PharmaSUG 2022 - Paper AI-116

Potential Variables Analysis Affects Covid-19 Spread

Yida Bao, Auburn University

Philippe Gaillard, Florida State University

ABSTRACT

The emergence of the Covid19 virus has deeply affected the world for two years. There are no government or effective means that can completely suppress its spread. On the other hand, we cannot deny that the spreading map of the Covid-19 virus is not completely average. There are always regions with fewer infections and deaths than others. In this paper, we will use different features to explore the reasons. We selected the epidemic data from 2,352 counties in the United States and updated the latest infection and death number in 2022. Meanwhile, we introduced 28 objective variables such as temperature, longitude and latitude, diabetes Ratio, smoker proportion, etc. for statistical analysis. This is a very meaningful research project, which can help health institutions to invest the right resources according to different situations in different regions. Basic and graceful Linear Regression will be used in our projects, telling us a lot of potential information. Besides, we use a series of machine learning algorithms such as neural networks, decision trees, etc. to predict the results. SAS® 9.4 and SAS® enterprise miner will be the main platforms for this project.

INTRODUCTION

In a few short years, many of the world's intrinsic rules were changed by the Covid-19 epidemic. As Laura Spinney writes in her book Pale Rider: "Many countries created or adapted health departments in the 1920s", Covid-19 has also affected a lot of aspects of the world. As a statistic programmer group, we don't have the opportunity and ability to study virus mutations or make vaccines like frontline scientists, we still feel interested in any project related to Covid-19. This is not a complicated project, and we are all using familiar statistic tools. Anyone can easily read and understand this paper. We will develop our ideas step by step according to the flow chart.

Dataset
Clean
idea
model
discussion

DATA SET AND DATA CLEAN

We have selected the data we need from the resource from the paper ‘Dataset of COVID-19 outbreak and potential predictive features in the USA’（hereinafter referred to as “DCOPPF” ）. This paper did not process the data or hold an opinion but did collect a significant amount of features in the dataset, which greatly reduced the workload of peers.

In ’DCOPPF’ paper, they collected 46 different variables that are closely related to Covid-19, and the information covered the 3142 counties in the USA. However, after checking, we found that there were several missing values in some of the factors we need. There are many ways of dealing with missing values, such as replacing them with Arbitrary Values or using a multivariate approach. Since this is a realistic project, so we decided that instead of using statistical methods to fill up for the missing value, we should just discard the counties to keep the prediction result more convincing. But these are not what we are going to do. This is a project-based on real data. Rather than simulating data to keep more sample size, we should use the intact sample size available to make a more robust data analysis. So, our dataset cover 2352 counties in the United States and 28 features from the original dataset. More, two important variables in the original dataset, cumulative Covid-19 cases and cumulative Covid-19 deaths per state, have only been updated to June 2020. We have taken it upon ourselves to capture the up to Jan 22 2022 confirmed Covid-19 cases and death number in these 2352 counties, which also make our project more timely.

The ’DCOPPF’ paper already gave a detailed description of each factor; we will briefly introduce those features here.

- **Target Variable**

Unsurprisingly, the confirmed case and death number will be the target variable in our article. Discussions and predictions based on these two variables will be presented throughout the article. In the 'DCOPPF paper, these two variables confirmed new cases and deaths on a daily basis, which facilitates statistical analysis by means of time series analysis. In this paper, we did not have in mind the use of time series analysis; our aim was to establish the relationship between the confirmed case, and death number with other objective factors. Therefore, we have used the cumulative total number of confirmed cases and death number for each county as our target variables.

From these two variables, we proceeded to clean the dataset. Since different counties have different populations, in Linear Regression, the confirmed case would be linearly dependent on the population of this county. So we used the confirmed case number to divide by population, that is

$$Case\ Ratio = \frac{Confirmed\ Case\ number}{Population},$$

The Death Ratio was obtained in the same way. From our calculations, we will easily find that when the Case Ratio is getting smaller, the epidemic is less severe in this area; In contrast, the Case Ratio is getting larger, which indicates the epidemic is more severe.

The next step is to prepare for the machine learning process and the classification process. Since the Case Ratio and the Death Ratio are both continuous variables, we expect two categorical variables based on the Ratio if we need to perform classification prediction. The approach we take is to split them according to the mean. A simple description of the data was made and, based on the results, two new categories were introduced, the new Case Ratio and the new Death Ratio. They have only two values, 0 and 1. We will not use them for now until the classification process.

