# **STA 261 Inferential Project**

**COVID-19 in Kentucky: An Examination of the Various Factors Affecting the Pandemic of Our Lifetimes**

**Introduction**

The dataset we are working with in this project looks at the number of COVID cases and deaths (both in their raw numbers and per thousand) that have taken place in Kentucky up to August 1, 2021. The dataset is broken down by the counties of Kentucky and includes information on population, population density, age, income, crime statistics (violent crime, property crime, and officers assaulted and killed), unemployment rate, mask usage, percent of college educated residents and percent of veterans in each of them.

The research questions we will be investigating are:

1. Does the income level of a county affect the number of COVID-19 cases per 1000 it reports?
2. Does the mask usage of a county affect the total number of COVID-19 cases per 1000?
3. Does the unemployment rate affect the total number of cases per 1000 for each county?
4. Does the percentage of college educated individuals living in a county affect the total number of cases per 1000 in that county?
5. Does the percentage of GOP identifying individuals in a county affect the number of cases per 1000 it experiences?

**Exploratory/Descriptive Analysis:**

The two primary variables that the first question focused on were income and total number of cases of COVID-19 per 1000 people (specifically in Kentucky for both of the variables). The income variable is quantitative in nature. It represents the median household income within each county of Kentucky. The distribution of income is skewed to the right (refer to Figure 1) meaning that the mean of the overall distribution is being pulled to the right/increasing in value because of some very large values that come up in that direction. After doing some research (in the Data Visualization Project), it was found that these counties that were showing up on the right were often home to the largest cities in the state and because cities generally have incomes, this would therefore explain the right skew that is seen in the histogram of income. The median income value of the 120 counties of Kentucky was ~$44,283 with a maximum income level of $99,128 and a minimum income level of $24,623 (refer to Table 1). Because the histogram showed that the data was skewed, using the median was better because it is resistant to outliers and extreme values. The total number of cases per 1000 is also a quantitative variable. It represents the total number of cases per 1000 people in each of the counties of Kentucky. It had a slightly right skewed distribution that seemed to have no outliers (refer to Figure 2). The median number of total cases of COVID19 per 1000 in Kentucky was ~104 with a maximum of ~191 cases in one county and a minimum of ~53 cases in another (refer to Table 2). Once again, because of the slight skew of the histogram, I chose to use the median because of the reasons mentioned previously.

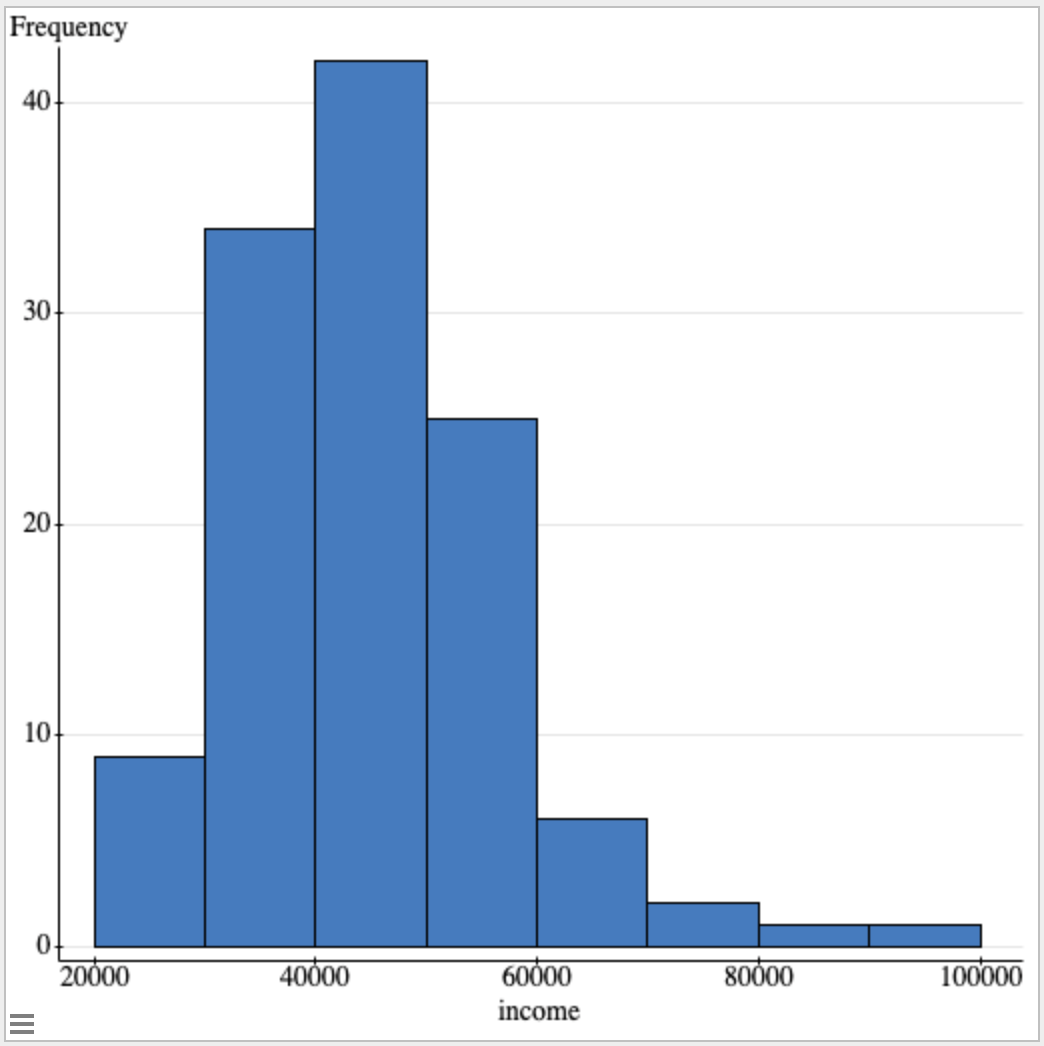
Figure 1: Distribution of Income (Kentucky)

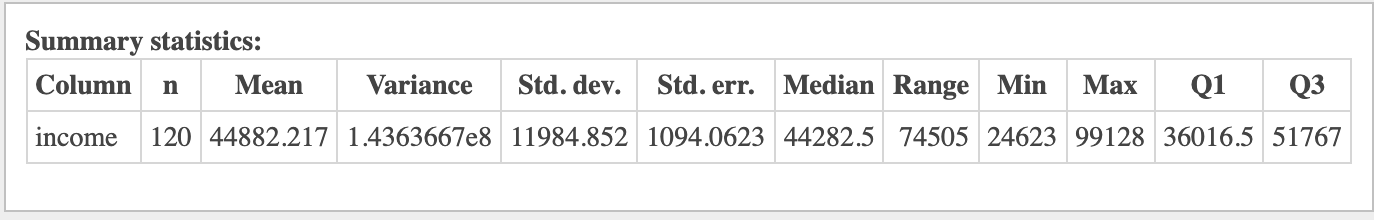
Table 1: Summary Statistics of Income (Kentucky)

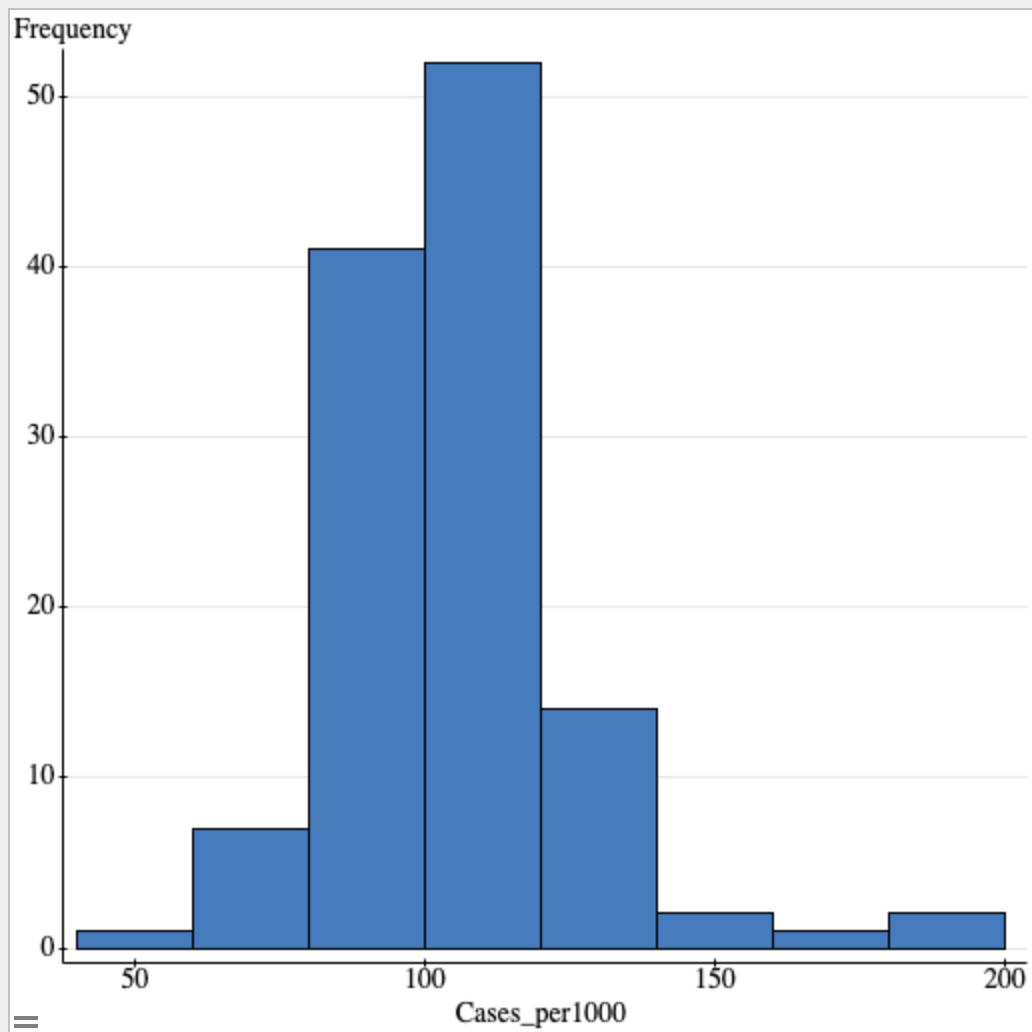
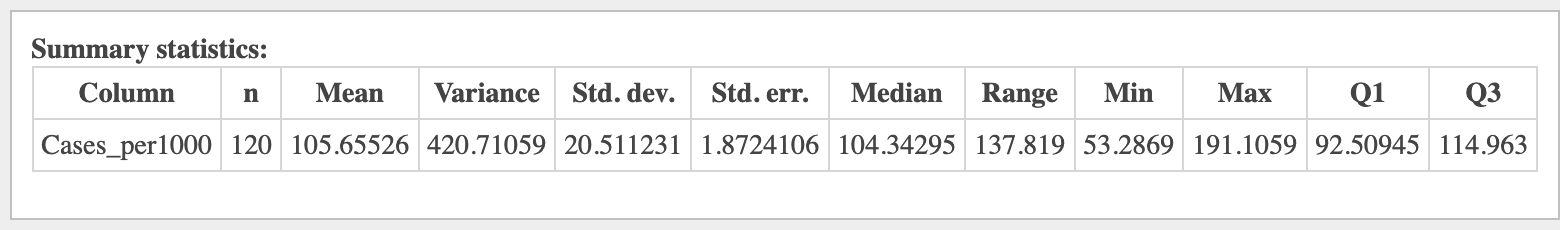
Figure 2: Distribution of Total Cases Per 1000 (Kentucky)

Table 2: Summary Statistics of Total Cases Per 1000 (Kentucky)

