Jessy Huang
Dec/20/22
Project 6
Professor Simonoff
Multivariate Regression Analysis

The Covid-19 lockdown Logistic Regression Model

Over the course of this class, I have used the theme of covid-19 to explore regression analysis from different angles. I've done a regression model for GDP, unemployment rate, and then a death per Capita ANCOVA model over covid-19. On this assignment I would like to study a covid-19 lockdown logistic regression model.

Lockdown is a binary category indicator which represents two categories: level 0 and level 1. The level 0 is defined as "recommended not to travel between regions/cities", and the level 1 is defined as "internal movement restrictions in place". Level 1 is much stricter than level 0. I believe lockdown is a function of the severity of this pandemic. It plays an important role in policy making and has impact on everybody's life. A better understanding of lockdown over the course of Covid-19 during the years 2020~2021, and the relation with these factors could help to effectively predict lockdowns in future pandemics.

The target variable here is "lockdown", a binary category variable, while the potential three predictors are "Covid-19 Cases", "Covid -19 deaths", and "Population Vaccinated percentage", representing the total confirmed covid-19 cases, deaths, and percentage of the population that were vaccinated. The analysis presented here is based on the data from the United States from March 2020 to June 2021 across all states. The data is averaged and divided into groups based on the lockdown level for each state. There are a total of 71 data points since some states have only one lockdown level group.

I chose this target and these three potential predictors mainly based on observations during covid-19 pandemic. While there are many factors which could impact lockdown, I think that the most pronounced factor are these three variables: "total cases", "total deaths", and "population vaccinated percentage". Lockdowns are the consequence of the severity of the pandemic, and Covid-19 Cases and Deaths are good indicators of the severity of pandemic. The percentage of vaccinated people has a direct connection with lockdowns since it is critical factor to determine if people can freely move.

Target variable: Lockdown

- 0 - recommend not to travel between regions/cities
- 1 - internal movement restrictions in place

Three predicting variables:

- Total Death = total Covid deaths of each state, averaged during 2020~June 2021 for each state, divided into groups by lockdown level
- Total Cases = total Covid cases of state, averaged during 2020~June 2021 for each state, divided into groups by lockdown level
- Population Vaccinated = people vaccinated/population*100%, averaged during 2020~June 2021 for each state, divided into groups by lockdown level

The covid cases, death data, percentage of population vaccinated, and lockdown data all come from covid-19-data github "nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. (github.com)". The data concerns all 50 states and DC in the year 2020~June, 2021

A Snapshot Of Data:

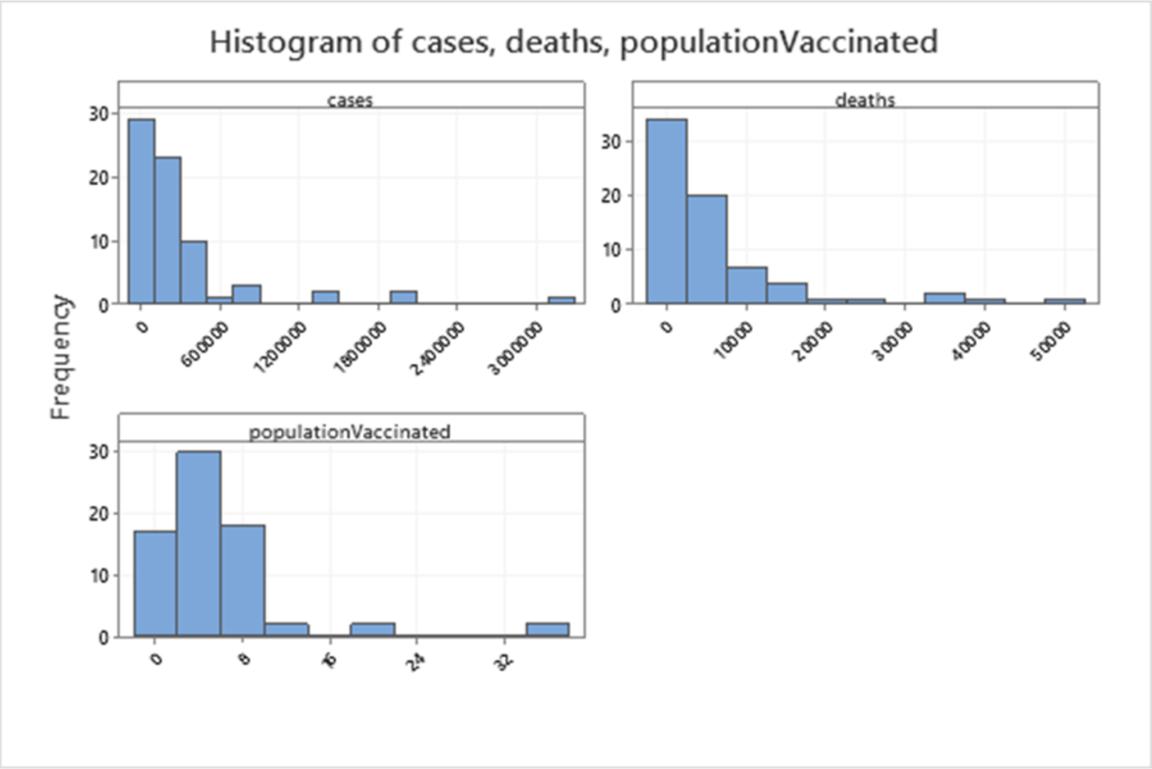
Descriptive Statistics: cases, deaths, populationVaccinated

