TS3

Dylan Hayashi

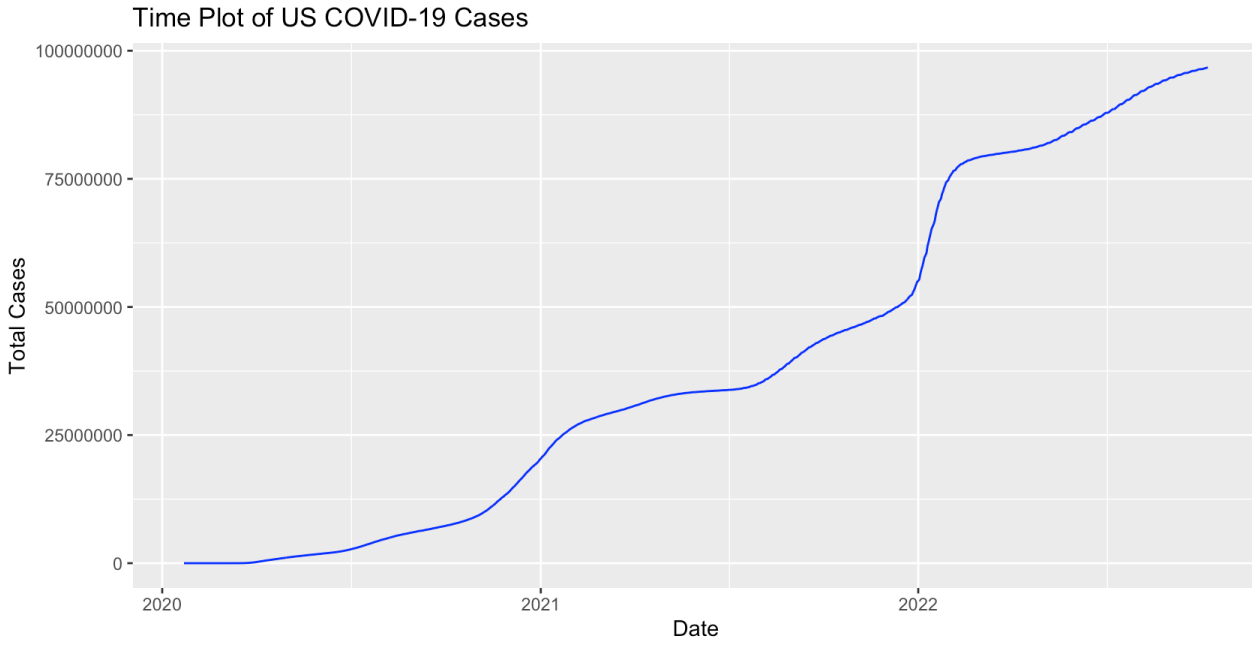
MSDS 413, Fall 2022, Section 55

Northwestern University, Time-Series Analysis & Forecasting

October 10, 2022

1. EDA

The filtered US COVID-19 case dataset contains two variables: date and total cases. They are plotted against one another below.



In order to be classified as time-series data, the Covid dataset must have three features: it must be indexed by time, with constant intervals, and the variable of interest must be stochastic.

To confirm the first condition, we can test the null hypothesis that the dataset contains identical counts of dates and total cases values, and that each date is unique. Thus, our null hypothesis is H10 : xit, $i \in \{1, 2\} t \in \{1, 2, .., n\}$, and our alternative hypothesis is H1a: not H10. Taking the length of the date, unique(date), and total cases vectors returns 991 in all cases. This means that there the date index has entirely unique values, and the total cases variable has a value for each date. Thus, we accept our null hypothesis.

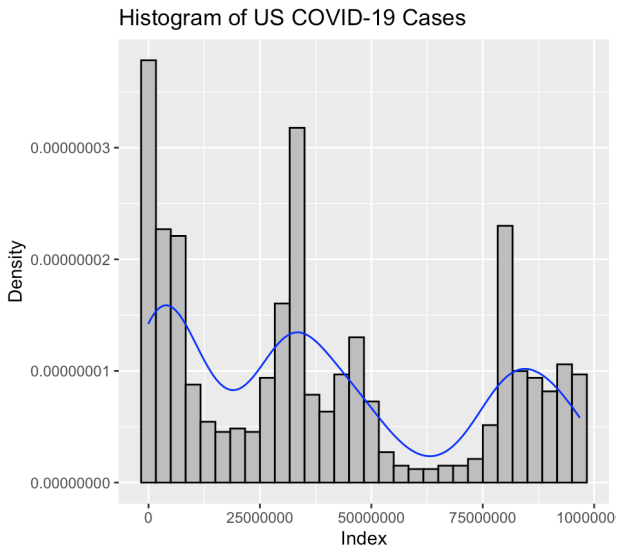
To confirm the second condition, we can check that the time interval between all dates stays constant. Our null hypothesis is that H20 : $(t + 1) - t = c, t \in 1,2,..,n$, and our alternative hypothesis is H2a: not H20. To test this, we can take the difference between each date and sum them. This should be equal to one less than the length of the data, as you cannot take the difference between the last and another observation. This calculation returned a value of 990, which is one less than the length of the data, 991 (as shown in the data summary to the right.) Thus, we accept our null hypothesis.

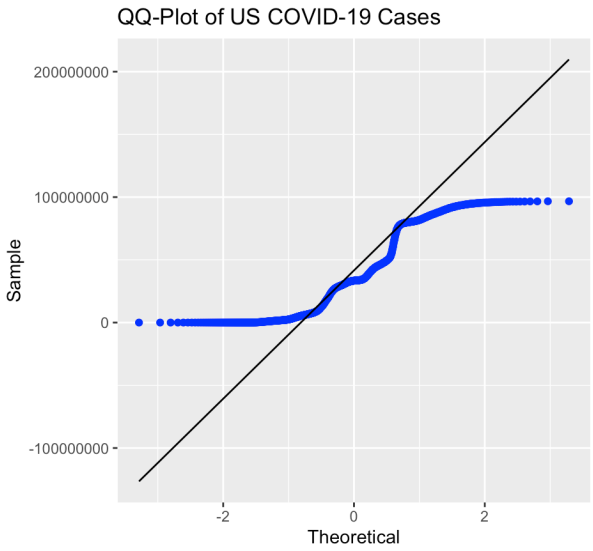
	X..X.total_cases
nobs	991.000000
NAs	0.000000
Minimum	1.000000
Maximum	96694214.000000
1. Quartile	7073055.000000
3. Quartile	76059440.500000
Mean	38930025.738648
Median	33389410.000000
Sum	38579655507.000000
SE Mean	1021193.281488
LCL Mean	36926073.724528
UCL Mean	40933977.752768
Variance	1033450196692654.625000
Stdev	32147320.210130
Skewness	0.423895
Kurtosis	-1.188560

The data summary to the right also shows that the total cases variable contains variance, indicative of the variable being stochastic. Each of the three conditions for time series data have been met, and we can confirm that the data are time series.

