COVID-19 Infection and Death Rates in America

December 13, 2020
Kasia Krueger
Lorenzo Gordon
Owen Applequist

Introduction

Since March of 2020, the people of the United States have been struggling to control the COVID-19 pandemic (“the virus”). Without much direction from the federal government, the states have enacted their own methods to try to control the virus (e.g. mask mandates, business restrictions, school shutdowns, etc.). Often, similar methods were adopted by states with similar parties in power; Democratic-led states typically had earlier and lengthier lockdowns, as well as mask mandates to try to control the virus, while Republican-led states typically did not shut down as quickly as Democratic states and had shorter lockdowns.

Regardless of governorship, the virus spread rapidly throughout the United States. The research team were interested in investigating, did some state’s protocols help temper the virus enough to see statistical significance? With each state government taking its own approach to try to control the pandemic, the following study explores if there is a correlation between the number of cases and deaths due to the virus, and the state governing party, or if there was no relationship at all.

This study piqued our interest as the conversation and controversy over lockdowns continues. For most of 2020, our governments did not know much about the virus or how to best control it. Do lockdowns and similar mandates work in helping to control the virus, or do all of the states have similar infection rates and deaths by COVID, regardless of the protocols set by their states’ party?

The target population is the number of citizens with COVID-19 in each state. The variables collected were the number of laboratory-tested cases and reported deaths in each state (Covid in the U.S.), the population of each state (US Census Bureau), the state political party by government trifecta (State government trifectas), and the state population as a proportion of the US total population.

Methods

The study is an observational study; most of the data came from the New York Times, who aggregate their totals daily from state and local health agencies. The research team collected data from the start of the pandemic (March 2020) through October 20, 2020. Each data point in the New York Times database falls under one of two categories: confirmed cases and deaths (diagnosis confirmed by a laboratory molecular test) or probable cases and deaths (tested using other methodologies, inferred diagnosis based on symptoms or other criteria) (Covid in the U.S.). Confirmed cases numbers are used in this study, so there is likely an

undercount of patients due to the use of the more strict inclusion criteria. The team is testing strictly reported cases (those confirmed by a lab test). The number of unreported cases and deaths may be much higher.

The sampling unit is one person with COVID-19. One interesting item the team may have found is that the deaths deviations do not follow the cases deviations. One would think that if a state has a high number of cases that it would have a high number of deaths, and vice versa. One reason for this is that the deaths may not necessarily be reported as COVID-19 deaths (for example, deaths can be reported as organ failure, disease, other complications, etc.) so the full data for COVID-19 deaths may be incomplete (Beusekom, 2020).

We included in the study, “state government by trifecta,” in order to better infer “the will of the people.” A trifecta is a derived nominal classification: A state government where a single political party holds the governor's office, the upper chamber of the legislature, and the lower chamber of the legislature is said to be in 'trifecta.’ The team included in the study which party holds the upper and lower chambers. This then creates three classifications of government (exclusively democratic, exclusively republican, or divided). See *Table 1.*

