Jessy Huang
11/1/22
Project 2
Professor Simonoff
Multivariate Regression Analysis

The Real GDP Percent Change over Covid-19

The COVID-19 pandemic triggered the sharpest downturn in the US economy since the Depression, with real GDP declining by 8.9 percent in the second quarter of 2020, 3.5 percent total in 2020 compared to a rise of 2.2 percent in 2019. I would like to take this opportunity to explore the impact of the COVID-19 pandemic on the real GDP in USA across all states over the course of 2020, specifically to study the impact on the real GDP Percent Change from Covid Cases per Capita, Covid Deaths per Capita, Unemployment Rate Change, Oil & Gas Extraction sector, Arts, entertainment, recreation, accommodation, and food services sector. By nature, I feel that these areas might have some relationship with the GDP change. The pandemic led to enforced lock downs, social distancing, job loss and business decline. I believe that these consequences will be reflected in unemployment rate, oil & gas extraction industry, entertainment, and tourism related industry. In this project, I'll try to use all the knowledge learned from the class to examine how much they are related.

I chose this topic because I want to better understand the impact the pandemic had, and in turn have a better understanding of future pandemic impacts. I understand that a subject like this can't be completely analyzed with the knowledge I currently have. However, I am still willing to try my best to explore it using simple regression, multivariable regression and, as the course progresses, more modelling techniques.

I chose the same subject as I did for Homework 2, which applied simple regression with the real GDP percent change as target and the Covid Case per Capita as the single predictor. In this project, I'll apply multivariable regression. I'll keep the same target and predictor from my previous project, but added four more predictors.

Data from all states in US are the basis of the following analysis, which covers the years 2019 and 2020

The target variable is real GDP Percent Change of each state, which is calculated by:

real GDP Percent Change = (real GDP of the state – previous year real GDP of the state)/ previous year real GDP of the state * 100%

Five predicting variables are the following and are calculated by:

1. Total Covid Case per Capita = Covid Case of state / population of state *1000
2. Total Death per Capita = Total Covid Deaths of state /population of state *10000
3. Unemployment Rate change=2020 Unemployment Rate of state – 2019 Unemployment Rate of state.

4. “Oil and gas extraction” GDP% = 2019 GDP in “Oil and gas extraction” category of state /2019 full industry GDP of state *100%
5. “Recreation Accommodation food services” GDP% = 2019 GDP in “Arts, entertainment, recreation, accommodation, and food services” category of state/2019 full industry GDP of state *100%

GDP data is from https://www.bea.gov/, website of “Bureau of Economic Analysis, US Department of Commerce”. Real GDP Percent Change in the year 2020 from all 50 states and DC.

The covid case data comes 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)”, Data were from all 50 states and DC in the U.S. in the year 2020

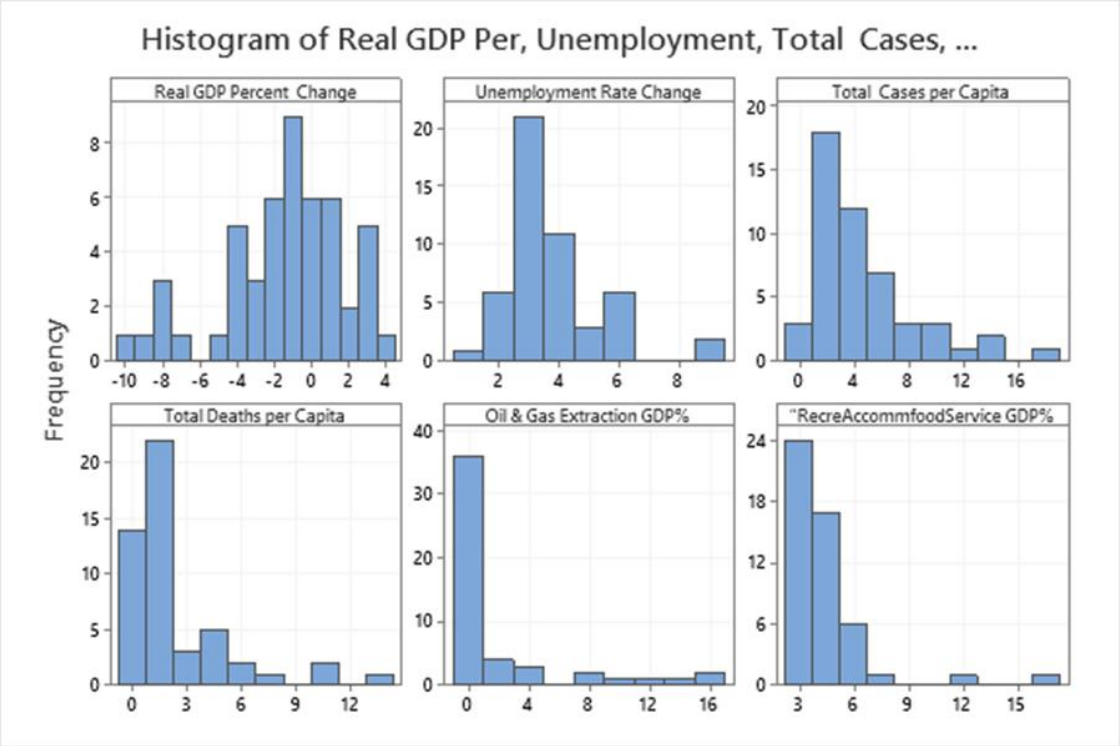
The unemployment rate data comes from US Bureau of Labor Statistics https://www.bls.gov/lau/lastch20.html

