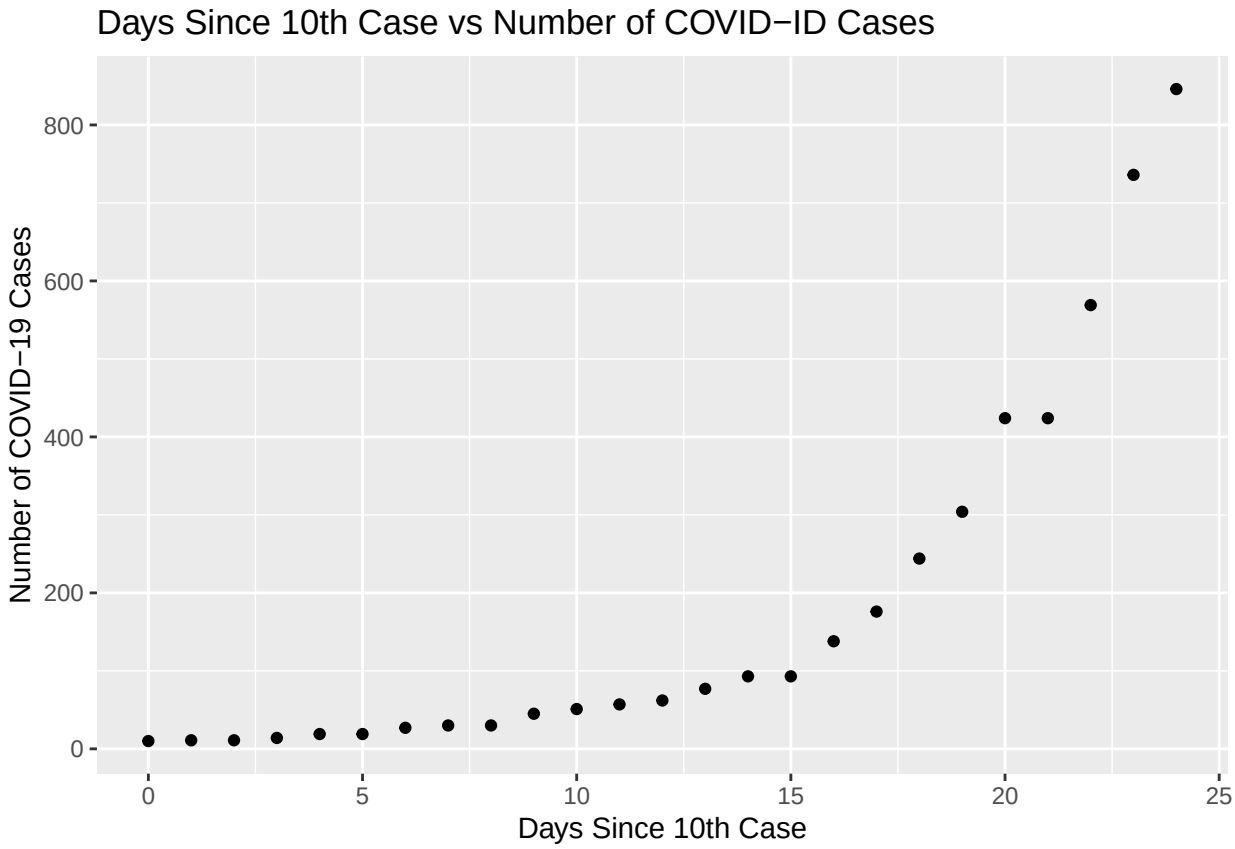
STA248 Final Exam Q1

Eric Zhu

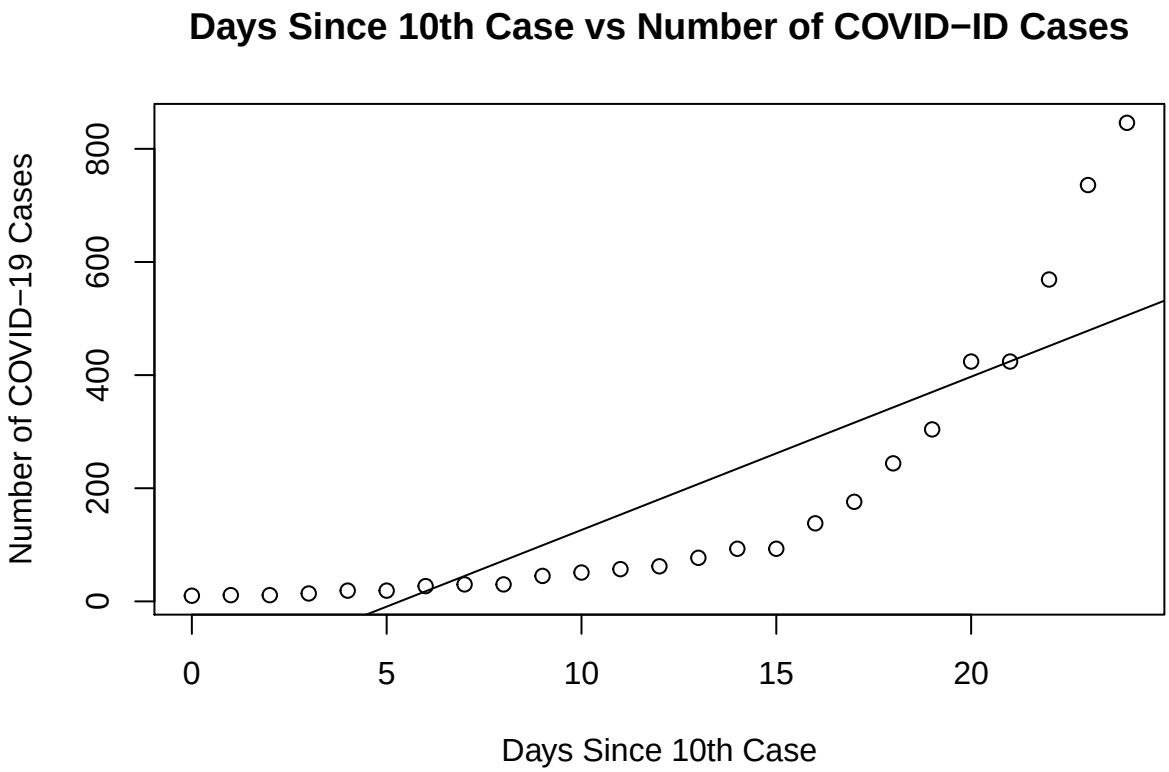
13/04/2020

Part a



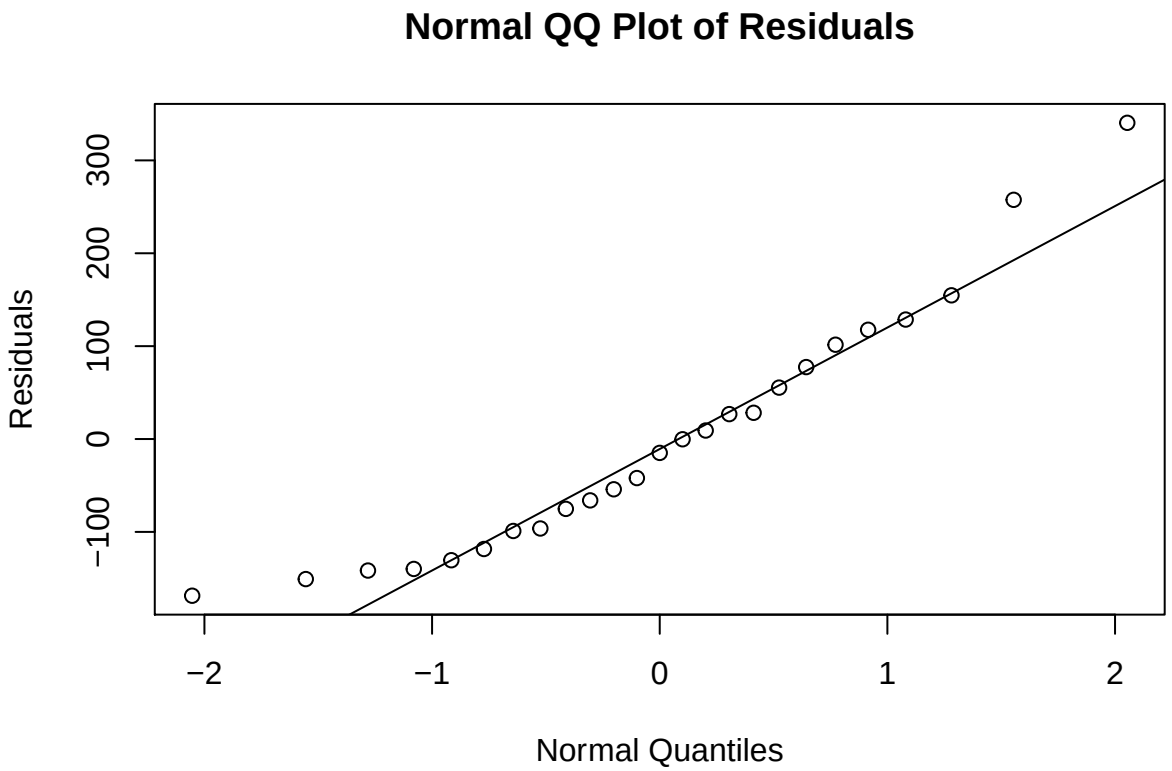
This relationship is not well represented using the SLR model because we see looking at the scatter plot that as we look at the relationship between days since the 10th case and cases between day 15 and day 25, we see a far steeper rate of growth in cases versus the growth of cases between days 0 to day 15.