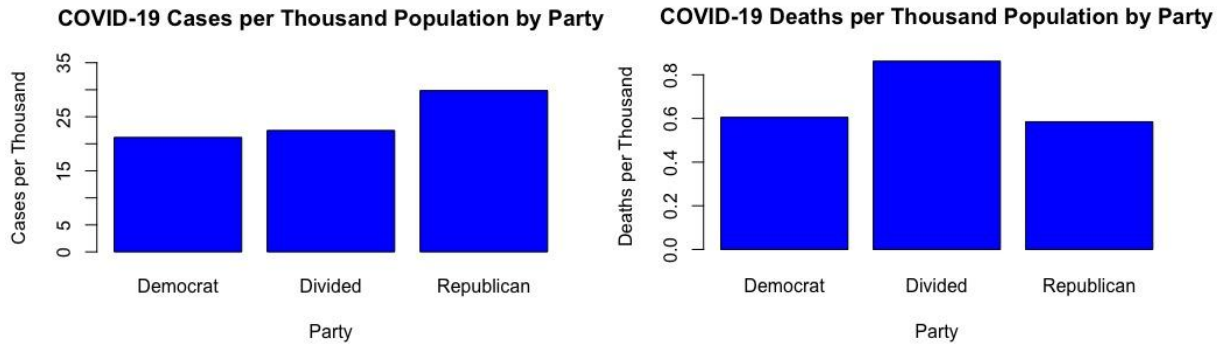
Table 1. States by Trifecta

States by Party

Parties	States
Democrat	CA CO CT DE HI IL ME NV NJ NM OR RI VA WA
Divided	KS KY LA MD MA MI MN MT NH NY NC PA VT WI DC
Republican	AL AK AZ AR FL GA ID IN IA MS MO NE ND OH OK SC SD TN TX UT WV WY

The team calculated cases per thousand for each of the trifectas, for cases and for deaths. *Figure 1* shows that there are more cases per thousand in Republican states, and democrat and divided states are fairly equal, while *Figure 2* shows that deaths are slightly higher in divided trifecta states, while democrat and republican states are fairly equal.

Figure 1. Cases per Thousand by Trifecta
Figure 2. Deaths per Thousand by Trifecta



To compare COVID-19 cases and deaths by state, the team normalized their counts by looking at each state's deviation from its expected count. States' expected count was calculated by distributing the U.S. total cases (or deaths) by the state's population. Then the deviation from the expected count was defined as a percentage of the expected count. i.e., deviation = (expected-count - actual-count)/expected-count.

Figure 3 and *Table 2* show how the actual cases deviate from the expected cases. The team found that the Republican states have a positive deviation, meaning the deviation of actual cases is higher than the expected cases for Republican states. Democrat and divided states show that they trend lower than the number of expected cases the team found.

However, *Figure 4* and *Table 3* show that each of the trifecta states trend lower for actual deaths than the team calculated for expected deaths. Again, refer back to the previous investigation that reported deaths may be lower than actual COVID-19 deaths.

Figure 3. Cases Deviation by Party
Figure 4. Deaths Deviation by Party

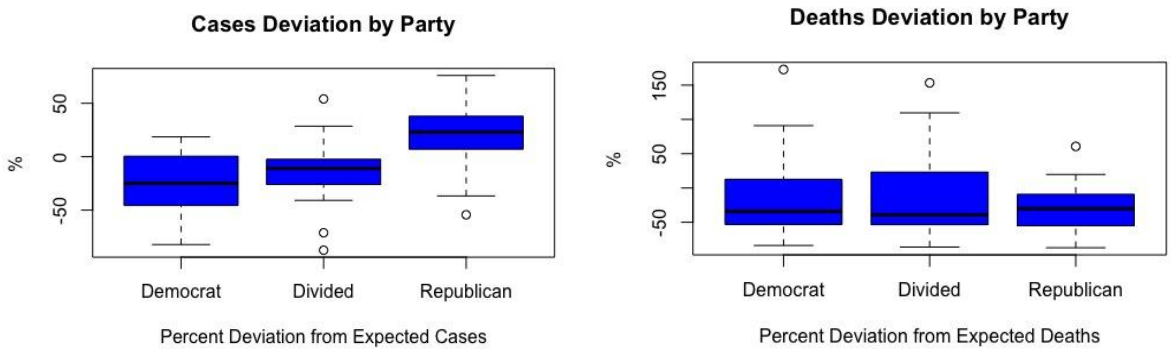


Table 2. Cases Deviation Mean and Standard Deviation
Table 3. Deaths Deviation Mean and Standard Deviation

Cases Deviation Mean & SD

Parties	Mean	SD
Democrat	-0.2449	0.3054
Divided	-0.1514	0.3472
Republican	0.1714	0.3275

Deaths Deviation Mean & SD

Parties	Mean	SD
Democrat	-0.0843	0.732
Divided	-0.0348	0.7091
Republican	-0.2816	0.3653

See Section *Citations*, for a list of data and sources.