Statistics

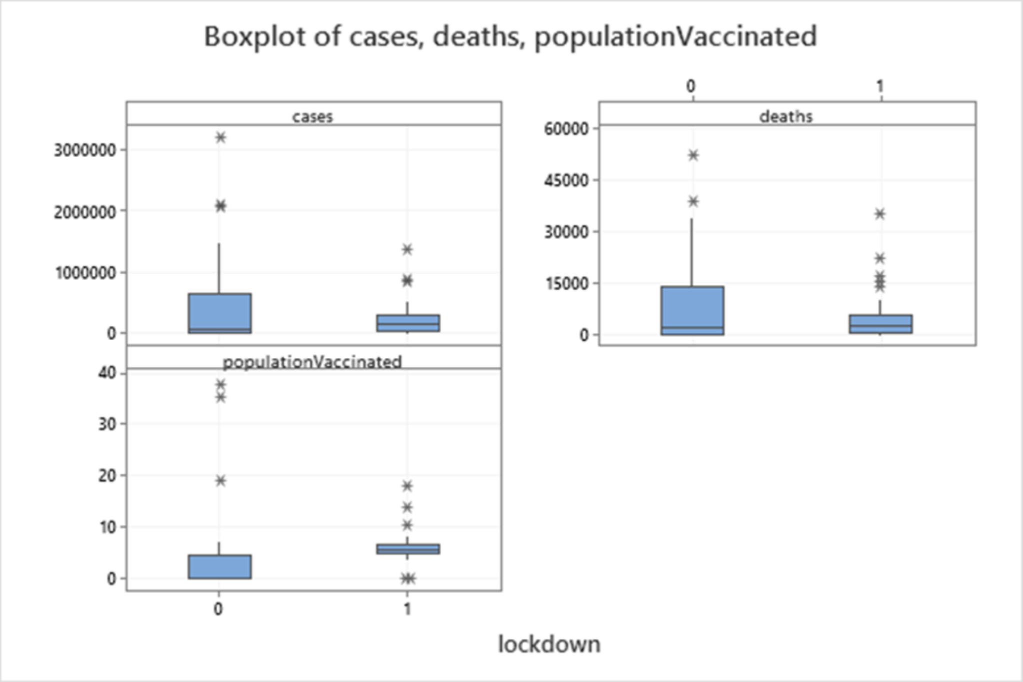
Variable	lockdown	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
cases	0	20	0	541269	206406	923078	4	11827	83069
	1	51	0	239284	35757	255355	7146	56523	170375
deaths	0	20	0	9778	3457	15458	0	311	2231
	1	51	0	4968	897	6408	109	912	2909
populationVaccinated	0	20	0	5.47	2.59	11.56	0.00	0.00	0.00
	1	51	0	5.931	0.367	2.620	0.000	5.008	5.694

Variable	lockdown	Q3	Maximum
cases	0	644707	3216529
	1	321757	1368674
deaths	0	14173	52387
	1	5844	35303
populationVaccinated	0	4.62	37.89
	1	6.525	18.068

First I plotted a histogram for each of the predictors. All of them show a long right tailed distribution, which suggests that a natural log transformation might be helpful.