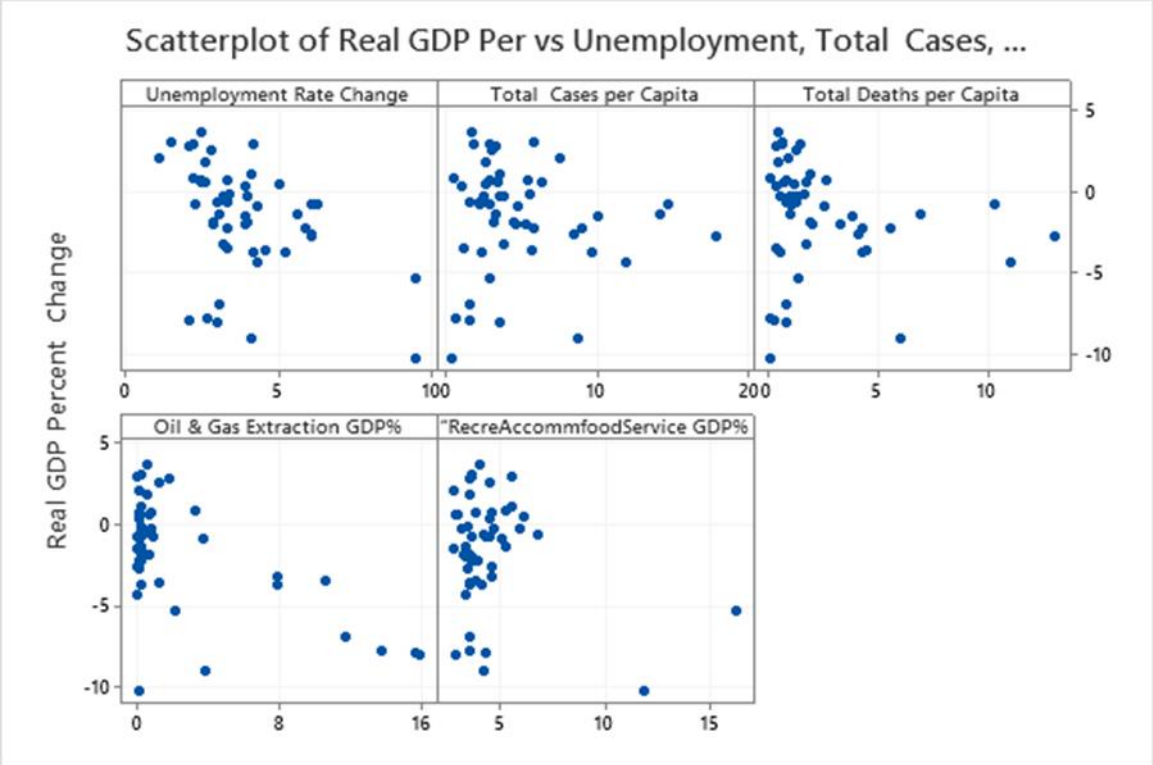
A snapshot of the data shows:

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Real GDP Percent Change	50	0	-1.528	0.464	3.282	-10.202	-3.255	-0.818	0.673
Unemployment Rate Change	50	0	3.816	0.241	1.701	1.100	2.675	3.300	4.300
Total Cases per Capita	50	0	4.767	0.542	3.834	0.428	2.309	3.467	5.774
Total Deaths per Capita	50	0	2.363	0.401	2.835	0.119	0.708	1.299	2.786
Oil & Gas Extraction GDP%	50	0	2.222	0.596	4.215	0.008	0.119	0.301	1.427
“RecreAccommfoodService GDP%	50	0	4.507	0.313	2.213	2.831	3.548	3.867	4.644