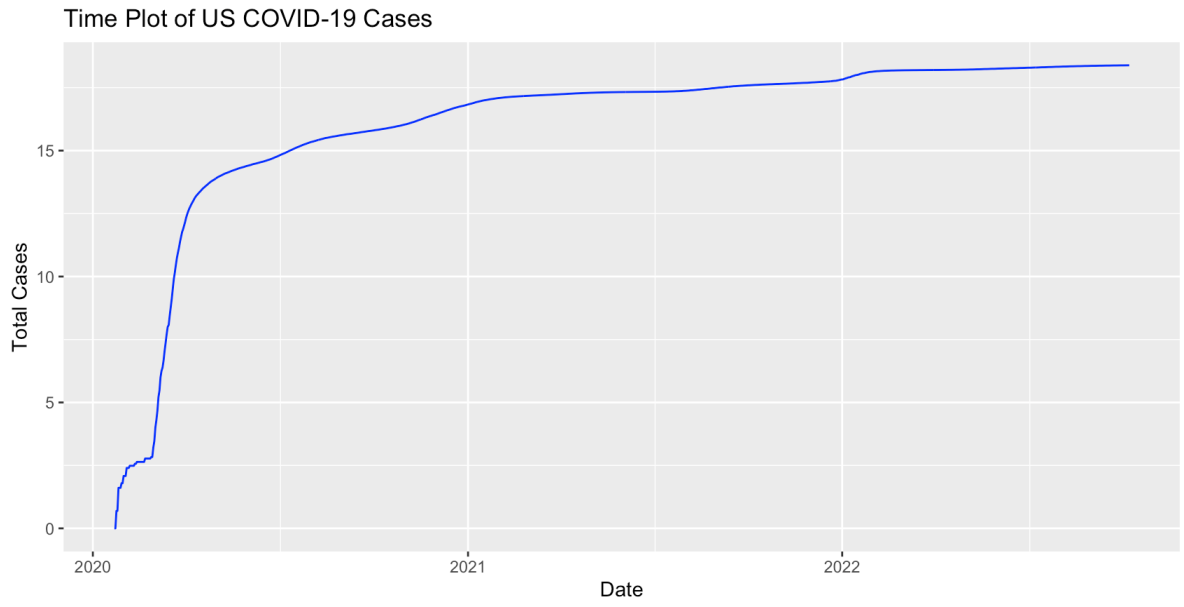
The data do have auto-correlation. A Box-Ljung test of the null hypothesis that the data are not auto-correlated produced a p-value that was within 15 decimal places of zero. That is sufficiently low for us to reject the null hypothesis and assume that the data has auto-correlation.

A histogram of the data, shown below on the left, indicates that it is not normally distributed. The distribution appears to be tri-modal. Non-normality is corroborated by a QQ plot, shown below on the right, which shows how the distribution's tails deviate from the normal line. Additionally, neither the 95% confidence intervals for skew (0.422, 0.428) nor excess kurtosis (-1.965, -1.876) include 0.

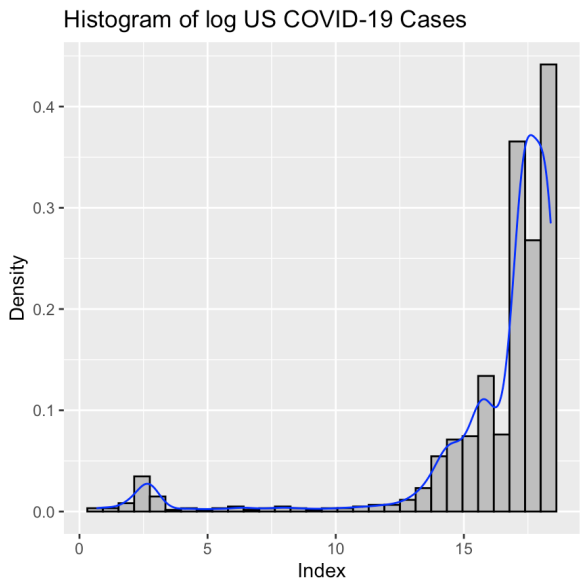


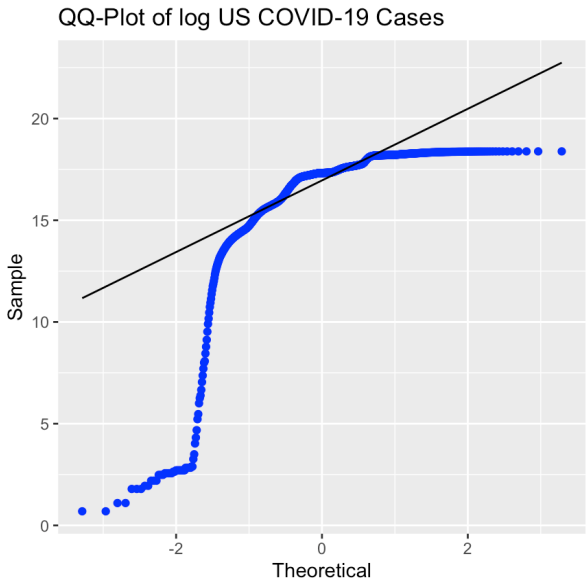


Log transformation of the total cases variable does not improve the variable's distribution enough to justify its use. The result of log transformation is shown as a time plot below.

