Regression and Multivariate Data Analysis
Professor Simonoff
9/29/22
Jessy Huang
Homework 2

Real GDP Percent Change and Covid Case per Capita

The COVID-19 pandemic triggered the sharpest downturn in the US economy since the Great Depression, with Real GDP declining by 8.9 percent in the second quarter of 2020, and an overall 3.5 percent in 2020 compared to an increase of 2.2 percent back in 2019. I would like to take this opportunity to explore the impact of the COVID-19 pandemic on the Real GDP in United States across the states over the course of 2020, specifically to study the impact on the Real GDP Percent Change by Covid Case per Capita.

Analyzing this event thoroughly can allow us to understand the effect of a pandemic on the economy, and thus allow us to predict and better prepare for future pandemics. However, I am aware that a simple regression cannot adequately represent the effect of the pandemic. I am still interested in understanding this topic through the lens of regression, first with simple regression and later multivariate regression as the course progresses.

The target variable is the 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%

The predicting variable is Covid Case per Capita of each state, which is calculated by:

Covid Case per Capita = Covid Case(s) of state / population of the state *100%

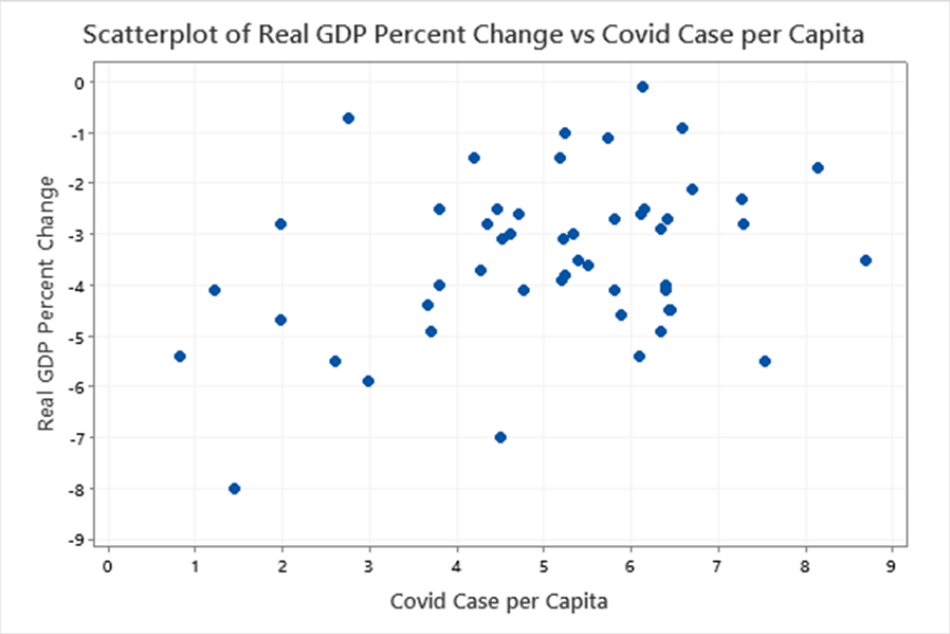
The following analysis is based on government data covering every state in the US over the years 2019 and 2020. GDP data is from the website of "Bureau of Economic Analysis, US Department of Commerce" (https://www.bea.gov/).

The Covid Case data comes from covid-19-data coverage from the New York Times (nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. (github.com)), which covers all 50 states and DC in the U.S. in the year 2020.

The regression formula is:

$$\text{Real GDP Percent Change} = B_0 + B_1 * \text{Covid Case per Capita} + \text{random error}$$

The following is a scatter plot of the two variables:



There seems to be a somewhat weak linear relationship between the two variables. To understand further, a least squares regression with Real GDP Percent Change as the dependent (target) variable and Covid Case per Capita as the independent (predicting) variable was run:

Regression Equation

Real GDP Percent Change = -4.934 + 0.293 Covid Case per Capita

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.934	0.655	-7.53	0.000	
Covid Case per Capita	0.293	0.122	2.39	0.021	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.51960	10.44%	8.62%	1.83%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	13.20	13.196	5.71	0.021
Covid Case per Capita	1	13.20	13.196	5.71	0.021
Error	49	113.15	2.309		

Total 50 126.35

Fits and Diagnostics for Unusual Observations

Obs	Real GDP Percent Change	Fit	Resid	Std Resid	
12	-8.000	-4.509	-3.491	-2.43	R
45	-0.100	-3.140	3.040	2.03	R
46	-5.400	-4.696	-0.704	-0.50	X
48	-0.700	-4.126	3.426	2.32	R
51	-7.000	-3.620	-3.380	-2.25	R