Variable	Maximum
Real GDP Percent Change	3.718
Unemployment Rate Change	9.500
Total Cases per Capita	17.884
Total Deaths per Capita	13.030
Oil & Gas Extraction GDP%	15.930
“RecreAccommfoodService GDP%	16.257

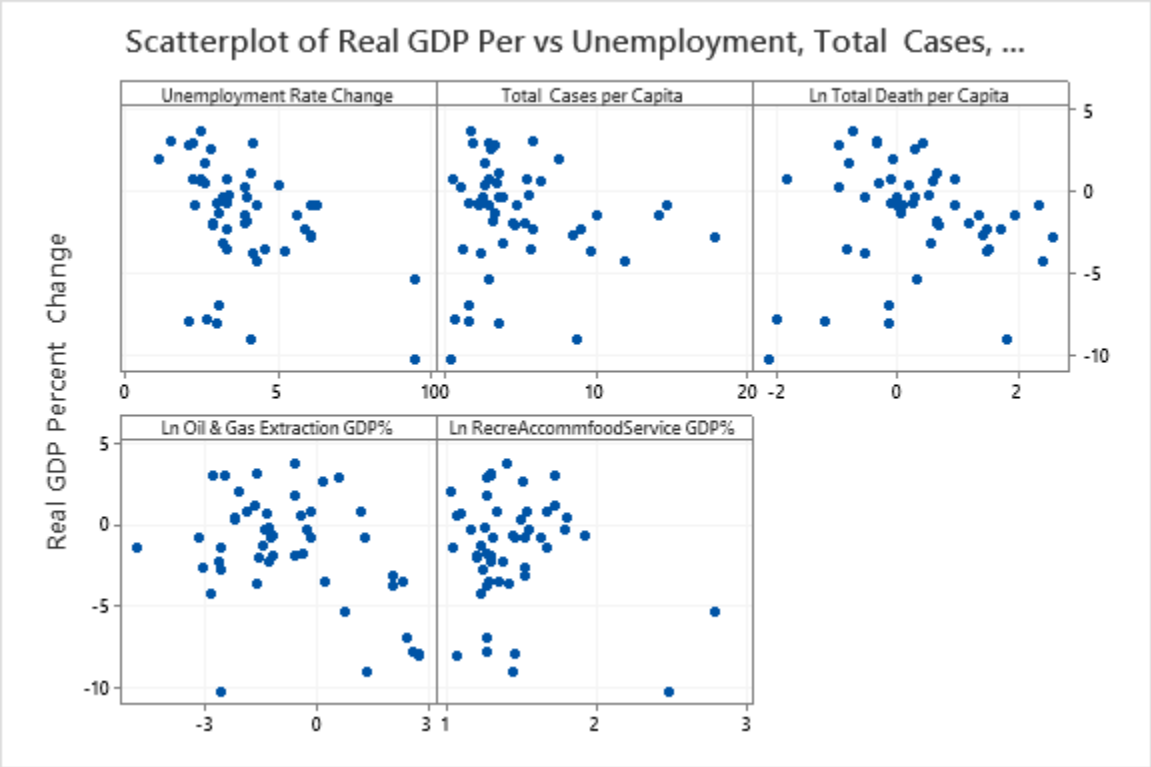
For a quick impression of the factors relationships we can look at the histogram of all variables including both target and predictors, and the scatterplots of the predictors and the response:





We can see that three predictors, “Total Deaths per Capita”, “Oil&Gas Extraction GDP%” and “RecreAccommFoodService GDP%” seem to have a long right-tailed distribution,

so I decided to take logs for these three variables. I plot the semi-log scatter plots with the above three variables taking natural log as below:



These graphs seem to show some sort of relationship between the target and predictors. I need to check the correlation to get a closer look.

Correlations

	Real GDP Percent Change	Unemployment Rate Change	Total Cases per Capita	Ln Total Death per Capita	Ln Oil & Gas Extraction GDP%
Unemployment Rate Change	-0.421				
Total Cases per Capita	-0.050	0.270			
Ln Total Death per Capita	-0.017	0.286	0.832		
Ln Oil & Gas Extraction GDP%	-0.403	-0.258	-0.485	-0.416	
Ln RecreAccommfoodService GDP%	-0.184	0.632	-0.224	-0.199	-0.028

First, we can see that the correlation between variables varies, with some exhibiting a higher correlation than others. A regression using all of the variables might exhibit multicollinearity, and will presumably include several redundant variables.