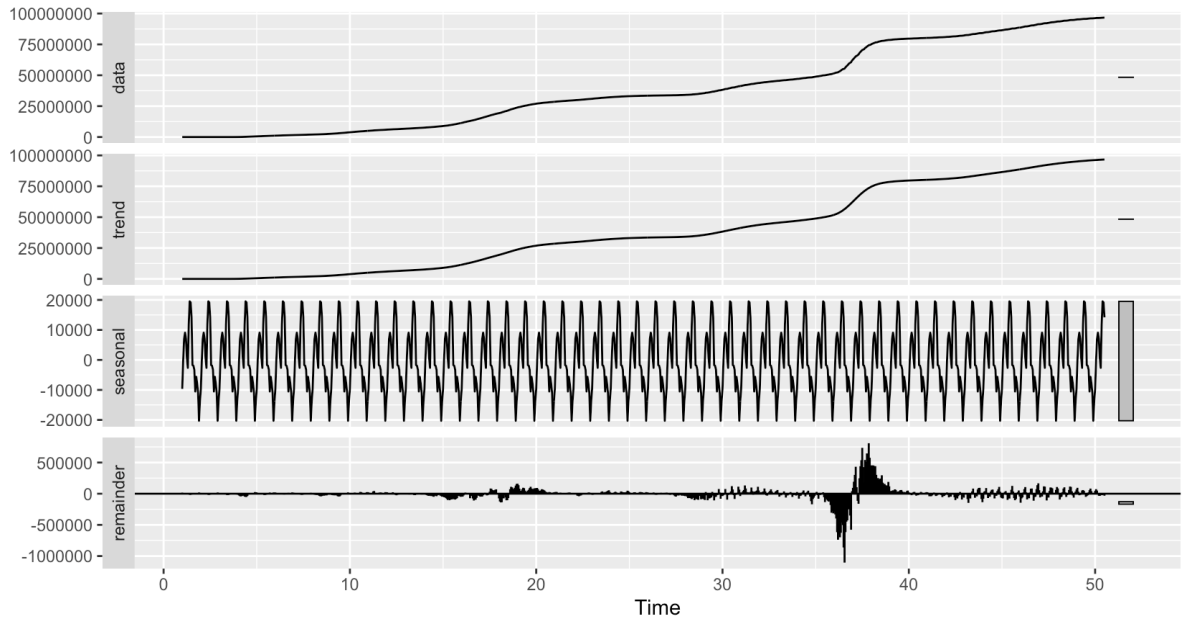


While log transformation does cause the histogram of total cases to more closely resemble a normal distribution by making it more uni-modal, other methods of confirming normality worsen. As the QQ plot below demonstrates, the tails of the distribution for log total cases are further from the normal line than before. Similarly, the 95% confidence intervals for skew (-2.952, -2.934) and excess kurtosis (8.486, 8.367) are further from zero. Given these features, I will not apply log transformation.



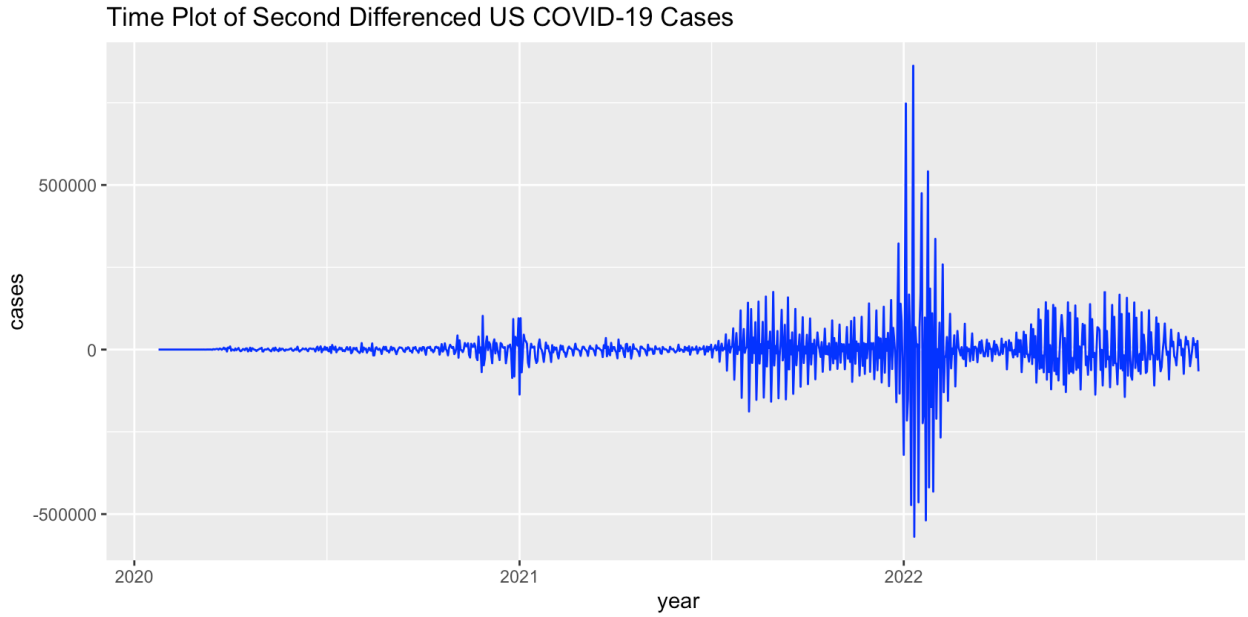


STL Decomposition of the time series produced the plots below. As one can see, the time plot and the trend line look nearly identical. While the plots do indicate that there is a seasonal component to the data, the variance it causes doesn't seem to surpass 20,000, which is equal to about 0.05% of the mean value of the total cases. So, it does not have that much of an effect.



The original time plot of the time series indicates that the time series is not stationary, as there is an obvious linear trend. The various methods we use to test stationarity all agree. As shown in the data summary on page 1, the mean of total cases was 38,930,025.73. A t-test of the null hypothesis that the true mean is 0 returned a 95% confidence interval of (36926074, 40933978) and a p-value that was within 15 decimal places of zero. Thus, we can reject the null hypothesis. ADF and KPSS tests provided similar results. The ADF test, for which the alternative hypothesis is random-walk stationarity, returned a p-value of 0.7124, not allowing us to reject the null hypothesis. Thus, random-walk stationarity is not present. The KPSS, for which the alternative hypothesis is linear-trend stationarity, returned a test-statistic of 1.433, far above the test-statistic to reject the null hypothesis at the 95% confidence level, which was 0.146. Thus, we cannot reject the null hypothesis and do not assume linear-trend stationarity.

Differencing the data twice produced a time series that was stationary. Below is a time plot of the twice differenced data, which no longer shows such an obvious linear trend.



Applying the same tests of stationarity produced positive results. As the data summary to the right indicates, the mean is now 7.39, which is close to zero. A t-test of the null hypothesis that the mean is zero returned a 95% confidence interval of (-4851.453, 4866.236) and a p-value of 0.9976. Thus, we fail to reject the null hypothesis that the mean is 0.

