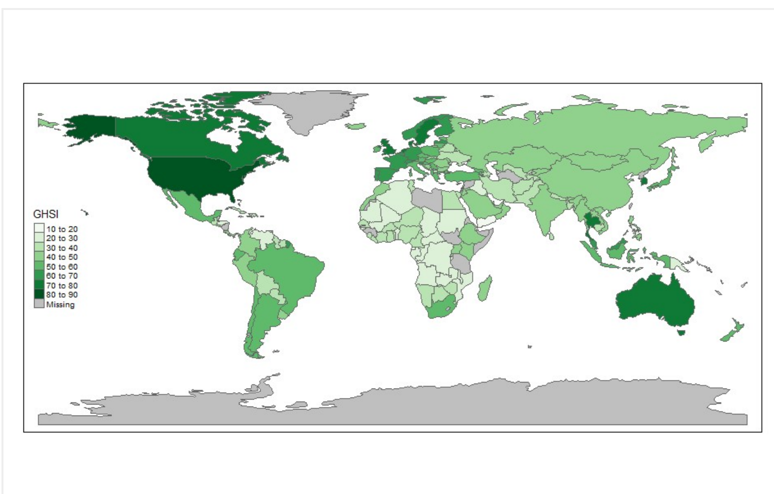
Anticipating Success

Looking at a disease's impact in countries around the world, there might be cost effective ways to attain better health success. The measures put in place around the world to prevent its spread are more disruptive and costly than any response policies used before. Government policies seek the best health security possible, and now that the policies have been tested, we can find out which ones work best.

Before the worst impacts of the epidemic were seen, policymakers looked for ways to prepare for a health security shock. An evaluation method was needed for federal decision makers around the world to determine paths to a more robust health sector. By adhering to normal healthcare guidelines, it was anticipated that healthcare systems would have more success mitigating the spread of an epidemic. In this report, the Global Health Security Index (GHSI) is used as a proxy for typical efforts to prepare for an epidemic. The index was released by Johns Hopkins, The Economist Intelligence unit, and the Nuclear Threat Initiative in 2019. The GHSI can be used as a way to rank countries by how prepared they were and how closely they followed establishment epidemiology norms. However, the factors and weightings were very carefully chosen, and so the GSHI can also be interpreted as a pre-COVID prediction of which policies would work best. The factors that are used to predict health security tend to relate to qualitative measures of healthcare infrastructure. Now that these measures have been tested, a re-evaluation can be done on whether COVID's attributes took establishment opinions on public health by surprise.

Figure 1

World Map With GHS Index



Although effort was taken by the creators of the GHSI to not allow too much influence from the wealth and quality of life of a country's inhabitants, many of the factors that make up the index seem to favour countries with big health budgets. Among these factors were "socio-economic resilience" and "infection control practices and availability of equipment". The context of historical handling of infectious diseases and things that worked in the past seem to influence the methodology of the index. Conversely, the precise takeaways from past epidemics are not so clear. Other aspects of the methodology are made more clear in the GHSI report. The indicators that were used are displayed clearly in the report. These indicators can be used to contextualize findings from COVID data by comparing results to a country's healthcare profile. Furthermore, their publicly available report gives qualitative viewpoints using healthcare terms, and this was helpful for finding practical implications for data insights. Overall the GHSI score was not a strong predictor in COVID models, but it serves as a good portal into the backdrop of health infrastructure situations before the pandemic. Looking at a correlation matrix, the relationship between the GHSI and real metrics of success can be seen.