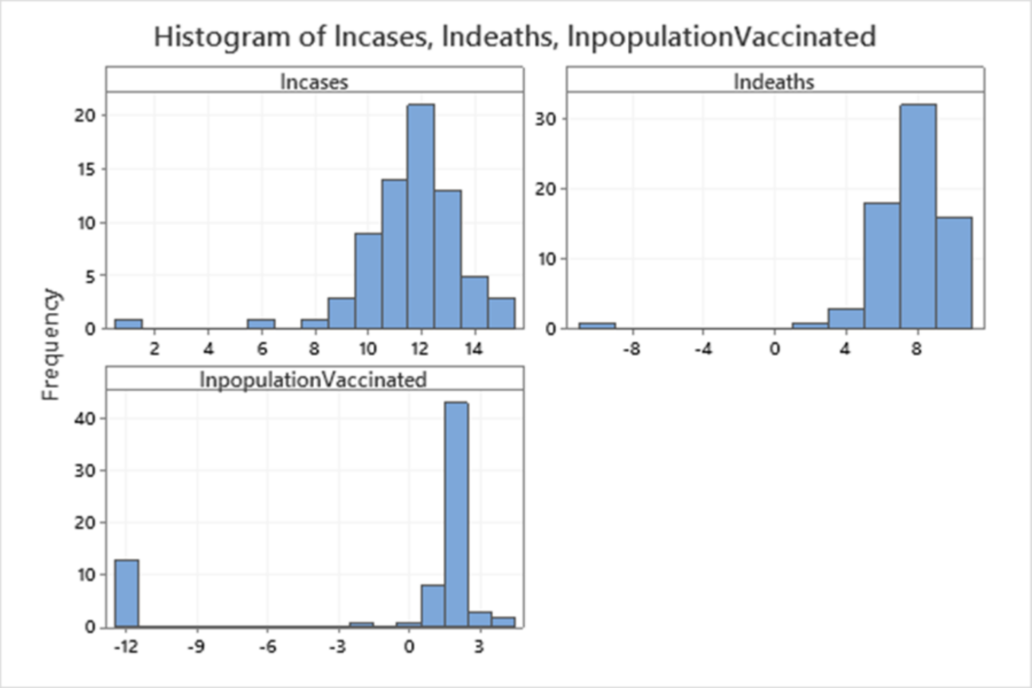


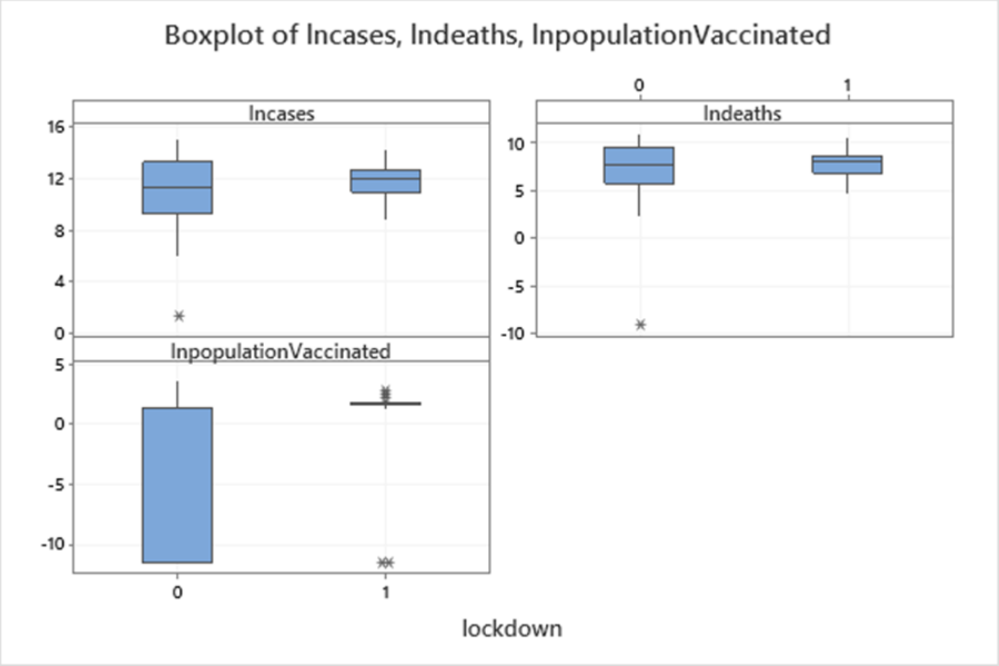
To first get an understanding of the predicative power of each predictor, I plotted side-by-side boxplots to see if there is separation between lock down level 0 and level 1 on the three predicters:



We can see that the predictor “populationVaccinated” shows a clear separation between lockdown level 0 and 1. The predictors covid-19 “cases” and “deaths” show less predictive power.

Because all three predictors are right-tailed, I took natural log for each predictor and replotted the histogram and boxplot.





I ran a logistics regression model with the three natural logged predictors: lncases, lndeaths and lnpopulationVaccinated. Here is the result.

Binary Logistic Regression: lockdown versus lncases, lndeaths, lnpopulationVaccinated

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	71

Response Information

Variable	Value	Count
lockdown	1	51 (Event)
	0	20
	Total	71

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 16.51 - 2.137 lncases + 1.311 lndeaths + 0.519 lnpopulationVaccinated

Coefficients

Term	Coef	SE Coef	VIF
Constant	16.51	6.17	
lncases	-2.137	0.995	26.22
lndeaths	1.311	0.808	17.63
lnpopulationVaccinated	0.519	0.137	3.69

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
lncases	0.1180	(0.0168, 0.8293)
lndeaths	3.7087	(0.7609, 18.0780)
lnpopulationVaccinated	1.6798	(1.2836, 2.1983)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve
39.37%	35.81%	59.19	59.80	68.24	0.9010

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	67	51.19	0.924
Pearson	67	64.54	0.562

Hosmer-Lemeshow	8	6.00	0.648

Analysis of Variance

				Likelihood Ratio	
Source	**DF**	**Adj Dev**	**Adj Mean**	**Chi-Square**	**P-Value**
Regression	3	33.235	11.0784	33.24	0.000
lncases	1	8.096	8.0955	8.10	0.004
lndeaths	1	6.006	6.0061	6.01	0.014
lnpopulationVaccinated	1	30.184	30.1837	30.18	0.000
Error	67	51.190	0.7640		
Total	70	84.425			

Measures of Association

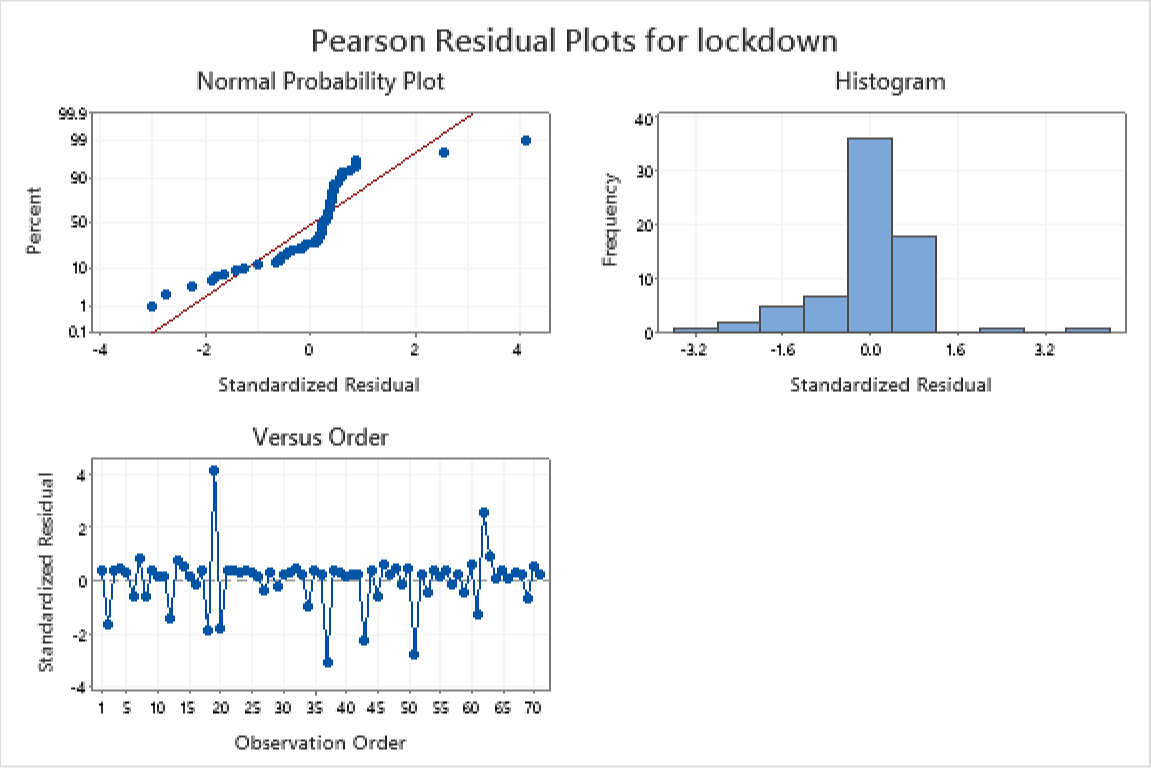
Pairs	Number	Percent	Summary Measures	Value
Concordant	912	89.4	Somers' D	0.80
Discordant	101	9.9	Goodman-Kruskal Gamma	0.80
Ties	7	0.7	Kendall's Tau-a	0.33
Total	1020	100.0		