**Inferential Results:**

**Does the income level of a county affect the number of COVID-19 cases per 1000 it reports? (Vimal Vinod)**

Inferential Procedure: Simple Linear Regression Analysis Test

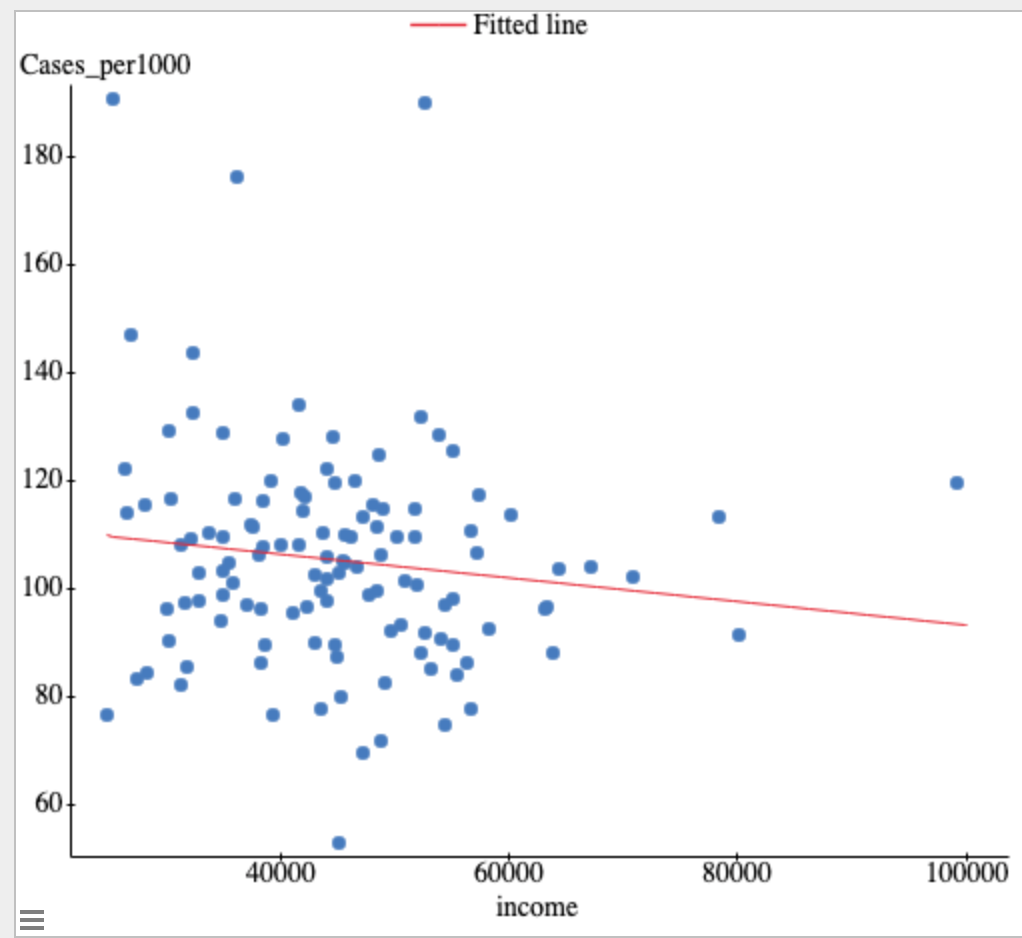
Hypotheses:

H0: β1 = 0 or the model is not useful in predicting y.

HA: β1 ≠ 0 or the model is useful in predicting y.

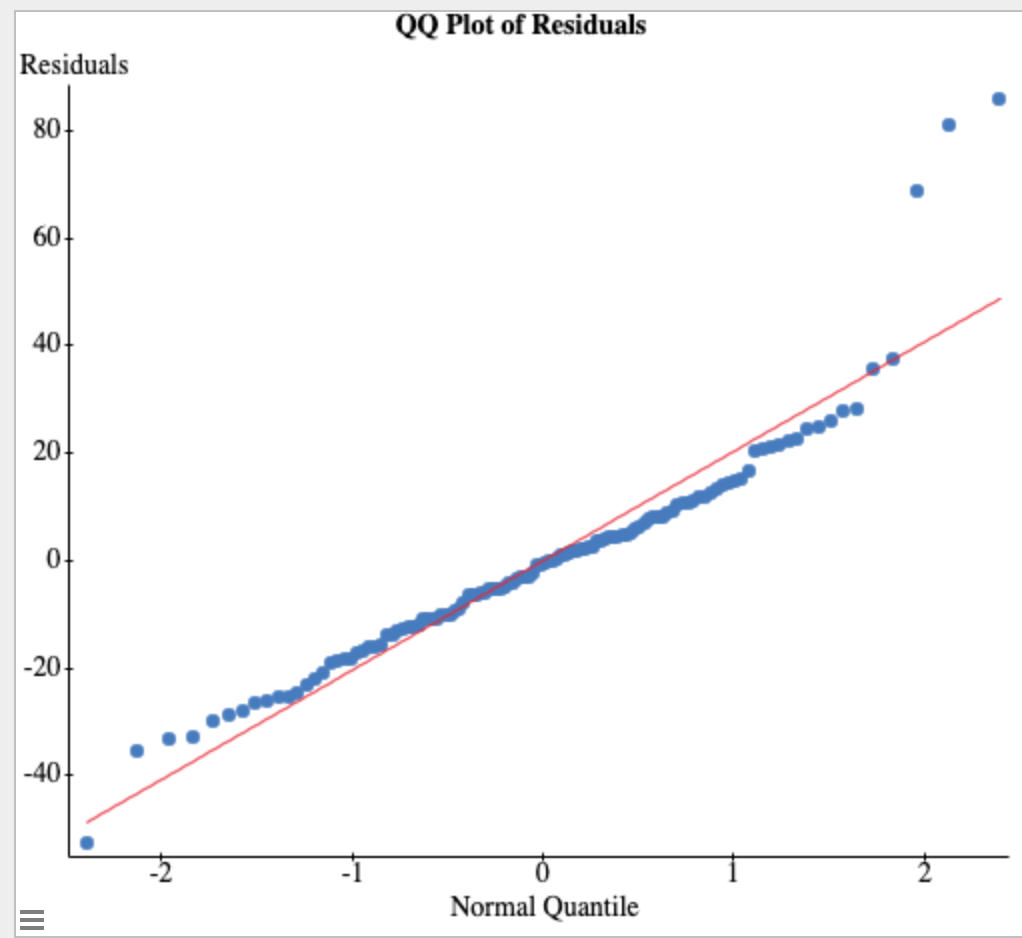
Assumptions:

**Linear relationship between x and y:** The relationship between the independent variable (income) and the dependent variable (total cases per 1000) is linear in nature. This can be proven by a scatterplot of the two variables as shown below. The two variables have a weak negative linear relationship. Thus, this assumption is met.

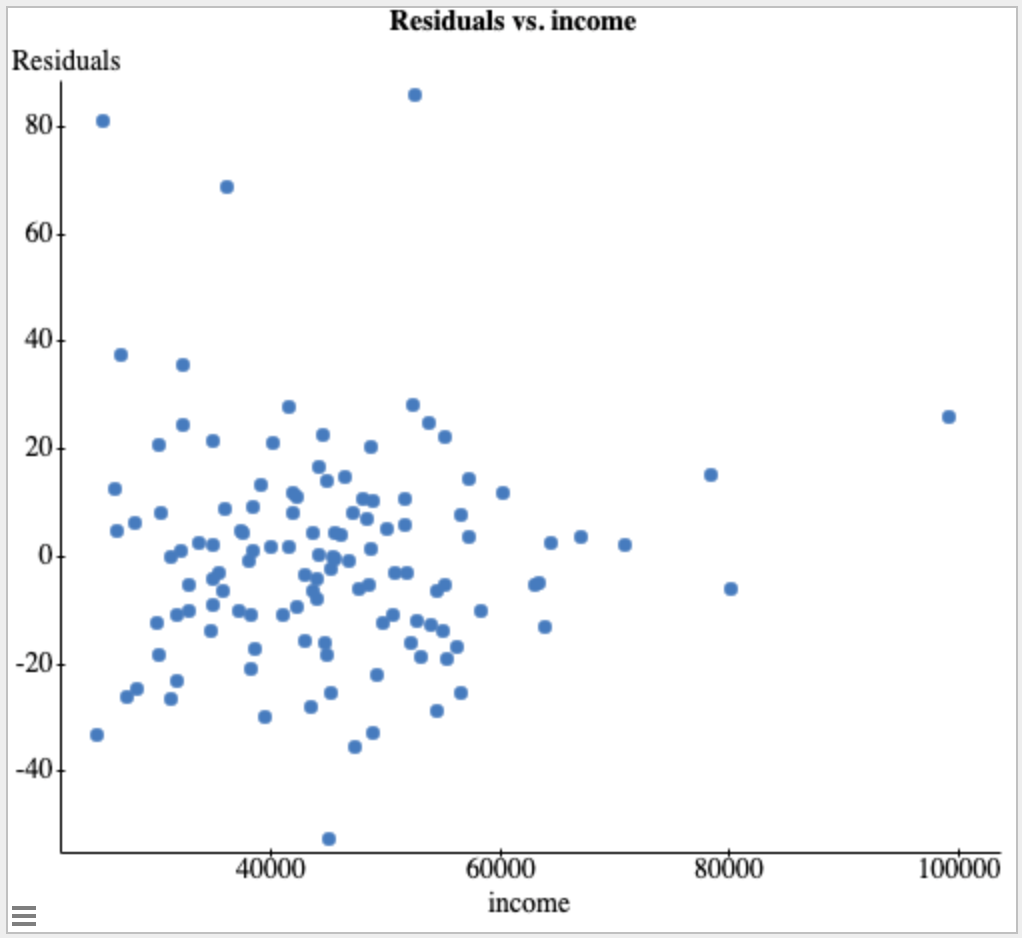


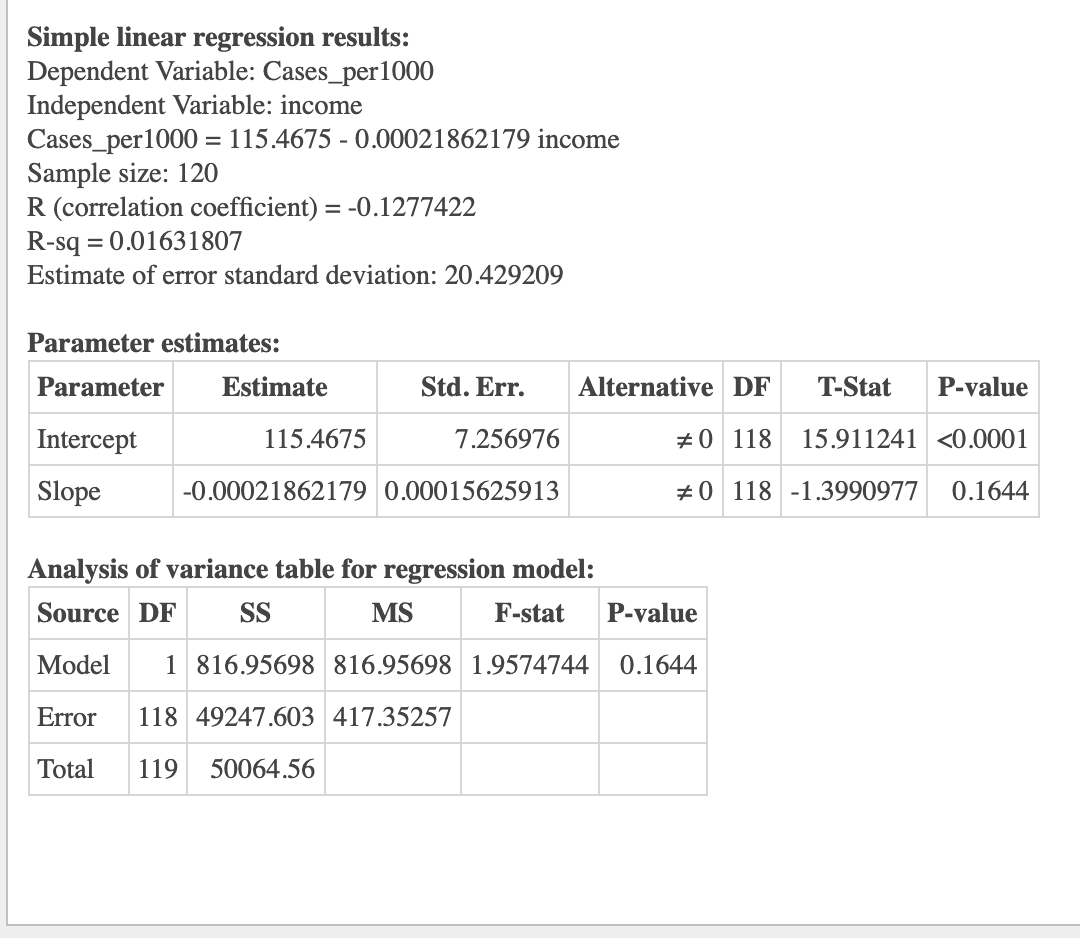
**Independence of observations and residuals:** This assumption is assumed to have been met in this class.

**Residuals are Normally distributed:** This assumption can be proven with a QQ plot that is approximately linear and this is shown to be the case in this test as seen below where a clear linear relationship is visible. Thus, this assumption is met.



**Y-values have Equal variance:** This assumption can be proven with a residual plot that shows no megaphoning or patterns and is just a broad range of points which is shown to be the case in the below graph. Thus, this assumption is met.