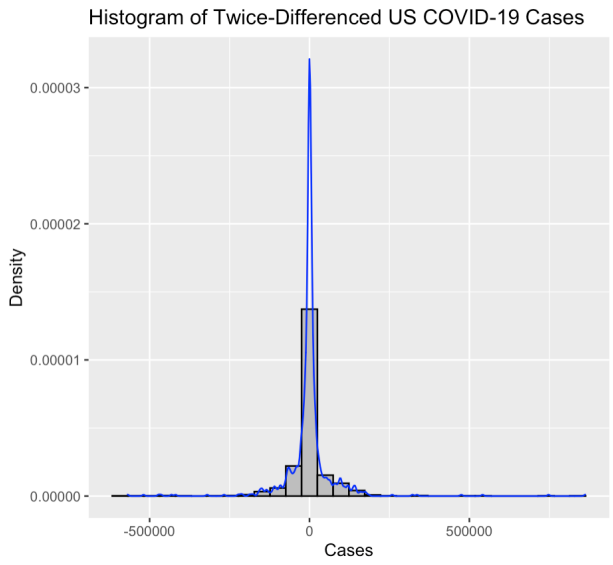
nobs	989.000000
NAs	0.000000
Minimum	-568791.000000
Maximum	862412.000000
1. Quartile	-12101.000000
3. Quartile	9356.000000
Mean	7.391304
Median	0.000000
Sum	7310.000000
SE Mean	2476.010884
LCL Mean	-4851.453129
UCL Mean	4866.235738
Variance	6063192966.924659
Stdev	77866.507350
Skewness	1.445716
Kurtosis	36.164594

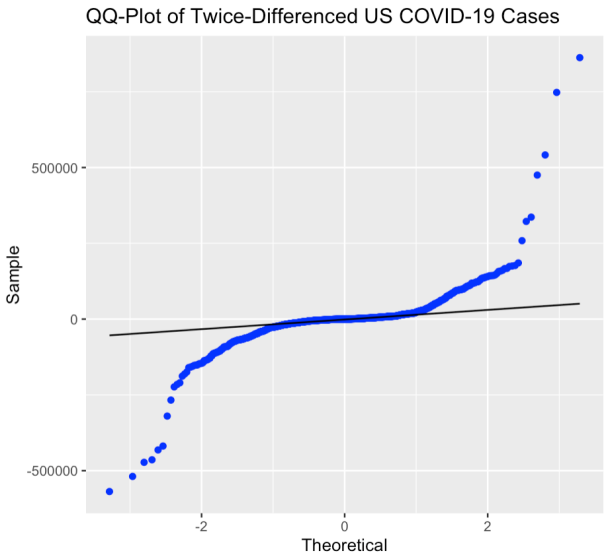
An ADF test with the alternative hypothesis of random-walk stationarity returned a p-value of 0.01, allowing us to reject the null hypothesis and accept random-walk stationarity. Similarly, a KPSS test with the alternative hypothesis of linear-trend stationarity returned a test-statistic of 0.0174, which

is below the test-statistic required to reject the null hypothesis at the 95% confidence level, which was 0.146. Thus, the test indicates linear-trend stationarity.

The second differenced data still has auto-correlation. A Box-Ljung test of the null hypothesis that the data are not auto-correlated returned the same value as the last test, which was within 15 decimal places of zero. Thus, we reject the null hypothesis and assume auto-correlation.

In addition to being stationarity, the twice-differenced data is also better approximated by a normal distribution than the untransformed data. The histogram, shown below on the left, is uni-modal. Similarly, the QQ plot shows that a greater number of the observations fall closer to the normal line, and the deviation of the tails has been reduced. That being said, skew and excess kurtosis both worsened. Kurtosis, which was previously -1.19, is now 36.164 with a 95% confidence interval of (37.906, 38.914). Skew, which was previously 0.423, is now 1.446, with a 95% confidence interval of (1.498, 1.692). Despite these statistics, I would still argue that, based on the histogram and QQ plot, the distribution is closer to normal than before.



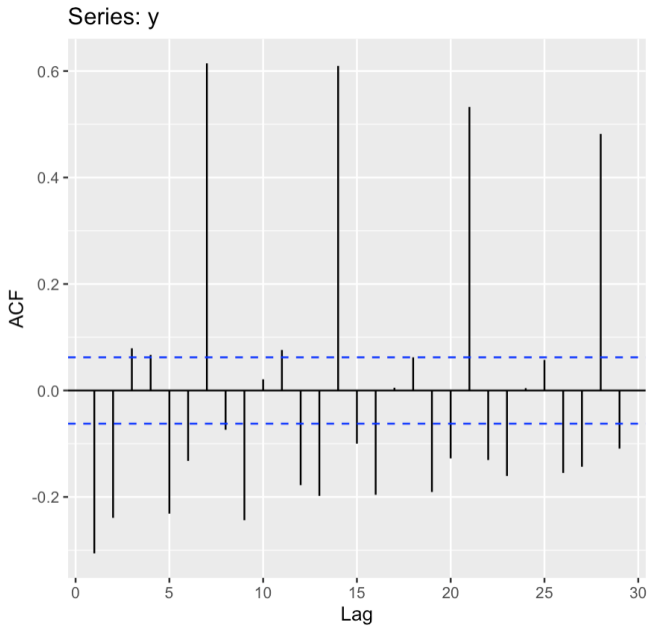


In conclusion, applying second differencing to the data resulted in achieving a stationary time series. In addition, it transformed the distribution to one that is more closely approximated by a normal distribution. This data is now ready for modeling.

2. Moving Average (MA) Models

2.1 ACF and Model Order

The diagram to the right is an ACF plot of the data. The plot demonstrates that lags 1 through 9 are significant. This indicates that an MA model of order 9 is appropriate.



2.2 Model Order via auto.arima

Applying the auto.arima function to the data produced the output on the right. This is an ARIMA(0,0,4) model, which has an MA component of the order 4.

```
Series: y
ARIMA(0,0,4) with zero mean

Coefficients:
          ma1      ma2     ma3     ma4
      -0.6926  -0.2875  0.0280  0.3913
s.e.   0.0290   0.0349  0.0467  0.0346

sigma^2 = 17.17:  log likelihood = -2808.41
AIC=5626.81   AICc=5626.88   BIC=5651.3
```

2.3 ARIMA(0,0,9) Model

Below is the ARIMA(0,0,9) model that I said was appropriate:

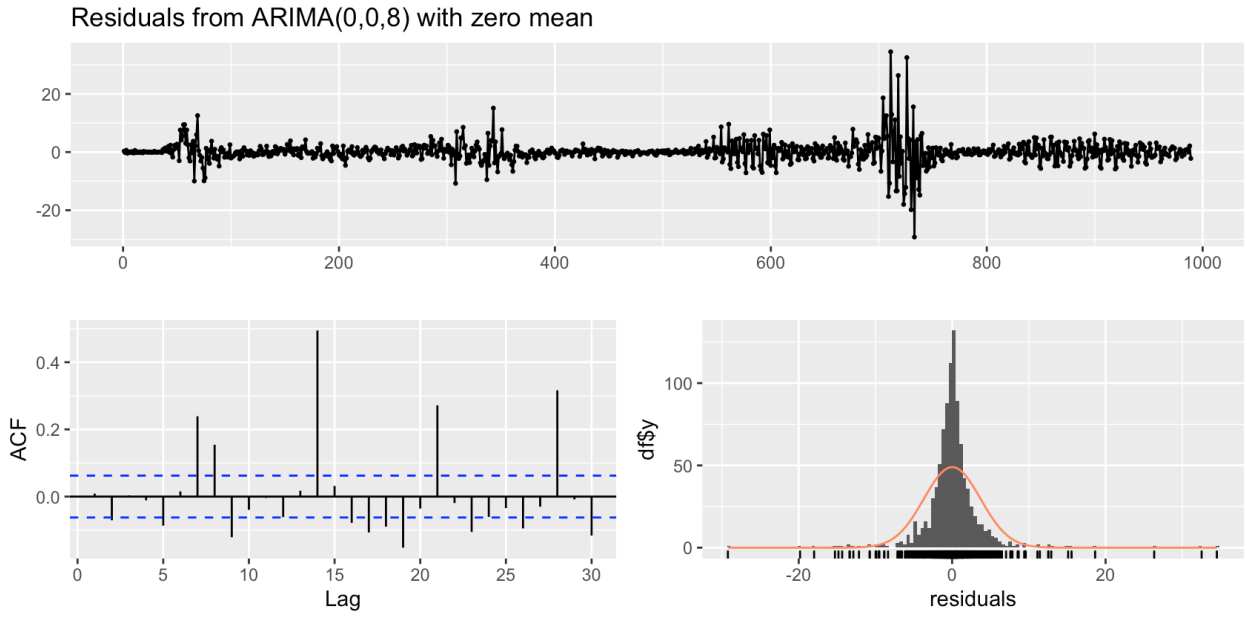
```
ARIMA(0,0,9) with zero mean

Coefficients:
          ma1      ma2     ma3      ma4      ma5     ma6     ma7      ma8      ma9
      -0.6767  -0.0809  0.1090  -0.0264  -0.0638  0.2543  0.3251  -0.3561  -0.0236
s.e.   0.0344   0.0375  0.0369   0.0344   0.0343  0.0430  0.0351   0.0440   0.0390

sigma^2 = 13.64:  log likelihood = -2692.56
AIC=5405.12   AICc=5405.35   BIC=5454.09
```

Compared to the ARIMA(0,0,4) model, all of the information criteria are lower in the ARIMA(0,0,9) model. This suggests that my model is a better fit, and so I will use it.

Applying the parameter test function the model indicated that the 4th, 5th, and 9th lags are not statistically significant at confidence levels equal to and above 95%. Removing these lags produced the following model:

```
Call:
arima(x = y, order = c(0, 0, 8), include.mean = F, fixed = c1)

Coefficients:
          ma1      ma2     ma3  ma4  ma5     ma6     ma7      ma8
      -0.6929  -0.0499  0.0706    0    0  0.2074  0.3190  -0.3856
s.e.   0.0289   0.0372  0.0301    0    0  0.0285  0.0322   0.0279

sigma^2 estimated as 13.64:  log likelihood = -2696.95,  aic = 5407.9
```

While the AIC did increase slightly, I will continue with this model, as it is simpler than the previous and overall not much different. Below is this model's residual panel.



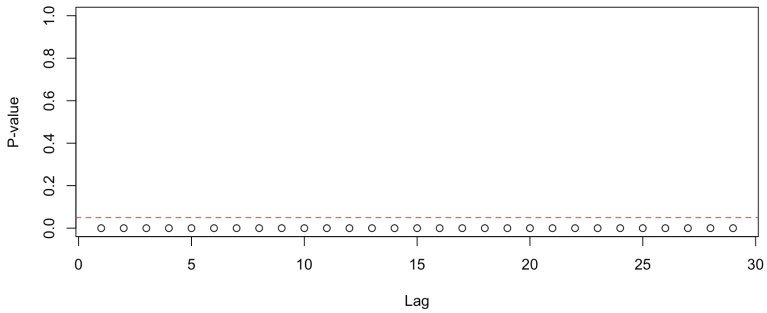
The time plot of the residuals lies close to y = 0. A data summary of the residuals, shown on the right, indicates the mean is -0.0052, which is quite close to zero, agreeing with the line plot. A t-test of the null hypothesis that the residuals' mean is zero produced a p-value of 0.9646 and a 95% confidence interval of (-0.225, 0.235). The p-value allows us to accept the null hypothesis that the residuals' mean is zero, which is also included in the confidence interval.

nobs	989.000000
NAs	0.000000
Minimum	-29.233700
Maximum	34.502525
1. Quartile	-1.160097
3. Quartile	1.129730
Mean	0.005222
Median	0.000744
Sum	5.164563
SE Mean	0.117495
LCL Mean	-0.225347
UCL Mean	0.235791
Variance	13.653321
Stdev	3.695040
Skewness	1.124252
Kurtosis	23.177086

Applying a Box-Ljung test of the null hypothesis that the residuals are non-auto-correlated produced a p-value less than 0.01, indicating we should reject the null hypothesis and assume that the

residuals do have auto-correlation. This is intuitive, as the line plot of the residuals shown in the panel above resembles that of the line plot of the second-differenced data, which also featured auto-correlation. The ACF graph in the panel also indicates auto-correlation, as it has numerous lags that feature spikes above the significance line.

While we accept that the mean of the residuals is zero, they do not feature constant variance, and therefore do not feature strict stationarity. This can be seen in the line plot, which shows an uneven distribution of larger and smaller residuals. A McLeod-Li test of the residuals produced the plot to the right. All of the lags have values beneath the critical line, indicating that we should reject the null hypothesis of constant variance.



However, ADF and KPSS tests of the residuals indicate that they are random-walk and linear trend stationary. Both of these tests have the alternative hypothesis of stationarity, and their test p-value and test statistic were both beneath the critical value to reject the null hypotheses ($0.01 < 0.05$ and $0.0347 < 0.146$, respectively.)

The residuals panel above also features a histogram that depicts the residuals as approximately normally distributed. The distribution is unimodal with good symmetry. However, the distribution is taller than normal, demonstrated by its excess kurtosis of 23.178 and 95% confidence interval equal to (23.56, 24.14).