Table 1

Correlation Between Indicators, by Country

	Cases in April	Cases in March	Cases per Capita	Deaths in March	Population	GHSI	Mortality in March	Health Spending
Cases in April	1	0.8452	0.0573	0.3882	0.1725	0.3551	0.0352	0.4715
Cases in March	0.8452	1	0.1785	0.7672	0.1207	0.3929	0.1369	0.5037
Cases per Capita, March	0.0573	0.1785	1	0.2125	-0.0747	0.082	0.0922	

								0.4655
Deaths in March	0.3882	0.7672	0.2125	1	0.0619	0.2532	0.2536	0.2574
Population	0.1725	0.1207	-0.0747	0.0619	1	0.0966	0.2288	-0.0511
GHSI	0.3551	0.3929	0.082	0.2532	0.0966	1	-0.0825	0.651
Mortality in March	0.0352	0.1369	0.0922	0.2536	0.2288	-0.0825	1	-0.0772
Health Spending	0.4715	0.5037	0.4655	0.2574	-0.0511	0.651	-0.0772	1

Note. A correlation value in this table describes the statistical relationship, and observed sensitivity, between the variable in the column title and a change in the variable in the row title. For indicators that correlate positively, there is a tendency to increase together. For negatively correlating indicators, there is a tendency for one to decrease as the other increases. As a rule of thumb, correlations above 0.7 can be considered strongly positive.

Looking at the effect of GHSI score and total monthly change in positive tests (from a correlation matrix), the correlation was 0.392. This might not be significant in terms of its effect on its own, but by looking at its combination with other indicators, it might help to tell a better story about why countries are getting better or worse over time. Since there is so much difference between the level to which different countries were hit early on in the pandemic, the only fair way to make an analysis is to take recordings over time in a specific country, and then find factors that separated the countries that were getting worse from the ones that were getting better. Comparing March to April, the average change in case numbers of 126% (or a factor of 2.26). The hardest hit countries were still the ones that are frequently talked about. The United Staes, Italy, Britain, Spain, Russia and Turkey had the highest number of cases without scaling to population size, but why were these countries hit harder than others?

We are going to look at diseases differently after this. The countries that saw high performance in real time success metrics only sometimes confirm prior expectations. Countries that had seen good statistics in the pandemic but had previously not adhered closely to health security norms might have done something that can be learned from. Trying to find factors that had the most effect on the disease spread was difficult, even with real COVID data. Healthcare infrastructure was not a strong predictor of how hard a country was hit (cases per capita). In the later stages of the pandemic, local disease spread within the borders of a country was much higher than cases coming from international disease hotspots.

Evaluating Success

Using subset selection techniques, it was determined that no model of the data more complex than a linear relationship from the number of cases in the previous had sufficiently low regression error or standard error to be accepted. It must be made clear that the effect of indicators other than infection rates from the previous month is only speculative. That being said, certain models with slightly higher complexities showed comparable regression error rates. The level of diminishing performance returns as complexity increased was disappointing, but the result of the modeling process are still relevant for policy decisions. The number of data points used in the analysis was only 107. To get a rolling monthly average, the original set of 16500 recordings was aggregated into sums and averages for the key indicators. Error rates with a more representative group of recordings might give the models more to work with, and it is worth exploring the possibility of a rolling weekly average. Furthermore, the removal of recordings with no data, or unusable data produced a substantial reduction of the number of rows in the datasets used for analysis. Priority was placed on creating consistency, and all of the analysis from this report was done on the same group of data. Missing values for health security, out of pocket health spending, country codes and population size were excluded, as well as values for monthly death rates or infection rates that were zero or negative, since they would prevent the ability to scale them, the ability to plot them or the ability to calculate mortality rates. Overall, the number of cases in March was the most positively amplifying factor for the number of cases in April. A country with 3,000 cases in March would be expected to have about 10,000 new cases in April. This makes sense, since the actual data shows an uptick in cases, with 4111 new cases per country on average in March versus 13422 new cases on average in April. The disease was in a growth stage during this time period, and the model reflects that. That being said there is a group of models which were strong performers, but couldn't be accepted because of the high standard error values. Using the "best" method, models, 4, 5, 6, and 7 performed better in certain tests (R^2, bic and cp tests can be used as proxies for regression error) than the accepted model, which only looks at the linear relationship between total new cases and cases in the previous month.