StatCrunch Output: 

Inferential Procedure Results/Conclusion:

I chose to use the simple linear regression analysis procedure (SLR) in analyzing the relationship between income and total cases of COVID-19 per 1000 because it was the only procedure that could show me if income levels (the x-variable) was useful in predicting (in a linear relationship) the values of the total cases of COVID-19 per 1000 residents in each county of Kentucky. Because I wanted to answer this direct question of whether one variable could explain/predict another, I had to choose the simple linear regression analysis procedure. Because all of the assumptions of this particular test were met as seen above, this procedure can be used to look at the relationship between the two variables I have chosen.

Upon seeing that all of the SLR assumptions were met, I set about analyzing the results of the test I had chosen to use to answer my question. After running the SLR test, I got back a p-value of 0.1644. The significance level which I ran this test on was 5% or 0.05 and because 0.1644 > 0.05, I failed to reject the null hypothesis which stated that the model was not useful in predicting the y-values (cases per 1000). Because I could not definitively show that there was a relationship between income levels and total COVID-19 cases per 1000, I cannot say that income has an appreciable effect on total cases per 1000. This means that I could answer my question whether the “income level of a county affects the number of COVID-19 cases per 1000 it reports?” with the fact that income levels were not a good predictor of the number of cases per 1000 of COVID-19 a county reports in Kentucky.

**Exploratory/Descriptive Analysis:**

The two main variables of the third question are high mask usage (x), and COVID-19 cases per 1000 (y), which are both being measured in counties of Kentucky. The x variable, high mask usage, is categorical because the participants of the experiment were asked if they wear a mask or not on a survey. The y variable, COVID-19 cases per 1000, is a quantitative variable since the number of cases recorded is a numerical value that can be measured. It can be seen from Figure 1 below that the distribution of total cases is majorly skewed to the right, which could be explained by mask usage being a good measure of COVID-19 cases, which causes COVID-19 cases to stay more on the lower side.

The high mask usage is a categorical variable, and the distribution slightly left skewed, as shown below by Figure 2. This means that high mask usage seems to be increasing, and is more on the higher side, which could explain why COVID-19 cases seemed to have stayed at a lower frequency, referencing Figure 1.

Figure 1: Distribution of Total Cases Per 1000

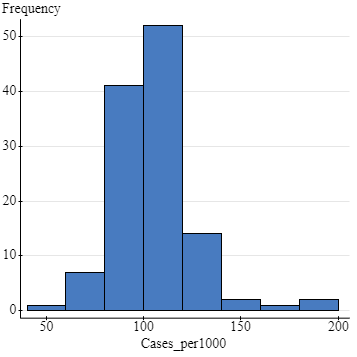
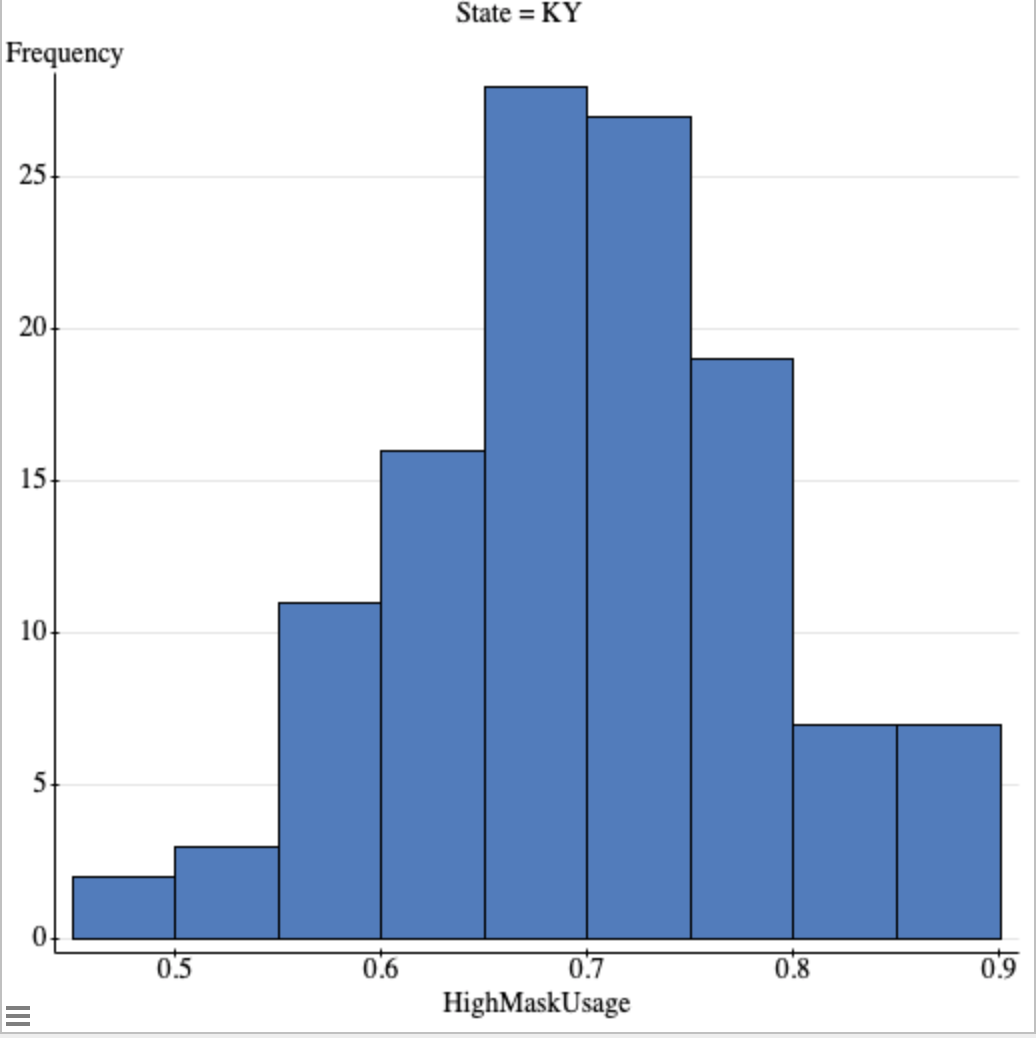
****

Figure 2: Distribution of High Mask Usage

**Inferential Results:**

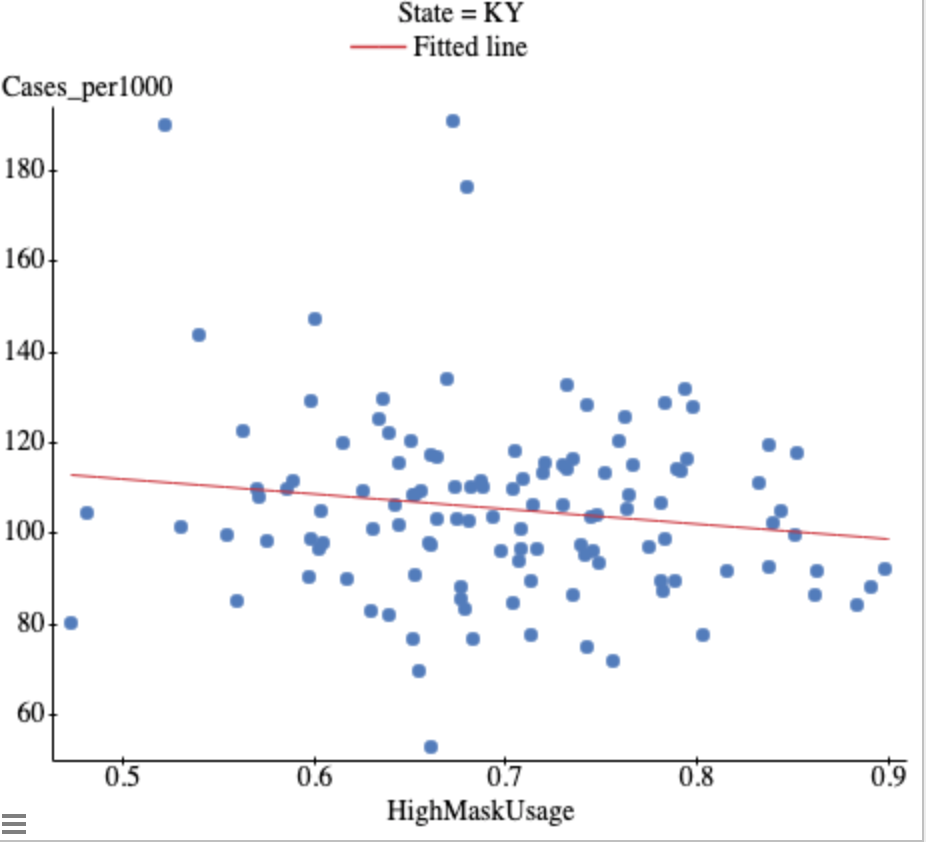
**Does the mask usage of a county affect the number of COVID-19 cases per 1000? (Anna Herrmann)**

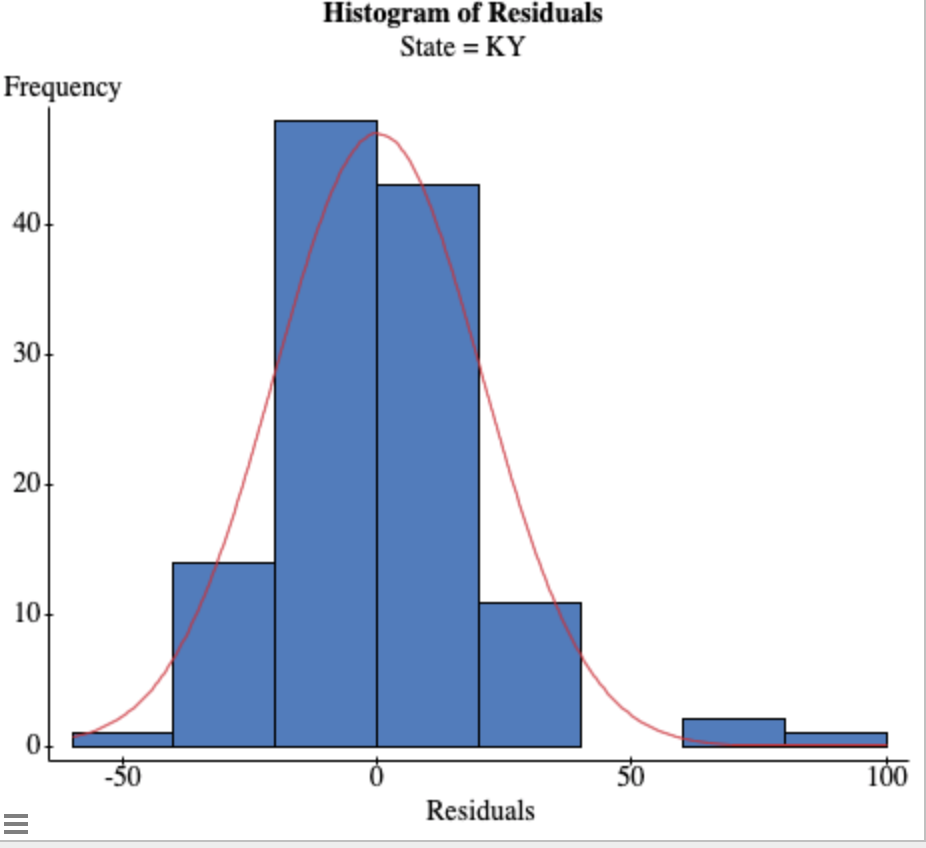
Hypotheses:

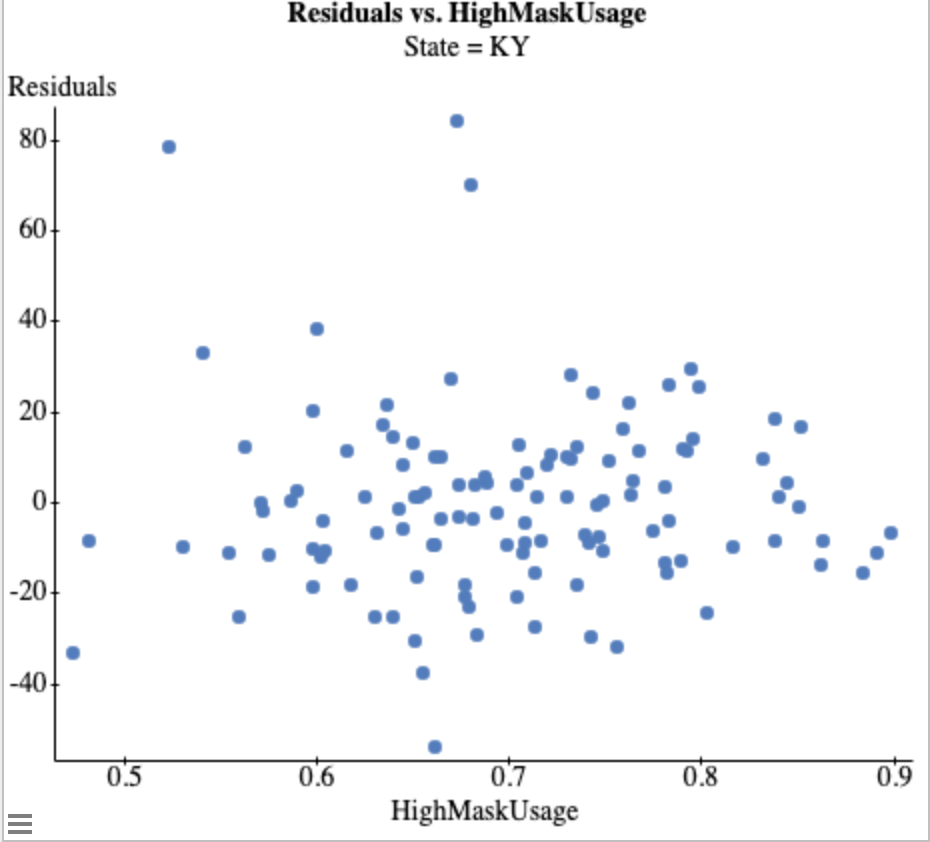
H0: β1 = 0 ---> This model is not useful for predicting y (Number of COVID-19 Cases)