Results

Analysis #1 (ANOVA)

Hypothesis

Cases: $H1_o$: $\mu_{Democrat}=\mu_{Republican}=\mu_{Divided}$ against $H1_a$: There exists some difference in the μs

Deaths: $H2_o$: $\mu_{Democrat}=\mu_{Republican}=\mu_{Divided}$ against $H2_a$: There exists some difference in the μs

There is no difference in the case or death deviation means of COVID-19 in Republican states, Democratic states, and Divided States vs.
there is a significant difference in the case or death deviation means of COVID-19 in Republican states, Democratic states, and Divided States.

Confidence level: 95%

Checking Assumptions of Linear Model

The team cannot assume homoscedasticity since the number of observations in the three groups is unbalanced. To verify that the deviations are normal and the population has the same variance, the research team plotted q-q plots for cases deviations (*Figure 5*) and deaths deviations (*Figure 6*), density plots to assess normality (*Figure 7, Figure 8*) and performed a Shaprio-Wilk test (*Table 4*).

For the cases deviation, the QQ plot shows a relatively straight line, the density plot has a single peak and equal tails, and the Shapiro-Wilk tests results in a large p-value, which all indicate that the cases deviations come from a normally distributed population. However, the deaths deviations fail these tests, so the research team needs to be more critical of death deviations in the ANOVA analysis. The large p-value for case deviation in the Shapiro-Wilk results indicates that the team can not reject the null hypothesis that the samples came from a normally distributed population.

Figure 5. Q-Q Plot of Cases Deviation
Figure 6. Q-Q Plot of Deaths Deviation

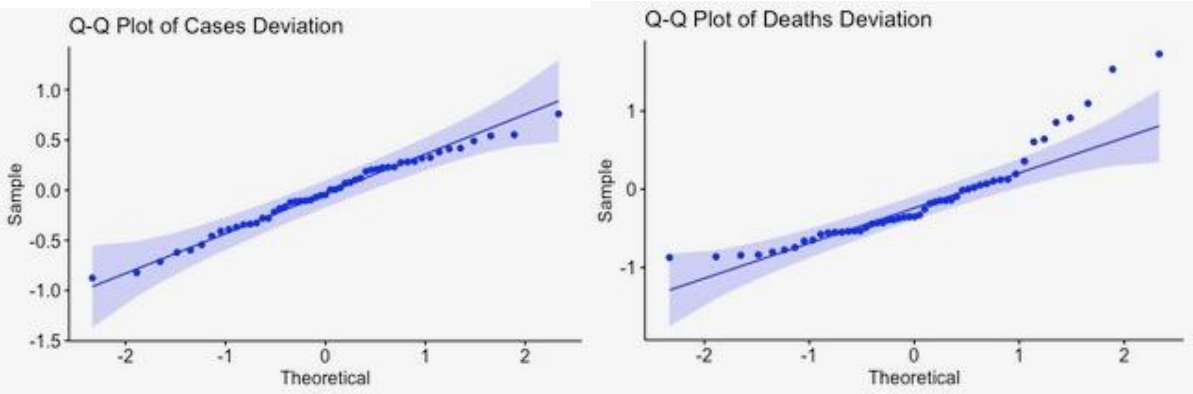


Figure 7. Density Plot of Cases Deviation
Figure 8. Density Plot of Cases deviation