Our functions identified ‘business cycles’ of lengths 5.48, 2.248, 3.075, and 22.797 in the residuals. While COVID cases and economic conditions can be related in the sense that shutting down to decrease cases also decreases economic activity, I cannot confidently conclude that this is what is being detected, as we have no economic data. Rather, I would argue that it indicates there is some pattern in the data that the model did not capture, hence why it is showing up in the residuals.

Backtesting this model produced the output shown below. I will compare these results with the results of the next model at the end of the next section.

```
> backtest(m,y,500,7)  # one week
[1] "RMSE of out-of-sample forecasts"
[1] 4.995425 6.136960 6.156417 6.161757 6.126806 6.136408 6.218816
[1] "Mean absolute error of out-of-sample forecasts"
[1] 2.796311 3.321131 3.314713 3.290366 3.281602 3.265276 3.263255
[1] "Bias of out-of-sample forecasts"
[1] -0.051840815 -0.002764626  0.001095243 -0.030008192 -0.015560197  0.004701540 -0.036245487
```

Overall, the residuals do not indicate that this model’s predictive ability is very strong. When plotted, the residuals appear very similar to the line plot of the data itself, indicating that the model failed to predict the data’s variance. The residuals are also auto-correlated, have non-constant variance, and contain patterns identified as business cycles. While it is true that

they do feature random-walk and linear trend stationarity, these were features of the underlying data, and the McLeod test indicated that not a single lag featured constant variance.

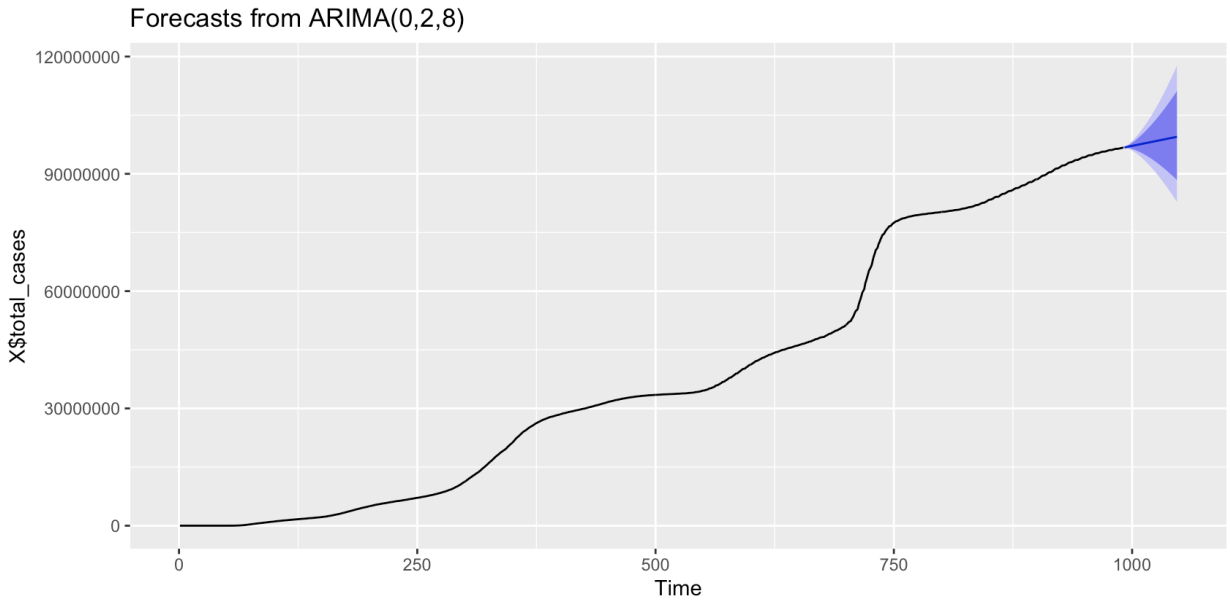
2.4 Step Forecasts

This model's 95% confidence interval predictions for the next seven days are shown on the right. The model predicts that for the next two days, the rate of new cases will increase. This is expressed as the confidence interval shifting upward. From three days on, the model predicts that the rate of new cases will decrease, expressed as the confidence intervals shifting downward.

Time Period	Lower Limit	Upper Limit
t+1	-5.732445	8.744801
t+2	-7.824868	9.788480
t+3	-8.787297	8.840834
t+4	-8.821424	8.836313
t+5	-8.794549	8.863187
t+6	-8.587184	9.070553
t+7	-10.467385	7.443873

2.5 Forecast

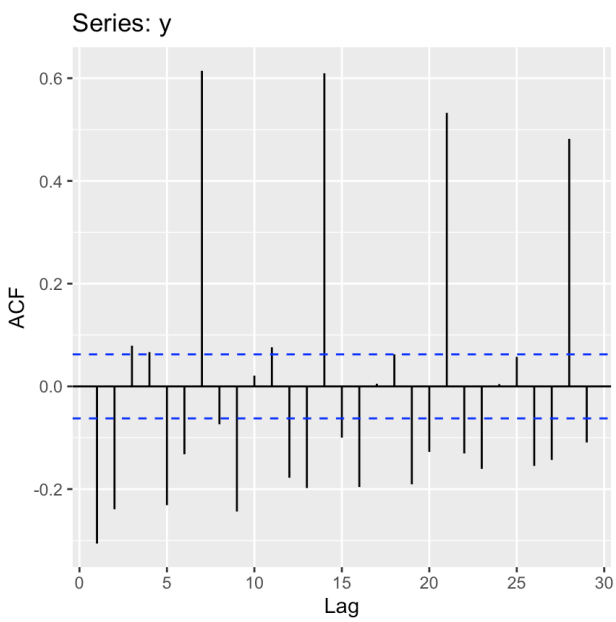
Below is a 95% confidence interval forecast that my model produced. The dark blue line, which represents the average of the forecast, has a positive slope. This indicates that the model predicts it is mostly likely that cumulative cases will rise. This is intuitive, as the cumulative number of cases wouldn't decrease and would only level out in the event of no new cases. A limitation of this model is that it doesn't properly capture that this is cumulative, as it predicts the possibility of decreasing cumulative cases. In that sense, I am inclined to only assess if the rate of new cases changes, which would be represented by change in the slope of the line. While the eye is an imperfect measuring device, it looks to me like the slope of the line was leveling off prior to the predictions, and the predictions of the next 7 days seem to match this pattern of leveling off by decreasing as time passes.