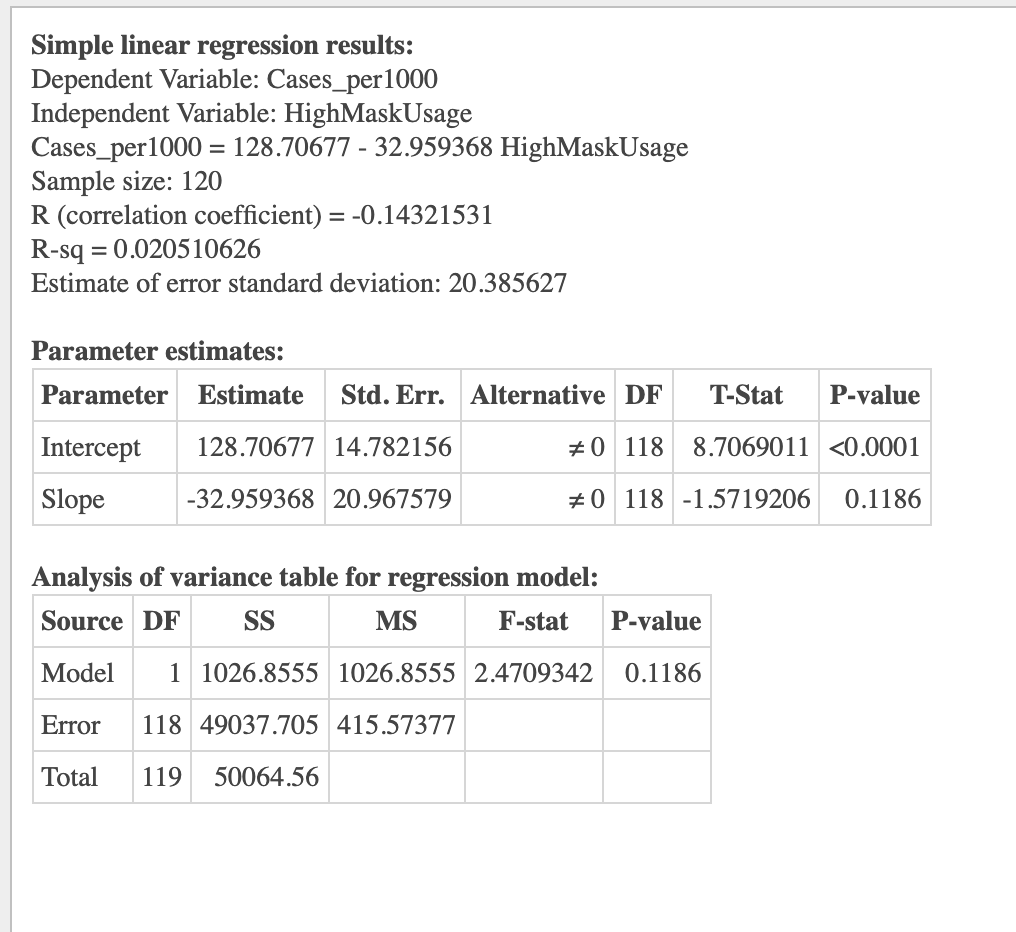
HA: β1 ≠ 0 --->This model is useful for predicting y (Number of COVID-19 Cases)

Assumptions for Simple Linear Regression Model:

1. **There is a Linear Relationship Between X and Y:** This can be modeled by a scatterplot of the data, which shows whether the correlation between the data is linear, whether that be a weak or strong linear relationship. It can be seen from the scatterplot below that there is a weak negative correlation between the variables x and y, modeled by the line of best fit. 
2. **There is Independence of Observations and the Residuals:** This assumption can reasonably be assumed.
3. **The Residuals are in a Normal Distribution:** This assumption can be modeled by either a QQ plot or histogram, to ensure that the data is following an approximately normal distribution. As modeled here by the histogram, the vast majority of the residuals are following a normal distribution, and that can be seen because the data fits almost perfectly under the normal curve on the histogram. Although there are a few outliers, most of the residuals are normally distributed, and therefore the assumption is met.



1. **Y Values Have Equal Variance:** This assumption is modeled by a residual plot, and this can tell us if there is any clumping or megaphoning of the data, which means the data is clumped together in a megaphone-like shape, and the assumption is not met. It can be seen in the residual plot below that the data is spread pretty equally, and there are no major clumps or megaphoning of the data, so this assumption has been met. 

Statcrunch Output: 

Inferential Procedure Results/Conclusion:

I chose to use a simple linear regression analysis because by using this analysis, it can give me a clear idea of the correlation between the percentage of mask usage, and whether that is causing COVID-19 cases per 1000 to increase or decrease in the counties of Kentucky. There are many graphs used to interpret a simple linear regression analysis, which can give me a clear idea if there is a positive or negative correlation present, and furthermore show me the normality of the data present. In conclusion, I decided to use a simple linear regression model because it is the most efficient way to show a simple relationship between my two variables of x (percentage of mask usage), and y (number of COVID-19 cases per 1000 in Kentucky).

After computing a simple linear regression model for the data, all assumptions were successfully met, and computing the test gave the data of whether to reject or accept the null hypothesis that high mask usage is not a good predictor of the number of COVID-19 cases per 1000 in the state of Kentucky. When reviewing data from the test, I used a significance level of 0.05, which means that I conducted a 95% confidence interval test in order to conclude whether to reject or fail to reject the null hypothesis. Additionally, I came to the conclusion that I failed to reject the null hypothesis that mask usage is a good predictor for the number of COVID-19 cases per 1000 in Kentucky. In the results, the data had a P-value of 0.1186, which is greater than the significance level of 0.05 which justified my conclusion in the previous sentence. In conclusion, there is sufficient evidence to conclude that high mask usage is not a good predictor for the number of COVID-19 cases per 1000 in the state of Kentucky.

**Exploratory/Descriptive Analysis**

The two variables used in this analysis are the unemployment rate in Kentucky and the total number of COVID-19 cases per 1000 in each county in Kentucky. The x variable, unemployment rate, is quantitative, because it gives us numerical values that can be measured. The y variable, total cases per 1000, is also quantitative because it gives us numerical values that can be measured.

The y variable seems to be skewed to the right, which means for the most part the cases are on the lower side. No outliers are present. This is shown on a histogram. (Refer to figure 1) The x variable, unemployment rate, is also skewed to the right. This shows that unemployment rates are fairly low in the distribution. No outliers are present. This can also be seen on a histogram. (Refer to figure 2)

Figure 1: Distribution of Total Cases Per 1000

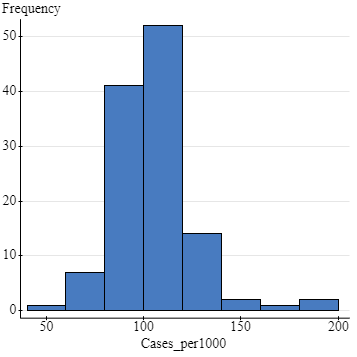
****

Figure 2: Distribution of Unemployment Rate

****

**Inferential Results:**

**Does the unemployment rate affect the total number of cases per 1000 for each county? (Taylor McKinney)**

Inferential Procedure:

I chose to use the simple linear regression analysis.

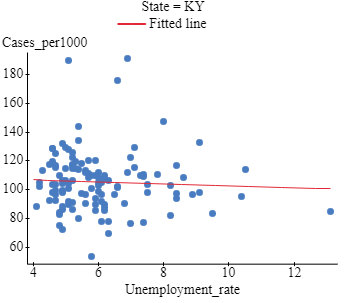
Hypotheses:

H0: β1 = 0 (The model is not useful for predicting y)

HA: β1 ≠ 0 (The model is useful for predicting y)

Assumptions:

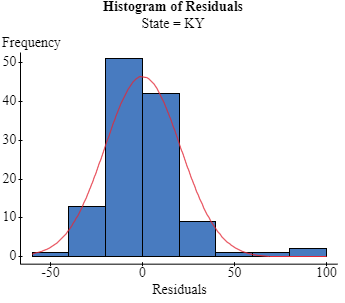
1. **Linear relationship between X and Y:**

This type of relationship can be shown with a scatter plot. I chose to use this method because it clearly shows the linear relationship between the two variables. This scatter plot shows that there is a weak negative correlation between X, the unemployment rate of the county, and Y, the total number of cases per county. 

1. **There is independence of observation and the residuals:**

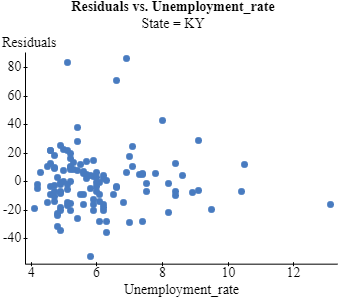
This assumption is assumed to have been met.

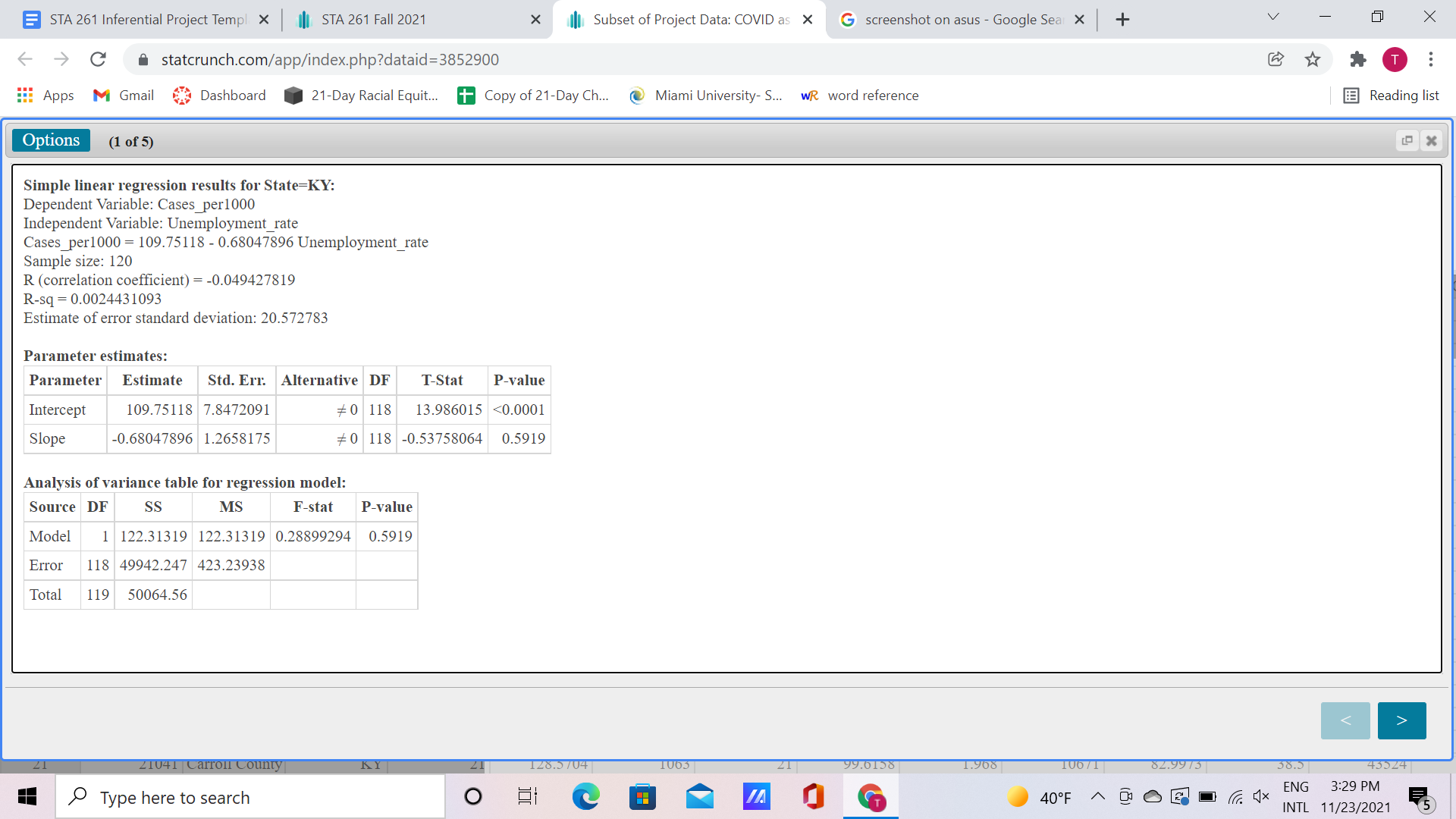
1. **Residuals are normally distributed:**

This can be modeled by either a QQ plot or a histogram to show that the distribution is approximately normal. This histogram shows that most of the residuals fall within the normal curve of the distribution. Because the residuals are normally distributed, the assumption is met. 