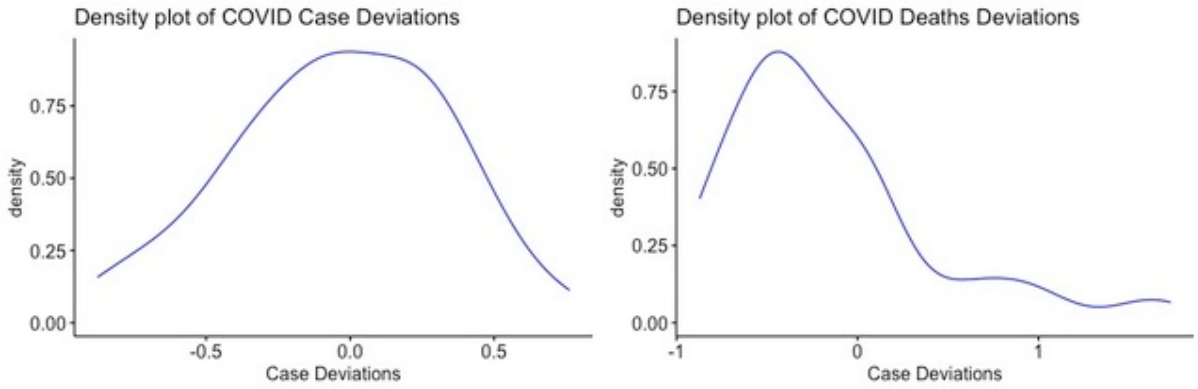


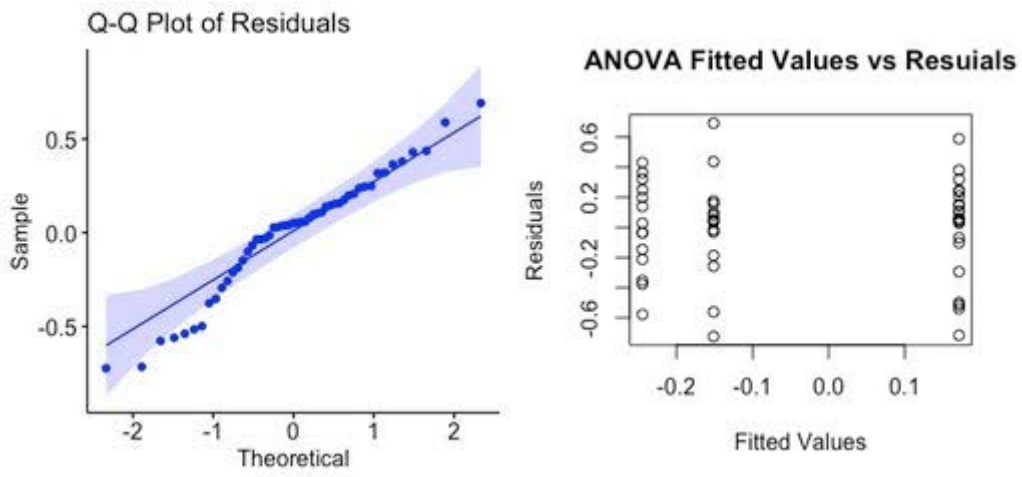
Table 4. Shapiro-Wilk Tests on Residuals

Cases Deviation	Deaths Deviation
W = 0.98742	W = 0.87478
p-value = 0.8617	p-value = 6.696e-05

Homogeneity of Variances

The team used the residuals versus fits plot to check the homogeneity of variances. In *Figure 9*, there is no evident relationship between residuals and fitted values (the mean of each group), and the residuals appear to be evenly distributed, therefore, the team can assume the homogeneity of variances.

Figure 9: ANOVA Fitted Values vs. Residuals Plot
Figure 10: ANOVA Fitted Values vs. Residuals Plot



One-way ANOVA Analysis

The one-way ANOVA test of the case deviations for the three groups (Democrat, Divided, Republican) produced an F-value of 8.1845 with 2,48 degrees of freedom, resulting in a very small p-value of .0009, showing significant evidence that the team can reject the null hypothesis for case means (See *Table 5)*. The null hypothesis that the means of the case deviations are equal can be rejected at the test level of alpha = .05 or even alpha = .001. However, the one-way ANOVA test did not find a difference in the mean deviation of the deaths deviations. The one-way ANOVA test produced a large p-value (see *Table 6*) showing there is no significant evidence and the research team would fail to reject the null hypothesis for the death deviations mean.