Association is between the response variable and predicted probabilities

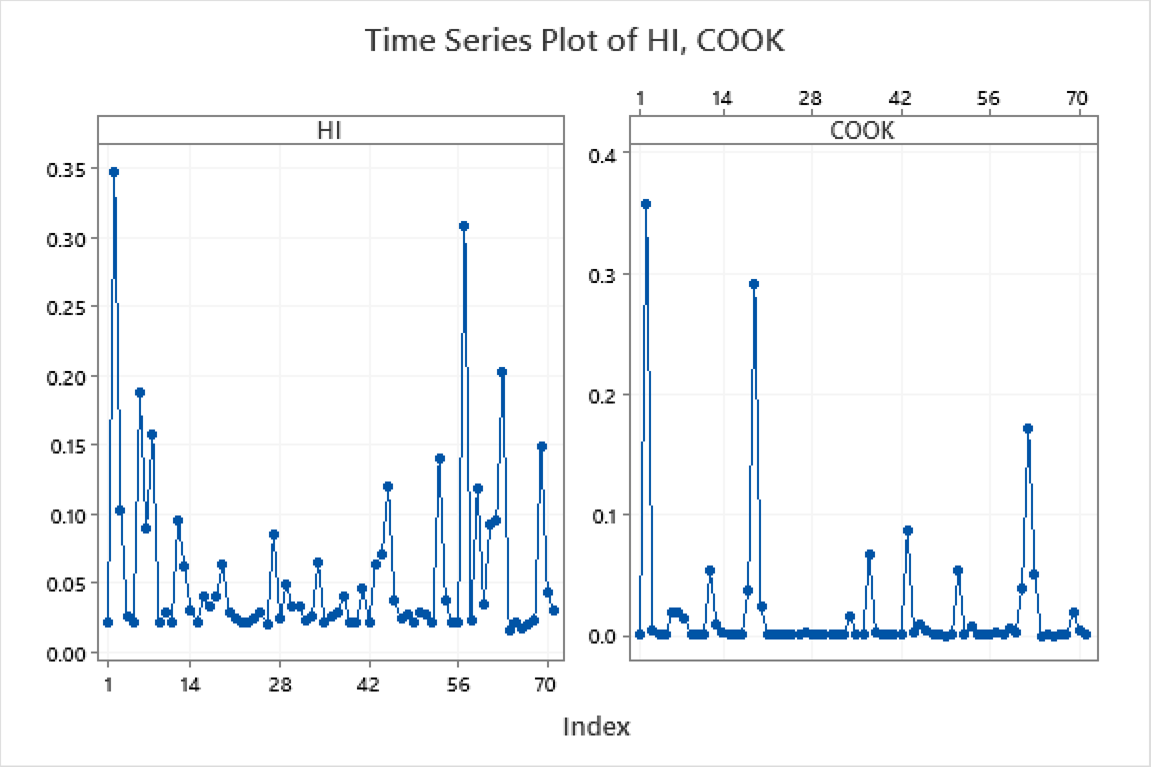
Fits and Diagnostics for Unusual Observations

	Observed				
Obs	**Probability**	**Fit**	**Resid**	**Std Resid**	
2	0.000	0.637	-1.324	-1.64	X
6	0.000	0.211	-0.517	-0.57	X
19	1.000	0.059	4.008	4.14	R
37	0.000	0.899	-2.981	-3.02	R
43	0.000	0.827	-2.190	-2.26	R
51	0.000	0.880	-2.704	-2.74	R
57	0.000	0.011	-0.106	-0.13	X
62	1.000	0.145	2.427	2.55	R
63	1.000	0.607	0.805	0.90	X

R Large residual
X Unusual X



Time Series Plot of HI, COOK



The results show that the overall regression test is highly statistically significant with a Chi-square value = 33.24 and a P-value = 0.000. Overall, the three predictors provide highly statistical significant information differencing lockdown level 0 and 1. We can see that all coefficients for the three predictors are highly statistically significant with Chi-square test(likelihood ratio tests) P-value = 0.004 for “lncases”, 0.014 for “lndeaths” and 0.000 for “lnpopulationVaccinated”. We can interpret that the coefficients mean a 1% change in “cases” is associated with a 2.137% decrease in the odds of the state going to the lockdown level 1 holding all else fixed, a 1% change in “deaths” is associated with a 1.311% increase in the odds of the state going to the lockdown level 1 holding all else fixed, and a 1% change in “populationVaccinated” is associated with 0.519% increase in the odds of the state going to the lockdown level 1 holding all else fixed.

The results show that the groups are identified very well with 89.4% concordant and 9.9% discordant. The Somers’ D value of .8 suggests that there is a lot of separation. The Hosmer-Lemeshow test with P-value = 0.648 indicates no evidence of lack of fit.