Here is the regression output. Just as suspected, there are several redundant variables with insignificant t-statistic values. The VIF values look suitable, suggesting that multicollinearity is not an issue.

Regression Equation

Real GDP Percent Change = 0.57 - 1.359 Unemployment Rate Change
- 0.249 Total Cases per Capita
+ 0.624 Ln Total Death per Capita
- 1.159 Ln Oil & Gas Extraction GDP%
+ 2.24 Ln RecreAccommfoodService GDP%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.57	1.98	0.29	0.774	
Unemployment Rate Change	-1.359	0.316	-4.30	0.000	2.47
Total Cases per Capita	-0.249	0.172	-1.45	0.155	3.72
Ln Total Death per Capita	0.624	0.577	1.08	0.285	3.38
Ln Oil & Gas Extraction GDP%	-1.159	0.221	-5.24	0.000	1.35
Ln RecreAccommfoodService GDP%	2.24	1.64	1.36	0.180	2.39

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.39195	52.29%	46.87%	21.24%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	275.908	55.182	9.64	0.000
Unemployment Rate Change	1	105.854	105.854	18.50	0.000
Total Cases per Capita	1	11.950	11.950	2.09	0.155
Ln Total Death per Capita	1	6.703	6.703	1.17	0.285
Ln Oil & Gas Extraction GDP%	1	157.191	157.191	27.47	0.000
Ln RecreAccommfoodService GDP%	1	10.643	10.643	1.86	0.180
Error	44	251.742	5.721		
Total	49	527.650			

I used the best subsets regression model to better identify the best model. Here is the output:

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Unemployment Rate Change	Total Cases per Capita	Ln Total Deaths per Capita	Oil & Gas Extraction GDP%	Real Accomm & Food Service GDP%
1	17.8	16.0	473.5	10.3	29.9	3.0069	256.465	261.679	1.000	X				
1	16.3	14.5	485.0	8.1	31.2	3.0339	257.361	262.575	1.000		X			
2	45.8	43.5	353.5	33.0	6.0	2.4664	237.965	244.725	1.694	X	X			
2	24.2	20.9	471.5	10.6	25.9	2.9178	254.772	261.532	2.884	X			X	
3	50.0	46.8	342.9	35.0	4.1	2.3944	236.401	244.597	5.184	X	X		X	X
3	49.5	46.2	375.6	28.8	4.6	2.4080	236.966	245.163	3.279	X		X	X	
4	51.0	46.7	377.2	28.5	5.2	2.3965	237.980	247.499	8.551	X	X		X	X
4	50.3	45.9	422.7	19.9	5.9	2.4147	238.737	248.256	14.331	X	X	X	X	
5	52.3	46.9	415.6	21.2	6.0	2.3919	239.380	250.097	14.413	X	X	X	X	X

I looked at Mallows Cp and AICc and chose to minimize Cp and AICc, while maximizing R-Sq value. There are three candidates, highlighted in red rectangles, that are in consideration. One is a two variable model, the others are three variable models. I examined all of them, and of these three candidates, the three variable model ,{ “Unemployment Rate Change, Total Cases per Capita, Ln Oil & Gas Extraction GDP%”}, turned out the best. The others have some sort of regression assumption violation at the end, both with residuals off the normal distribution. To save space, I only show the “best model” regression result.

Regression Equation

Real GDP Percent Change = 2.382
- 1.019 Unemployment Rate Change
- 0.189 Total Cases per Capita
- 1.184 Ln Oil & Gas Extraction GDP%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.382	0.892	2.67	0.010	
Unemployment Rate Change	-1.019	0.212	-4.80	0.000	1.10
Total Cases per Capita	-0.189	0.104	-1.82	0.075	1.35
Ln Oil & Gas Extraction GDP%	-1.184	0.222	-5.33	0.000	1.34