Table 5. One-way ANOVA of Cases Deviation by Trifecta
Table 6. One-way ANOVA of Deaths Deviation by Trifecta

One-Way ANOVA of Cases Deviation by Party

Source	df	SS	MS	F	P-Value
Between Groups	2	1.757	0.8785	8.1845	9e-04
Within Groups	48	5.1522	0.1073		
Total	50	6.9092			

One-Way ANOVA of Deaths Deviation by Party

Source	df	SS	MS	F	P-Value
Between Groups	2	0.6391	0.3195	0.9126	0.4083
Within Groups	48	16.8073	0.3502		
Total	50	17.4464			

Post Hoc Tests

Fisher's Least Significant Difference test results listed in *Table 7* presents a very small p-value, p = 0.0005 for Democrat vs. Republican, and Divided vs. Republican, p = .005. This indicates that the Republican group was the odd group, confirming the earlier boxplots (*Figure 3 and Figure 4*).

Table 7. Fisher's Least Significant Difference

Pairwise comparisons using t tests with pooled SD		
data: CovidByState$CasesDeviation and Trifecta		
	Democrat	Divided
Divided	0.44595	-
Republican	0.00053	0.00500

Welch two-sample t-tests were run to test the null hypothesis that the mean cases and deaths deviations between Democrat and Republican states are equal.
Cases: H_0: $u_{Democrat} = u_{Republican}$ against H_a: $u_{Democrat} \neq u_{Republican}$
Deaths: H_0: $u_{Democrat} = u_{Republican}$ against H_a: $u_{Democrat} \neq u_{Republican}$

Table 8 lists the results of the cases deviation test run with an alpha level of 0.05. The results of the t-test show that T = -3.876, p-value = 0.001, indicating there is significant evidence to reject the null hypothesis for the cases deviation means. Again, *Table 9* shows no significant evidence to reject the hypothesis that the means of the death deviations are equal, therefore the test fails to reject the null hypothesis for death deviations.

Table 8. Republican vs. Democratic States Cases Deviation t-test
Table 9. Republican vs. Democratic States Deaths Deviation t-test

t-test COVID Cases, Rep vs Dem

Statistic	Value
Method	Welch Two Sample t-test
Conf Level	0.95
t	-3.876
df	29.28
p-value	0.001
Conf Int Low	-0.636
Conf Int High	-0.197
Mean Dem	-0.245
Mean Rep	0.171

t-test COVID Deaths, Rep vs Dem

Statistic	Value
Method	Welch Two Sample t-test
Conf Level	0.95
t	0.937
df	17.18
p-value	0.362
Conf Int Low	-0.247
Conf Int High	0.641
Mean Dem	-0.084
Mean Rep	-0.282

Conclusion

The team found statistical significance (at 95%) that the case deviation means differed between the states categorized by government in power. This confirmed the earlier boxplots, showing a difference in deviations between the three governments in power. There is significant evidence that a state's government has an effect on the expected vs. actual case deviation. The research team were not able to find statistical significance (at 95%) that the death deviations means differed between states categorized by the government in power. If there was a full accounting of death data, then the deaths should likely trend like the cases. However, there is evidence that COVID-19 cases are not necessarily reported as COVID-19 deaths, whereas all infections found by lab test were reported as COVID-19 cases (Beusekom, 2020). There may be a discrepancy in the collection of data which is reflected in the mixed results. The team needs more data to repeat the test.

Analysis #2 (Linear regression)

Following the ANOVA study, the research team decided to look at the question, "are death rates affected by case rates"? *Table 10* shows the mean and standard deviations of the case deviations and death deviations for all states.