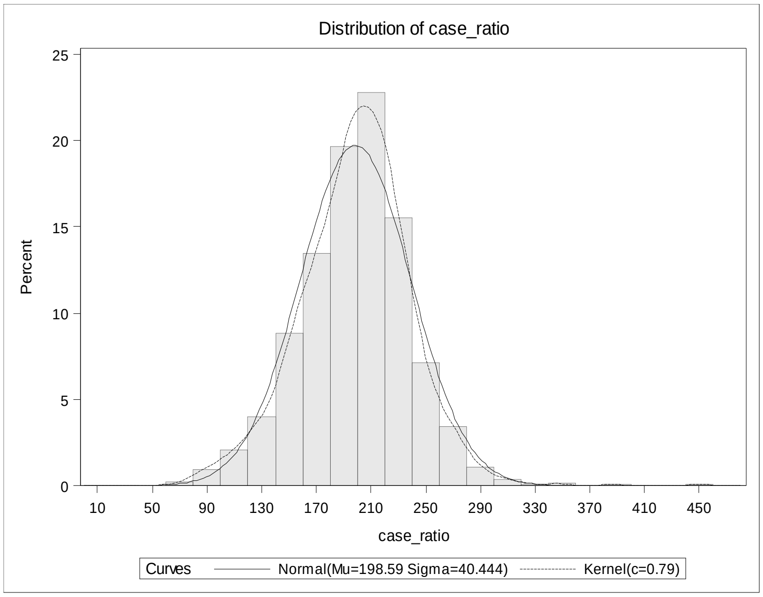


Figure 1 . Distribution for Case Ratio

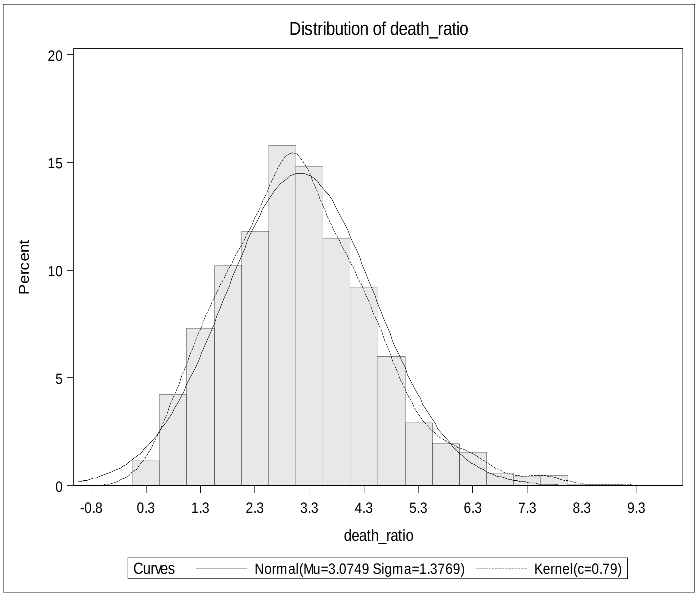


Figure 2 . Distribution for Death Ratio

	New Case Ratio		***New Death Ratio***
Case Ratio < 198	***0***	***Death Ratio < 3.07***	***0***
Case Ratio > 198	***1***	***Death Ratio > 3.07***	***1***

Here’s a table for the dependent variables

• **Demographic Variable**	Description
Gender Percentage	Gender percentage
Population Density	We use population divide area for each county
Lcu Beds Ratio	Icu beds Ratio
Less High School Diploma	Education background
High School Diploma	Education background
Some College or Higher	Education background
Religious Congregation Ratio	Total number of active members of a county’s religious congregations.
Immigratnt Student Percentage	Immigrant students are those who enrolled in the fall of 2018 in any institution in the county but reside in another state

• **Temporal Variable**	Description
Precipitation	Daily precipitation
Temperature	Average temperature each day
Social Distance	We take the mean of three variable, social distancing travel distance grade, social distancing visitation grade and social distancing grade
Virus Pressure	Explain below

The virus pressure at county x_i and day t, denoted by V (x_i, t), is defined based on the number of COVID -19 cases in the neighboring counties:

$$V\ (x_i, t) = \frac{\sum_{x_k \in N(x_i)} C(x_k, t)}{|N(x_i)|} ,$$

where $C(x_k, t)$ denotes the number of COVID-19 cases in county x_k at day t, and $N(x_i)$ is the set of all adjacent counties that share a border with county x_i , excluding x_i itself.

• **Geographic Variable**	Description
Latitude	Latitude information
Longitude	Longitude information
Airport Distance	Distance to the nearest international airport with average daily

	passenger load more than ten.
Passenger Load Ratio	Average daily passenger load of that nearest international airport divided by the total population.

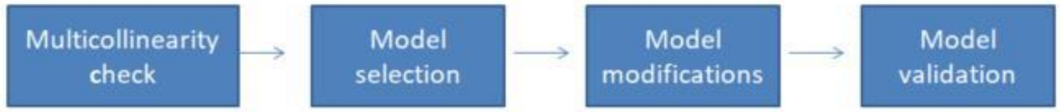
• **Economic and Health Variable**	Description
House Density	House density information
Smoker Percentage	Smoker percentage of total population
Diabetes Percentage	Diabetes percentage of total population.
Political Party	The political party of the governor of each state (0 for Republican and 1 for Democratic)
Meat Plants	Number of meat processing plants
Median House Income	Median house income information
Insured Percentage	Percentage of health insured residents
GDP	Gross Domestic Product per capital

Each of these variables is described in detail in the 'DCOPPF paper and this article focuses on the use of these data for data processing.