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.40796	49.45%	46.15%	28.82%

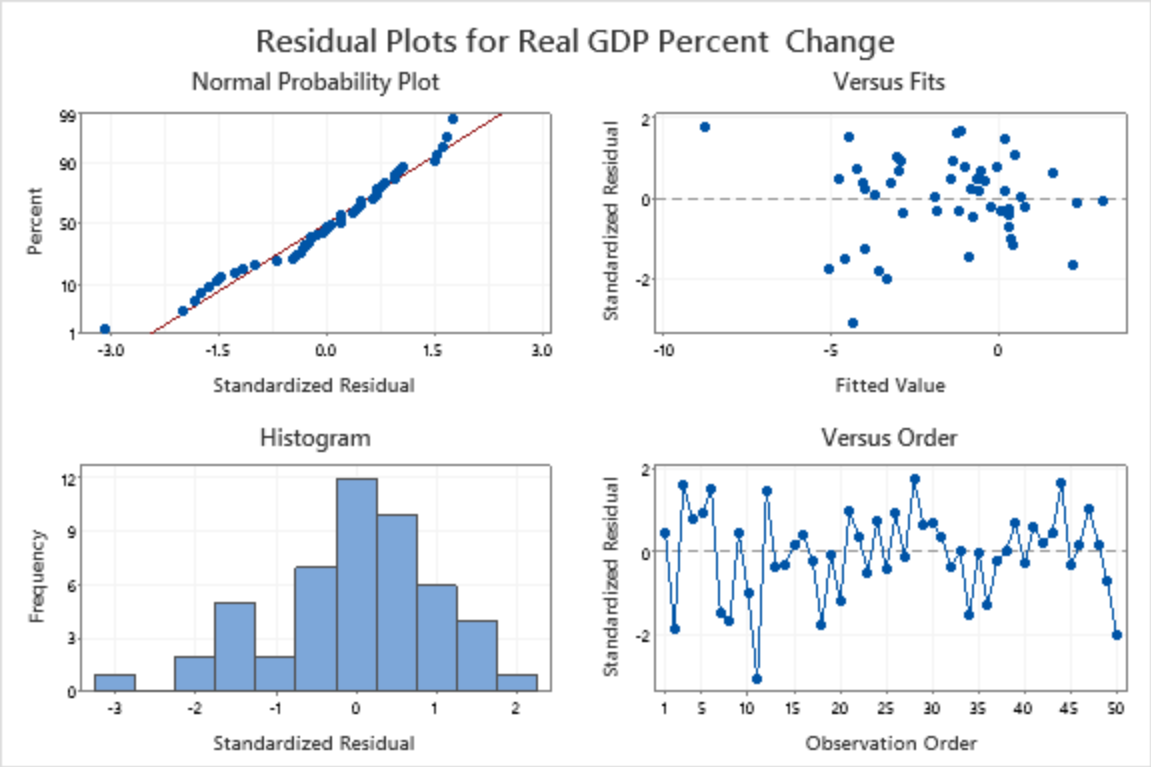
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	260.93	86.976	15.00	0.000
Unemployment Rate Change	1	133.42	133.419	23.01	0.000
Total Cases per Capita	1	19.19	19.189	3.31	0.075
Ln Oil & Gas Extraction GDP%	1	164.97	164.969	28.45	0.000
Error	46	266.72	5.798		
Total	49	527.65			

Fits and Diagnostics for Unusual Observations

Obs	Real GDP Percent Change	Fit	Resid	Std Resid	
11	-10.202	-4.330	-5.872	-3.07	R X
28	-5.282	-8.778	3.495	1.76	X
30	-2.752	-4.211	1.459	0.71	X

R Large residual
X Unusual X



We can see that there seems to be three unusual points: #11 (Hawaii), #50 (Wyoming), and #2 (Alaska). Alaska and Hawaii are relatively isolated areas with lower populations, one being an island and the other close to the North Pole. Their Covid-19 cases are very low compared to the other states as a result. In addition, Hawaii has a Cook's distance of 1.38. Wyoming holds "Total Oil & Gas Extraction GDP%" as high as 15%, which is very high compared to other states. Therefore I decided to omit them and rerun the regression and the diagnostic plots, here are the results:

Regression Equation

Real GDP Percent Change = 2.103
- 0.676 Unemployment Rate Change
- 0.3174 Total Cases per Capita
- 1.118 Ln Oil & Gas Extraction GDP%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.103	0.800	2.63	0.012	
Unemployment Rate Change	-0.676	0.212	-3.19	0.003	1.17
Total Cases per Capita	-0.3174	0.0950	-3.34	0.002	1.51

Ln Oil & Gas Extraction GDP%	-1.118	0.202	-5.53	0.000	1.32

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	47-fold S	47-fold R-sq
2.00621	52.42%	49.10%	41.08%	2.13550	41.08%

