NU Time Series 413, Assignment 3

ARMA Models (TS3)

Reed Ballesteros

Northwestern University SPS, Fall 2022
MSDS-413-DL
Instructor: Dr. Jamie D. Riggs, Ph.D
2022-10-10

1. EDA (20 points)

Consider the data set of daily total number of Covid-19 cases confirmed after positive test. The data file is ___https://covid.ourworldindata.org/data/owid-covid-data.csv___ with column names *date*, *iso_code*, *total_cases*, and *population*. Use the columns *date* and *total_cases*. Use your EDA from Assignment 1 to obtain and justify a stationary total cases time series. You may need to log-transform or first difference or both.

Validate data as a time series:

```
## [1] 987
```

```
## [1] 987
```

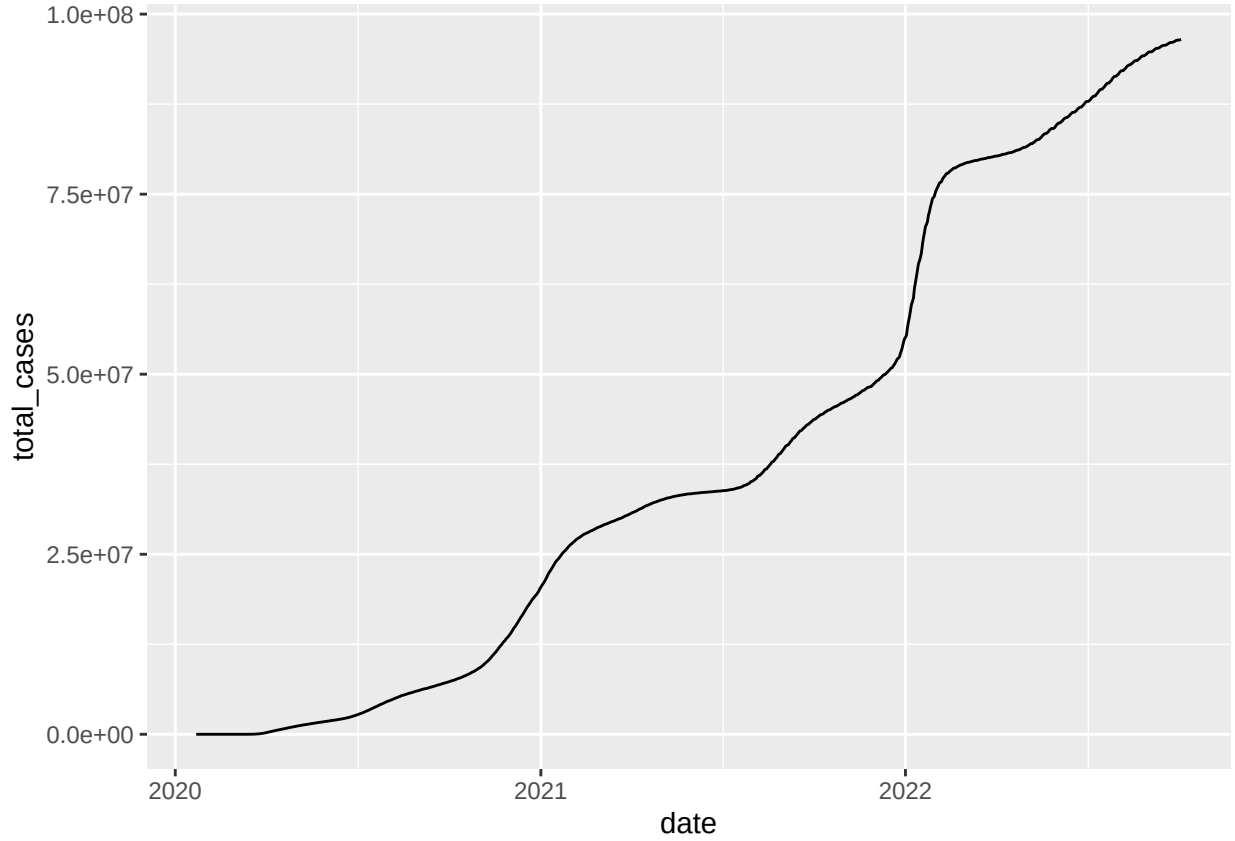
As seen in Assignment 1, we have 987 unique dates in 987 observations, which meets the $H_{10} : x_{it},\ i \in \{1, 2\},\ t \in \{1, 2, ..., n\}$ requirement for time series validation.

```
## [1] 987
```

```
## dif
##   1
## 986
```