Part b



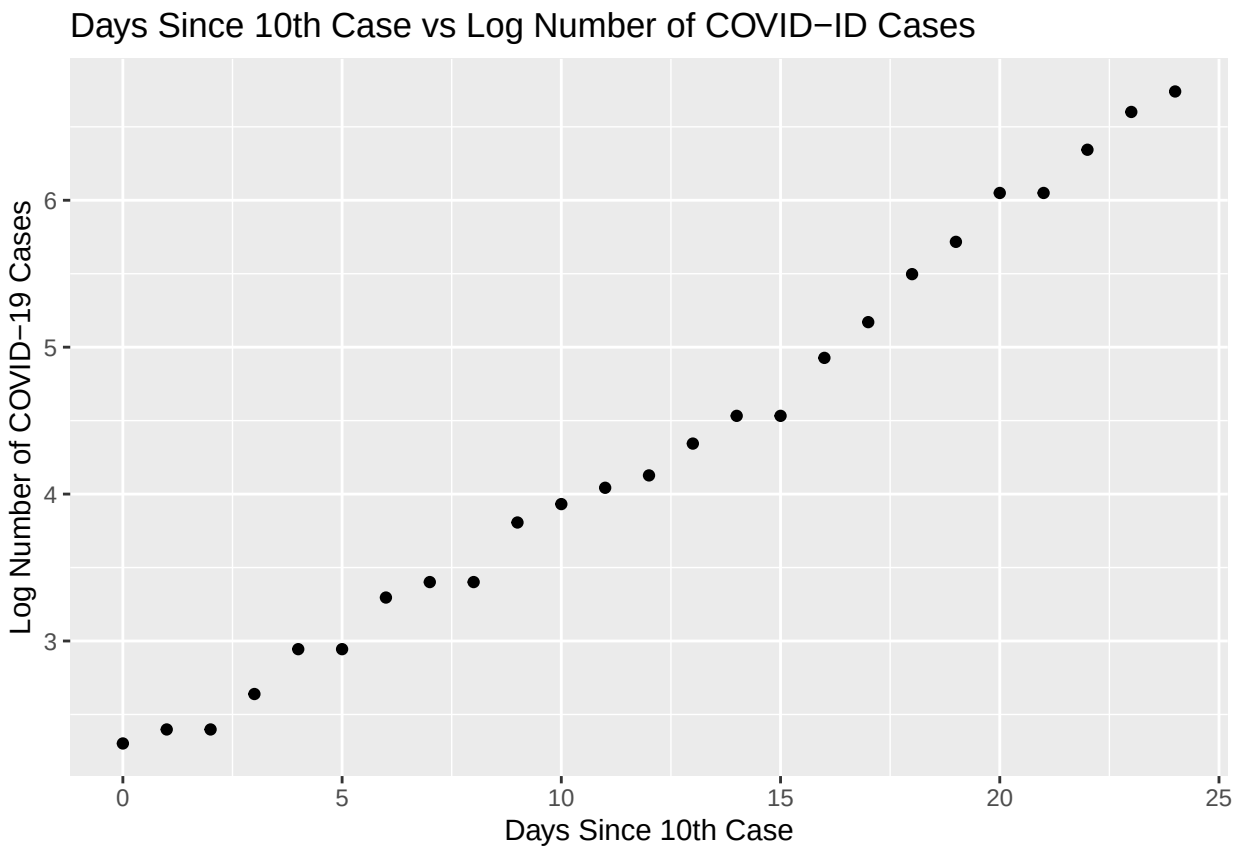
We see from fit of the model that it does not capture much of the relationship of the data because the fitted line is clearly far less steep than the relationship is from day 15 and onward, and far too steep for the relationship from day 0 to day 15. Additionally, almost every data point will have a large residual except for a couple points around day 20 and a couple points around day 6-7, which suggests that the fitted line does not capture the relationship of days since the 10th case and the number of COVID-19 cases.