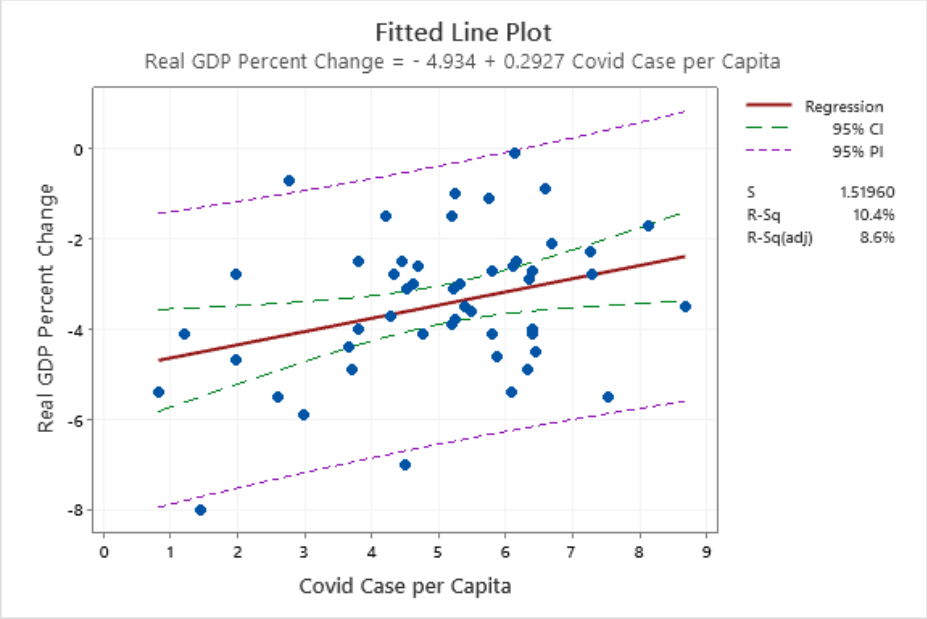
R Large residual
X Unusual X

The least square regression shows a weak regression with an R-sq of 10.44%, and a statistically significant F-statistic of 5.71. The slope coefficient shows that one percentage change in Covid Case per Capita is associated with an estimated expected .293 percentage point change in Real GDP Percent Change.

The t-statistic tests the null hypothesis of B_1=0. The t-value given the output for Covid per Capita is 2.39 with a p-value of .021 (t 95% = 2.008). Therefore the null hypothesis is strongly rejected and there is a relationship between Real GDP Percent Change and Covid Case per Capita. The t value that tests the null hypothesis for B_0=0 for B_0 is given as -7.53, which is also statistically significant and rejects the null hypothesis, meaning that when a state has Covid Cases not equal to 0, there would be -4.934 added to the value of B_1*Covid Case per Capita to predict the impact on Real GDP Percent Change.

The standard error of the estimate of 1.51960 tells us that this model could be used to predict the real GDP Percent Change within ±3.0392 percentage points, roughly 95% of the time.

The following simple regression scatter plot includes the confidence interval and prediction intervals:



It can be seen that the pointwise prediction interval is much wider than the pointwise confidence interval. This is because the prediction interval expresses more uncertainty; It expresses inherent uncertainty in the particular data point as well as sampling uncertainty. It is also noted that the confidence interval is narrowest in the center of the plot and gets wider at the extremes, which shows that predictions become progressively less accurate as the predicting value gets more extreme compared with the bulk of the points.

We can also see that three points seem to be slightly unusual – Washington on the upper left corner and Hawaii and Wyoming on the lower left side are all slightly outside of 95% prediction interval.

To gain a more detailed view of the confidence and prediction intervals, here is an example of the specific values of Real GDP Percent Change for Indiana when the Covid Case per Capita is 5.23014:

Prediction for Real GDP Percent Change

Regression Equation

Real GDP Percent Change = -4.934 + 0.293 Covid Case per Capita

Settings

Variable	Setting
Covid Case per Capita	5.23014

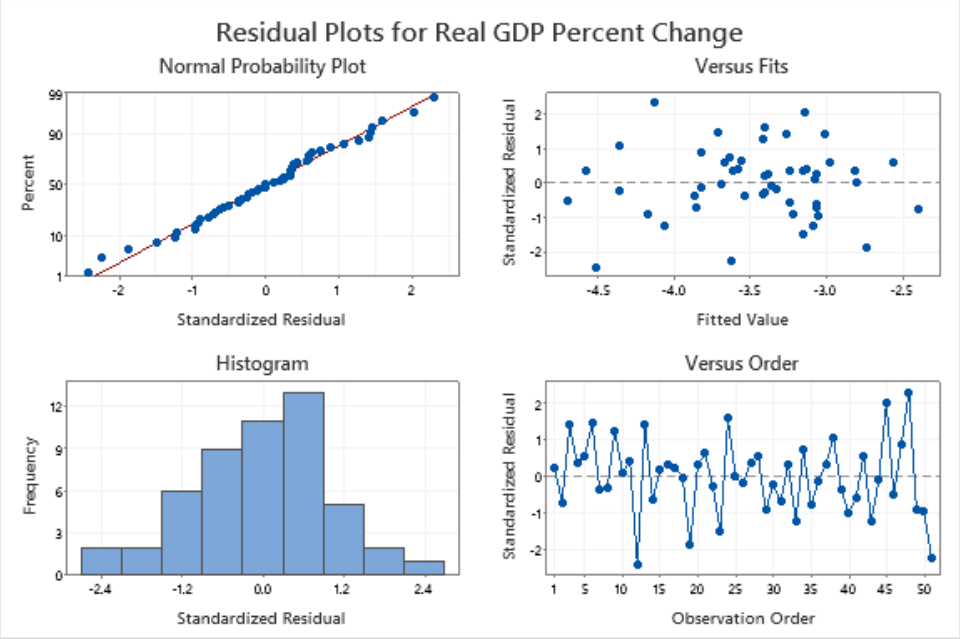
Prediction

Fit	SE Fit	95% CI	95% PI

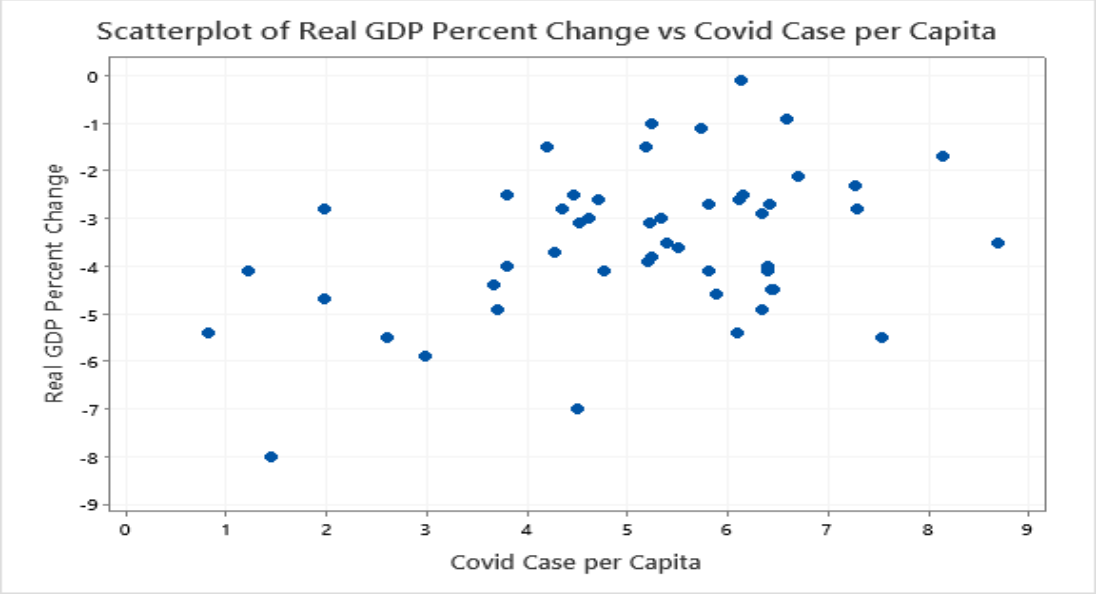
	- 0.213783	(-3.83329, -2.97406)	(-6.48750, -0.319842)
3.40367			

We can again see that the prediction interval is much wider than the confidence interval. The confidence interval predicts the average Real GDP Percent Change for all states with Covid Case per Capita of 5.23014 , which is (-3.83329, -2.97406). The prediction interval predicts the Real GDP Percent Change for one state with Covid Case per Capita equal to 5.23014, which would be (-6.48750, -0.319842).