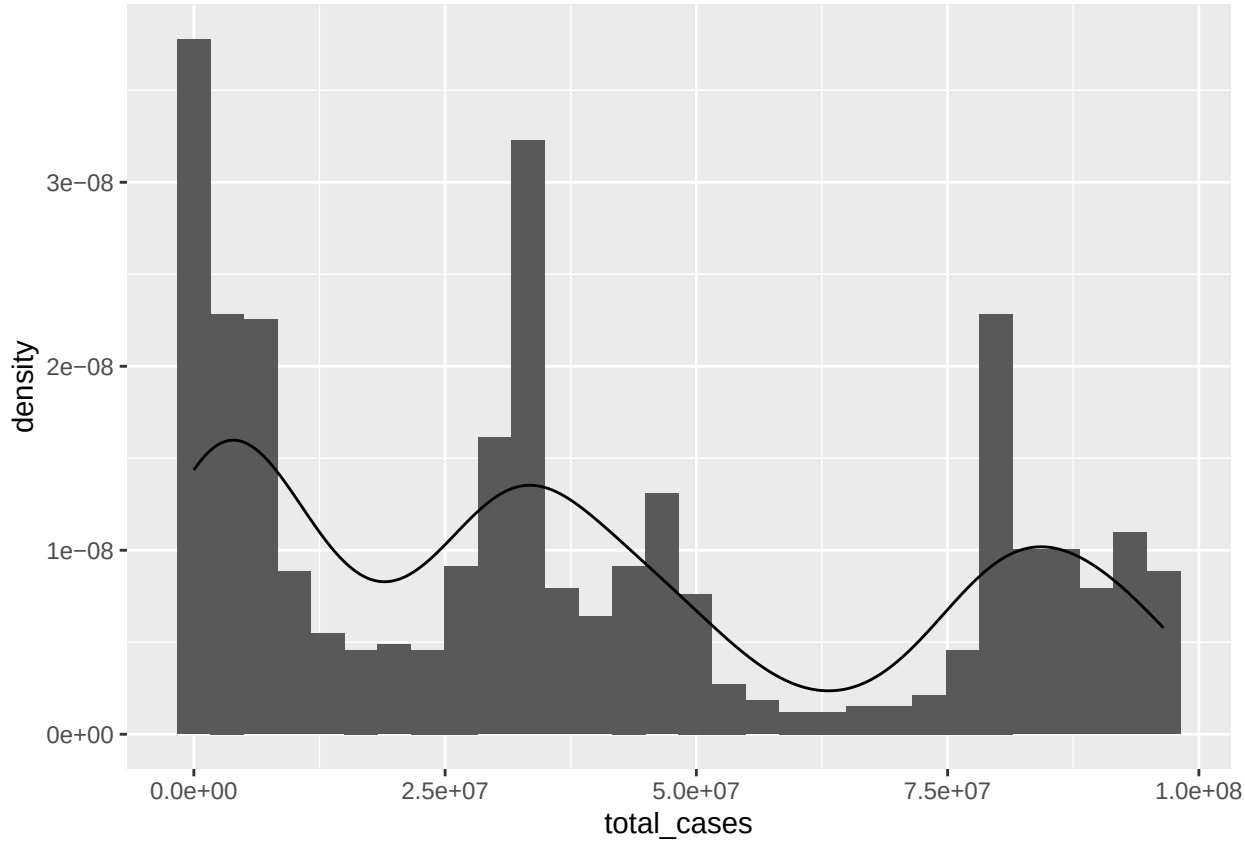
From the test above, we can verify that the constant time span between each date is only one day, denoted by the single value 1. This meets the $H_{20} : (t + 1) - t = c, t \in \{1, 2..., n\}$ requirement for time series validation.

Plotting the time series:

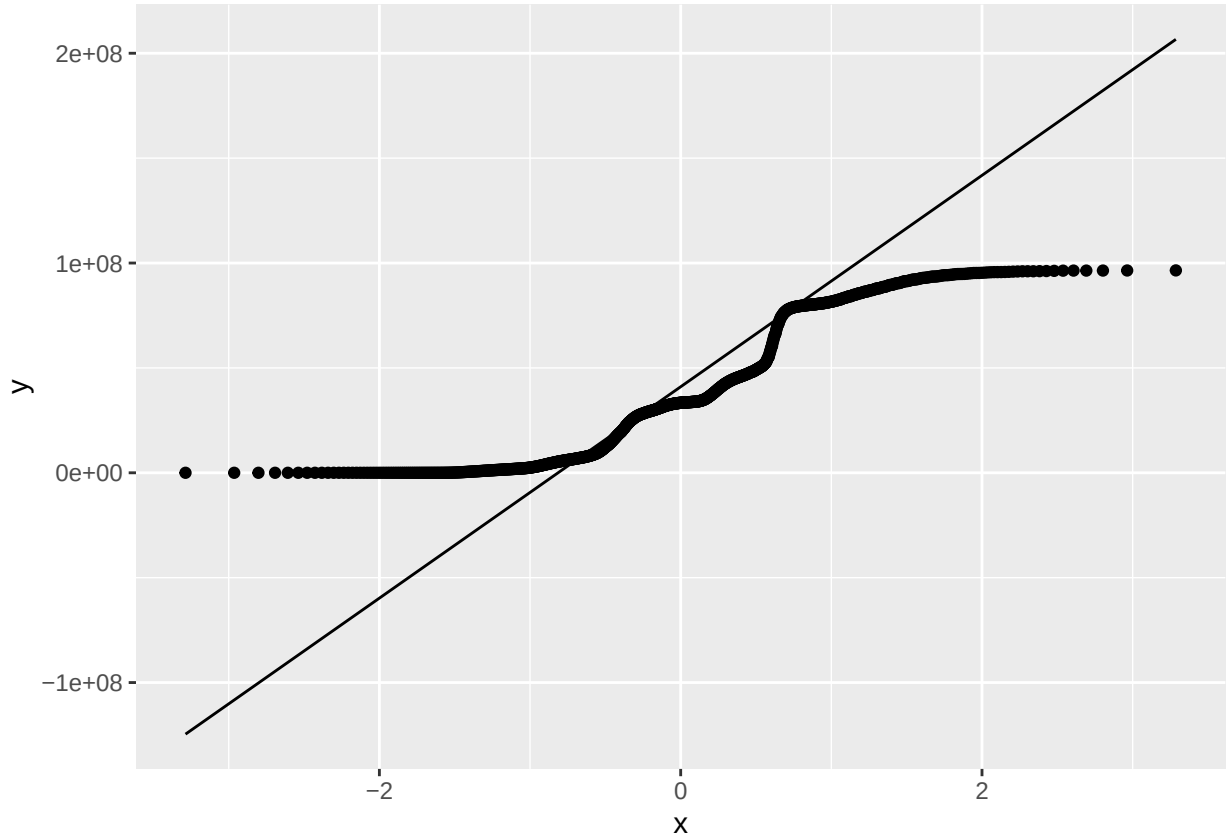


From the dataset we can observe an overall upward trend in total_cases, but with some distinct trend-cycles

as the growth rate of total_cases varies throughout the time period. There are steep trend-cycle rises in total_cases such as during the fall season of 2020 and the beginning of 2022. With this upward increasing trend we can also say the time series is not stationary, as the mean is increasing over time.



The total_cases histogram displays a trimodal plot, exhibiting non-normal distribution, thus not conforming to a Gaussian PDF.



Much of the total_cases data in the Q-Q plot deviate from the ideal normal distribution line on opposite sides with long tails exhibiting very tall kurtosis, demonstrating non-normality and non-conformity to a Gaussian PDF.

Normal test with Skewness:

```
##      skew    lwr.ci    upr.ci
## 0.4295922 0.4251327 0.4311414
```

```
## [1] FALSE
```

The total_cases data is calculated to have some right skewness, as the 95% CI does not contain zero, making the distribution not normal.

Normal test with (excess) Kurtosis:

```
##      kurt    lwr.ci    upr.ci
## -1.178420 -1.187016 -1.178429
```

```
## [1] FALSE
```

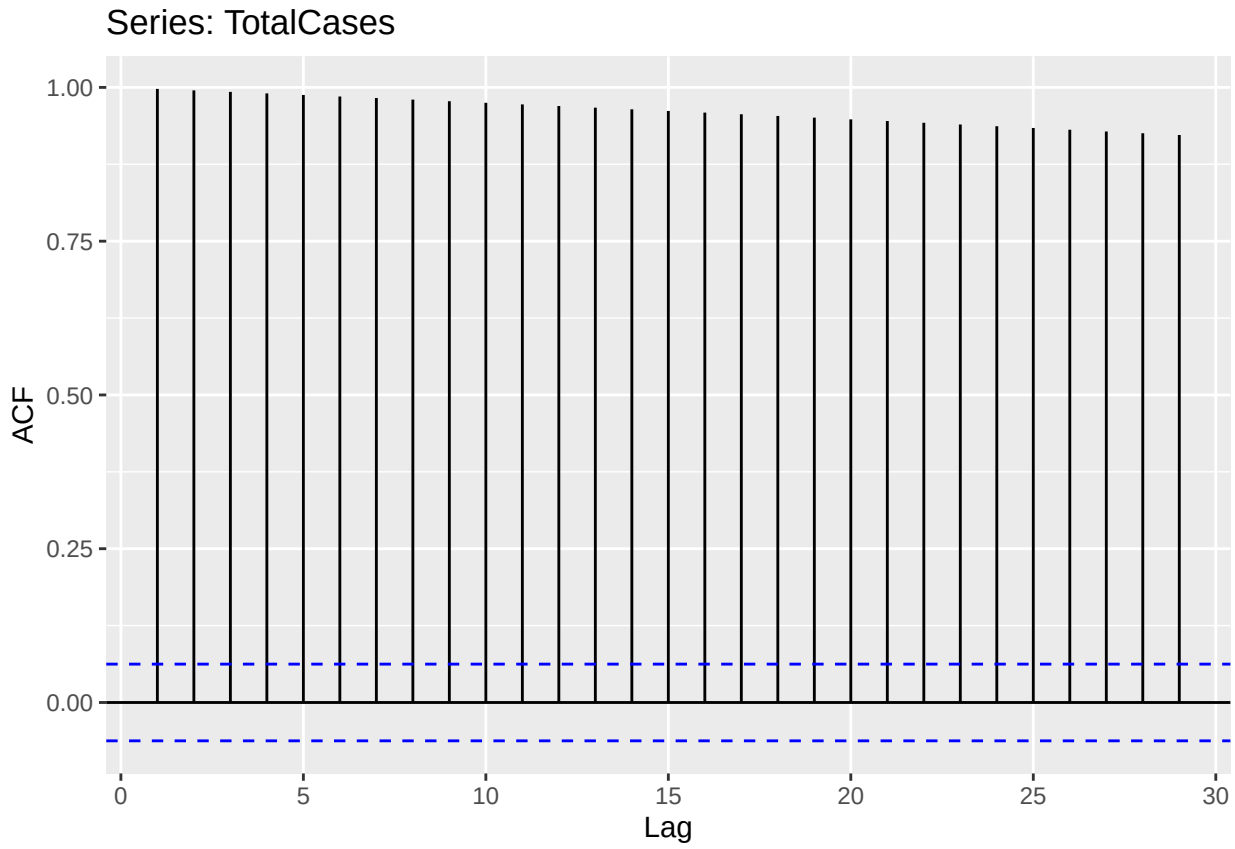
The total_cases data is calculated to have highly-peaked (excess) Kurtosis, making the distribution not normal.