METHOD AND MODEL

Linear Regression

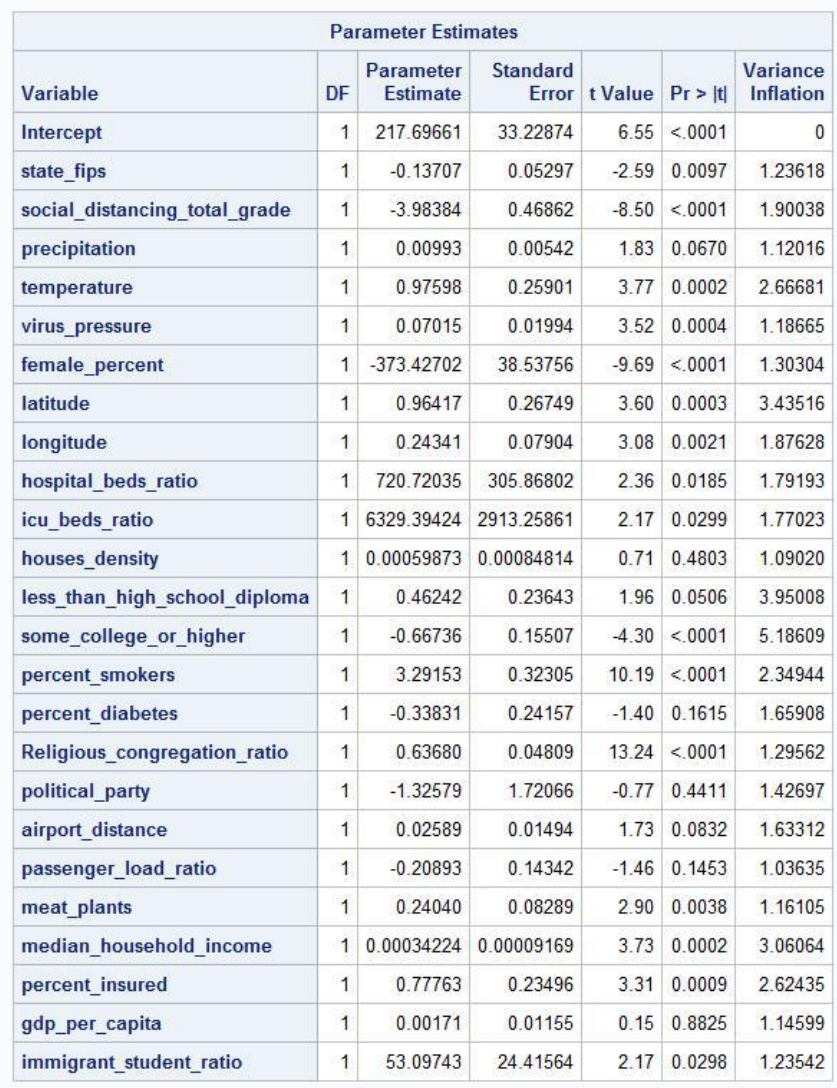
If we need to present the correlation between the target variable and independent variable in a most understandable statistical way, Linear Regression is our first choice. We would like to fit a Linear Regression model to use for those variables.



The assumptions of Linear Regression do not require that there cannot be multicollinearity between the independent variables. We perform the multicollinearity diagnosis for reasons of model stability. The term 'instability' warrants a significant change in the model results when the variable is slightly changed. So it is necessary to diagnose multicollinearity before processing a multiple Linear Regression.

```
proc reg data = Covid191 ;
  fullmodel: model Case_Ratio = social_distancing_total_grade
                                recipitation
                                virus_pressure /*population_density*/
                                ....../ vif;
run;
```

We use VIF for diagnostics, which stands for variance inflation factor. VIF is a measure of the increase in variance caused by covariance, also known as variance inflation. When VIF > 10, it indicates that the variable may have a high Multilinearity with some other variables, which may cause the model to be unstable. When a very high VIF value is found for some variable, we will use /* variable */ to remove the factor. Once we have removed the variables High School_Diploma_Only, Population_Density, and Ventilator_Capacity_Ratio, the VIF values for all variables are less than 10, which means we can proceed to the next step of model selection.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	217.69661	33.22874	6.55	<.0001	0
state_fips	1	-0.13707	0.05297	-2.59	0.0097	1.23618
social_distancing_total_grade	1	-3.98384	0.46862	-8.50	<.0001	1.90038
precipitation	1	0.00993	0.00542	1.83	0.0670	1.12016
temperature	1	0.97598	0.25901	3.77	0.0002	2.66681
virus_pressure	1	0.07015	0.01994	3.52	0.0004	1.18665
female_percent	1	-373.42702	38.53756	-9.69	<.0001	1.30304
latitude	1	0.96417	0.26749	3.60	0.0003	3.43516
longitude	1	0.24341	0.07904	3.08	0.0021	1.87628
hospital_beds_ratio	1	720.72035	305.86802	2.36	0.0185	1.79193
icu_beds_ratio	1	6329.39424	2913.25861	2.17	0.0299	1.77023
houses_density	1	0.00059873	0.00084814	0.71	0.4803	1.09020
less_than_high_school_diploma	1	0.46242	0.23643	1.96	0.0506	3.95008
some_college_or_higher	1	-0.66736	0.15507	-4.30	<.0001	5.18609
percent_smokers	1	3.29153	0.32305	10.19	<.0001	2.34944
percent_diabetes	1	-0.33831	0.24157	-1.40	0.1615	1.65908
Religious_congregation_ratio	1	0.63680	0.04809	13.24	<.0001	1.29562
political_party	1	-1.32579	1.72066	-0.77	0.4411	1.42697
airport_distance	1	0.02589	0.01494	1.73	0.0832	1.63312
passenger_load_ratio	1	-0.20893	0.14342	-1.46	0.1453	1.03635
meat_plants	1	0.24040	0.08289	2.90	0.0038	1.16105
median_household_income	1	0.00034224	0.00009169	3.73	0.0002	3.06064
percent_insured	1	0.77763	0.23496	3.31	0.0009	2.62435
gdp_per_capita	1	0.00171	0.01155	0.15	0.8825	1.14599
immigrant_student_ratio	1	53.09743	24.41564	2.17	0.0298	1.23542