Part c



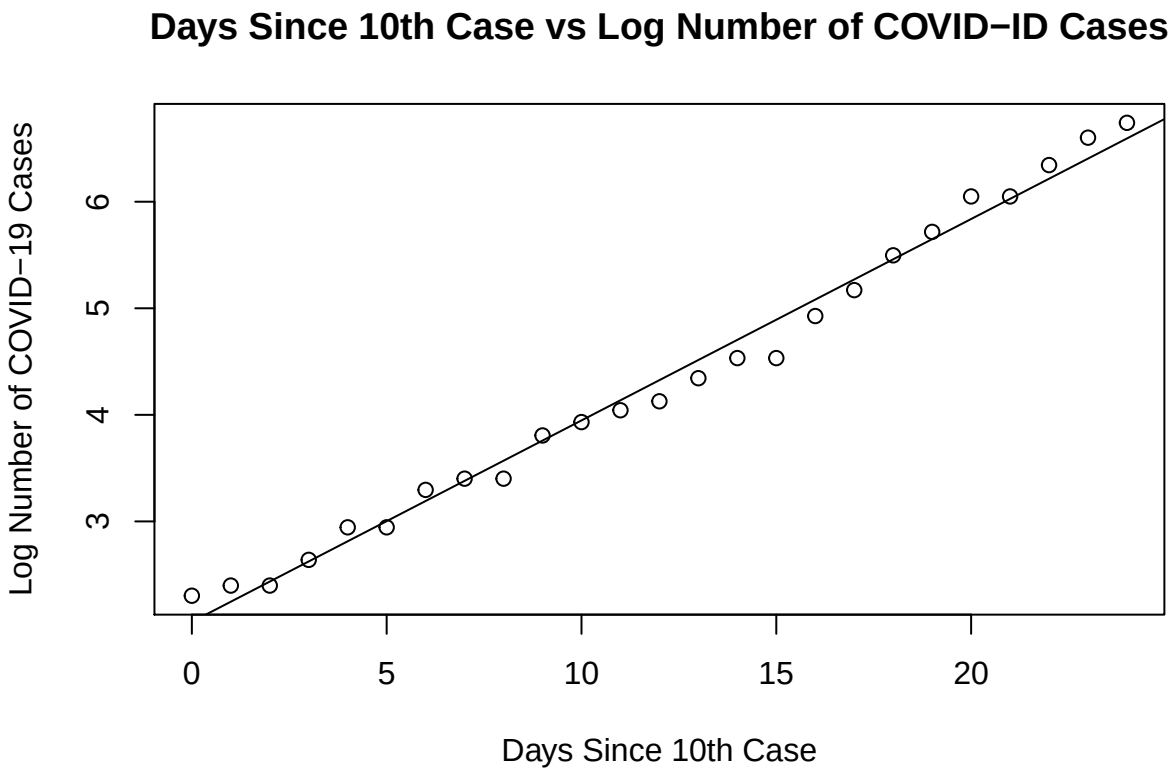
We see that the normal qq plot for residuals shows that we have cause for concern that the residuals do not follow the normal distribution because the points do not follow the line, i.e, points past -1 do not fall on the line and points past about 1.5 do not fall on the line.