Table 2

Closeness of Fit of Model Predictions to Actual Cases in April, by Country

	1	4	5	6	7
R^2	0.71435	0.8789	0.88156	0.88336	0.88366

Note. An R^2 value in this table indicates the statistical tendency of a value in the data set to match the predicted value in the model number in the column title. All of the models are attempting to predict the number of cases in April, by country, from data available before the end of March. As a rule of thumb, a value for R^2 above 0.35 is considered meaningful in the social sciences. Table 3 gives coefficients appearing in these models.

What the first model does not communicate is when the growth stage will die down. The number of possible new cases has an upper bound at the total population size of a country, since the tests are done on people within the country's borders (although it's possible that some people have been tested accidentally twice in one month). In the models that included population size, there was not the expected negative correlation. Using model number 5, Canada's population size would explain an increase of 900 new cases in April out of the 44,000 total that were seen in that month.

Table 3

Coefficients from Five Models That Performed Well

Model Number	Intercept	Monthly Cases	Monthy Deaths	Country population	Cases per Capita	Health Spending
1	-187.514	3.439176				
4	-21.9767	5.413295	-39.6063			
5	-1107.7	5.3757	-39.363	0.0000239		
6	0.382196	5.385825	-38.88	0.000022	-4001613	
7	900.9042	5.452804	-39.4266	0.0000213	-3135385	-1.15539

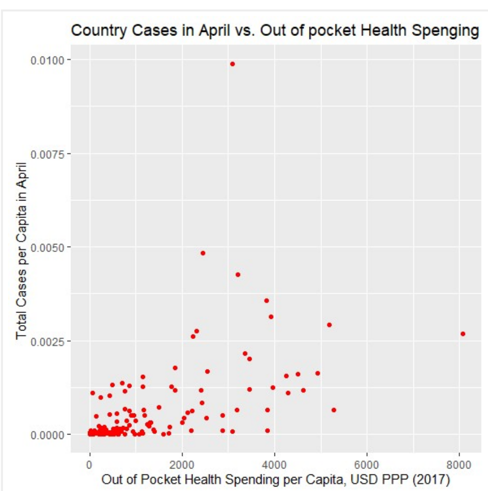
Note. The coefficients of the variables in the column titles can be used to make an equation that predicts the number of cases in April. The variables were made using information available before the end of March. The models were made using a statistical technique called the "best" method. The goodness of fit between these models and real data is shown in Table 2.

Population size could bolster the number of new cases if most of the susceptible population hasn't been infected yet. Once infected people are not exposed to uninfected, or susceptible people in most of their interactions, the possibility for new cases are diminished. A positive coefficient for population size can be explained by the fact that infection rates were growing quickly overall. Long before the point where everyone, or even half of a population, is infected, the number of recorded cases is limited by the number of tests done. Iceland, an exception to a trend of low testing rates around the world, has practically tested everyone within its borders, but it still does not test all of them every month and the number of recorded new cases will be less than the true new cases. The disparity of testing rates between countries is likely worthy of inclusion in a more complex model. Health spending per capita was included in the analysis as a proxy for many things, including tests per capita. According to an interactive article by "Our World in Data", countries characterized by somewhat or substantially lower than average gdp per capita did not come close to the testing rates that were typical for higher than average gdp per capita countries. It is still possible that, in a country with low testing rate, everyone who had the virus was tested. The symptoms of coronavirus are well documented and easy to find, and some confidence might be placed in expecting people that are sick to find a way to get tested. However, if the same 100,000 actual cases in Canada, or 2600 cases per million, were present in in Ethiopia, only 7.42 cases per million would be recorded.