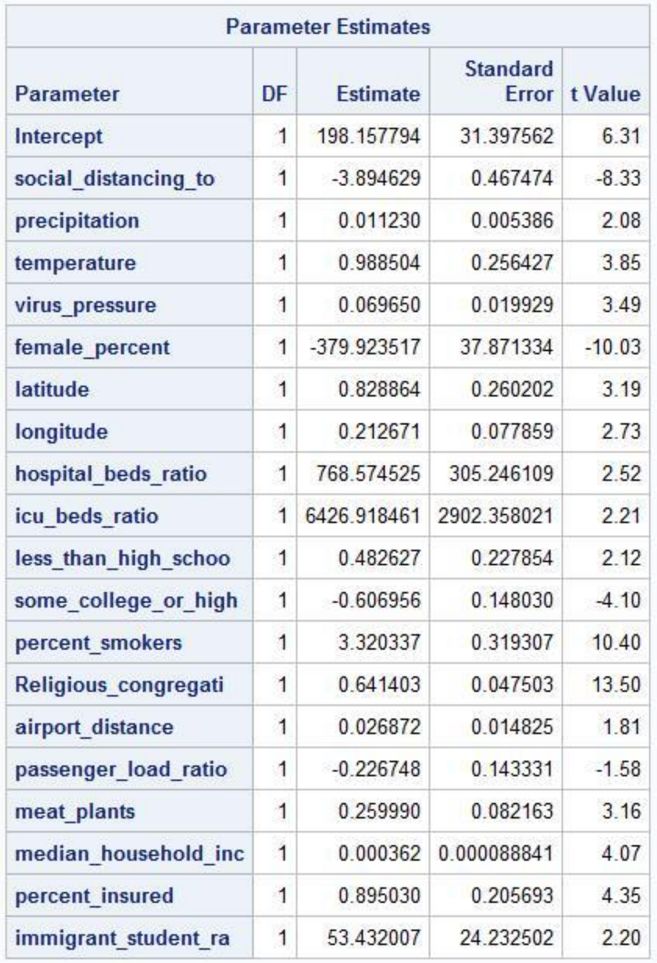
Figure 3. Figure for VIF

The GLMSELECT procedure implements the selection of statistical models within the framework of a general linear model. The approach includes not only extensions to GLM-type models, but also the familiar procedure proc reg.

Here’s the result for **GLMSELECT** procedure for Case Ratio.

```
proc glmselect data=Covid191 plots=asePlot;
     model case_Ratio =   social_distancing_total_grade  precipitation
                          temperature  virus_pressure female_percent
                           /*population_density*/
                           ....................
                          / selection= stepwise (select=SL);
```

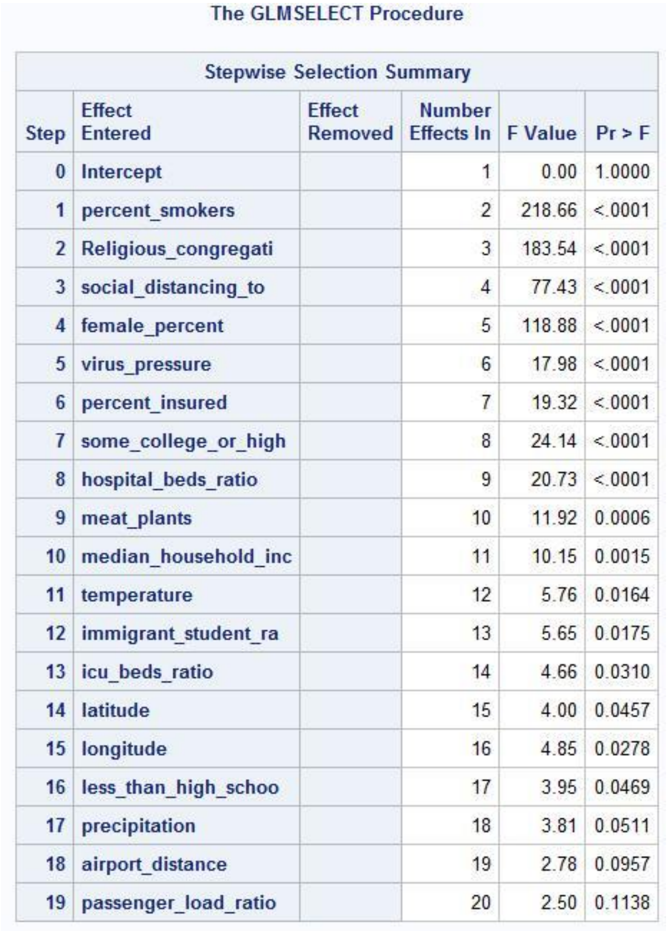
```
Run;
```



Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	198.157794	31.397562	6.31
social_distancing_to	1	-3.894629	0.467474	-8.33
precipitation	1	0.011230	0.005386	2.08
temperature	1	0.988504	0.256427	3.85
virus_pressure	1	0.069650	0.019929	3.49
female_percent	1	-379.923517	37.871334	-10.03
latitude	1	0.828864	0.260202	3.19
longitude	1	0.212671	0.077859	2.73
hospital_beds_ratio	1	768.574525	305.246109	2.52
icu_beds_ratio	1	6426.918461	2902.358021	2.21
less_than_high_schoo	1	0.482627	0.227854	2.12
some_college_or_high	1	-0.606956	0.148030	-4.10
percent_smokers	1	3.320337	0.319307	10.40
Religious_congregati	1	0.641403	0.047503	13.50
airport_distance	1	0.026872	0.014825	1.81
passenger_load_ratio	1	-0.226748	0.143331	-1.58
meat_plants	1	0.259990	0.082163	3.16
median_household_inc	1	0.000362	0.000088841	4.07
percent_insured	1	0.895030	0.205693	4.35
immigrant_student_ra	1	53.432007	24.232502	2.20

Figure 4. Coefficients for Case Ratio model



The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	percent_smokers		2	218.66	<.0001
2	Religious_congregati		3	183.54	<.0001
3	social_distancing_to		4	77.43	<.0001
4	female_percent		5	118.88	<.0001
5	virus_pressure		6	17.98	<.0001
6	percent_insured		7	19.32	<.0001
7	some_college_or_high		8	24.14	<.0001
8	hospital_beds_ratio		9	20.73	<.0001
9	meat_plants		10	11.92	0.0006
10	median_household_inc		11	10.15	0.0015
11	temperature		12	5.76	0.0164
12	immigrant_student_ra		13	5.65	0.0175
13	icu_beds_ratio		14	4.66	0.0310
14	latitude		15	4.00	0.0457
15	longitude		16	4.85	0.0278
16	less_than_high_schoo		17	3.95	0.0469
17	precipitation		18	3.81	0.0511
18	airport_distance		19	2.78	0.0957
19	passenger_load_ratio		20	2.50	0.1138

Figure 5 P-value for Case Ratio model

Using a traditional stepwise select regression analysis, we get a model with a 25.96% R square Model. We can see that the proportion of smokers, social distance grade, virus pressure, educational background, proportion of hospital beds, proportion of ICUs, house prices, climate , and geographical location all have an effect on the number of confirmed Covid-19 diagnoses.