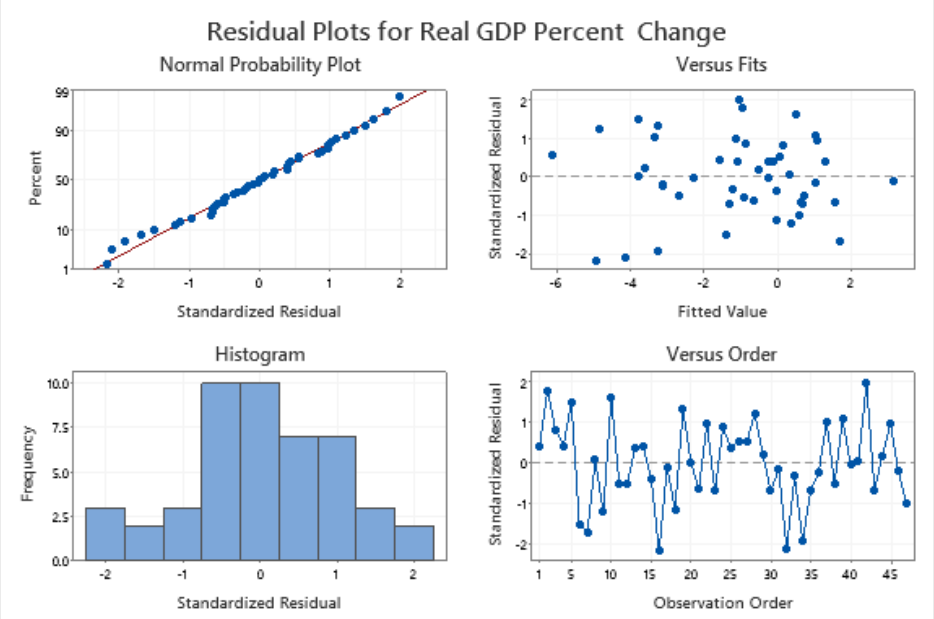
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	190.70	63.568	15.79	0.000
Unemployment Rate Change	1	40.99	40.994	10.19	0.003
Total Cases per Capita	1	44.93	44.927	11.16	0.002
Ln Oil & Gas Extraction GDP%	1	123.23	123.226	30.62	0.000
Error	43	173.07	4.025		
Total	46	363.77			

Fits and Diagnostics for Unusual Observations

Obs	Real GDP Percent Change	Fit	Resid	Std Resid	
16	-8.985	-4.941	-4.044	-2.17	R
26	-5.282	-6.118	0.836	0.56	X
28	-2.752	-4.854	2.103	1.24	X
32	-8.068	-4.133	-3.935	-2.11	R
42	2.847	-1.043	3.889	2.00	R

R Large residual
X Unusual X



By omitting these unusual points, we see that there are still slight violations of regression assumptions, as the normal plot of the residuals seems slightly off the straight line. There does not seem to be a problem with collinearity as VIFs are very small. All variables are statistically significant since their p-values are very small, therefore the null hypothesis is strongly rejected and there is a relationship between the target and all predictors in the regression equation.

Omitting three data points means changing the data set, therefore the "best model" might change as well. We'll need to rerun the best set regression on the new data set, here is the result:

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Unemployment Rate Change
1	17.8	16.0	318.1	12.6	41.8	2.5780	226.913	231.905	1.000	X
1	16.1	14.2	330.6	9.1	43.5	2.6041	227.860	232.853	1.000	
2	53.7	51.6	198.6	45.4	6.7	1.9560	202.298	208.746	2.356	
2	41.2	38.5	248.9	31.6	19.7	2.2057	213.590	220.039	2.930	
3	58.3	55.3	185.4	49.0	4.0	1.8792	199.963	207.751	4.174	X
3	53.9	50.7	216.9	40.4	8.6	1.9751	204.644	212.431	2.402	
4	60.2	56.4	188.1	48.3	4.0	1.8558	200.317	209.318	8.731	X
4	58.4	54.4	191.3	47.4	5.9	1.8988	202.470	211.471	17.200	X
5	60.2	55.4	194.7	46.5	6.0	1.8782	203.083	213.162	17.335	X

Vars	Total Cases per Capita	Ln Total Death per Capita	Oil & Gas Extraction GDP %	Crea Acco Serv GDP %
1				
1		X		
2		X	X	
2	X		X	
3		X	X	
3		X	X	X
4		X	X	X
4	X	X	X	
5	X	X	X	X

Using the same criteria as before, there are two candidates that stand out. One is two variables model {"Ln Total death per capita" and "Oil & Gas Extraction GDP %"}. The other is three variables model { "Unemployment Rate Change, " "Ln Total death per capita" and "Oil & Gas Extraction GDP %"}.