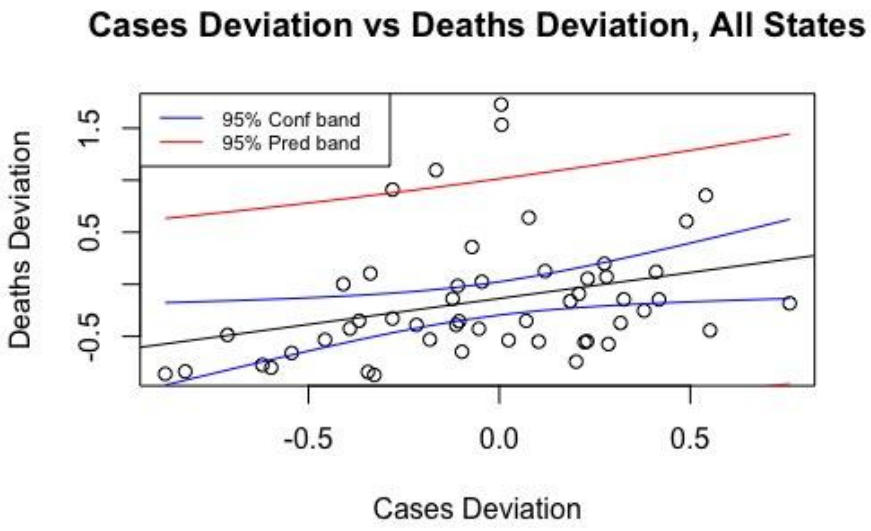
Table 10. Cases Deviation vs. Deaths Deviation, All States

Deviations Mean & SD

Variables	Mean	SD
Cases Deviation	-0.0378	0.3717
Deaths Deviation	-0.1549	0.5907

The scatter plot of Death Deviations vs. Case Deviations for all states is presented in *Figure 11*, along with the confidence band and prediction band. This chart suggests there may be a positive relation. The team used regression analysis to test if there is a linear relationship between the two variables.

Figure 11. Cases Deviation vs. Deaths Deviation, All States



Hypothesis

H_o: $\beta_1 = 0$ against H_a: $\beta_1 \neq 0$.

There is no relationship between death deviations to case deviations in all states vs.
there is a significant relationship between death deviations to case deviations in all states.

Confidence level: 95%

Checking Assumptions of Linear Model

The team used a q-q plot of the regression residuals to assess the normality assumption. The residual outliers shown in *Figure 12* indicate a possible problem with normality. The boxplot of the regression residuals shown in *Figure 13* indicates some outliers, and the histogram in *Figure 14* again indicates the outliers and a strong right skew. Since the sample size is large, the team can be less concerned about a non-normal population of error terms. There are no serious violations of the normality assumption, but further analysis of the data would still be useful here.