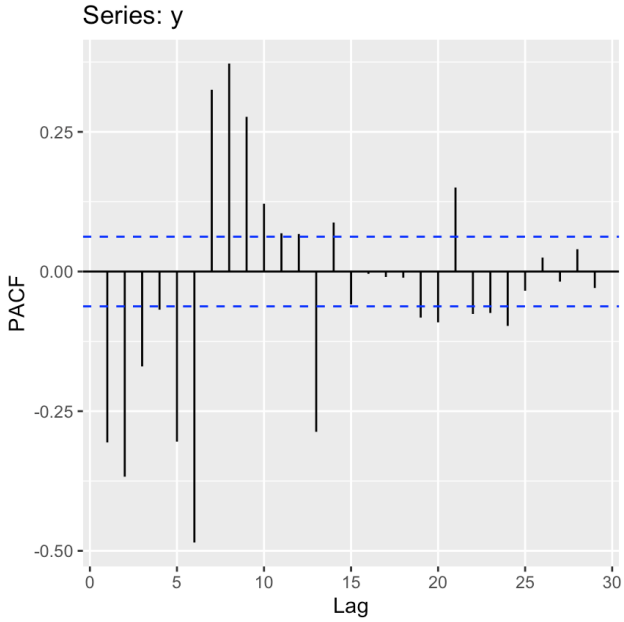


3. AutoRegressive Moving Average (ARMA) Models

3.1 ACF, PACF, and Model Order

Below are the ACF and PACF plots of the data. As in section 2, I argue that the ACF plot shows that lags 1 - 9 are significant. Thus, an MA component of order 9 is appropriate. The PACF plot indicates that lags 1 - 14 are significant, which indicates an AR model of order 14 is appropriate. Together, this would be an ARIMA(14,0,9) model.





3.2 Model Order via auto.arima

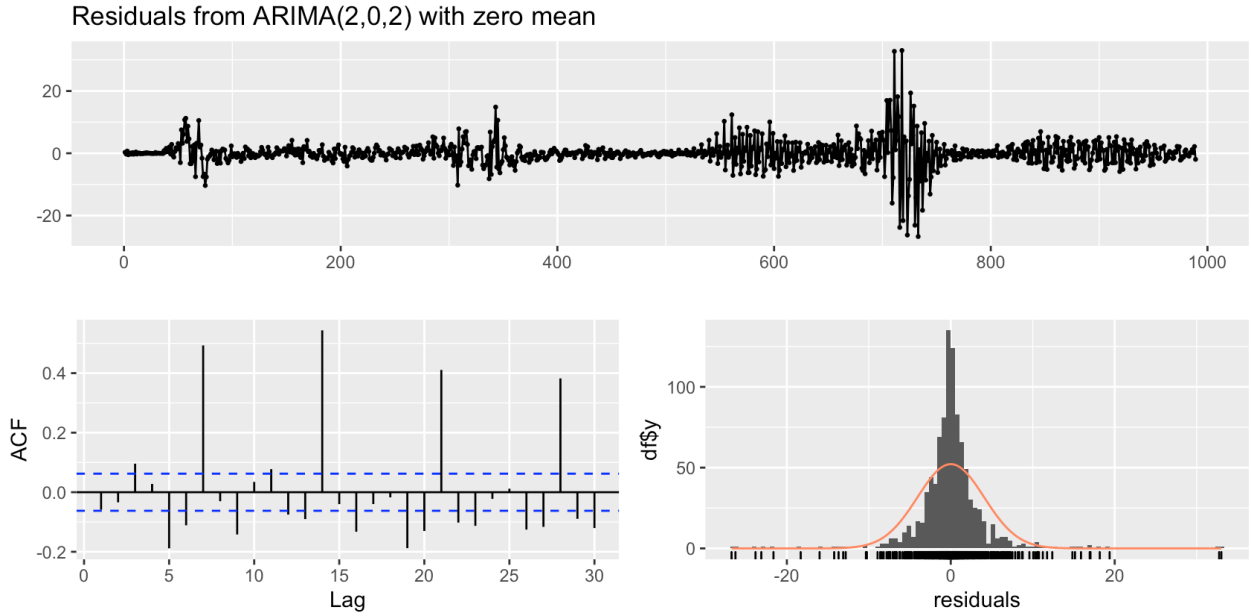
The model that auto.arima suggested is shown on the right. It is an ARIMA(2,0,2) model. This is considerably more simple than the model I proposed, which has a total of 23 variables. Given the simplicity of the recommended model, I will proceed with it rather than mine.

```
Series: y
ARIMA(2,0,2) with zero mean

Coefficients:
         ar1      ar2      ma1     ma2
      0.9512  -0.5472  -1.6099  0.8737
s.e.  0.0306   0.0350   0.0167  0.0171

sigma^2 = 16.64:  log likelihood = -2792.98
AIC=5595.97   AICc=5596.03   BIC=5620.45
```

3.3 ARIMA(2,0,2) Model

The structure of this model is shown on the previous page. A parameter test produced the output on the right, which indicates that all of the variables are statistically significant, even at the 99% confidence level. So, I will keep the model as is. The model's residual panel is displayed below.

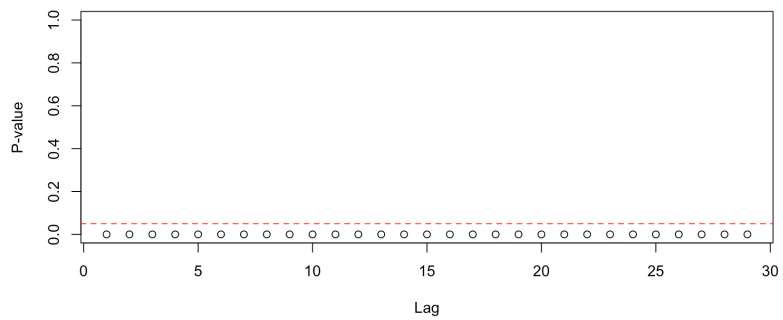
	t	pval_t	pval_z	Pr(>|t|)
ar1	31.08382	0	0	***
ar2	-15.62947	0	0	***
ma1	-96.67792	0	0	***
ma2	51.12447	0	0	***



Tests indicate that the residuals' mean is approximately zero. As the data summary on the right shows, the mean is 0.0054 with a 95% confidence interval produced by a t test of (-0.249, 0.26).