But the diagnostic results also show two unusual points, #2 and #57. Point #2 is Alaska with only total 10 deaths and 438 cases during lockdown level 0, and #57 is South Dakota, which has zero deaths and a total 4 cases during lockdown level 0. Both are leverage points, so I decided to take these two points out and rerun the logistic regression.

Binary Logistic Regression: lockdown versus lncases, lndeaths, lnpopulationVaccinated

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	69

Response Information

Variable	Value	Count
lockdown	1	51 (Event)
	0	18
	Total	69

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 20.15 - 2.39 lncases + 1.253 lndeaths + 0.550 lnpopulationVaccinated

Coefficients

Term	Coef	SE Coef	VIF
Constant	20.15	7.45	

lncases	-2.39	1.17	25.83
lndeaths	1.253	0.951	15.44
lnpopulationVaccinated	0.550	0.152	4.64

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
lncases	0.0920	(0.0093, 0.9065)
lndeaths	3.4995	(0.5424, 22.5772)
lnpopulationVaccinated	1.7328	(1.2870, 2.3329)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve
39.31%	35.52%	56.07	56.70	65.01	0.8943

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	65	48.07	0.943
Pearson	65	64.04	0.510
Hosmer-Lemeshow	8	8.92	0.349

Analysis of Variance

				Likelihood Ratio	
Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	3	31.136	10.3785	31.14	0.000
lncases	1	5.184	5.1835	5.18	0.023
lndeaths	1	1.862	1.8622	1.86	0.172
lnpopulationVaccinated	1	28.114	28.1141	28.11	0.000
Error	65	48.072	0.7396		
Total	68	79.207			

Measures of Association

Pairs	Number	Percent	Summary Measures	Value
Concordant	820	89.3	Somers’ D	0.79
Discordant	95	10.3	Goodman-Kruskal Gamma	0.79
Ties	3	0.3	Kendall’s Tau-a	0.31
Total	918	100.0		

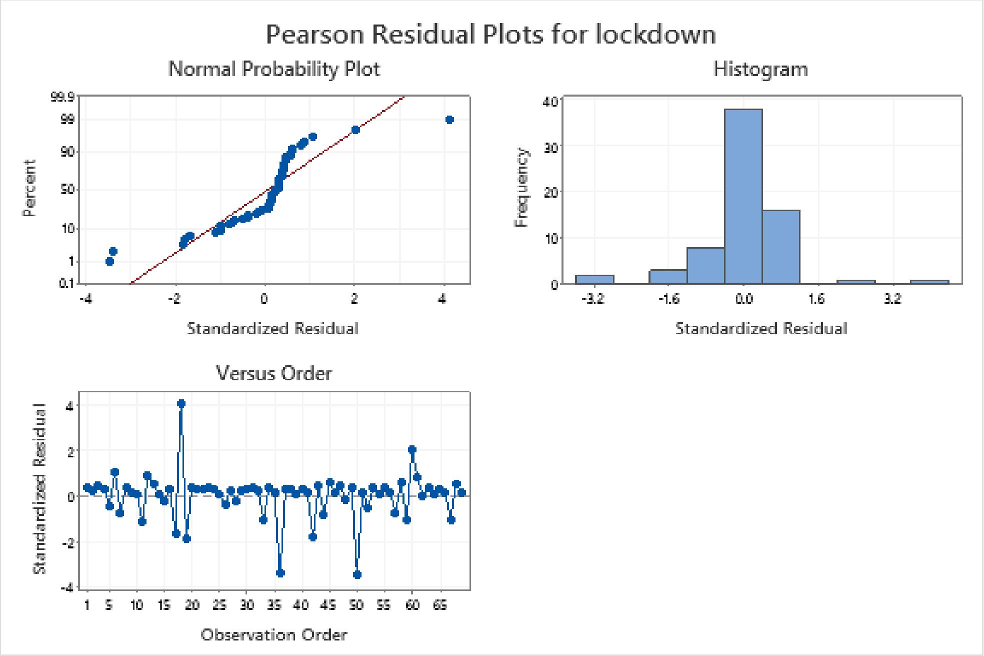
Association is between the response variable and predicted probabilities

Fits and Diagnostics for Unusual Observations

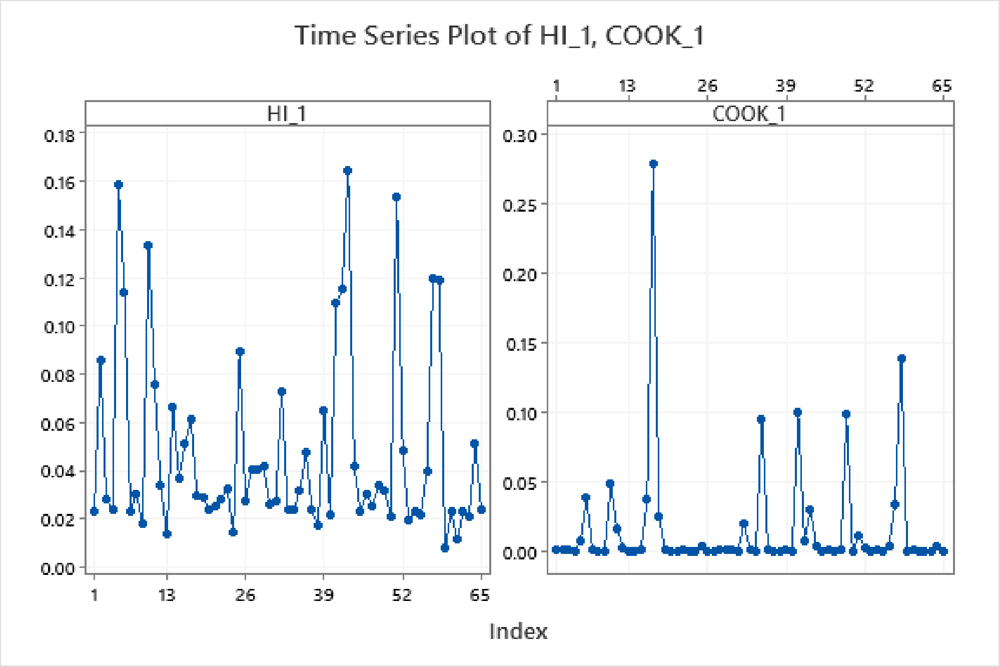
Obs	Observed Probability	Fit	Resid	Std Resid	
7	0.000	0.282	-0.627	-0.69	X
18	1.000	0.059	3.998	4.13	R
36	0.000	0.918	-3.340	-3.39	R
50	0.000	0.921	-3.411	-3.47	R
57	0.000	0.274	-0.615	-0.72	X

```
60          1.000 0.217   1.902        2.03 R
61          1.000 0.670   0.702        0.82     X
67          0.000 0.440  -0.887       -0.99     X
```