1. **Y values have equal variance:**

This residual scatter plot shows this assumption. This assumption is met when there is no megaphoning of the data. There is not any megaphoning in this plot, which means that the assumptions have been met.



Statcrunch Output:

Inferential Procedure Results/Conclusion:

All of the assumptions were successfully met in using a simple linear regression analysis. I chose to do this type of analysis because it is the easiest way to see if one variable is a predictor of the other. That way I could test if my x-value, unemployment rate per county in Kentucky, was an accurate predictor for my y-value, total number of cases per one-thousand per county. I chose to compare these two variables because I know that unemployment went up last year, so I wanted to see if COVID-19 cases were the cause of this. A simple linear regression analysis seemed to be the only direct way to conduct this test.

I decided to test for the unemployment rate compared to COVID-19 cases by 1000 across counties in Kentucky. When conducting a test for this data, I used a significance level of .05. This data set has a p-value of .5919, which is greater than the significance level of .05. This high p-value indicates that we failed to reject the null hypothesis. Therefore, there is sufficient evidence to conclude that the unemployment rate is not a good indicator of the number of cases per county per one-thousand in Kentucky.

**Exploratory/Descriptive Analysis:**

The variables used in this simple linear regression analysis are → **X (independent variable): number of cases per 1000** and **Y (dependent variable): percentage of college educated individuals**.

Both variables are quantitative variables, the X variable giving a number of cases per 1000 and the Y variable giving a percentage of college educated individuals. Figure 1 shows the distribution of the X variable, the number of cases per 1000, using a histogram graph. The data is skewed slightly right, meaning that number of cases per 1000 tends to stay on the lower side. Figure 2 shows the distribution of the Y variable, the percentage of college educated individuals, on a histogram as well. This data is greatly skewed to the right, meaning that many counties have a lower percentage of college educated individuals.

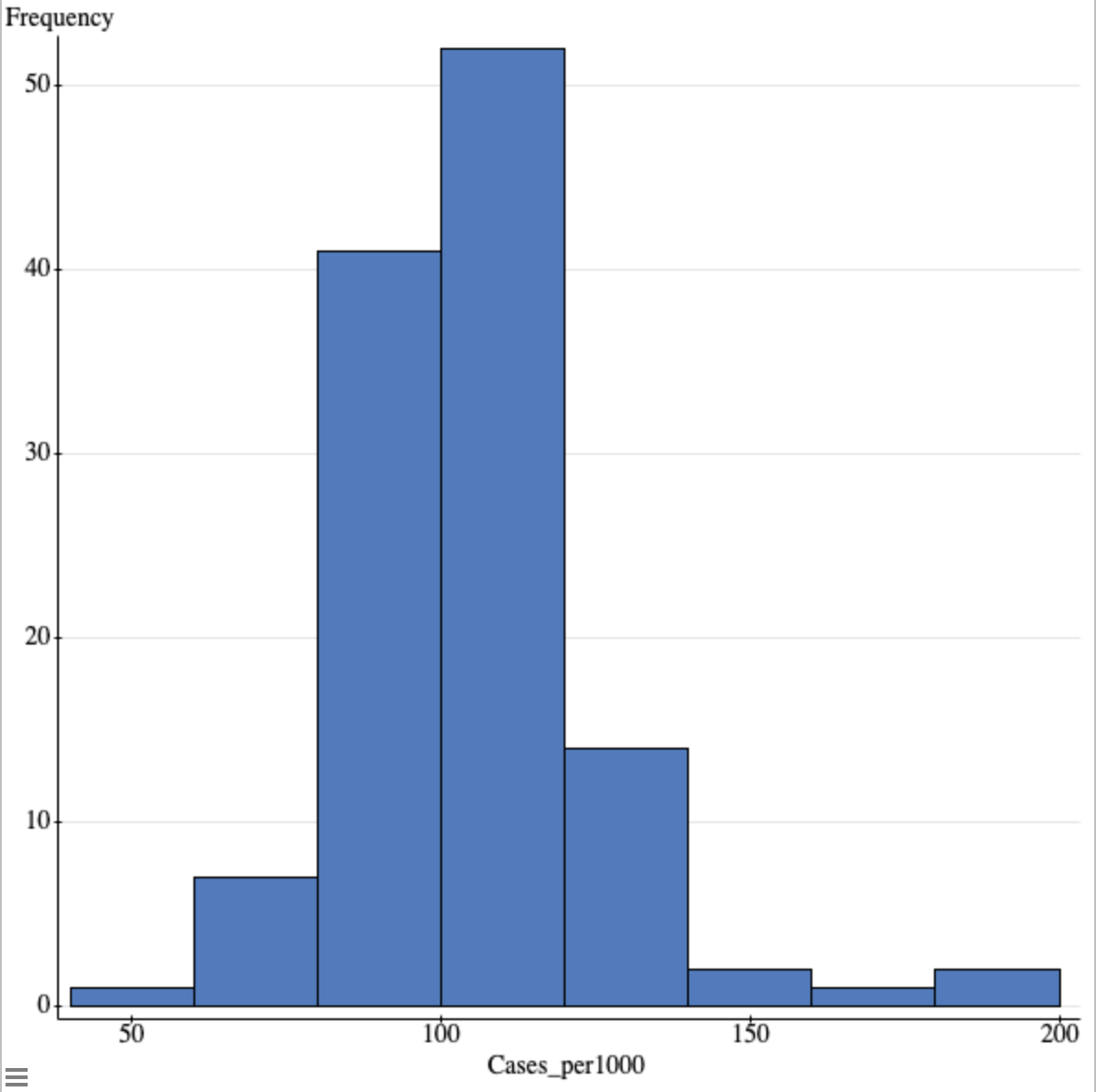
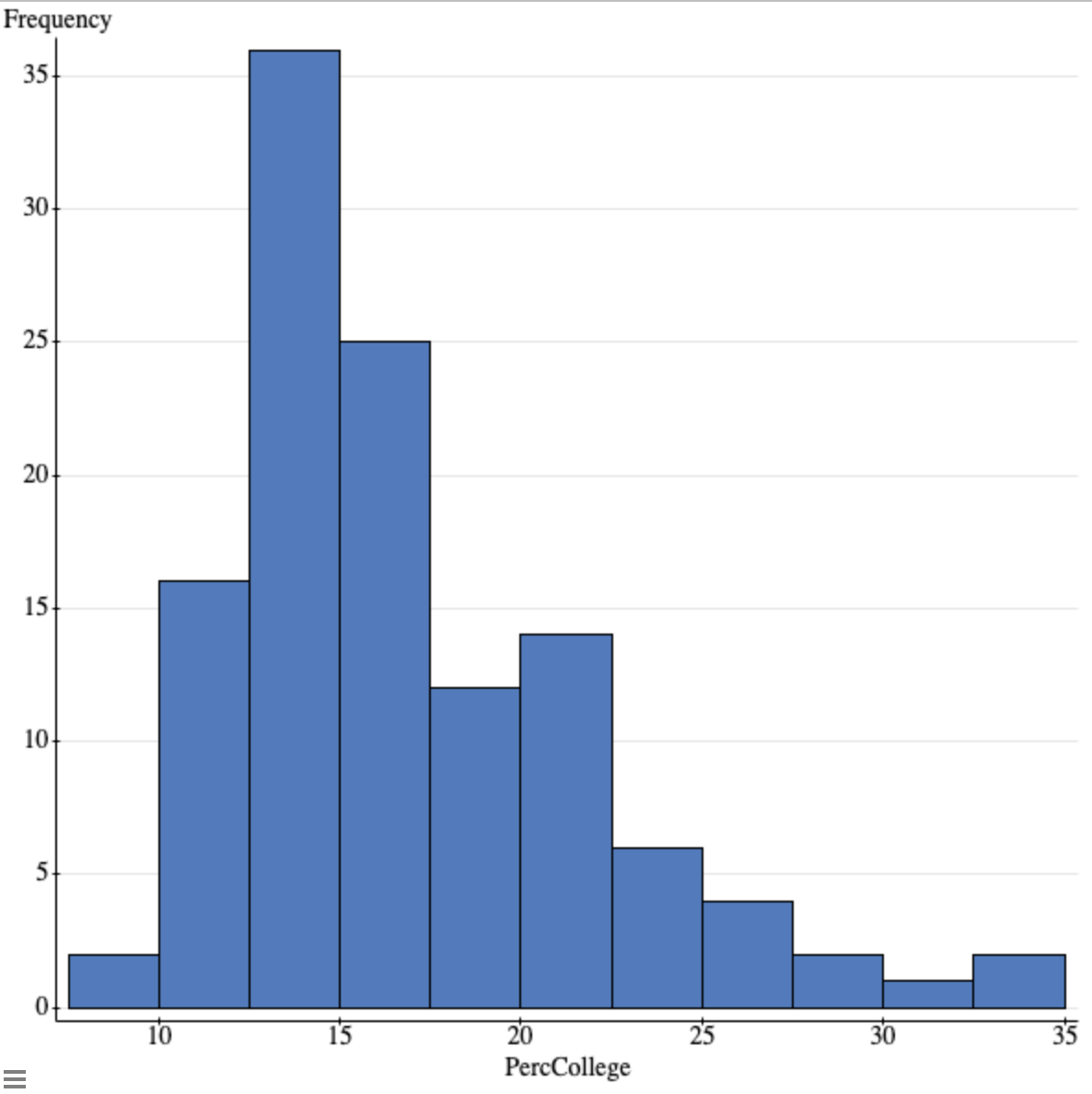
Figure 1: Distribution of Number of Cases Per 1000

Figure 2: Distribution of Percentage of College Educated Individuals



**Inferential Results:**

**Does the percentage of college educated individuals living in a county affect the total number of cases per 1000 in that county? (Allie Kleber)**