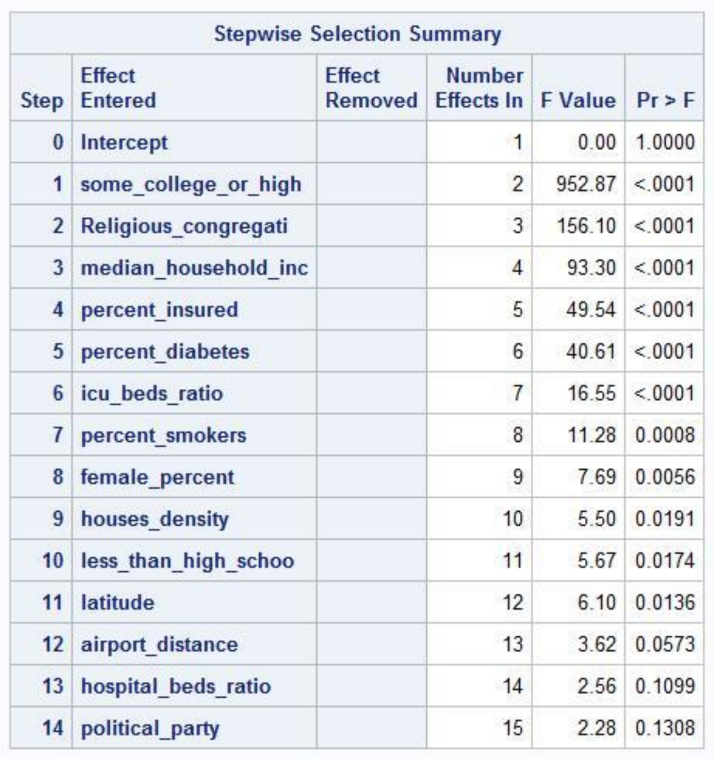
The results are almost identical to our common sense Covid-19 spread, in that having a better social distance, better access to health care, better personal life habits ,and a higher median house price can suppress the increase in confirm number. University towns with large numbers of students, and relatively hot counties, also increase the likelihood of Covid-19 spreading.

The results also remove the effects of political affiliation, GDP ,and diabetes rates on the number of diagnoses.

Here's the result for the GLMSELECT procedure for Death Ratio.

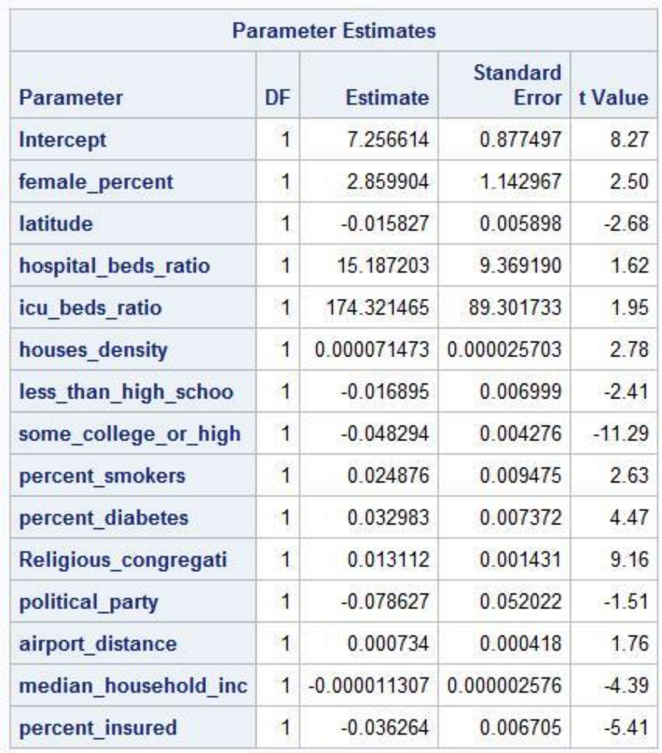
```
proc glmselect data=Covid191 plots=asePlot;
model death_Ratio =   social_distancing_total_grade  precipitation
```

```
                                        temperature  virus_pressure female_percent
                                       /*population_density*/

                                         ....................

                                       / selection= stepwise (select=SL);

run;
```



Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	some_college_or_high		2	952.87	<.0001
2	Religious_congregati		3	156.10	<.0001
3	median_household_inc		4	93.30	<.0001
4	percent_insured		5	49.54	<.0001
5	percent_diabetes		6	40.61	<.0001
6	icu_beds_ratio		7	16.55	<.0001
7	percent_smokers		8	11.28	0.0008
8	female_percent		9	7.69	0.0056
9	houses_density		10	5.50	0.0191
10	less_than_high_schoo		11	5.67	0.0174
11	latitude		12	6.10	0.0136
12	airport_distance		13	3.62	0.0573
13	hospital_beds_ratio		14	2.56	0.1099
14	political_party		15	2.28	0.1308

Figure 6. p-value for Death Ratio Model



Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	7.256614	0.877497	8.27
female_percent	1	2.859904	1.142967	2.50
latitude	1	-0.015827	0.005898	-2.68
hospital_beds_ratio	1	15.187203	9.369190	1.62
icu_beds_ratio	1	174.321465	89.301733	1.95
houses_density	1	0.000071473	0.000025703	2.78
less_than_high_schoo	1	-0.016895	0.006999	-2.41
some_college_or_high	1	-0.048294	0.004276	-11.29
percent_smokers	1	0.024876	0.009475	2.63
percent_diabetes	1	0.032983	0.007372	4.47
Religious_congregati	1	0.013112	0.001431	9.16
political_party	1	-0.078627	0.052022	-1.51
airport_distance	1	0.000734	0.000418	1.76
median_household_inc	1	-0.000011307	0.000002576	-4.39
percent_insured	1	-0.036264	0.006705	-5.41

Figure 7 Coefficients for Death Ratio Model

The results for the Death Ratio of 39.46% are roughly the same as the confirmed Ratio, higher insurance percentage, a smaller percentage of smoking, and a smaller housing density would result in a decrease in deaths. However, unlike the previous confirm Ratio model, precipitation, temperature, social distance, geographic information even virus pressure does not have an effect on the increase in deaths.