R Large residual
X Unusual X



Time Series Plot of HI_1, COOK_1



After omitting unusual points #2 and #57, the diagnostic results show that points #7, #57, #61, #67 are suspected to be potential unusual points. However, the regression results don’t change much after omitting them, so I believe they aren’t really unusual points and kept them in the data set. I used a similar methodology as before to determine unusual points, so to save space the results are not shown.

The model remains strong. The overall regression test on all three coefficients is highly significant with the Chi-square tests (likelihood ratio tests) P-value = 0.000, both “lncases” and “lnpopulationVaccinated” remain highly statistically significant with Chi-square tests (likelihood ratio tests) P-value =0.023 for “Incases”, 0.000 for “lnpopulationVaccinated”, but predictor “lndeaths” became statistically insignificant with Chi-square tests (likelihood ratio tests) P-value = 0.172.

The results show that the groups remained identified very well with 89.3% concordant and 10.3% discordant. The Somers’ D value of .79 still suggests that there is a lot of separation. The Hosmer-Lemeshow test with P-value=0.349 still indicates no evidence of lack of fit.

To further identify the best model, I used ordinary least subsets regression to sort out the models. Although it is not technically valid, I still can use it to get the guidance.

Best Subsets Regression: lockdown versus lncases, lndeaths, lnpopulationVaccinated

Response is lockdown

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	lncases	lndeaths	lnpopulationVaccinated

								e d
1	32.0	31.0	27.3	10.9	0.36752			X
1	0.2	0.0	0.0	46.3	0.44517		X	
2	40.3	38.4	33.3	3.6	0.34703	X		X
2	37.5	35.6	30.3	6.7	0.35488		X	X
3	41.7	39.0	33.7	4.0	0.34539	X	X	X

The best subset output suggests that a model with two predictors, “lncases” and “populationVaccinated”, is the best candidate. I ran a logistic regression with these two predictors.

Binary Logistic Regression: lockdown versus lncases, lnpopulationVaccinated

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

Method

Link function	Logit
Residuals for diagnostics	Pearson
Rows used	69

Response Information

Variable	Value	Count
lockdown	1	51 (Event)
	0	18
	Total	69

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 13.37 - 0.989 lncases
+ 0.451 lnpopulationVaccinated

Coefficients

Term	Coef	SE Coef	VIF
Constant	13.37	4.64	
lncases	-0.989	0.371	2.77
lnpopulationVaccinated	0.451	0.118	2.77

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
lncases	0.3721	(0.1798, 0.7701)
lnpopulationVaccinated	1.5702	(1.2463, 1.9782)

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve
36.96%	34.43%	55.93	56.30	62.64	0.8856

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	66	49.93	0.929
Pearson	66	73.34	0.250
Hosmer-Lemeshow	8	12.44	0.133

Analysis of Variance

				Likelihood Ratio	
Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	2	29.273	14.6367	29.27	0.000
lncases	1	9.508	9.5078	9.51	0.002
lnpopulationVaccinated	1	29.219	29.2190	29.22	0.000
Error	66	49.934	0.7566		
Total	68	79.207			