T-Test:

```
##
##  One Sample t-test
##
## data:  c
## t = 37.989, df = 986, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  36697165 40694926
## sample estimates:
## mean of x
##  38696046
##
## [1] "T-Test: mean not zero, linear trend present -> reject H0"
```

```
## [1] FALSE
```

We can tell from the plot the mean of the time series is not zero, and the t.test formally confirms it since the 95%CI does not contain 0, and also confirms linear trends are present.



The high lag spikes in the ACP plot show the total_cases data is not stationary. We can observe this from the basic time series plot as total_cases is ever-growing.

Independence Box-Ljung test:

```
##
##  Box-Ljung test
##
```

```
## data:  c
## X-squared = 27773, df = 30, p-value < 2.2e-16
```

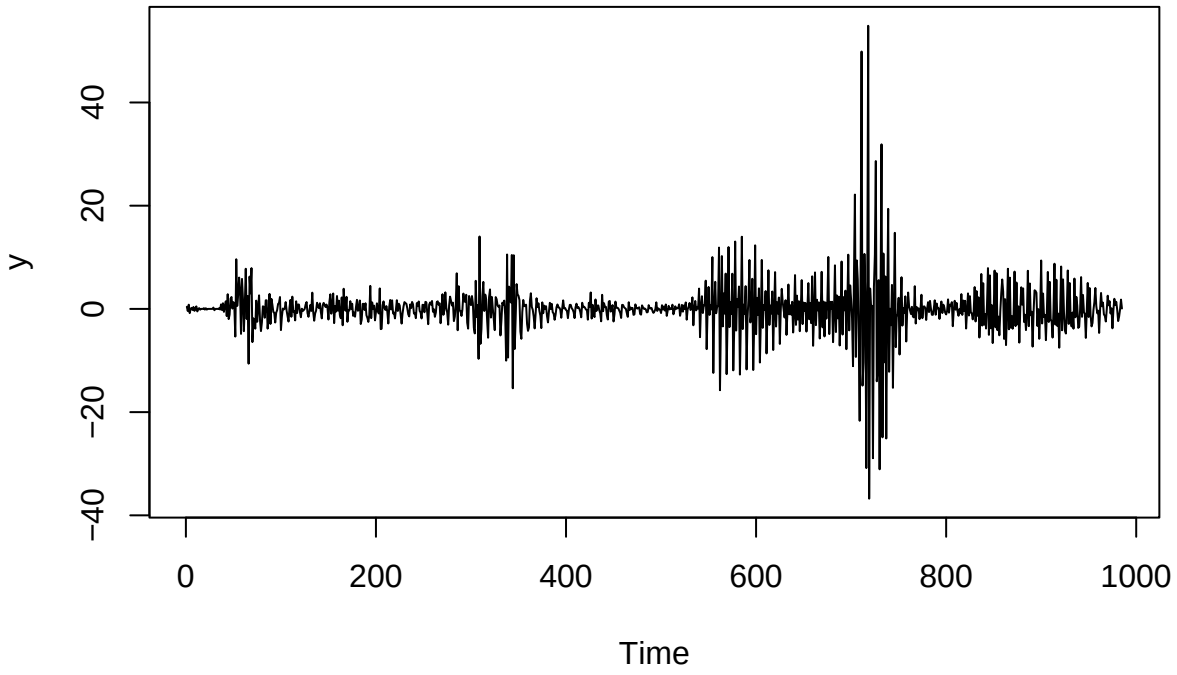
```
## [1] FALSE
```

We fail to reject null hypothesis of independent lags of the total_cases data, as the Box Ljung test implies dependency in the data over 30 lags.

We want to transform the total_cases data in a way where the mean is closer to zero and have a more normal-like distribution. The transformation might not meet all the expectations of a Gaussian PDF, but we can attempt to try to meet some of those requirements, as well attempt to make the data more stationary.