Variable diabetes enters into our model, which does not enter into the model of confirming Ratio. This is very much in line with common sense, as most Covid-19 patients die from complications. And diabetes is undoubtedly the most common one of these.

PCA

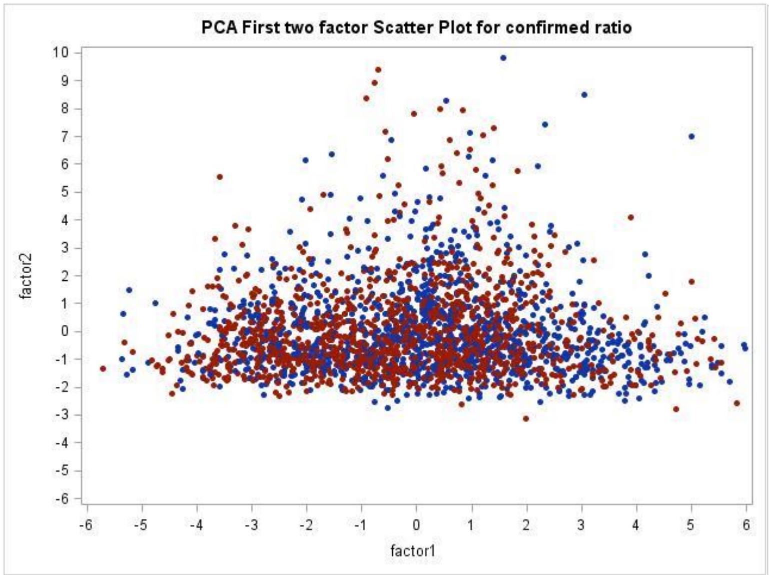
PCA is one of the most common statistical methods and, in fact, we will present a practical method to visualize our data. We will use the princomp procedure for principal component analysis and we will obtain a dataset with eigenvectors, eigenvalues of the Correlation Matrix, and standardized and unstandardized principal component scores. In other terms, we can understand this correlation matrix as a compression of the original data set, which contains part of the information from the previous data.

```
proc princomp data = Covid191  out = outcan ;

Var                  social_distancing_total_grade  precipitation

                          temperature  virus_pressure female_percent
```

```
                         .................... ;
run;
```

Afterward, we will use a different color to control the categorical target variable new Case Ratio which only contains values 0 and 1. We will use the first two eigenvectors, and use the score to draw a scatter plot. This is a very effective way to see how different values vary based on the target variable vary. We use this method before we do machine learning, and we can also have an expectation of the classification rate.

```
proc sgplot data = outcan;
scatter x=prin1 y=prin2/ colorresponse = new_case_Ratio
        markerattrs=(symbol=CircleFilled size=5) ;
         keylegend / title="Storm Classification";
         xaxis label="factor1 " values=(-6 to 6) ;
         yaxis label="factor2" values=(-6 to 10) ;
         title "PCA First two factor Scatter Plot for confirmed Ratio";
run;
```



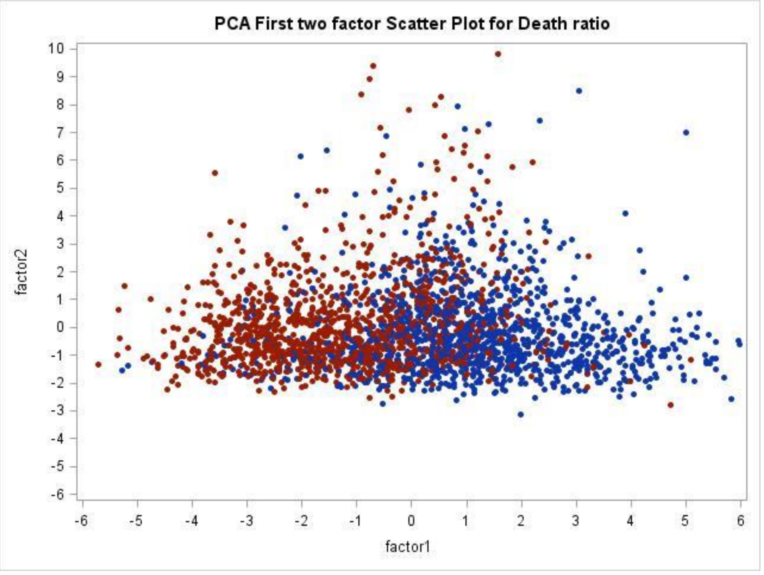


Figure 8. PCA Factor for Case Ratio

Figure 9 PCA Factor for Death Ratio

In the above diagram, both factor1 and factor 2 are the scores of the original eigenvector, red represents the new Case Ratio = 0, and blue represents the new Case Ratio = 1. From the left diagram, the red and blue dots almost mix together, which is not a good signal for classification. Meanwhile, the blue and red dots in the right plot have a distinctly different distribution.

Machine learning classification process

SAS® 9.4, known for its simplicity and accuracy, does not seem to give us more options when it comes to doing machine learning. So when doing machine learning opeRations, we will use SAS ENTERPRISE MINER, a newer platform of SAS®, to do so. Decision Tree, Random Forest, and Neural Network will be applied for the machine learning analysis. In SAS Enterprise Miner, the data mining process is driven by a process flow diagram.