The death rates in March were the most plausible outside influence on how much cases would grow between the two months. It is notable that the influence of death rates on the change of infection rates was negative. Since cases grew by a factor of 3.3 between the two months, any relationships that tend to decrease the number of cases are unexpected. The expected effect of high death rates is an occurrence of reactionary policies from local health services that double down on COVID protocol. It would be plausible that once a country is hit hard by the virus, health services will do more testing, and the recorded cases go up. If this anticipated effect was driving infection rates, the relationship would be positive, so this must be ruled out when considering policy takeaways. It is anticipated that higher death rates within a country lead people to be more careful about not getting infected in the first place. Furthermore, if people were dying in March, there's more incentive for a local government to enforce stricter guidelines in April. This could be seen as evidence that strict guidelines should be enforced before a country is hit hard. Once a country's healthcare system is overwhelmed, there is much more political incentive to find ways to combat the disease, but at this point many more resources are required to have the same effect. Another point is that there by the time a COVID death is recorded, there has already been a period of several weeks when that person was infected. If the toughest guidelines are only enforced once deaths are spiking, it might be the case that the reaction was too late.

The group of countries with disease success was not exclusive to countries with expensive disease prevention infrastructure. In fact, the correlation between health spending and cases per capita in the month of April was positive. This could indicate that rich countries have special vulnerability to infectious diseases that others don't. The chart below shows monthly infection rates (per capita) for 170 countries during the month of April.

Figure 2



Could it be the case that policymakers in countries with less means have become resourceful, and found strategies that the rest of the world could learn from? Looking at the countries who fit into the bottom 20% of healthcare spending in 2019 as well as the bottom 20% of infection rates, nineteen top performers were found. By chance alone, only 7 would be expected in this list, but this is particularly surprising since all of them except Laos were rated as having more vulnerable than average healthcare systems by the health security index in 2019. The top performers are listed in the following table:

Table 4

Countries With Lowest Infection Rates Among Those Who Spend Little on Health

Countries	Health Spending per Capita	Cases per 100,000, April

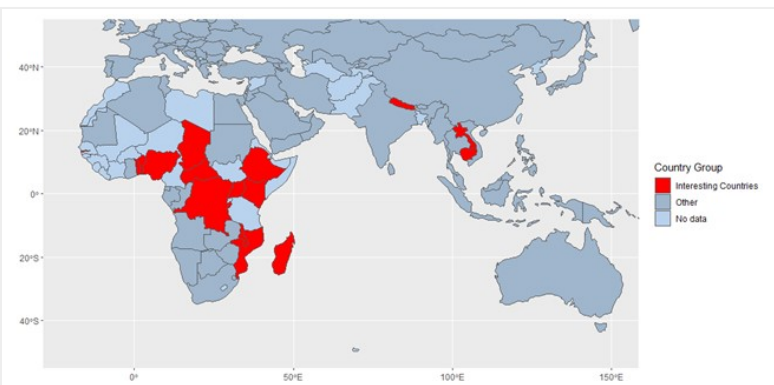
Burundi	15	0.13422365
Benin	17	0.54853928
Central African Republic	4	0.94291567
Democratic Republic of the Congo	4	0.47818381
Ethiopia	19	0.09796332
Gambia	14	0.30700381
Haiti	15	0.54840452
Kenya	52	0.64989383
Cambodia	50	0.08000099
Laos	50	0.14161283
Madagascar	43	0.3198493
Mozambique	58	0.23054003
Malawi	32	0.19842019
Nigeria	28	0.81531697
Nepal	29	0.18513329
South Sudan	28	0.30976902
Chad	18	0.29073991
Togo	20	0.95067951
Uganda	19	0.1123513

Note. A lower number of cases per 100,000 of one country compared to another is interpreted as a relative success. Quintiles were used to identify which countries among the 20% with the lowest out of pocket health spending as well as being among the 20% with the lowest infection rates. The purpose of selecting these countries was to show outliers from expected trends. There are other methods of identifying outliers such as clustering, but this method was deemed sufficient for the purposes of informing policy decisions. Having more than the expected value from chance alone (4%) of the countries appearing in this category is an additional insight that underscores the positive correlation between health spending and April infection rates.