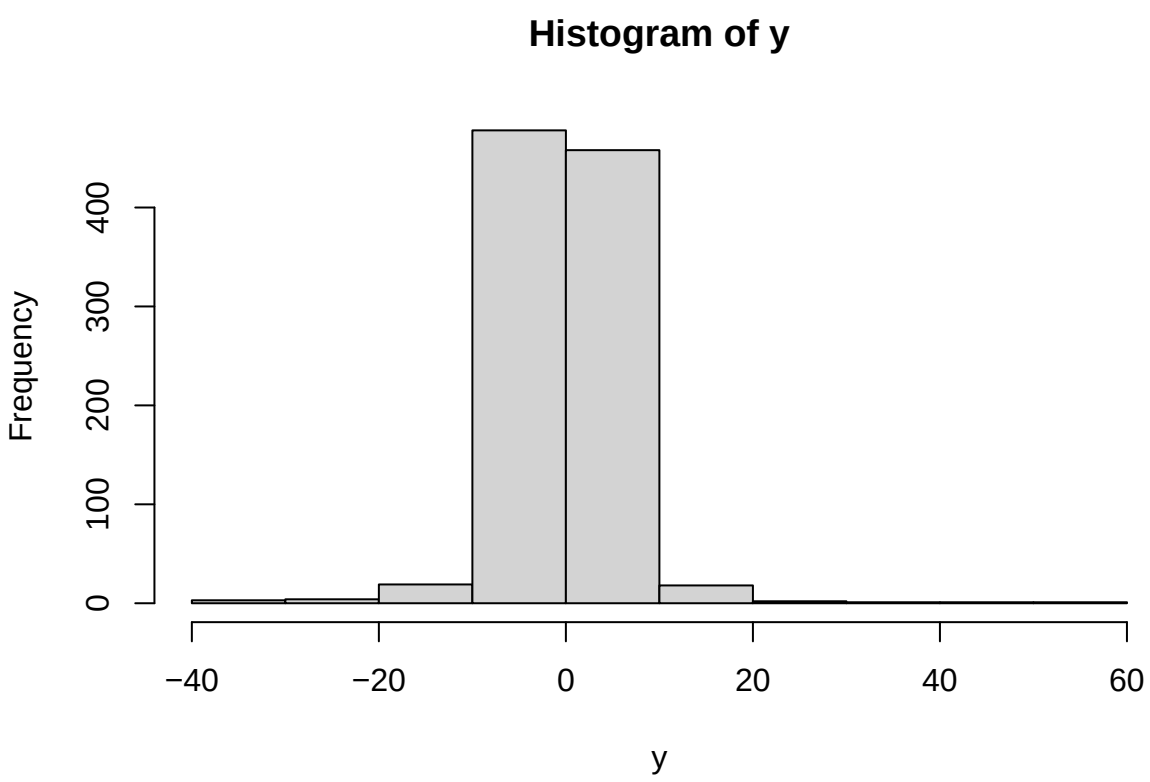
The best we are able to do to make the time series data more 'normal' is taking the difference of the square root of total_cases twice: diff(diff(sqrt(total_cases))).

Plot diff(diff(sqrt(total_cases)))



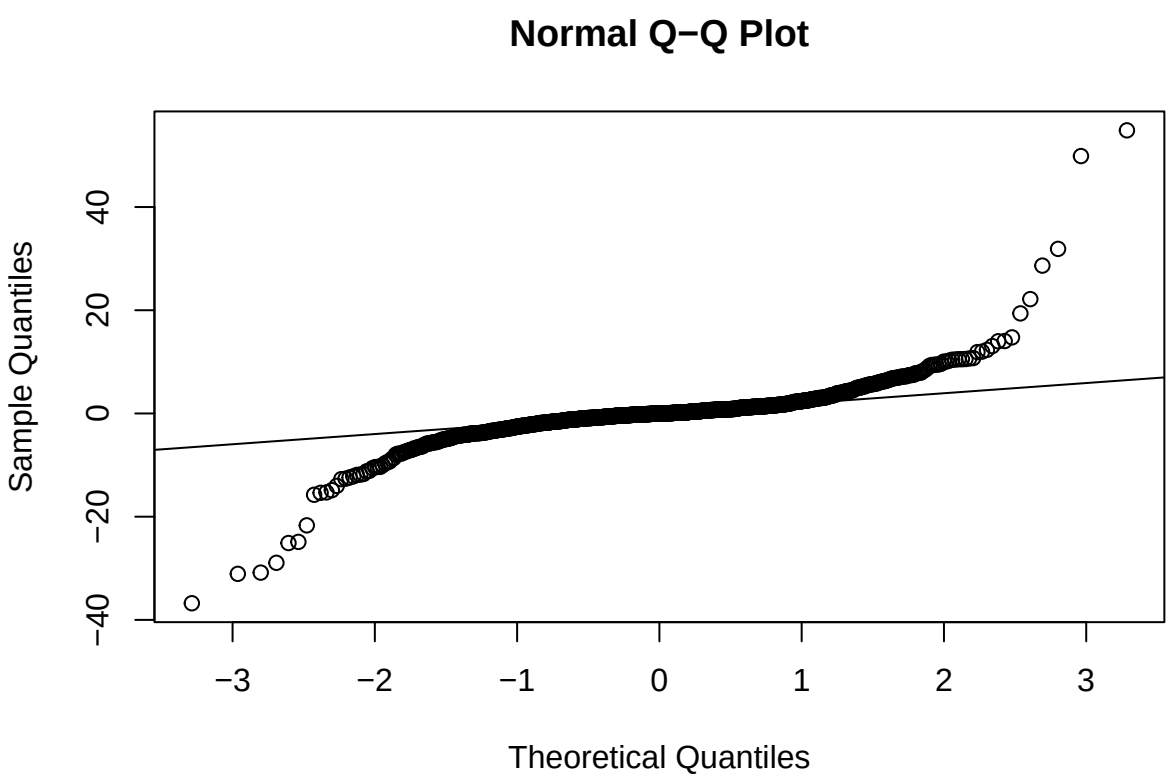
We can infer from the plot of diff(diff(sqrt(total_cases))) that the mean could be close to zero and not constant variance.