However, the residuals do not feature strict stationarity. A McLeod-Li test of the residuals (pictured below) showed that they do not have constant variance. All lags up to thirty showed non-constant variance, thus we can exclude strict stationarity. ADF and KPSS tests do indicate that the residuals have random-walk and linear trend stationarity, as they both produced a p-value and test statistic beneath the critical values required to reject the null hypotheses of random-walk and linear trend stationarity respectively. The p-value for the ADF test was 0.01 and the test-statistic of the KPSS test was 0.032, with the critical value of the 95% confidence level being 0.146.

nobs	989.000000
NAs	0.000000
Minimum	-26.747700
Maximum	33.017547
1. Quartile	-1.354610
3. Quartile	1.341307
Mean	0.005455
Median	-0.025715
Sum	5.395023
SE Mean	0.129519
LCL Mean	-0.248709
UCL Mean	0.259619
Variance	16.590638
Stdev	4.073161
Skewness	0.302524
Kurtosis	16.830187



Performing a Box-Ljung test of the null hypothesis that the residuals are non-auto-correlated produced a p-value of less than 0.01, indicating that the residuals do feature auto-correlation. This is corroborated by the ACF plot within the residuals panel, which shows numerous lags with significant auto-correlation.

The histogram in the residuals panel demonstrates that the residuals are uni-modal and have good symmetry. Additionally, they have close to zero skew, with a value of 0.303 and a 95% confidence interval of (0.27, 0.38). However, the distribution is quite tall, demonstrated by its excess kurtosis of 16.83, which has a 95% confidence interval of (17.28, 17.55). Given this kurtosis, I do not feel that the residuals are approximately normal.

Our functions identified one 'business cycle' in the residuals, with a duration of 6.329. As with the previous model, I am inclined to believe this indicates a pattern in the data not identified by the model, rather than an indication of the relationship between economic activity and cumulative COVID cases.

Performing a back test with this model produced the output below.

```
> backtest(m,y,500,7)  # one week
[1] "RMSE of out-of-sample forecasts"
[1] 5.932858 6.901534 7.119526 7.026873 6.988955 6.941336 6.938421
[1] "Mean absolute error of out-of-sample forecasts"
[1] 3.410936 4.014552 4.025943 3.937409 3.892671 3.892195 3.888043
[1] "Bias of out-of-sample forecasts"
[1] -0.084419260  0.006460787  0.023938458 -0.048139675 -0.052057893 -0.024699748 -0.003326057
```

Overall, the residuals of this model seem similar to those of the previous, indicating a similar lack of strength in this model's predictive ability. While both models' residuals are mean zero, they do not feature strict stationarity, are auto-correlated, non-normally distributed, and contain patterns identified as 'business cycles.' Further, the plots of the residuals are remarkably similar to that plot of the data itself, indicating that neither model was able to predict much of the data's variance.

However, the results of the first model's backtest were better than the results for the second model. The RMSE, MAE, and Bias for the first model were lower for all periods than in the second model. This is indicative of a more predictive model. Thus, I conclude that the previous model is preferable to the second. However, given their similarity, I will include insights from both in the report while emphasizing the relative strength of the first model.

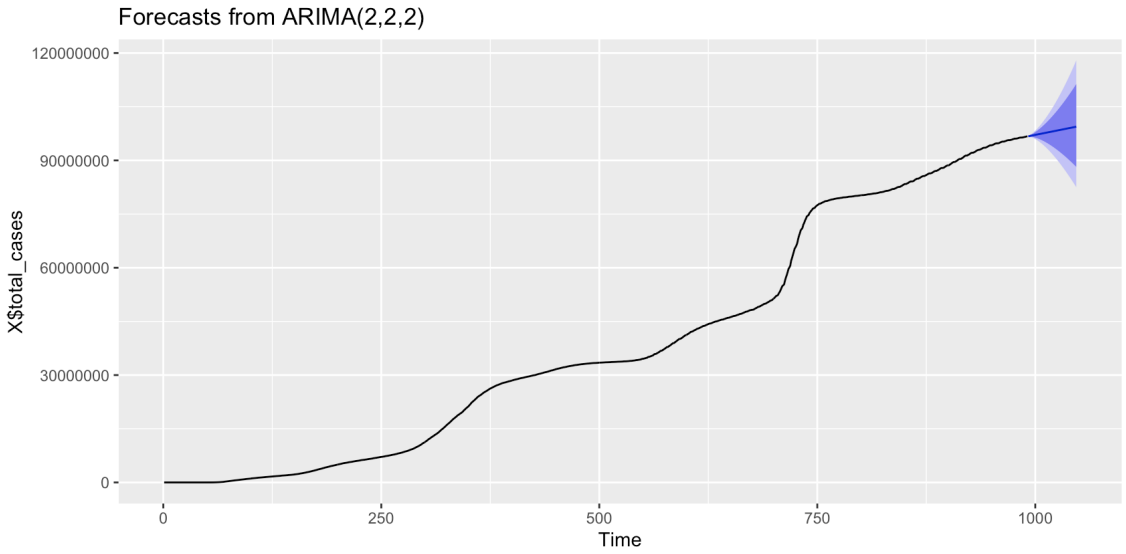
2.4 Step Forecasts

This model's forecasts of the next seven days are shown on the right. Unlike the previous model, this model's predictions feature back and forth decreases and increases in daily cases, rather than an eventual decrease. Overall, this has a similar effect of leveling out, as the values stay close, but with some more fluctuations.

Time Period	Lower Limit	Upper Limit
t+1	-6.884313	9.106788
t+2	-8.381223	10.768331
t+3	-9.343500	10.398049
t+4	-10.040520	9.737485
t+5	-10.499396	9.634161
t+6	-10.501156	9.843967
t+7	-10.254820	10.103125

3.5 Forecast

Below is this model's 95% confidence level forecast. Similar to the previous model, this model suggests the possibility of decreasing cumulative cases, which isn't possible. However, this forecast features a line with a steeper slope than the previous model, indicating that this model expects cumulative cases to rise more quickly. This matches the step forecasts, in which the first model predicts leveling off and the second predicts a consistent number of new cases.



4. Report

I have created two models that are designed to predict the number of cumulative COVID-19 cases over the course of the next two months. Both of these models function by identifying patterns in previous numbers of cumulative cases and making predictions based on these patterns. Statistically, both of these models were similar to one another, with one showing slightly better test results than the other.

Naturally, both of these models predicted increases in the number of cumulative cases. Since June, the number of new daily cases has been decreasing, resulting in a cumulative count that is leveling off. My better model predicted a continuation of this leveling off, the other predicted a continuation of the current rate of new cases. Given the recent trend and the prediction of my better model, I will predict that the number of daily cases will decrease and the cumulative count will level off.

That being said, one additional factor must be considered. My models did not account for seasonality, meaning that if there is a relationship between the time of year and cases, my models did not account for it. While we have witnessed a leveling off of cumulative cases between August and October, I advise to remain cognisant of the upcoming winter and the potential for a spike in cases due to cold weather.