To check whether the assumptions for a least squares regression hold, diagnostic plots can be plotted. Diagnostic plots can also be used to determine whether or not the three points noted earlier, Washington, Hawaii, and Wyoming are unusual. The following are plots of residual versus fitted values, residuals in observation order, a normal plot of the residuals, and a histogram of the residuals:



In the diagnostic plots, we can see that the assumptions of least squares regression seem to hold well. In residuals versus fitted values, the point that seems most unusual is also the point that was noted as most unusual earlier. Washington is considered an outlier point as it has an unusual response value given its predictor value. This can be a problem as outliers can have an affect on the fitted regression as well as measure of fit such as R-sq, t-value and F-value. Though unclear of the factors that cause Washington to be an outlier, the point was removed from the data set in order to mitigate influencing the results too much away from the bulk of the datapoints. The following is the scatter plot with Washington omitted:



Regression Equation

Real GDP Percent Change = -5.279 + 0.347 Covid Case per Capita

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-5.279	0.641	-8.24	0.000	
Covid Case per Capita	0.347	0.119	2.92	0.005	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.44875	15.07%	13.30%	7.05%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	17.87	17.871	8.51	0.005
Covid Case per Capita	1	17.87	17.871	8.51	0.005
Error	48	100.75	2.099		
Total	49	118.62			

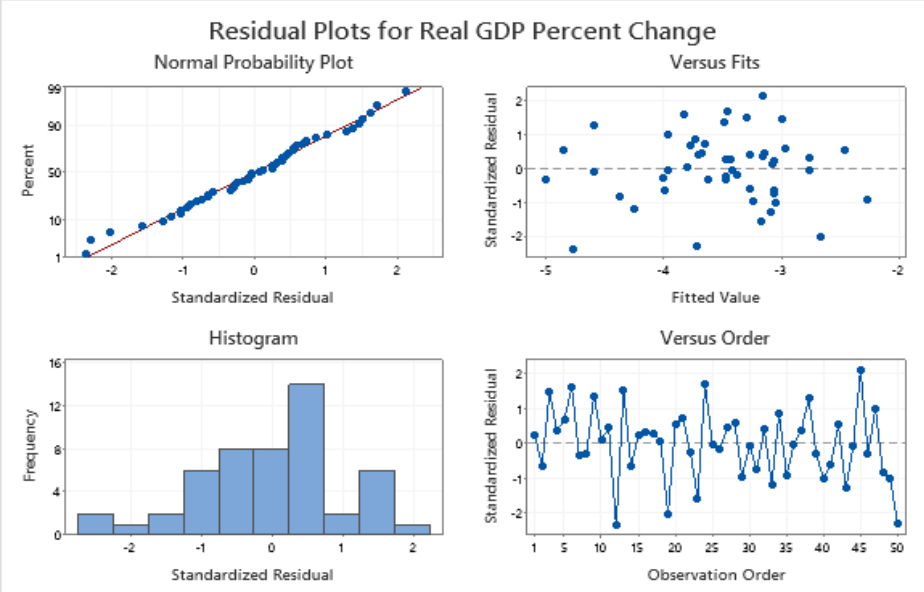
Fits and Diagnostics for Unusual Observations