Histogram: diff(diff(sqrt(total_cases)))



We see a more normal-like distribution of diff(diff(sqrt(total_cases))) with much less skewness before tranformation. But we also see a highly-peaked distribution, implying veryt high (excess) Kurtosis.

Q-Q Plot: diff(diff(sqrt(total_cases)))



The Q-Q plot shows most of the data on the ideal natural slope, with the ends sharply deviating away, indicating high (excess) Kurtosis.

Normal test diff(diff(sqrt(total_cases))) with Skewness:

```
##     skew   lwr.ci   upr.ci
## 1.307542 1.351752 1.517941
```

```
## [1] FALSE
```

The skewness 95% CI does not contain zero, therefore showing signs of right skewness.

Normal test diff(diff(sqrt(total_cases))) with (excess) Kurtosis:

```
##     kurt   lwr.ci   upr.ci
## 29.61356 31.59286 32.51996
```

```
## [1] FALSE
```

The (excess) Kurtosis 95% CI does not contain zero, but large values, therefore showing signs highly peaked Kurtosis.

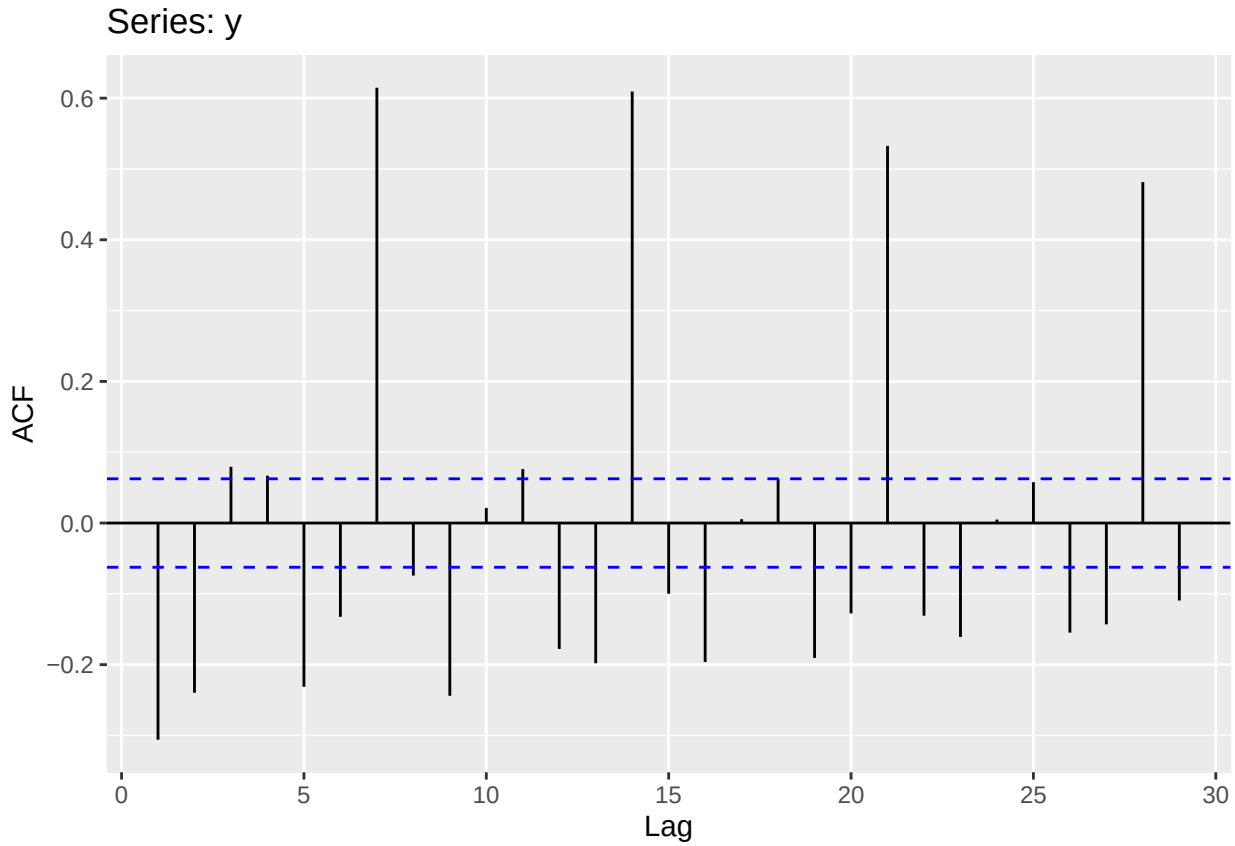
T-Test diff(diff(sqrt(total_cases))):

```
##
##  One Sample t-test
```

```
##
## data:  c
## t = 0.013166, df = 984, p-value = 0.9895
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.3240983  0.3284765
## sample estimates:
##   mean of x
## 0.002189097
##
## [1] "T-Test: mean is zero, removed linear trend -> fail to reject H0"

## [1] TRUE
```

As we can see from the plot of diff(diff(sqrt(total_cases))) the mean might be close to zero, and the t-test shows that the 95% CI does contain zero, thus the mean is statistically zero.