Part d



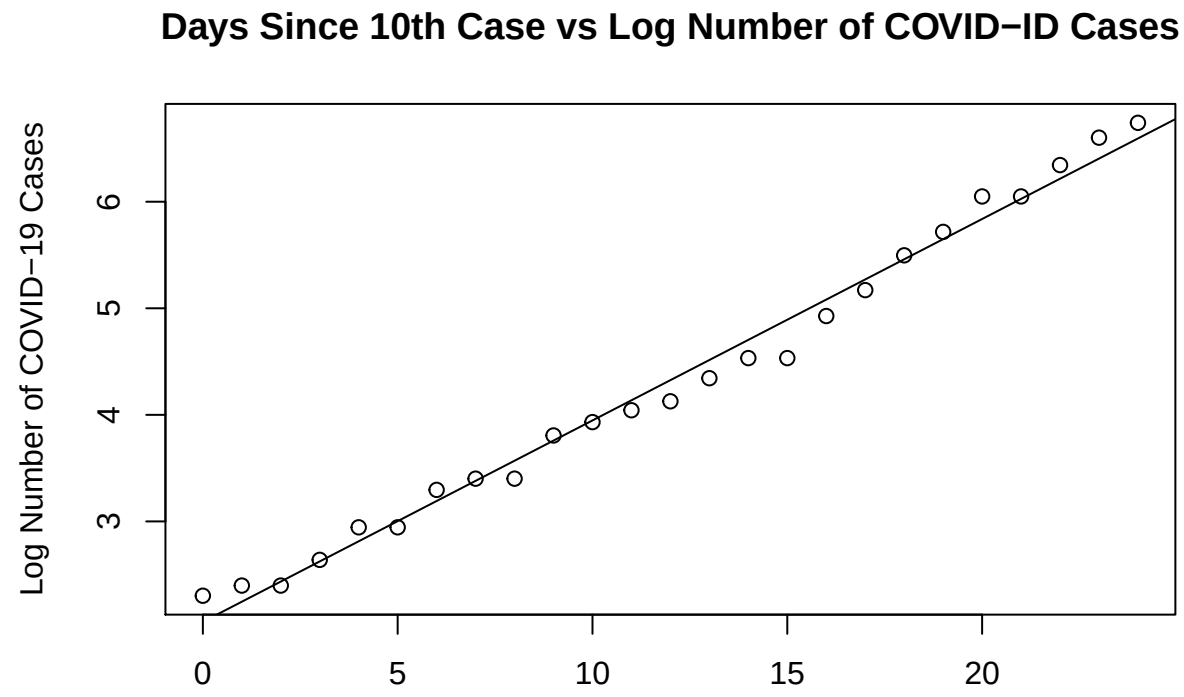
We see that this new scatterplot likely would be better represented by a SLR model because unlike the plot from part a, we see that the rate of growth of cases is more or less constant throughout the 25 days. Specfically unlike part a, we see that there isn't a drastic difference in rate of growth from day 0 to day 15 and from day 15 to day 25.