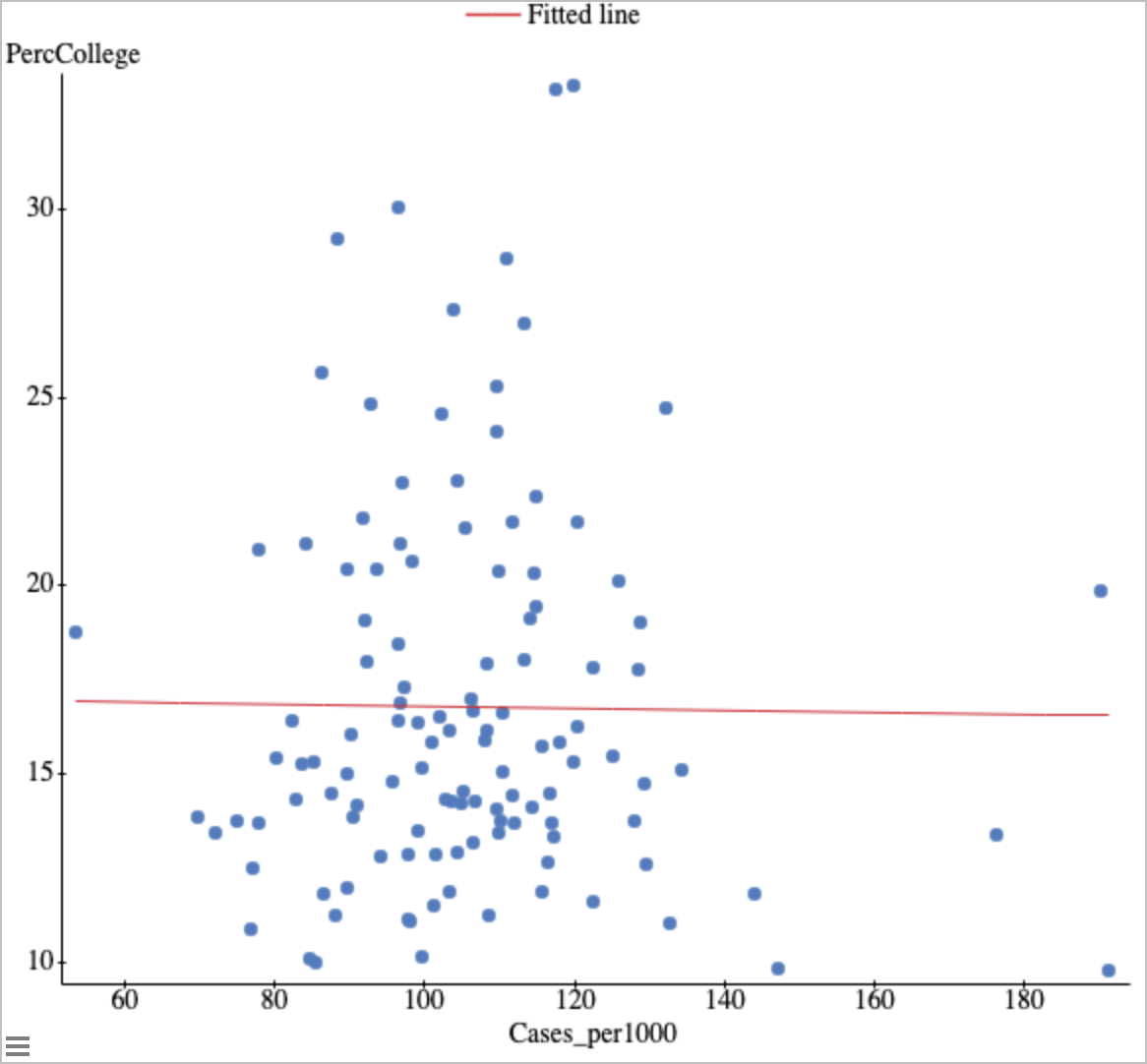
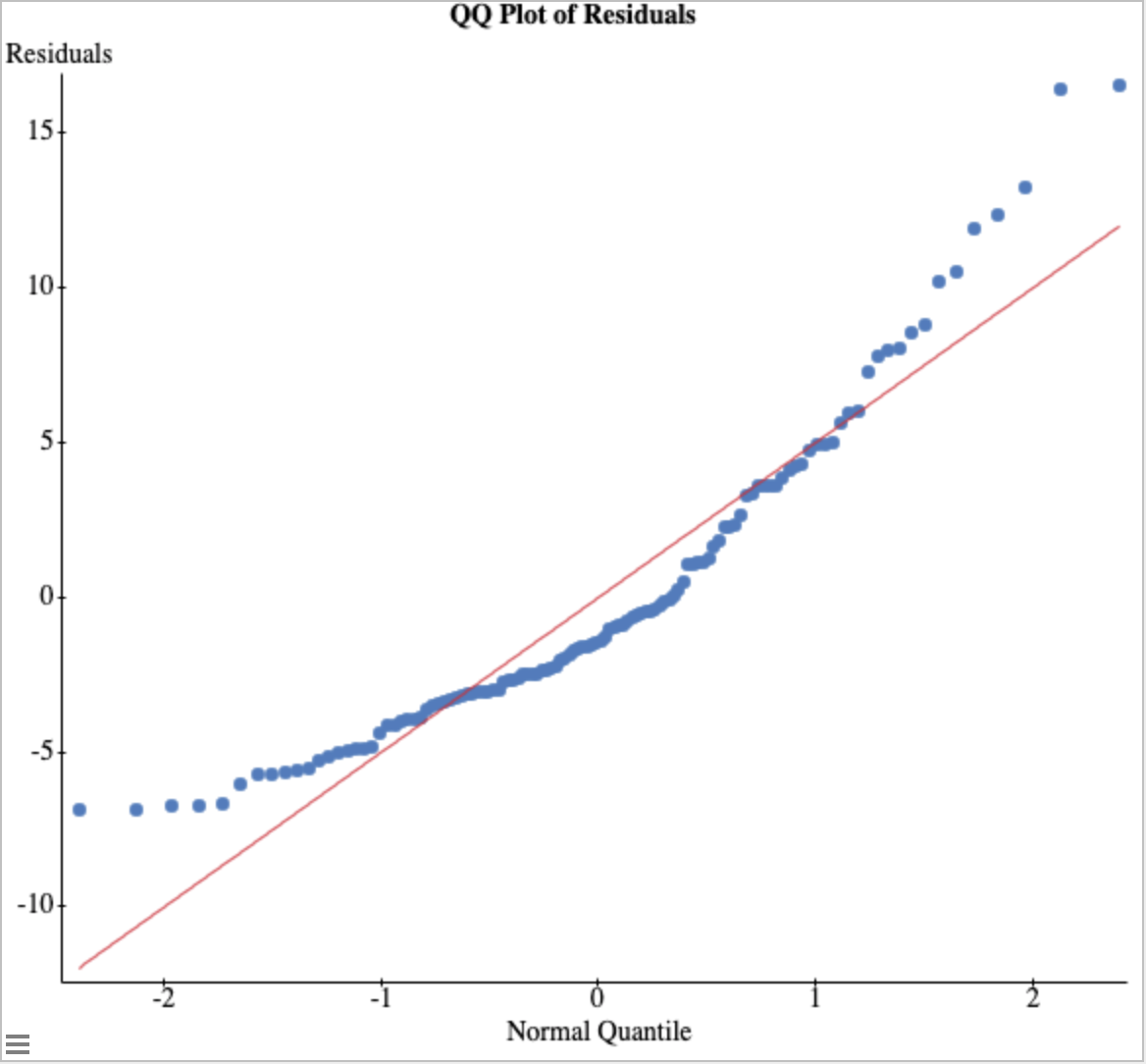
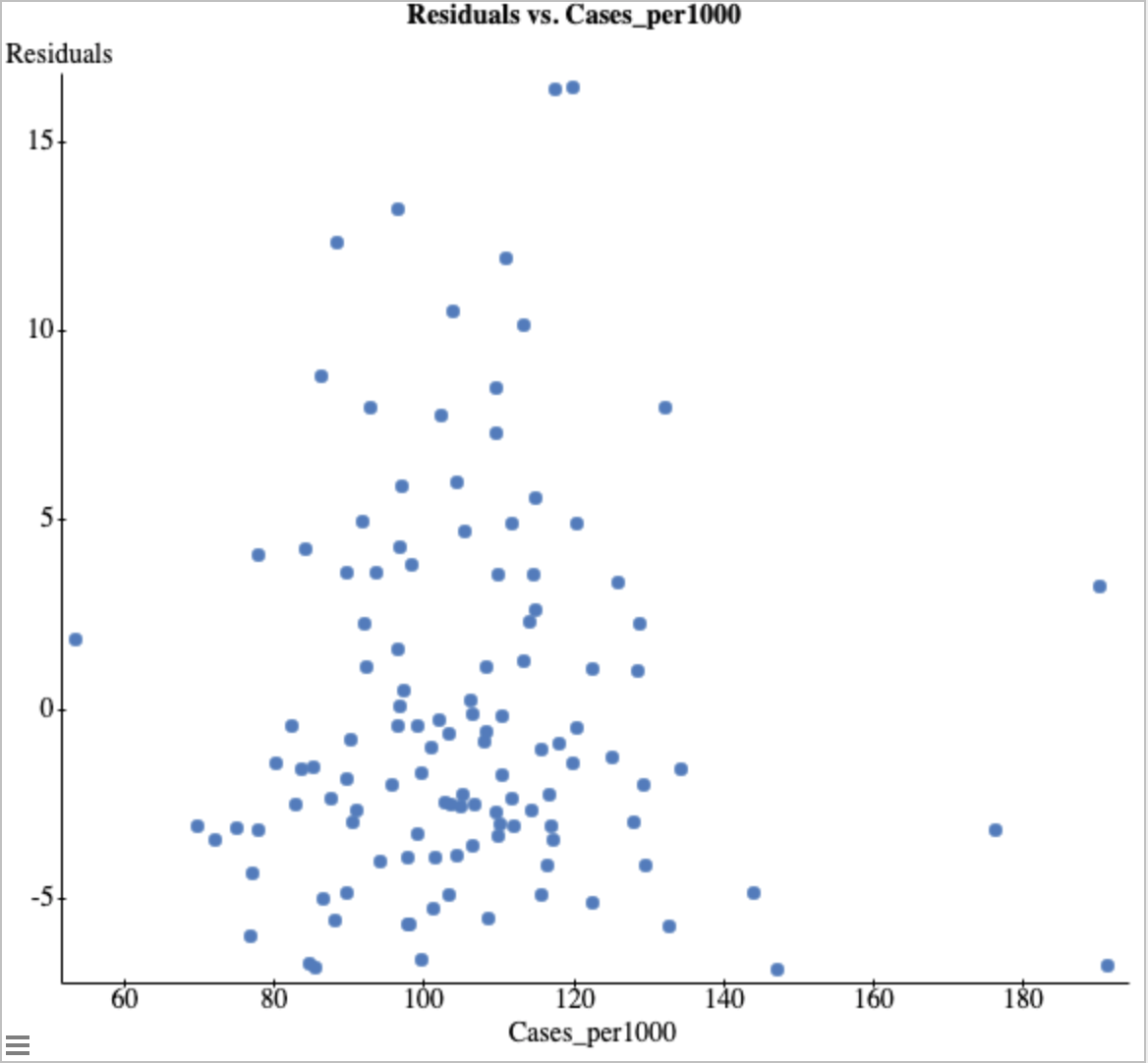
Inferential Procedure: Simple Linear Regression Analysis Test

Hypotheses:

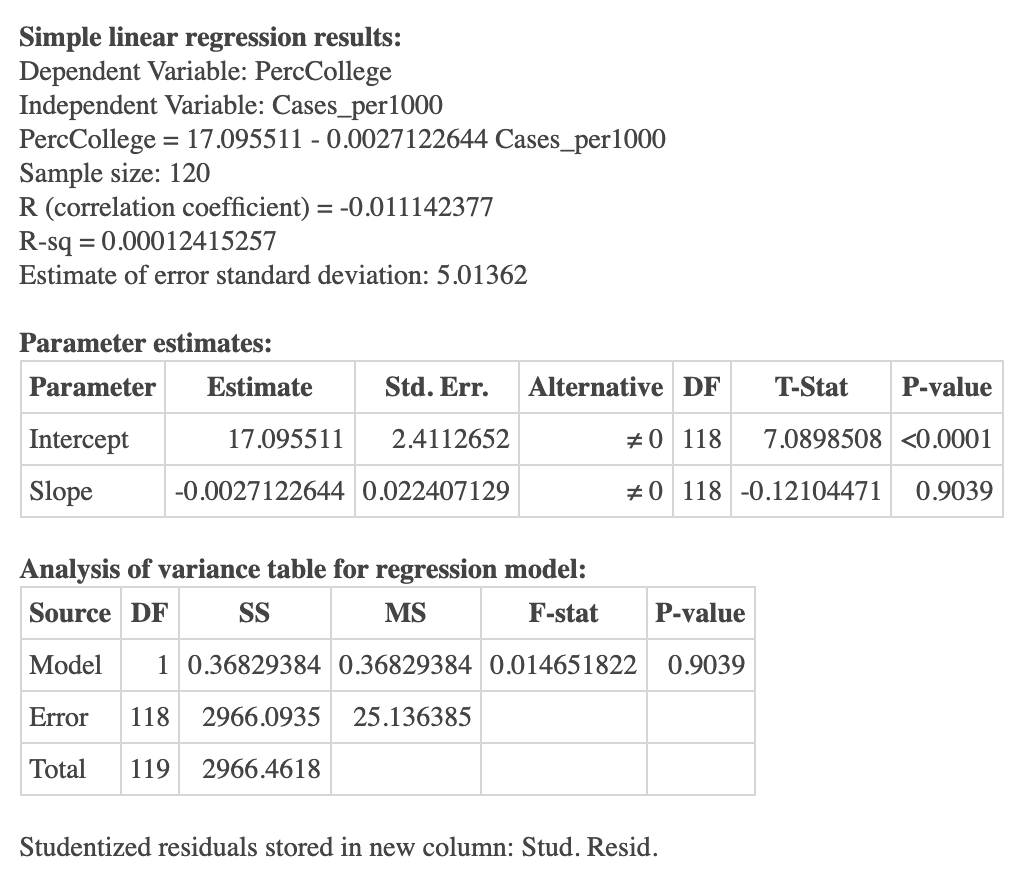
H0: β1 = 0 or the model is not useful in predicting y (number of cases per 1000)

HA: β1 ≠ 0 or the model is useful in predicting y (number of cases per 1000)

Assumptions:

1. **Linear relationship between X and Y:** This assumption can be modeled by a scatterplot with a line of best fit. The line of best fit here indicates that the X (percent college educated) and Y (cases per 1000) variables have a weak negative linear correlation, therefore this assumption is met.
2. **There is Independence of Observations and the Residuals:** This assumption can reasonably be assumed in this course.
3. **Residuals are normally distributed:** This assumption can be modeled by a QQ plot where all plotted data points are approximately linear. This QQ plot is relatively linear, meaning the assumption is met.
4. **Y Values have equal variance:** This assumption can be modeled by a residual plot that shows no megaphoning of data points. There is no megaphoning shown on this residual plot, meaning the assumption is met.