I examined both. It turned out two variables candidate seems the "best". The three variable model doesn't follow regression assumptions as strong as the two variable model, with R-Sq only 4% higher than two variable model. Since this model is for prediction purpose, simplicity is the key. Therefore I consider two variable model to be the "best" model. To save space, I only show the "best model" regression result.

Regression Equation

Real GDP Percent Change = -1.195 - 1.935 Ln Total Death per Capita - 1.121 Ln Oil & Gas Extraction GDP%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.195	0.329	-3.63	0.001	
Ln Total Death per Capita	-1.935	0.323	-6.00	0.000	1.20
Ln Oil & Gas Extraction GDP%	-1.121	0.187	-5.98	0.000	1.20

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	47-fold S	47-fold R-sq
1.95600	53.72%	51.62%	45.41%	2.05557	45.41%

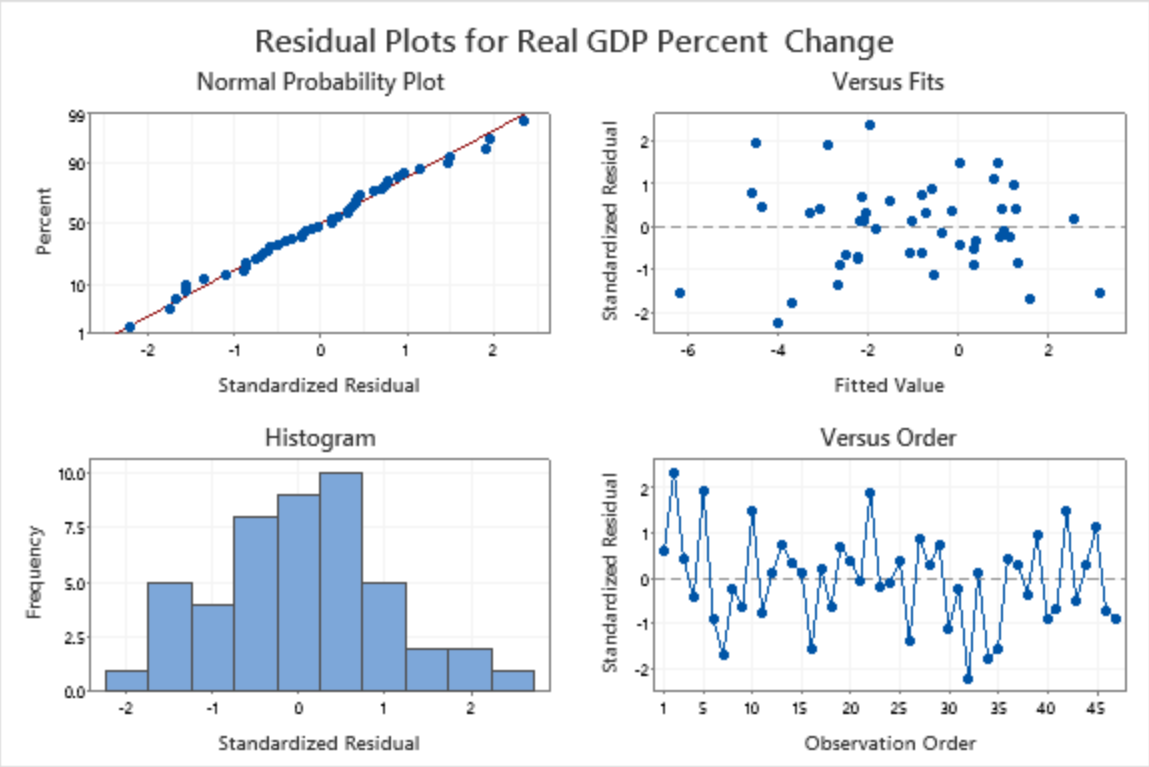
Analysis of Variance

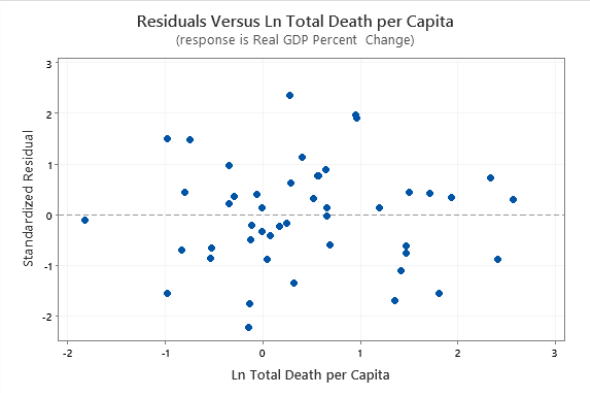
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	195.4	97.716	25.54	0.000
Ln Total Death per Capita	1	137.6	137.588	35.96	0.000
Ln Oil & Gas Extraction GDP%	1	136.8	136.816	35.76	0.000
Error	44	168.3	3.826		
Total	46	363.8			