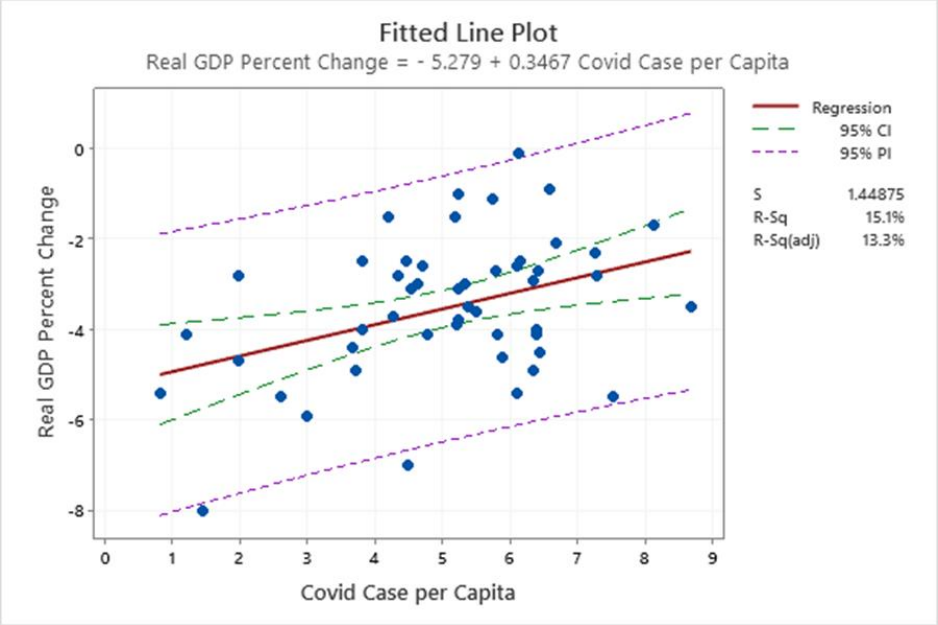
Obs	Real GDP	Fit	Resid	Std Resid

	Percent Change				
12	-8.000	-4.775	-3.225	-2.36	R
19	-5.500	-2.666	-2.834	-2.02	R
20	-4.100	-4.859	0.759	0.56	X
45	-0.100	-3.153	3.053	2.14	R
46	-5.400	-4.996	-0.404	-0.30	X
50	-7.000	-3.721	-3.279	-2.29	R

R Large residual
X Unusual X

With the unusual point Washington removed, we can see a large increase in the regression relationship as R-sq increased from 10.44 to 15.07. Moreover the intercept changed from -4.934 to -5.279 and the slope from 0.293 to 0.347. To see the effects on the diagnostic and confidence and predicting intervals after Washington was omitted, diagnostic plots and a regression plot with diagnostic and confidence and predicting intervals were run on the new data set:





After removing the point for Washington, there seems to be no outstanding outliers or leverages that strongly affect the regression. We can see that only Wyoming is a little off prediction interval, but not significantly. Therefore it seems that the "best" model is the one with Washington omitted.

The outlier Washington had very low response value given the predictor value; it had a lower Real GDP Percent Change than predicted by its Covid Case Per Capita. In other words, Washington was less vulnerable to the pandemic's economic consequences than others. Why would Washington's GDP suffer significantly less than other states? I find that areas with economies that rely on the movement of people—for example, Las Vegas and Hawaii rely heavily on tourism—faced substantially higher GDP change at the end of 2020 than cities with core industries based on the movement of information.

There seems to be two major factors playing a role here. First, Washington, specifically Seattle, benefits a lot from the technology industry, which may have benefitted from COVID-19. Seattle is the well-known birthplace of Microsoft and the home of Amazon. Employment in Seattle is most concentrated in their largest occupational groups, computer and mathematical occupations, which are both over two times (2.36 and 2.14 respectively) higher than the national average. Secondly the performance of their core industry spills over to support other industries in the proximity and therefore affects the regional economy; restaurants and retail stores do better when the core industry is booming and struggle when it is not. These two factors may explain why Washington appears more resilient to the pandemic.

The outlier shows that there are deficits in this model as there might be many outstanding variables that affect the Real GDP Percent Change for states in the US besides Covid Case per Capita. These variables could include industry sectors; For example, accounting for the weight of various industries in a state's economy and being able to quantify how vulnerable said industry is in the face of a pandemic would allow for a more nuanced and sophisticated analysis. Other important variables, such as the day of lockdown, the wealth of the area, the amount of federal aid, and

population density are all meaningful factors that influence the effect a pandemic can have on any given economy. These would all be interesting variables to include in further analysis.

However, after omitting this outlier to try to see the true regression model, we can see that there is a statistically significant relationship between the Real GDP Percent Change and Covid Case per Capita. With a R-sq of 15.07 and a statistically significant t-value, there seems to be support that there is a relationship between the Real GDP Percent Change and Covid Case per Capita.

Although it seems that this is a true regression model, the regression trend is unexpected. The trend direction suggests that the higher Covid Case per Capita, the less negative Real GDP Percent Change. One possibility is that there might be lurking variable of wealth of an area, since in 2020 COVID cases were concentrated in wealthy areas on the east and west coasts. I believe that including more variables, such as those previously mentioned, in a multivariate regression model would better represent the situation at hand and allow for a more in depth analysis.

Data Sources

Real GDP Percent Change from the years 2019 and 2020 across all states

Bureau of Economic Analysis. (2021, March 26). *News release*. Gross Domestic Product by State, 4th Quarter 2020 and Annual 2020 (Preliminary) | U.S. Bureau of Economic Analysis (BEA). Retrieved September 29, 2022, from https://www.bea.gov/news/2021/gross-domestic-product-state-4th-quarter-2020-and-annual-2020-preliminary

Covid Cases across all states

Nytimes. (n.d.). *Covid-19-data/US-states.csv at master · Nytimes/covid-19-DATA*. GitHub. Retrieved September 29, 2022, from https://github.com/nytimes/covid-19-data/blob/master/us-states.csv