Certain similarities might appear between the countries. Many of them are close together, along the equator on the African continent. The low infection rates might be explained by the hot weather in these countries. Even Nepal has a relatively warm climate despite the effects of high altitude on weather. If seasonality is one of COVID's attributes, hot weather would slow the disease's spread. The following map shows their locations in the world.

Figure 3

International Map with Interesting Countries Highlighted in Red



Note. The countries highlighted in red are the same ones that appear in Table 4.

Geography could influence infection rates in other ways. If the countries are less physically or economically connected to disease hotspots, then they minimize the number of disease-carrying visitors that come in and infect others. Inbound flight data was examined from the last ten years, and the most recent recording was observed if there was one present. Inbound flight volumes ranged from 38,000 in the Central African Republic to 3,207,000 flying into Laos (although only 6 of the countries had a datapoint from the last 10 years). The median number of flights into a country in 2018 was 84,730,004 among the 119 countries that had data for that year (264 countries were present in the overall dataset). Although this is not a per-capita figure, this leads to a picture of a group of isolated countries where tourism is relatively infrequent. One other possibility is that the weaker healthcare infrastructure in these countries might lower the life expectancy, resulting in a much lower number of elderly people. Hospital admission rates are much higher for elderly people. If most people in a country are following physical distancing guidelines in everyday, then hospitals could become infection hotspots when patients are brought close together. This is one way that elderly people could be more infectious than younger people.

Another explanation is that their testing rates were lower than countries with more resources. This is the factor that I believe is most likely to be causing the unexpected results. An interactive article by "Our World in Data" show that while Canada tests 99 people for every one person who tested positive, Nigeria tests 5.25, Nepal tests 15.34, Kenya test 24.59, and Ethiopia tests 29.58. Many of these countries did not release enough data to make it into this data set. Without merging the data sets for statistical analysis, it seemed as though countries with higher health budgets such as the United States, Norway and Luxembourg tended to appear in the data set and tended to have higher testing rates.

Creating Success

Despite an increase in cases between March and April, it was determined that this time frame was a late stage of the pandemic. This report suggests that control methods needed to be put in place in early on in order to have the most impact. By April, there were no factors that conclusively affected the absolute rate of new recorded cases. However, other factors may have determined each country's rate of change over the month. This report suggests that a disparity in testing rates between countries is likely to be masking trends in true infection rates, especially among countries with weaker health infrastructure. The models that were tested showed that the inherent infectivity of the virus was the most likely reason that the average country saw a threefold increase in new recorded cases. The tendency for countries with high death rates to see lower infection rates after a one month period should be investigated. This report recommends that policy should not react to death rate spikes, since the opportunity to create the most success comes earlier.

There is is not enough information to discredit the GHSI as a predictor for COVID success, despite it showing correlation with higher infection rates. Likewise, there is not enough information to analyse whether funding for health risk mitigation was well spent overall, despite correlation between high health spending and increased infection rates. Further investigation should be done on the true infection rates from countries who have below average out of pocket healthcare expenditures and substantially lower than average testing rates. Further investigation should also be done on measures of country interconnectivity during key time frames early on in the epidemic's early stages. This could include inbound flight volumes from disease hotspots before the outbreak and the speed with which border crossings tapered off in early stages of the pandemic.