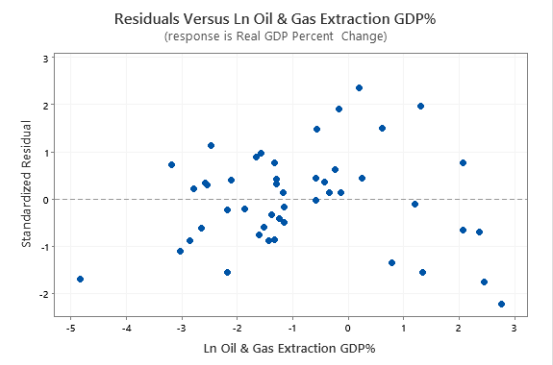
Fits and Diagnostics for Unusual Observations

Obs	Real GDP Percent Change	Fit	Resid	Std Resid	
2	2.599	-1.946	4.545	2.36	R
32	-8.068	-4.008	-4.060	-2.22	R

R Large residual







All of the variables are significant, and there does not seem to be a problem with collinearity as VIFs are very small. Diagnostics and residual plots look okay here. There is no pattern on the plots, so regression assumption does not seem to be violated.

Regression Equation

Real GDP Percent Change = -1.195
- 1.935 Ln Total Death per Capita
- 1.121 Ln Oil & Gas Extraction GDP%

The coefficient for Total Death per Capita (Ln Total Death per Capita) in the equation is −1.935. This means given that "Oil & Gas Extraction GDP%" is held fixed, a one unit increase in the natural logged Total Death per Capita is associated with a 1.935 unit decrease in Real GDP Percent Change.

The coefficient for "Oil & Gas Extraction GDP%" (Ln Oil & Gas Extraction GDP%) in the equation is −1.121. That means given that "Total Death per Capita" is held fixed, a one unit increase in the natural logged Oil & Gas Extraction GDP% is associated with a 1.121 unit decrease in Real GDP Percent Change.

There are only total 50 points available in the original data set, one point per state, so it's not possible to validate the model selection process and chosen model on new data. To validate the predictive power of the chosen model, I chose to use leave-one-out cross-validation method in Minitab, which estimates the standard deviation of the errors using the standard deviation of the set of prediction errors of the observations as each is left out. Here is the result:

S	R-sq	R-sq(adj)	R-sq(pred)	47-fold S	47-fold R-sq
1.95600	53.72%	51.62%	45.41%	2.05557	45.41%

"47-fold S", the leave-one-out cross-validation estimate of standard error σ is 2.05557, roughly 5% larger than S=1.956, the value given in the regression output. Thus, a better rough 95% prediction interval for "real GDP Percent Change" when using this model is ±(2)(2.05557) = ±4.11114, rather than ±(2)(1.956) = ±3.912, with the former roughly 5% wider than the latter.

The regression equation does show the underline process. It implies that as the pandemic continues, the number of deaths increase and real GDP will get negatively impacted. It also implies that the higher the Oil & Gas extraction industry GDP percent, meaning the higher proportion oil and gas makes up of the full GDP of the state, the real GDP of the state will face more negative impact. This is because the pandemic brought enforced lockdown and social distancing, causing gas and oil consumption to lower significantly, which decreases gas & oil extraction GDP leading to such impact on state GDP.

While I call this the "best" model, found through the best subsets regression model, the regression is not strong (R-sq (pred) 45.41%). With such a wide range of variability in using this model, this model has lots of room to improve. Future analyses might try to analyze other factors that might be associated with the pandemic's impact on real GDP percent change such as federal spending, including money sent as covid aid relief, as well as local government spending. Future analyses can also consider the state of healthcare and social assistance, as these factors contribute to how prepared a state would be to handle a pandemic. It would also be interesting to quantify how much of any given state's economy is made up of small businesses versus large corporations, as small businesses are more vulnerable in the face of the pandemic.