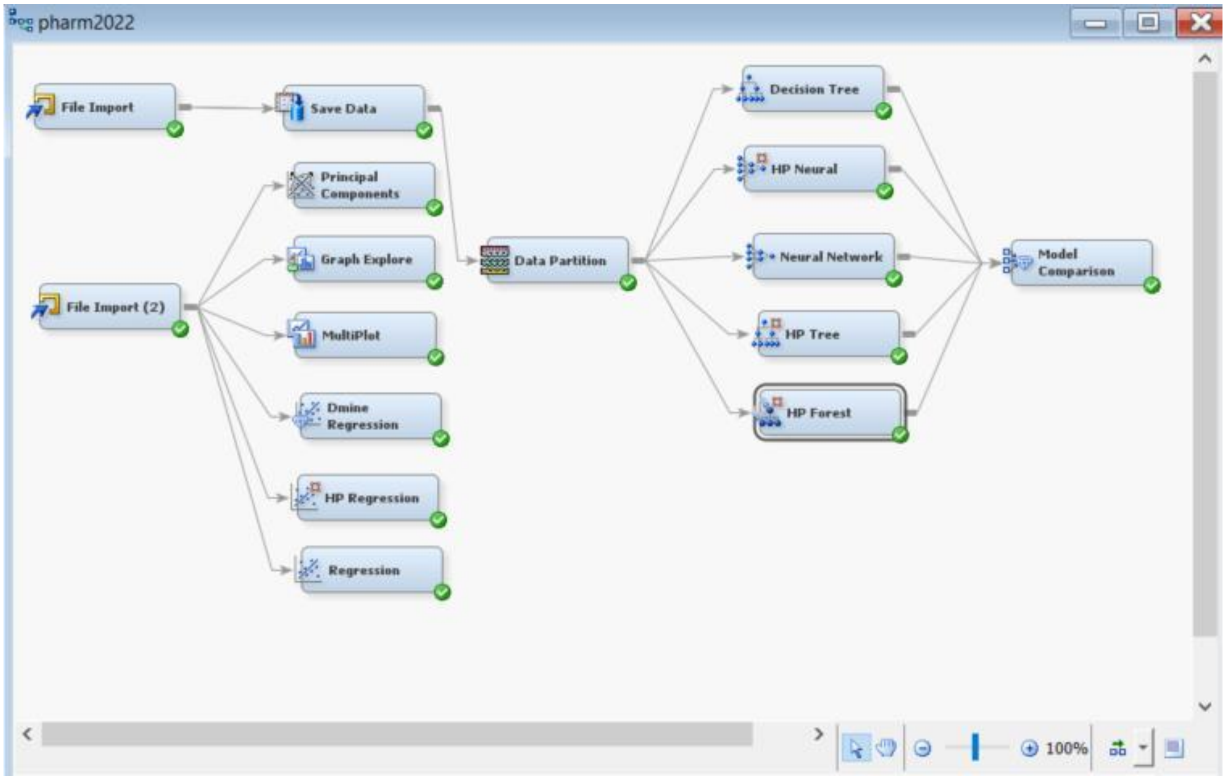
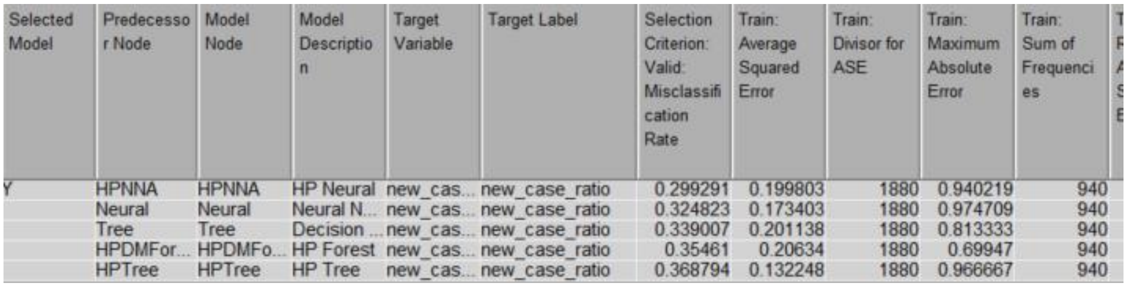
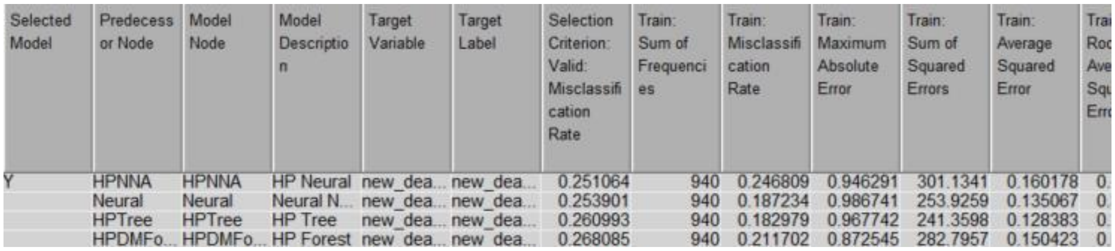


Figure 10 Flow Chart for SAS Enterprise Miner



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	T R A S E
Y	HPNNA	HPNNA	HP Neural	new_cas...	new_case_ratio	0.299291	0.199803	1880	0.940219	940	
	Neural	Neural	Neural N...	new_cas...	new_case_ratio	0.324823	0.173403	1880	0.974709	940	
	Tree	Tree	Decision ...	new_cas...	new_case_ratio	0.339007	0.201138	1880	0.813333	940	
	HPDMFor...	HPDMFo...	HP Forest	new_cas...	new_case_ratio	0.35461	0.20634	1880	0.69947	940	
	HPTree	HPTree	HP Tree	new_cas...	new_case_ratio	0.368794	0.132248	1880	0.966667	940	

Figure 11 Classification Error for Case Ratio model



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Trai Roo Ave Squ Erro
Y	HPNNA	HPNNA	HP Neural	new_dea...	new_dea...	0.251064	940	0.246809	0.946291	301.1341	0.160178	0.
	Neural	Neural	Neural N...	new_dea...	new_dea...	0.253901	940	0.187234	0.986741	253.9259	0.135067	0.
	HPTree	HPTree	HP Tree	new_dea...	new_dea...	0.260993	940	0.182979	0.967742	241.3598	0.128383	0.
	HPDMFo...	HPDMFo...	HP Forest	new_dea...	new_dea...	0.268085	940	0.211702	0.872545	282.7957	0.150423	0.