Appendix
Subset Selection

i) Using the "best" Method

```
best = regsubsets(cases.april~., data = geoNoName, nvmax=20)
```

```
         cases percapita deaths popData2018 ghsi mortality healthSpending
1  ( 1 ) "*"   " "       " "    " "         " "  " "       " "
2  ( 1 ) "*"   " "       "*"    " "         " "  " "       " "
3  ( 1 ) "*"   " "       "*"    "*"         " "  " "       " "
4  ( 1 ) "*"   "*"       "*"    "*"         " "  " "       " "
5  ( 1 ) "*"   "*"       "*"    "*"         " "  " "       "*"
6  ( 1 ) "*"   "*"       "*"    "*"         " "  "*"       "*"
7  ( 1 ) "*"   "*"       "*"    "*"         "*"  "*"       "*"
```

	1	2	3	4	5	6	7
$rsq	0.71435	0.8789	0.88156	0.88336	0.88366	0.88366	0.88367
$cp	185.208	3.3231	2.3512	2.3463	4.0132	6.0068	8
$bic	-163.06	-276.56	-274.7	-271.88	-267.3	-262.38	-257.46

```
bestII = regsubsets(cases.april~poly(cases, 2)+poly(deaths, 2)+poly(popData2018, 2)+
            +poly(ghsi, 2)+poly(mortality, 2)+ poly(healthSpending, 2)+poly(percapita, 2),
           data = geoNoName, nvmax = 10)
```

factor ranking table:

Factors	1	2	3	4	5	6	7	8	9	10
poly(cases, 2)1	“*”	“*”	“*”	“*”	“*”	“*”	“*”	“*”	“*”	“*”
poly(cases, 2)2	“ “	“*”	“*”	“*”	"*"	“*”	“*”	“*”	“*”	“*”
poly(deaths, 2)1	“ “	“ “	“ “	“*”	“*”	“*”	“*”	“*”	“*”	“*”
poly(deaths, 2)2	“ “	“ “	“*”	“*”	“ “	“*”	“*”	“*”	“*”	“*”
poly(popData2018, 2)1	“ “	“ “	“ “	“ “	“*”	“ “	“*”	“*”	“*”	“*”
poly(popData2018, 2)2	“ “	“ “	“ “	“ “	“*”	“*”	“*”	“*”	“*”	“*”
poly(ghsi, 2)1	“ “	“ “	“ “	“ “	“ “	“*”	“*”	“*”	“*”	“*”
poly(ghsi, 2)2	“ “	“ “	“ “	“ “	“ “	“ “	“ “	*”	“*”	“*”
poly(mortality, 2)1	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “
poly(mortality, 2)2	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “
poly(healthSpending,2)1	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“*”
poly(healthSpending,2)2	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“*"	“*"
poly(percapita, 2)1	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “
poly(percapita, 2)2	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “	“ “

Model strength table:

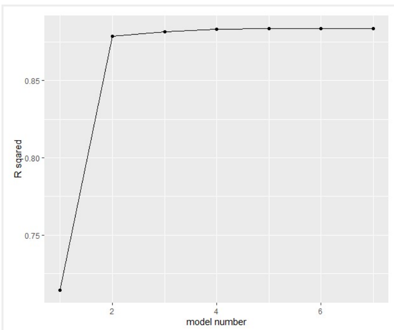
Column1	1	2	3	4	5	6	7	8	9	10
$rsq	0.714	0.887	0.94	0.947	0.953	0.957	0.958	0.958	0.959	0.959
$cp	734.176	210.035	52.162	31.954	17.118	7.351	6.188	6.405	7.083	8.636
$bic	-163.058	-286.675	-368.69	-381.704	-392.125	-399.038	-397.474	-394.48	-391.004	-386.57

ii) Using the "forward" Method:

forward = regsubsets(cases.april~.,method = "forward", data=geoNoName)

```
        cases percapita deaths popData2018 ghsi mortality healthSpending
1  ( 1 ) "*"   " "       " "    " "         " "  " "       " "
2  ( 1 ) "*"   " "       "*"    " "         " "  " "       " "
3  ( 1 ) "*"   " "       "*"    "*"         " "  " "       " "
4  ( 1 ) "*"   "*"       "*"    "*"         " "  " "       " "
5  ( 1 ) "*"   "*"       "*"    "*"         " "  " "       "*"
6  ( 1 ) "*"   "*"       "*"    "*"         " "  "*"       "*"
7  ( 1 ) "*"   "*"       "*"    "*"         "*"  "*"       "*"
```