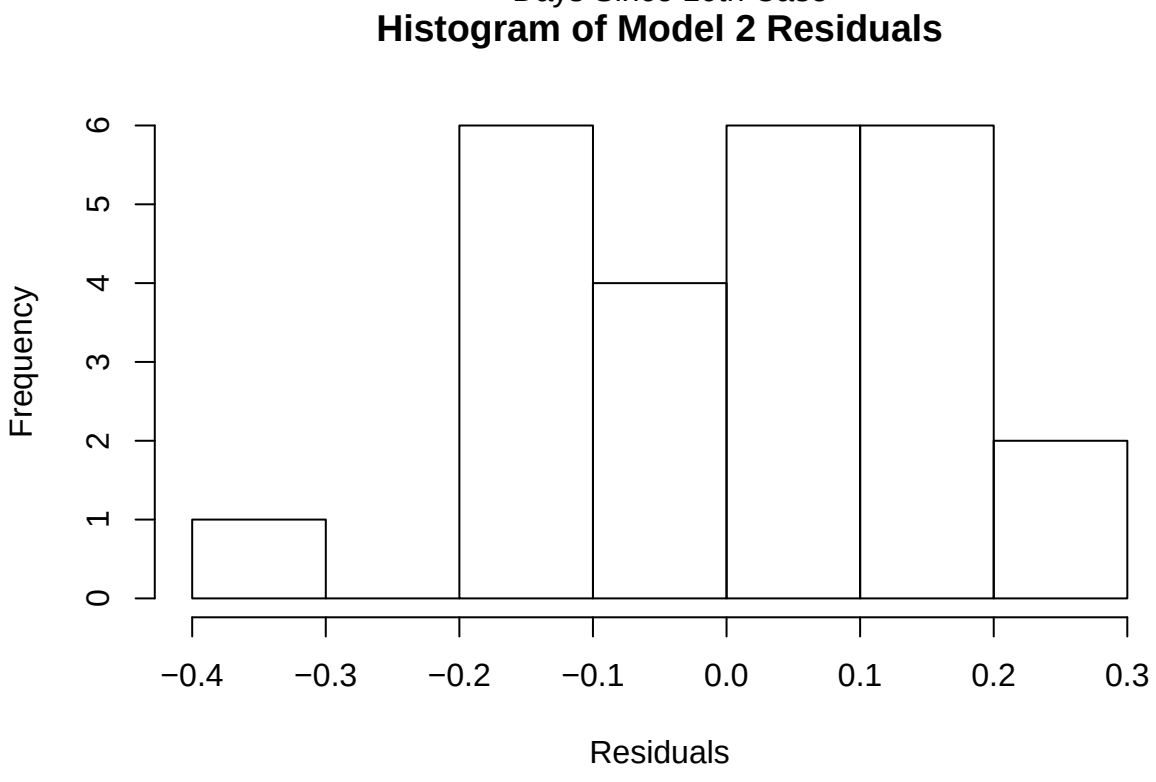
Part e

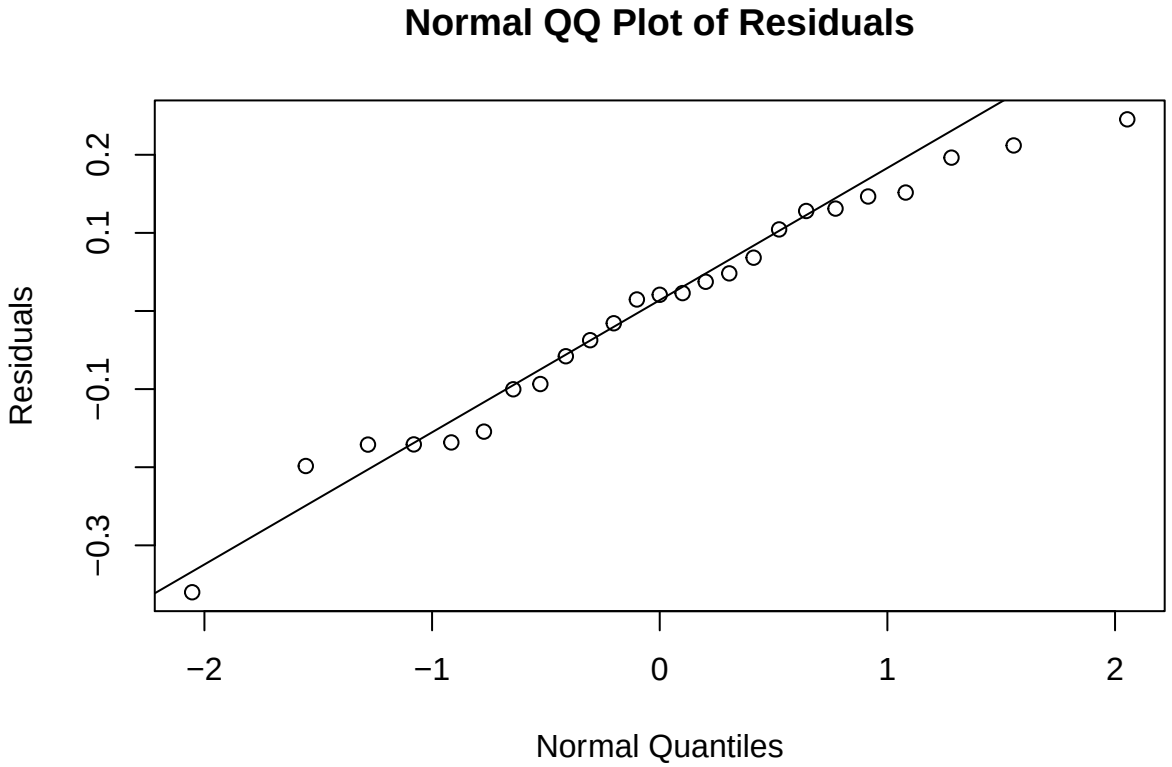


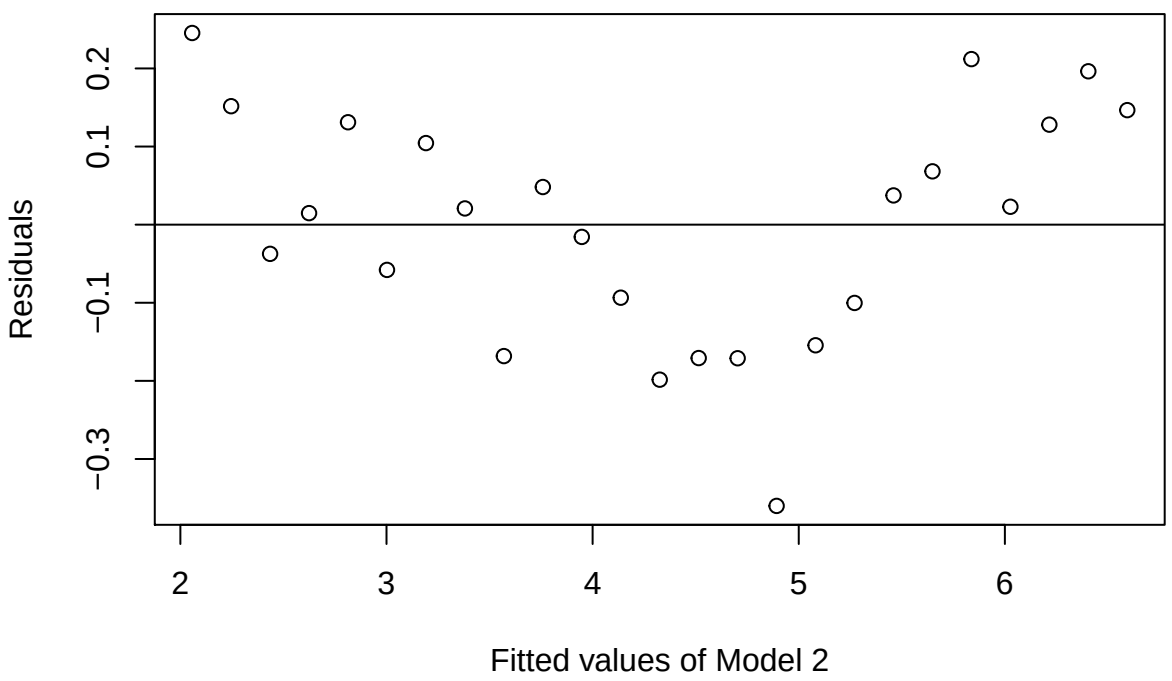
We see that the fitted line captures much of the relationship between days since the 10th case and the log number of COVID-19 cases because the fitted line goes through many of the data points. Further, regarding the points of which the line doesn't pass through, the residuals of those points are small, and suggests that the fitted line well captures the relationship between days since the 10th case and the log number of COIVD-19 cases.