StatCrunch Output:



Inferential Procedure Results/Conclusion:

I chose a simple linear regression test for this data because this type of inferential test allows me to compare the variables directly. I can look at the relationship between them in terms of correlation and find out if the percent of college educated people (Y variable) could predict the number of cases per 1000 (X variable) in each county. All of the assumptions for this test were met, as shown above, so I could proceed with the test.

The p-value given for this data is 0.9039. This is greater than the significance level of 5% or 0.05 used in this test, meaning that we fail to reject the null hypothesis. There is sufficient evidence to conclude that percent of college educated individuals in a county is not a useful predictor of the total number of cases per 1000 in that county. Therefore the answer to the question “Does the percentage of college educated individuals living in a county affect the total number of cases per 1000 in that county?” can be answered with the fact that the percentage of college educated individuals cannot predict the number of cases per 1000 in that county.

**Exploratory/Descriptive Analysis:**

The variables used for this linear regression analysis were percent of GOP in counties in Kentucky and number of cases per 1000 in counties in Kentucky. The x variable is the percent of GOP in counties in Kentucky and it is quantitative because it gives numerical values. The y variable is the number of cases per 1000 and it is quantitative because it also gives numerical values. As seen in Figure 1, the percentage of GOP is heavily left skewed. As seen in Figure 2, the cases per 1000 are right skewed.

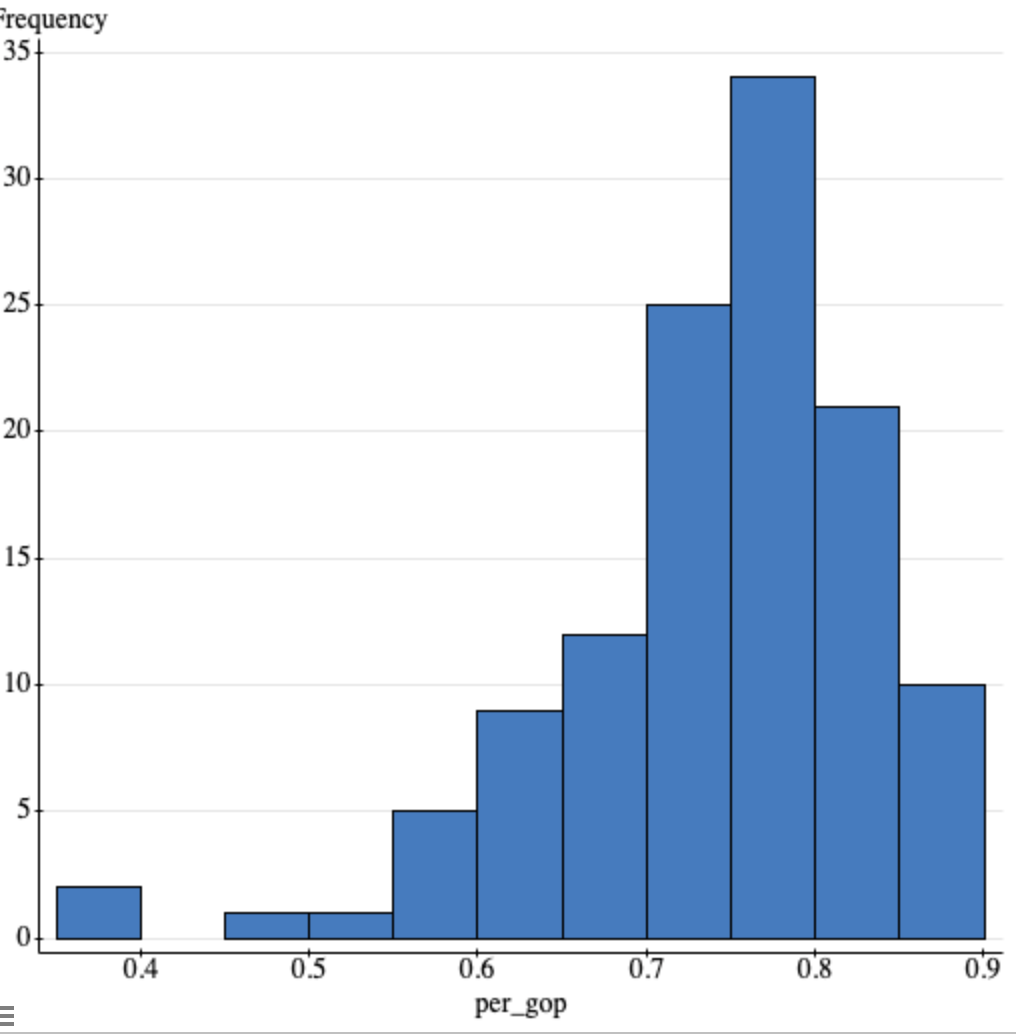
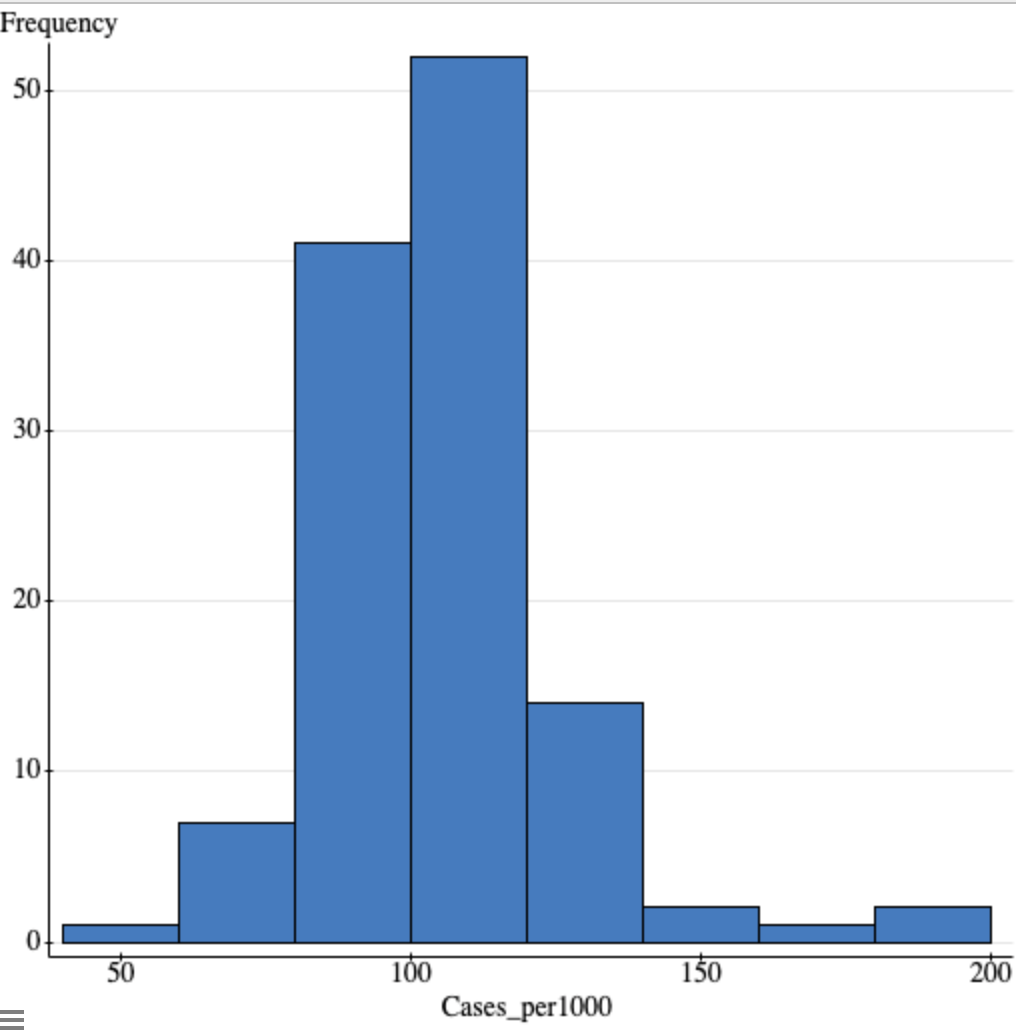
Figure 1: Distribution of Percent GOP

Figure 2: Distribution of Cases per 1000

**Inferential Results:**

**Does the percentage of GOP identifying individuals in a county affect the number of cases per 1000 it experiences? (Dhruv Iyengar)**

Inferential Procedure: Simple Linear Regression Analysis Test

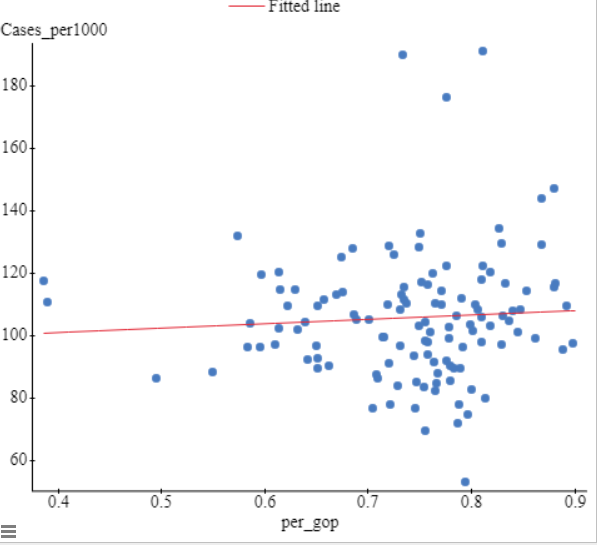
Hypothesis:

H0: β1 = 0 the percent of GOP is not useful in predicting the number of cases.

HA: β1 ≠ 0 the percent of GOP is useful in predicting the number of cases

Assumptions:

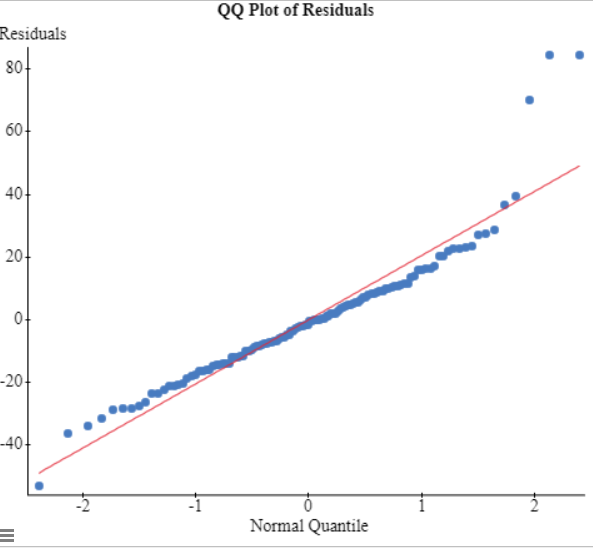
1. **Linear Relationship Between X and Y:**

The linear relationship between the amount of mask usage and the number of cases can be demonstrated through the scatterplot below. 