	1	2	3	4	5	6	7
$rsq	0.7143502	0.8789035	0.8815629	0.883357	0.8836552	0.8836609	0.883667
$cp	185.20833	3.323101	2.35123	2.346333	4.01315	6.006791	8
$bic	-163.0579	-276.5593	-274.6965	-271.8757	-267.3017	-262.3812	-257.4611



iii) Using the "backward" Method:

backward = regsubsets(cases.april~.,method = "backward", data=geoNoName, nvmax = 12)

Factor ranking table:

```
        cases percapita deaths popData2018 ghsi mortality healthSpending
1  ( 1 ) "*"   " "       " "    " "         " "  " "       " "
2  ( 1 ) "*"   " "       "*"    " "         " "  " "       " "
3  ( 1 ) "*"   " "       "*"    "*"         " "  " "       " "
4  ( 1 ) "*"   "*"       "*"    "*"         " "  " "       " "
5  ( 1 ) "*"   "*"       "*"    "*"         " "  " "       "*"
6  ( 1 ) "*"   "*"       "*"    "*"         " "  "*"       "*"
7  ( 1 ) "*"   "*"       "*"    "*"         "*"  "*"       "*"
```

Model strength table:

	1	2	3	4	5	6	7
$rsq	0.71435	0.8789	0.88156	0.88336	0.88366	0.88366	0.88367
$cp	185.208	3.3231	2.35123	2.34633	4.01315	6.00679	8
$bic	-163.06	-276.56	-274.7	-271.88	-267.3	-262.38	-257.46

Correlation Matrix:

	cases.april	cases	percapita	deaths	popData2018	ghsi	mortality	healthSpending
cases.april	1	0.8452	0.0573	0.3882	0.1725	0.3551	0.0352	0.4715
cases	0.8452	1	0.1785	0.7672	0.1207	0.3929	0.1369	0.5037
percapita	0.0573	0.1785	1	0.2125	-0.0747	0.082	0.0922	0.4655
deaths	0.3882	0.7672	0.2125	1	0.0619	0.2532	0.2536	0.2574
popData2018	0.1725	0.1207	-0.0747	0.0619	1	0.0966	0.2288	-0.0511
ghsi	0.3551	0.3929	0.082	0.2532	0.0966	1	-0.0825	0.651
mortality	0.0352	0.1369	0.0922	0.2536	0.2288	-0.0825	1	-0.0772
healthSpending	0.4715	0.5037	0.4655	0.2574	-0.0511	0.651	-0.0772	1

backwardII = regsubsets(cases.april~poly(cases, 2)+poly(deaths, 2)+poly(popData2018, 2)+
+poly(ghsi, 2)+poly(mortality, 2)+ poly(healthSpending, 2)+poly(percapita, 2),
data = geoNoName, nvmax = 20, method = "backward")

Using which.max(summary(backwardII)$rsq),
we get: 14
Using which.min(summary(backward)$cp),
we get: 7
which.min(summary(backward)$bic),
we get: 6

coef(backwardII, 14) gives:

(Intercept)	**17005.449**
poly(cases,2)1	752927.461
poly(cases,2)2	346483.904
poly(deaths,2)1	-128900.223
poly(deaths,2)2	-148117.873
poly(popData2018,2)1	41413.87
poly(popData2018,2)2	-83651.291
poly(ghsi,2)1	38921.175
poly(ghsi,2)2	22799.353
poly(mortality,2)1	-11214.812
poly(mortality,2)2	1892.344
poly(healthSpending,2)1	46790.443
poly(healthSpending,2)2	-22648.996
poly(percapita,2)1	-23560.273
poly(percapita,2)2	20318.525

coef(backwardII, 7) gives:

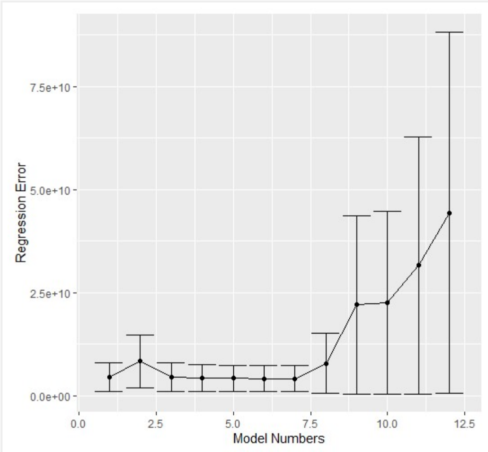
(Intercept)	**17005.45**
poly(cases, 2)1	770714.66
poly(cases, 2)2	343963.24
poly(deaths, 2)1	-148594.39
poly(deaths, 2)2	-149787.66
poly(popData2018, 2)1	29961.83
poly(popData2018, 2)2	-73846.04
poly(ghsi, 2)1	62908.65

coef(backwardII, 6) gives:

(Intercept)	**17005.45**
poly(cases, 2)1	769627.01
poly(cases, 2)2	348137.8
poly(deaths, 2)1	-144554.9
poly(deaths, 2)2	-154352.4
poly(popData2018, 2)2	-71443.4
poly(ghsi, 2)1	65222.57

model 1: cases.april~cases
model 2: cases.april~healthSpending   # another strong performer in correlation matrix
model 3: cases.april~cases+healthSpending #two strong performers in the correlation matrix
model 4: cases.april~cases+deaths   #model 2 from best ranking table
model 5: cases.april~cases+deaths+popData2018   #model 3 from best ranking table
model 6: cases.april~cases+percapita+deaths+popData2018 #model 4 from best ranking table
model 7: cases.april~cases+percapita+deaths+popData2018+healthSpending   #model 5 from best
model 8: cases.april~poly(cases, 2)"   #model 1 from bestII
model 9: cases.april~poly(cases, 2)+ poly(deaths, 2)" #model 3 from bestII ranking table
model 10: cases.april~poly(cases, 2)+ poly(deaths, 2)+poly(popData2018, 2)   #model 5, bestII
model 11: cases.april~poly(cases, 2)+ poly(deaths, 2)+poly(popData2018, 2) + poly(ghsi, 2)   #model 6 from bestII ranking table, also model 7 in backwardII

model 12: cases.april~poly(cases, 2)+poly(deaths, 2)+poly(popData2018, 2)+poly(ghsi, 2)+poly(mortality, 2)+poly(healthSpending, 2)+poly(percapita, 2)"   #model 14 in backwardII



x	avgRegEr.k	se
1	4452755877	3492497014
2	8372382008	6348931509
3	4560719226	3462880513
4	4291616269	3238283498
5	4233503189	3172083339
6	4170117672	3120337318
7	4180422456	3135105180
8	7845729976	7326962454
9	22093026832	21605162077
10	22601668855	22136720093
11	31593862803	31151474681
12	44356765740	43846459788

Model 6 has the least error, at 4170117672. Models 1, 3, 4, 5, 6, 7 are within one the standard error range from the mean standard error of model 6. Among these options, model 1 is the simplest by far, it is just a linear relationship with the cases in the previous month. Therefore, model 1 is accepted. Taking another look at this model, the total cases in April for a particular country are best modeled as 3.439 times the total cases in March minus 187.514 extra cases. A country with 76.891 cases (theoretically) in March will not increase in their case count in April, since 76.891(3.439)-187.514 = 76.881.

Posted by Luke Wedgwood's Random School Work at 7:59 PM

No comments:

Post a Comment

Enter your comment...

Comment as: Luke Wedgwoc

Sign out

Publish Preview

Notify me