Part f


Days Since 10th Case vs Log Number of COVID−ID Cases
Log Number of COVID−19 Cases
3
4
5
6
0
5
10
15
20
Days Since 10th Case


Histogram of Model 2 Residuals
Frequency
0
1
2
3
4
5
6
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Residuals





In this regression analysis, we will examine 4 plots:

1. The scatter plot with the fitted line over it
2. The histogram of the model residuals
3. The normal qq plot of the residuals
4. The plot of fitted values versus residuals

First, we see from the scatter plot with the fitted SLR model that the fitted line captures the relationship between days since the 10th case and the log number of COVID-19 cases. And that the residuals of the data points are quite small, i.e, they are close to the fitted line, which we can further examine in the histogram of the model residuals.

Second, the histogram of the model residuals is approximately normal despite being a little left skewed, centered around 0, which is what we expect to be the center of distribution of the error of the SLR model. Thus, the normality of the residuals is met.

Third, the data points of the normal qq plot of the residuals follows the line, further confirming that the residuals are likely following a normal distribution, and so normality of the residuals doesn't appear to be violated

Fourth, the plot of the fitted values versus the residuals has a slight parabolic pattern towards the extremes of the plot, which are slightly concerning regarding the independence of the residuals. However, the pattern isn't very noticeable, and there appears to be constant variance of the residuals, and further the residuals seem to be centered around 0. The weak parabolic trend could suggest that an SLR model doesn't perfectly capture the relationship between the days since the 10th case and the log number of cases of COVID-19.

Thus, through the analysis of the 4 plots, we see that a SLR model is appropriate for the transformed dataset.

Part g

```
##
## Call:
## lm(formula = log.y ~ days_since_10th_case, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36006 -0.10023  0.02077  0.12803  0.24536
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.057228   0.059803   34.40   <2e-16 ***
## days_since_10th_case 0.189029   0.004272   44.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.154 on 23 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.9879
## F-statistic:  1958 on 1 and 23 DF,  p-value: < 2.2e-16
```

We see from the summary that the coefficient B_1 (ie the coefficient of days since 10th case) is 0.189029, and so we see that for everyday that passes after the day which marks the 10th case, we will get a 0.189 increase in log units of the number of COVID-19 cases. We also see for the coefficient of days since the 10th case, that the p value is less than $2 \cdot 10^{-16}$, so it is very likely that there is an association between days since the 10th case and the number of cases. We also see from the summary that the coefficient of the intercept (i.e B_0) is 2.057228, and so we see that we expect to see using this estimated model 2.057228 log number of cases of COVID-19 on the day there is 10 cases of COVID-19.

Part h

On April 11th, 2020 the number of COVID-19 cases was recorded in Canada to be: 23,195. Using model 2, we find that we predict a $\hat{y}$ of 10.563533 log number of cases for 45 days (number of days since February 26th) after the 10th case was discovered. Thus, the number of cases predicted is $e^{10.563533} = 38697.605356098$ cases. Currently there is 23,195 cases and so current measures are more effective than predicted. The prediction is likely not very valid because the SLR model is too simple, i.e., it won't consider consider other factors such as testing, treatment, closed borders, etc. . . , so the model will likely not accurately predict the number of COVID-19 cases.

Part i

There are two cases of when a log transformed model would be useful:

1. When the relationship is best represented by an exponential function, i.e, suppose the relationship's function is $f(x) = e^x$.
2. When the relationship is best represented by a polynomial function, i.e, suppose the relationship's function is $f(x) = x^n$, where $n \in \mathbb{R}$

We see in the first case that taking the log of both sizes would get rid of the exponential, which is what we explored in this question, since we could see that the data was likely exponential. Many rates of growth in populations are exponential, especially when it comes to pandemics, where we see after some value of the independent value the value of the response variable grows extremely fast.

We see in the second case that we can similarly log transform it because if we the log of both sizes we get $\log f(x) = n \cdot \log x$, which is now linear, and the logic is similar to the first case. For example, if n were 2 (the relationship is quadratic), we'd see that we could likely transform the positive side similarly to an exponential function to a linear relationship. Additionally, If n were smaller than 1 we would also be able to linearize it since it's basically just "turning" n into a constant.