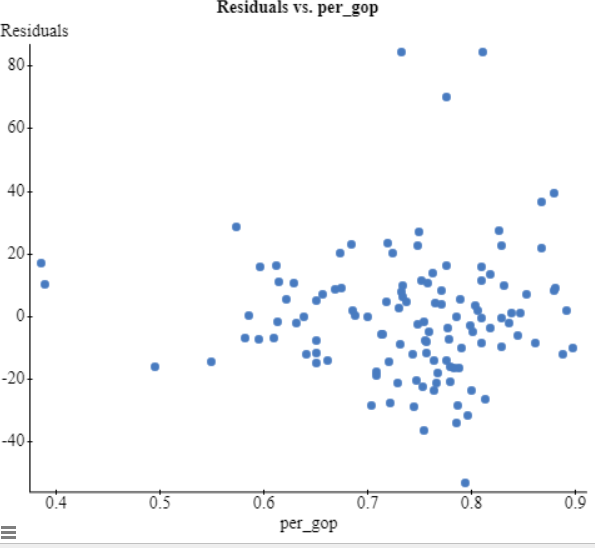
1. **Independence Assumption:**

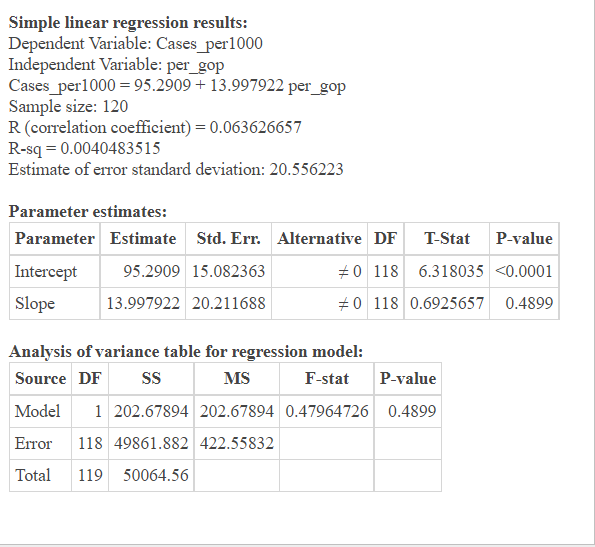
This assumption can be assumed to have been met.

1. **Normality Assumption:**

This assumption can be demonstrated through a QQ plot that is roughly linear. Since the QQ plot is linear, this can be interpreted as the normality assumption being met. 

1. **Equal Variance Assumption:**

This can be useful in modelling the variance of the variables. This residual scatter plot shows this assumption. This assumption is met when there is no megaphoning of the data. Since there is no megaphoning this assumption has been met. 

Statcrunch Output:

Inferential Procedure Results/Conclusion:

I chose to do a linear regression analysis because in this scenario it was the best way of comparing the two variables. I wanted to explore whether the percentage of people who identify as members of the GOP in a county is a useful predictor of cases per 1000 in that county. A linear regression model is the simplest way to show a relationship between the percentage of GOP and number of cases per 100.

The regression equation for the data is as follows: Cases per 1000 = 95.2909+13.997922(GOP) The correlation coefficient is 0.0634 and the coefficient of determination is 0.004. This means that there is a very weak linear relationship between the data. The P-Value for the slope is less than 0.4899 and since that is greater than a significance level of 0.05 we can determine that we cannot reject the null hypothesis. This means that percent GOP per county is not a useful predictor of the cases per 1000 in the counties of Kentucky.

**Conclusion**

Wrap up the paper. Summarize the answers to each of your research questions. If you were unable to answer your question(s) indicate why that was the case.

After choosing to examine the to see if there was a correlation between high mask usage and cases per 1000, I was interested to see how much of an effect, and what type of an effect, mask usage had on case numbers. I decided to work with the data only from counties in Kentucky, since there was such a broad database, and states could differ from each other in mask policies. After having examined the data, I came to the conclusion that I was not able to reject the null hypothesis that high mask usage is not a good predictor of COVID-109 cases per 1000 (meaning there is no linear relationship between the two variables). This means that by measuring the high mask usage in Kentucky, it is not a good indicator and predictor of the amount of COVID cases per 1000 in the state(Anna Herrmann).

I chose to analyze the relationship between the percentage of people who identified as GOP in counties in Kentucky with the cases per 1000 in those counties because through the pandemic I had noticed that people who tend to identify as right wing or GOP party members typically are dismissive of the effects of masks and coronavirus so I wanted to see if there was a statistically significant pattern in this. To find this I chose to use a linear regression analysis to analyze the data and set out with my hypotheses: my null hypothesis was that there would be no relationship between the variables, and my alternative hypothesis was that there would be a linear relationship between the variables. After this test, I came to the conclusion that I could not reject the null hypothesis, and that there is likely not a linear relationship between these two variables. (Dhruv Iyengar).

I chose to analyze the relation ship between income levels in a county and the total number of cases of COVID-19 per 1000 people. I chose to do a simple linear regression analysis to answer this question because I wanted to see if there was a causal relationship between income levels and cases per 1000 and this can be best accomplished with the aforementioned test. After conducting my test, I came to the conclusion that income levels could not predict the total cases per 1000 in a county because the P-value I ended up getting was greater than the significance level that I had chosen to use and thus, I had to conclude that there was no linear relationship that could state that income levels could confidently predict the number of cases per 1000 of COVID-19 that the counties of Kentucky experienced.

The biggest reason why I was interested in looking at the relationship between the income level and number of COVID-19 cases per 1000 was due to the fact that I have always been curious about the varied impacts that a person’s income has on their life. As someone who has family that lives in a developing country where the average person makes far less money than anyone in the USA, I have also been able to see the impact that income inequality has on the way that COVID-19 has been handled throughout the world. I really wanted to see if the richer areas of Kentucky, in this case, had better results in terms of COVID-19 cases because I intuitively thought that would be the case, seeing as a richer area would have better healthcare infrastructure, better doctors, and better resources to deal with and minimize the occurrence of COVID-19 cases when compared to poorer areas of the state. Even though that was not shown to be the case with the data that I was given for this project which says that income levels do not affect cases per 1000, I would be interested in getting a larger dataset to compare and see if differences that were hidden in the data I was provided might become more obvious because I really do think that there is a relationship (maybe not linear) between these two variables even though that was shown not to be the case with the data I had. (Vimal Vinod).

I chose to analyze the relationship between the percentage of college educated individuals in a county and the number of cases per 1000 in a county. I used a simple linear regression test to analyze the data because it allowed me to see if the percentage of college educated individuals was a good predictor of the cases per 1000. Using a significance level of 0.05, I concluded that the p-value of the data was greater than that so I failed to reject the null hypothesis, and so there is not sufficient evidence to conclude that the percentage of college educated people in a county is a good predictor of the number of cases per 1000 in that county Thus, I was able to answer that the percentage of college educated individuals in a county does not have any appreciable (linear) effect on the total number of cases per 1000 it experienced. The reason I was interested in looking at how these two variables relate is because I knew that mask usage was a good predictor of cases per 1000 and thinking about the relationship between college education and mask usage to then college education and cases was of interest to me. (Allie Kleber)

I chose to analyze the relationship between the unemployment rate and the total number of COVID-19 cases per 1000 in counties across Kentucky. I used a simple linear regression test in order to analyze my data and determine whether or not to accept the null hypothesis. I was attempting to determine if there was a relationship between the unemployment rate in counties in Kentucky and the total number of cases per 1000 in counties in Kentucky. I used a significance level of .05 for this test. I got a p-value of .5919, which is larger than my significance level of .05. I would fail to reject the null hypothesis in this case. This tells me that I can conclude there is no linear relationship between the number of cases per 1000 per county in Kentucky and the unemployment rate in the state of Kentucky.(Taylor McKinney)

Are you able to generalize your findings to a larger population? Why or why not?

I am not able to generalize my findings to a larger population because COVID-19 cases and states are independent from each other, meaning that the amount of COVID cases in one state does not directly affect COVID cases in another state. Additionally, different states could also have different mask policies, which would also affect the data. (Anna Herrmann)

The percentage of GOP was not found to be a useful predictor in the number of cases per 1000 in the counties of Kentucky and as a result I cannot generalize my findings to a larger population. (Dhruv Iyengar)

I am not able to generalize my findings to a larger population because Kentucky is not representative of the entirety of the USA and the many other states that are within it. For example, if I were to look at the state of California in this analysis, I might see a clear relationship where because income levels are on average higher than in Kentucky, we see an entirely different relationship between income levels and total cases per 1000 that is linear in nature. Furthermore, there could be other factors like a better public health system which exists in other states that would make generalizing the results I got in this particular analysis to a larger population like the entire USA, or the entire world if you tried to generalize these findings to their extreme, very hard to do (Vimal Vinod).

I am not able to generalize my findings to a larger population because both variables have far too much variation on a larger scale. Data for both of these variables changes drastically from state to state in the US, so data from the state of Kentucky is not representative of the rest of the American population, let alone a larger population like the whole world (Allie Kleber).

I am not able to generalize my findings to a larger population. This is because these two variables are only representative of Kentucky. The data for these two variables would be different for every state, and especially different if trying to generalize them to the entire country of the United States. There could very well be other variables in different states and areas that are not the same as Kentucky as far as unemployment rate and number of cases goes. In conclusion, it is not a good idea to generalize these findings to a larger population of any kind. (Taylor McKinney)

Did any new questions surface? What suggestions do you have for further research?

A new question that surfaced after my research is if there are any lurking variables that would also affect the relationship between mask usage and COVID cases. I also wonder what the percentage of people that still get COVID even when they still wear their masks very frequently. For further research, I would suggest making high mask usage an easier variable to measure, because there are some places you are required to wear a mask, and some places where it is not required to wear one, so specifying mask usage policies in specific places where they are required to, would make the data less broad, and easier to determine. By doing this, one could also examine the effectiveness of mask mandates in terms of their effect on mask usage and also COVID-19 cases too (Anna Herrmann).

A question that surfaced was the idea that there could have been any lurking variables between percentage GOP and number of cases per 1000 that could have interfered in the data. A suggestion for further research could be to compare this linear regression analysis with one that uses percent Democrats instead of percent GOP in order to see the differences between the two and whether or not that analysis shows a useful predictor for the number of cases per 1000 (Dhruv Iyengar).

A question that surfaced for me was whether there was any relationship between the income level and the level of mask usage you saw within that county. I only wonder this because although I was not able to establish a relationship between income and total cases per 1000, I still think there might be a way to connect income levels with a variable related to COVID-19 cases in a more circuitous manner. I also suggest that we do further research into the relationship between the amount of days a state spent in lockdown and the number of cases that state experienced. I would also be curious to see the relationship between the vaccination rate in a state/county and how that tracks with cases and deaths from COVID-19 (Vimal Vinod).

A question that came about for me when looking into this data and the correlations between it was what the relationship is between college education and vaccination rate. The cases per 1000 and percentage of college educated individuals does not have a linear relationship, but I wonder if the vaccination rate had any relationship to the college education rate and if it had any effect on the data and also the number of cases per 1000 (Allie Kleber).

One question that I came across while studying this data was whether or not the unemployment rate vs. total cases per 1000 would have different results in 2020 and 2021 separately. Obviously the job market is constantly changing, and I thought it would be interesting to see how the data would change between the two different years that Kentucky has been affected by COVID-19. The problem of lurking variables also posed a question to me, and I wondered if this data set had any that would have interfered with the data (Taylor McKinney).

General Conclusion:

In general, the questions that we asked in this project were all related by the fact that we wanted to see how different variables were affecting the case totals per 1000 in the state of Kentucky. These questions were of huge interest to us because of the fact that COVID-19 has been a constant part of all of our lives for the past two years. Even though this was the case, many of us were still not very well versed in what things were making the pandemic worse and what things were helping in slowing it down and this project allowed us to examine those factors/variables which have been such a large part of our lives for so long. The main takeaway from this project and the questions we asked and answered was that all of the factors that we examined from unemployment rate, percent of mask usage, percent GOP, income level, to percent college educated did not have a linear relationship with the cases per 1000 in Kentucky. This is not to say that there might not be some other non-linear relationship that exists between these variables and cases per 1000. Because we were narrowly looking at seeing if there was only a linear relationship between these variables, our whole group thinks that the next logical step if we were to continue this project would be to conduct further tests on our data to see if there might be some other relationship between these variables that is not obvious to us yet.

**References:**

Bureau, U. S. C. (2021, October 8). *U.S. Census Bureau Releases 2014-2018 ACS 5-Year Estimates*. Census.gov. Retrieved November 24, 2021, from https://www.census.gov/programs-surveys/acs/news/updates/2019.html.

Nytimes. (2020, July 28). *Covid-19-data/mask-use-by-county.csv at master· Nytimes/covid-19-DATA*. GitHub. Retrieved November 23, 2021, from https://github.com/nytimes/covid-19-data/blob/master/mask-use/mask-use-by-county.csv.

Tonmcg. (n.d.). *TONMCG/US\_COUNTY\_LEVEL\_ELECTION\_RESULTS\_08-20: United States general election presidential results by county from 2008 to 2020*. GitHub. Retrieved November 23, 2021, from https://github.com/tonmcg/US\_County\_Level\_Election\_Results\_08-20.

U.S. Bureau of Labor Statistics. (2021, November 3). *Unemployment rates for metropolitan areas*. U.S. Bureau of Labor Statistics. Retrieved November 23, 2021, from https://www.bls.gov/web/metro/laummtrk.htm.

*US covid-19 cases and deaths by State*. USAFacts.org. (2021, November 23). Retrieved November 23, 2021, from https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.