Measures of Association

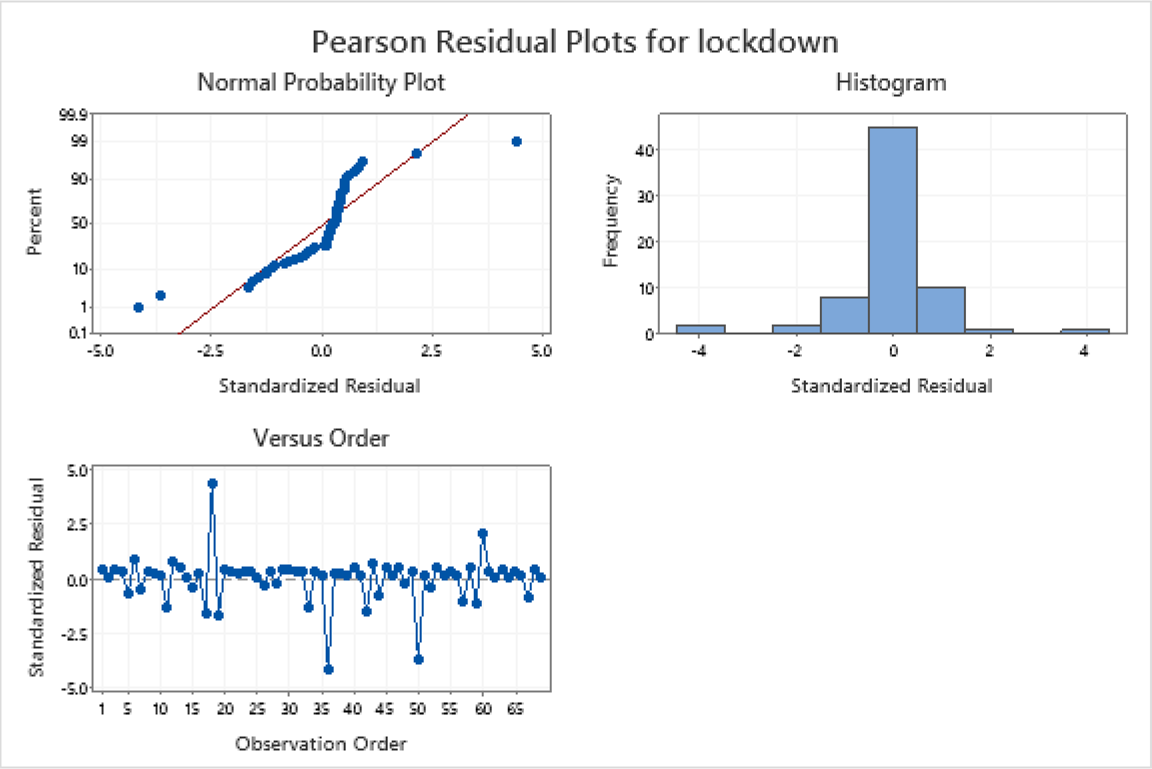
Pairs	Number	Percent	Summary Measures	Value
Concordant	813	88.6	Somers' D	0.77
Discordant	103	11.2	Goodman-Kruskal Gamma	0.78
Ties	2	0.2	Kendall's Tau-a	0.30
Total	918	100.0		

Association is between the response variable and predicted probabilities

Fits and Diagnostics for Unusual Observations

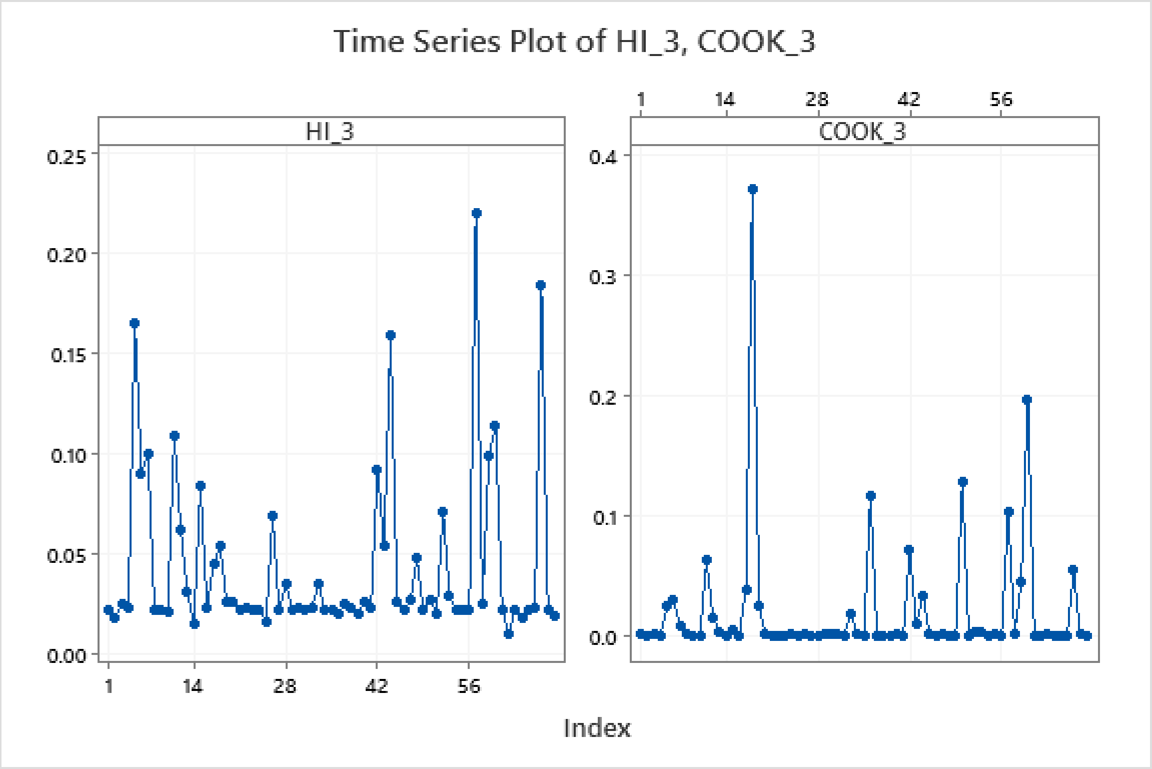
Obs	Observed Probability	Fit	Resid	Std Resid	
5	0.000	0.240	-0.562	-0.61	X
18	1.000	0.051	4.298	4.42	R
36	0.000	0.944	-4.089	-4.13	R
44	0.000	0.309	-0.668	-0.73	X
50	0.000	0.929	-3.611	-3.66	R
57	0.000	0.460	-0.923	-1.04	X
60	1.000	0.198	2.014	2.14	R
67	0.000	0.371	-0.768	-0.85	X

R Large residual
X Unusual X



COPY FROM 2020_2021_FIST_SIX_MONTHRESULT_NEW.CSV(W6)

Time Series Plot of HI_3, COOK_3



Compared with the three predictors model, the overall regression test remains highly statistically significant with a Chi-square test (likelihood test) P-value = 0.000 while the coefficients become much more statistically significant: “lncase” has Chi-square tests (likelihood ratio tests) P-value change from 0.023 to 0.002 and “populationVaccinated” remain highly statistically significant with P-value = 0.000. The Somers’ D value decreased from .79 to .77, but it is still very close, a relatively small difference. The AIC value decreased from 56.07 to 55.93.