Figure 12. QQ plot of Regression residuals

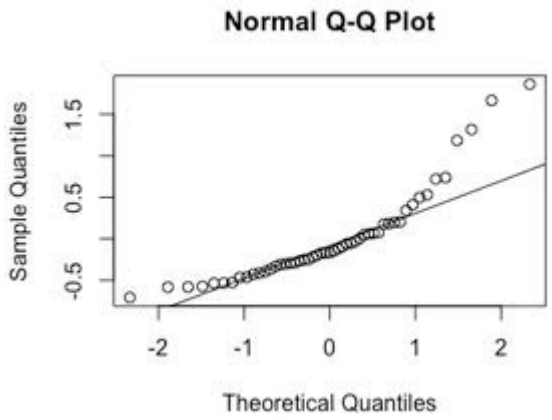


Figure 13. Boxplot of Regression Residuals

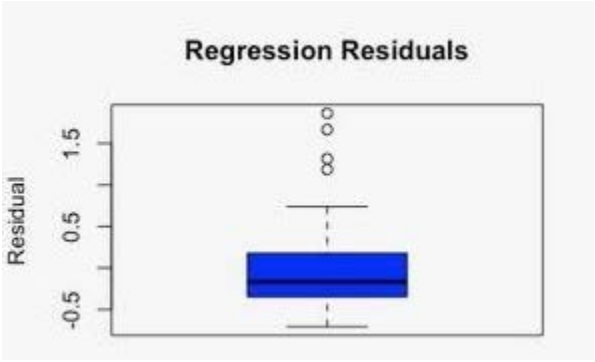


Figure 14. Histogram of Regression Residuals

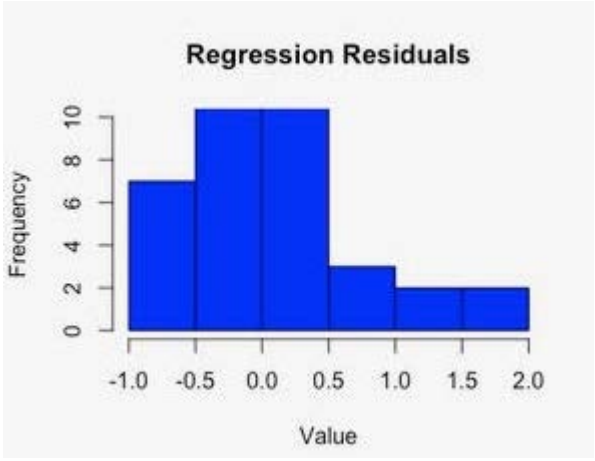


Table 11 shows the coefficients of the regression analysis for case deviations.

Table 11: Regression Coefficients

(Intercept)	CasesDeviation
-0.1359898	0.4991377

Table 12 shows the linear model summary statistics for case deviations.

Table 12: Cases Deviation Linear Model Summary Statistics

Coefficients:						
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**		
(Intercept)	-0.13599	0.07974	-1.705	0.0945	.	
CasesDeviation	0.49914	0.21552	2.316	0.0248	*	
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1
Residual standard error:	0.5665 on 49 degrees of freedom					
Multiple R-squared:	0.09866,	Adjusted R-squared:	.08027			
F-statistic: 5.364 on 1 and 49 DF, **p-value**: 0.02479						

Interpretation of the Correlation Coefficient

The sample correlation coefficient (r^2 value) is 0.09866. This is interpreted as 9.866% of the states' death deviations can be attributed to case deviation. While it's not an overwhelming relationship, it is still a significant relationship with a p-value of 0.0248.

Interpretation of the Slope

The p-value of the linear model is 0.02479 (See *Table 12*). The p-value is less than the chosen alpha of 0.05, therefore, the evidence to reject the null hypothesis is statistically significant and there is strong evidence that β_1 differs from zero.There is strong evidence of a real relationship between the two variables; this means that there is strong evidence that the deviation in the number of deaths in is related to the number of cases and would not have occurred by chance.

Conclusion

The regression analysis of the relationship of death deviations to case deviations indicates significant evidence to reject the null hypothesis, $\beta_1 \neq 0$. There is enough evidence to show that there is a relationship of death deviations to case deviations with a p-value of 0.0248. The coefficient of determination, $R^2 = 0.0987$, indicates that the case deviations can only explain 9.87% of the variation in death deviations. Also, due to the possible problems shown in the plots, further exploration (better data) is necessary.

Citations

US Census Bureau. (2019, December 30). State Population Totals: 2010-2019. The United States Census Bureau. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html

State government trifectas. (2019, November 19). Ballotpedia. https://ballotpedia.org/State_government_trifectas

Covid in the U.S.: Latest Map and Case Count. (2020, October 20). The New York Times. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html#states

Beusekom, Mary Van. 2020, July 1. About 30% of COVID deaths may not be classified as such. Center for Infectious Disease Research and Policy (CIDRAP), University of Minnesota. https://www.cidrap.umn.edu/news-perspective/2020/07/about-30-covid-deaths-may-not-be-classified-such

Tupper, Seth. 21 Nov. 2020. Two Rural States With GOP Governors And Very Different COVID-19 Results. NPR. https://www.npr.org/2020/11/20/936800456/two-rural-states-with-gop-governors-and-very-different-covid-19-results