Figure 12 Classification Error for Death Ratio model

We use the default setting to do the dataset partition process, and the Ratio between training, validation, and testing will be 4:3:3. We can find that the test accuracy of the neural network, decision tree, and random forest are 70.1%, 67.6% and 64.5% in confirm number model. Meanwhile, In the death number model, the test accuracy of the neural network, decision tree, and random forest were 74.9%, 73.9%, and 73.2%. We can easily see that the neural network has an advantage over the other two algorithms on this dataset and that the accuracy of the DEATH model is higher than on the CONFIRMED number, which was already expected when we did the PCA scatter plot.

CONCLUSION

We performed statistical calculations using different methods. First, we used Linear Regression. We use the VIF method to avoid multicollinearity, the result of this is that we have to drop three variables, which will greatly reduce the R square of the model, but will increase the reliability of the model. Next, we performed two Linear Regression modeling for Case Ratio and Death Ratio using stepwise selection. The significant results is obtained. The R square is 0.26 and F-statistic is 44.03 for Case Ratio Model, meanwhile, the R square is 0.39 and F-statistic is 110.45 for Case Ratio Model. It can be easily found that both the results of F-statistic and R square, the Death Ratio model is better than the Case Ratio model. This result is also verified by the scatter plot of the PCA **Figure 8** and **Figure 9.** We obtained regression results that were consistent with common sense.

According to the results of the model, better social distance, higher temperatures, higher smoking rates, including college towns with university students, and higher rates of uninsured people in certain area would be more serious for the spread of Covid-19. At the same time, a higher proportion of smoking, accompanied by a higher proportion of diabetes, would result in a higher proportion of deaths. Social distance, temperature, and college students had no effect on Death Ratio.

Next, we used SAS Enterprise Miner for common sense machine learning. We used several methods of machine learning and neural networks were the best of those.

Dataset download address: https://github.com/yzb0010/Pharmsug2022

REFERENCES

Jack Shostak,2014. SAS Programming in the Pharmaceutical Industry, Second Edition 2nd Edition Cary,North Carolina: SAS Institute.

Haratian, A., Fazelinia, H., Maleki, Z., Ramazi, P., Wang, H., Lewis, M.A., Greiner, R. and Wishart, D., 2021. Dataset of COVID-19 outbreak and potential predictive features in the USA. Data in Brief, 38, p.107360.

Ayine, P., Selvaraju, V., Venkatapoorna, C.M., Bao, Y., Gaillard, P. and Geetha, T., 2021. Eating behaviors in relation to child weight status and maternal education. Children, 8(1), p.32.

Guo, J., Bao, Y., Davis, R., Abebe, A., Wilson, A.E. and Davis, D.A., 2020. Application of meta‐analysis towards understanding the effect of adding a methionine hydroxy analogue in the diet on growth performance and feed utilization of fish and shrimp. Reviews in Aquaculture, 12(4), pp.2316-2332.

Cohen, R.A., 2006, March. Introducing the GLMSELECT procedure for model selection. In Proceedings of the Thirty-First Annual SAS Users Group International Conference (pp. 4770-4792). Citeseer.

Available at: https://facweb.cdm.depaul.edu/sjost/csc423/documents/glmselect-summary.pdf

Holland, M., Hudson, J., Bao, Y. and Gaillard, P., 2020. Aortic to caudal vena cava Ratio measurements using abdominal ultrasound are increased in dogs with confirmed systemic hypertension. Veterinary Radiology & Ultrasound, 61(2), pp.206-214.

H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238, doi: 10.1109/TPAMI.2005.159 .

Bao, Y. and Gaillard, P., 2019. “Summarizing some conventional methods to classify a binary target.” *Proceedings of SESUG 2019 Conference, Williamsburg,VA, lexjansen.com*

Available at: https://www.lexjansen.com/sesug/2019/SESUG2019_Paper-200_Final_PDF.pdf

American Hospital Association Annual Survey. https://www.ahadata.com/aha- annual- survey-database . Accessed May 11, 2020.

County Health Rankings and Roadmaps. https://www.countyhealthrankings.org/app/ . Accessed May 11, 2020.

P. Ramazi, Z. Maleki, H. Fazelinia, A. Haratian, USA Covid-19 data, figshare, 2020. doi: 10.6084/m9.figshare.12986069. v1 .

Bao,Y, Wang,W, Guo, J. and Gaillard, P., 2020. “Special Plots methods with diabetes disease data.” *Proceedings of PharmSug 2020 Conference, San Francisco,CA, lexjansen.com*

Available at: https://www.lexjansen.com/pharmasug/2020/DV/PharmaSUG-2020-DV-163.pdf

United States Census Bureau, USA Counties: 2011. https://www.census.gov/library/publications/2011/compendia/ usa- counties- 2011.html#LND . Accessed May 6, 2020.

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS® For Dummies®*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yida Bao
Auburn University
yzb0010@auburn.edu

Philippe Gaillard
Florida State University
pgaillard@fsu.edu