The coefficients can be interpreted such that a 1% change in “cases” is associated with a 0.989% decrease in the odds of the state going to the lockdown level 1 holding all else fixed and a 1% change in “populationVaccinated” is associated with a 0.451% increase in the odds of the state going to the lockdown level 1 holding all else fixed.

The Hosmer-Lemeshow test with P-value = 0.133 still suggests no evidence of lack of fit.

I also performed a chi-squared test to compare this two-variable model to the three-variable model.

Cumulative Distribution Function

Chi-Square with 1 DF

x	P(X ≤ x)
1.87	0.828525

The logistic regression test statistic Chi-square (likelihood ratio) for the three-variable model is 31.14 and for the two-variable model is 29.27, meaning the test statistic is 31.14 - 29.27 = 1.87. With a test statistic of 1.87 with one degree of freedom we can calculate a P-value of .1715. This implies that you cannot reject the two-variable model in favor of the three-variable model at a .10 level. After considering the principles of parsimony, I chose the two-variable model as the “best” model.

Finally, I created a classification matrix based on whether the estimated probability is above or below .5.

Rows: lockdown   Columns: predictors1

	0	1	All
0	10 14.49	8 11.59	18 26.09
1	2 2.90	49 71.01	51 73.91
All	12 17.39	57 82.61	69 100.00

Cell Contents
Count
% of Total

The two omitted points, #2 and #57, are leverage points which belong to the “lockdown level 0” group. The results show that the fitted probability is 89.7% for #2 and 99.88% for #57. Both are not correctly classified by the model. The results are below.

Prediction for lockdown

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = 13.37 - 0.989 lncases + 0.451 lnpopulationVaccinated

Settings

Variable	Setting
lncases	6.0822
lnpopulationVaccinated	-11.5129

Prediction

Fitted Probability	SE Fit	95% CI
0.897127	0.141824	(0.300228, 0.994390)

Settings

Variable	Setting
lncases	1.3862
lnpopulationVaccinated	-11.5129

Prediction

Fitted Probability	SE Fit	95% CI
0.998897	0.0034610	(0.657330, 1.00000)

The new classification table after counting the omitted leverage points #2, #57 as incorrect classification points is below.

	0	1	All
0	10 14.08	10 14.08	20 28.16
1	2 2.82	49 69.01	51 71.83

All	12	59	71
	16.9	83.09	100.00

The classification table suggests that 83.09% of the states were correctly classified, which is higher than the Cmax of 71.83% and slightly lower than the Cpro of (1.25)(0.169*0.169+0.8309*0.8309) = 89.86%. It indicates that these predictors do have the ability to classify states into lockdown level 0 and level 1 groups. I would have liked to verify these models on new data, but I did not have additional data because the second doses of vaccines had started rolling out in the second half of 2021, so the situation became complicated.

The final model is therefore a logistic regression with two predicators, “lncases” and “lnpopulationVaccinated”. Overall, the regression test remains highly statistically significant with a Chi-square test (likelihood test) P-value = 0.000 and the coefficients of each predictor are also highly statistically significant; “lncase” has a Chi-square tests (likelihood ratio tests) P-value = 0.002 and populationVaccinated has a P-value =0.000. Moreover, the Somers’ D value is 0.77 which shows that there is a lot of separation , and the Hosmer-Lemeshow test with P-value = 0.133 indicates no evidence of lack of fit. A classification matrix shows that the model does have an ability to classify states into lockdown level 0 and level 1 groups.

Regression Equation

```
P(1)  =   exp(Y')/(1 + exp(Y'))
Y'   =  13.37 - 0.989 lncases + 0.451 lnpopulationVaccinated
```

The coefficients can be interpreted such that a 1% change in “cases” is associated with a 0.989% decrease in the odds of the state going to the lockdown level 1 holding all else fixed, and a 1% change in “populationVaccinated” is associated with 0.451% increase in the odds of the state going to the lockdown level 1 holding all else fixed.

The relation logic seems strange. I would expect that as more of the population got vaccinated, the more likely lockdown restrictions were loosened or removed, and that the more covid-19 cases were present, the higher chance of strict lockdown. However, the results appear to suggest that as more people got vaccinated, the more severe lockdown became. The results also imply that as covid cases increased, the less severe lockdown became. The reason might be because the data is not accurate. For example, there wasn’t a clear diagnostic procedure in the beginning of the pandemic, which could lead to incorrect amounts of diagnoses. There might also be a lurking variable issue which hides the true relationship. Perhaps I did not correctly identify the best variables to predict lockdown enforcement. Maybe, for example, a better variable through which to understand lockdown severity would be the politics of the state. Republican states are far less likely to believe COVID to be a public health hazard, and so are less likely to enforce policies like lockdown (Tyson).

In the future, I would spend more time to understand the data to gain a better understanding of what is the most important factors which could impact the lockdown. I hope that through gaining more experience in the future, I will get a better intuition for the relationship between variables, and to improve my data pre-processing skill to avoid data set bias. I’ve learned a lot about various models used for multi regression analysis from this course through these hands-on practices, which provides an invaluable basis for statistics application. Thanks for teaching this course.

Works Cited

Nytimes/COVID-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. GitHub. (n.d.). Retrieved December 20, 2022, from https://github.com/nytimes/covid-19-data

Tyson, A. (2020, July 28). *Republicans remain far less likely than Democrats to view covid-19 as a major threat to public health*. Pew Research Center. Retrieved December 20, 2022, from https://www.pewresearch.org/fact-tank/2020/07/22/republicans-remain-far-less-likely-than-democrats-to-view-covid-19-as-a-major-threat-to-public-health/