From the ACF plot above we can notice a 7-lag cycle counting from the highly-positive peaks. While lags 3,4, and 8 are near statistically 0, lag 10 is the first lag to show a statistically 0 dropoff.

We can see from the ACF plot that the transformation is not stationary.

Constant variance for diff(diff(sqrt(total_cases))): Breusch-Pagan test

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(c ~ seq(1, length(c)))
```

```
## BP = 10.933, df = 1, p-value = 0.0009446
```

```
##    BP
## FALSE
```

We observed non-constant variance in the plot of diff(diff(sqrt(total_cases))), and the Breuch-Pagan test formally confirms it quantitatively.

Independence via Box-Ljung test:

```
##
##  Box-Ljung test
##
## data:  c
## X-squared = 1874.3, df = 30, p-value < 2.2e-16
```

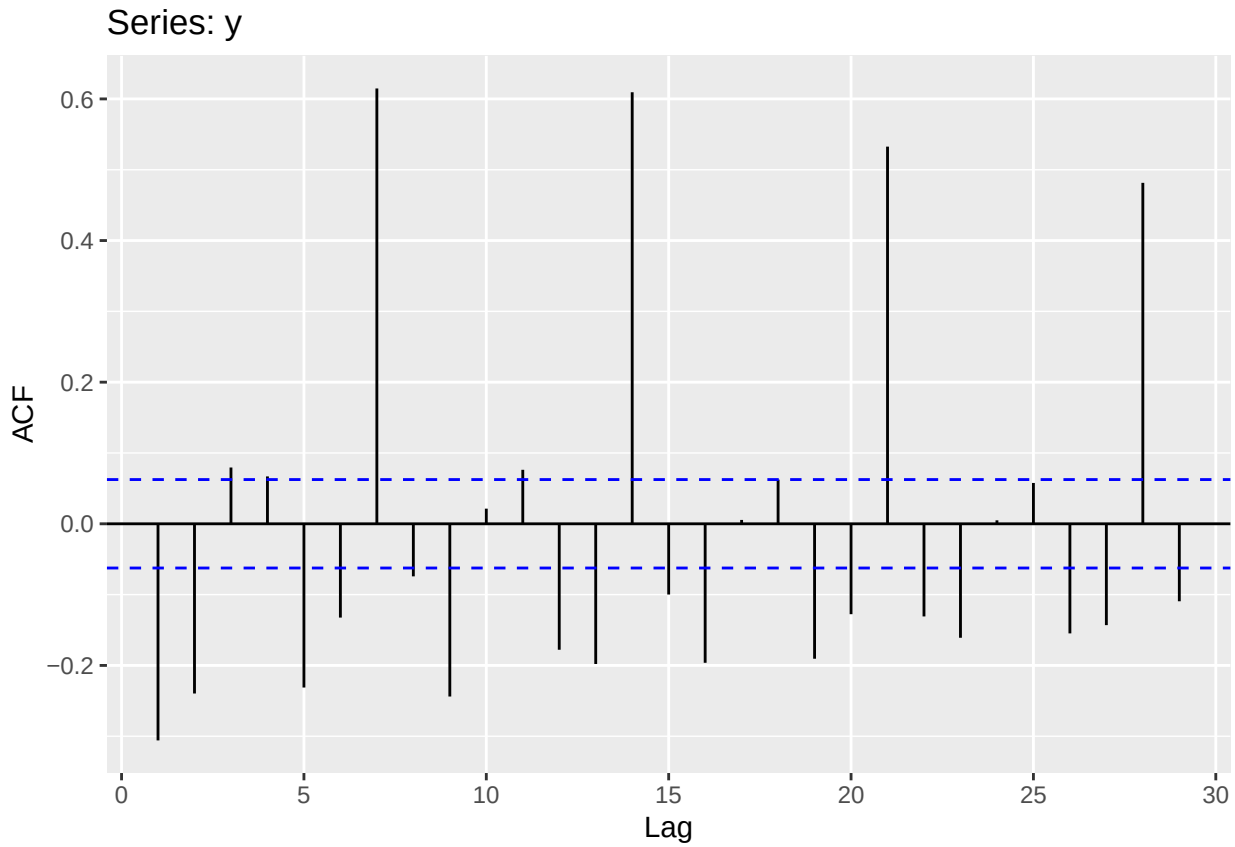
```
## [1] FALSE
```

We fail to reject null hypothesis of independent lags of diff(diff(sqrt(total_cases))), as the Box Ljung test implies dependency in the data over 30 lags.

2. Moving Average (MA) Models (20 points)

2.1. Use the ACF to determine the order to fit a MA model. Justify your choice of order.

ACF Plot of diff(diff(sqrt(total_cases))):



From the rule of thumb document, with an ACF plot 'the function drops off to 0 after lag q (i.e. D(q)).'

Based on the ACF plot above, the dropoff to 0 is at lag 10; lag 10 is after lag 9 (with q=9, D(q)=D(9)), therefore we could propose a 9th-order moving-average model, or MA(9).

A summary of MA(9) is as follows:

```
## Series: y
## ARIMA(0,0,9) with zero mean
##
## Coefficients:
##           ma1      ma2     ma3      ma4      ma5     ma6     ma7      ma8
##       -0.6765  -0.0812  0.1089  -0.0266  -0.0636  0.2540  0.3256  -0.3559
## s.e.   0.0345   0.0375  0.0370   0.0345   0.0343  0.0431  0.0352   0.0440
##           ma9
##       -0.0236
## s.e.   0.0391
##
## sigma^2 = 13.69:  log likelihood = -2683.28
## AIC=5386.55   AICc=5386.78   BIC=5435.48
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE         ACF1
## Training set 0.003743469 3.682904 2.041794 NaN  Inf 0.4429655 -0.003562706
```

2.2. Use the command auto.arima(y, $d = 0$, $max.p = 0$, $stationary = TRUE$), where y is your stationary *total_cases* time series, to find the MA order. Interpret and compare with your ACF choice. Give the model degrees of freedom (df).

auto.arima() proposes using a 4th-order moving-average model, MA(4).

```
am <- auto.arima(y,d = 0,max.p=0,stationary=TRUE)
summary(am)
```

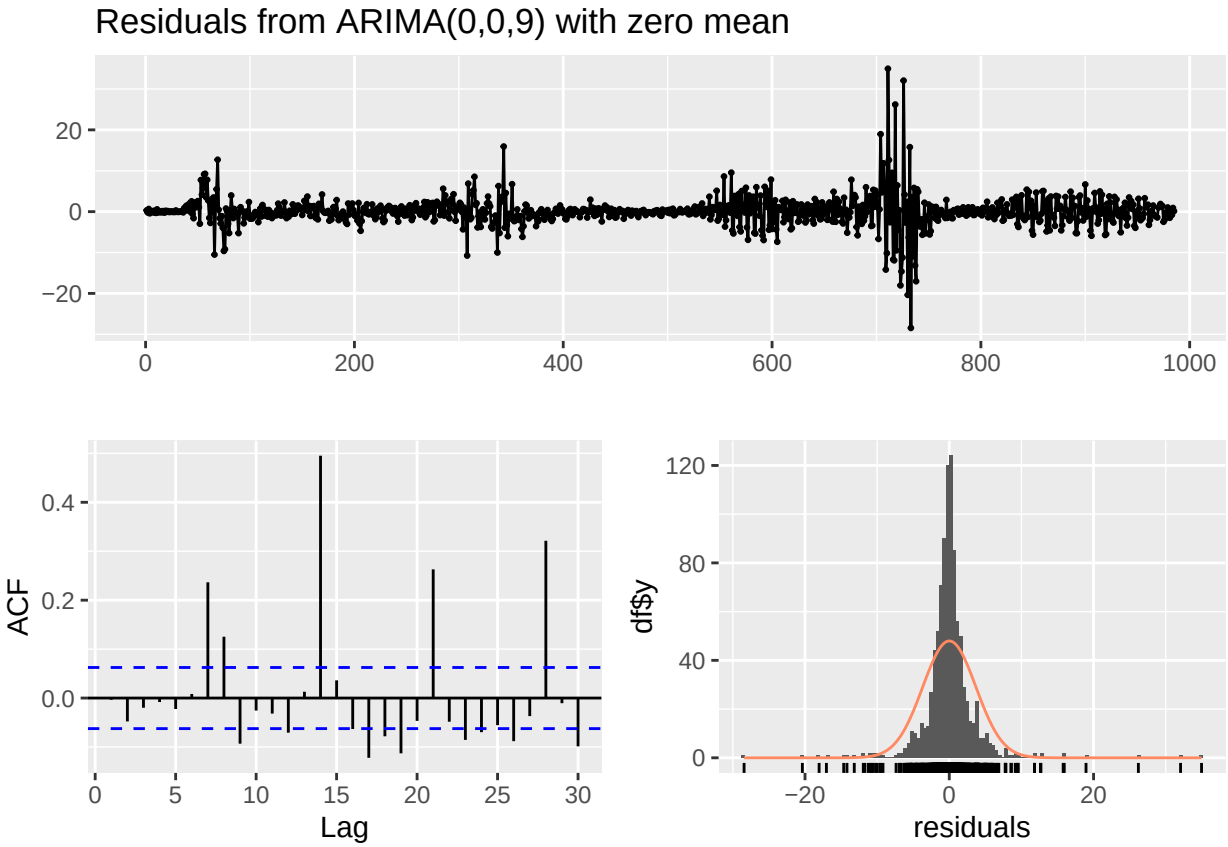
```
## Series: y
## ARIMA(0,0,4) with zero mean
##
## Coefficients:
##           ma1      ma2     ma3     ma4
##       -0.6927  -0.2875  0.0284  0.3914
## s.e.   0.0291   0.0349  0.0468  0.0347
##
## sigma^2 = 17.23:  log likelihood = -2798.68
## AIC=5607.35   AICc=5607.41   BIC=5631.82
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE       ACF1
## Training set 0.001857826 4.142563 2.436954 NaN  Inf 0.5286952 0.01289817
```

auto.arima() MA(4) degrees of freedom: number of parameters, df(MA(4)) = 4

MA(9) degrees of freedom: number of parameters, df(MA(9)) = 9

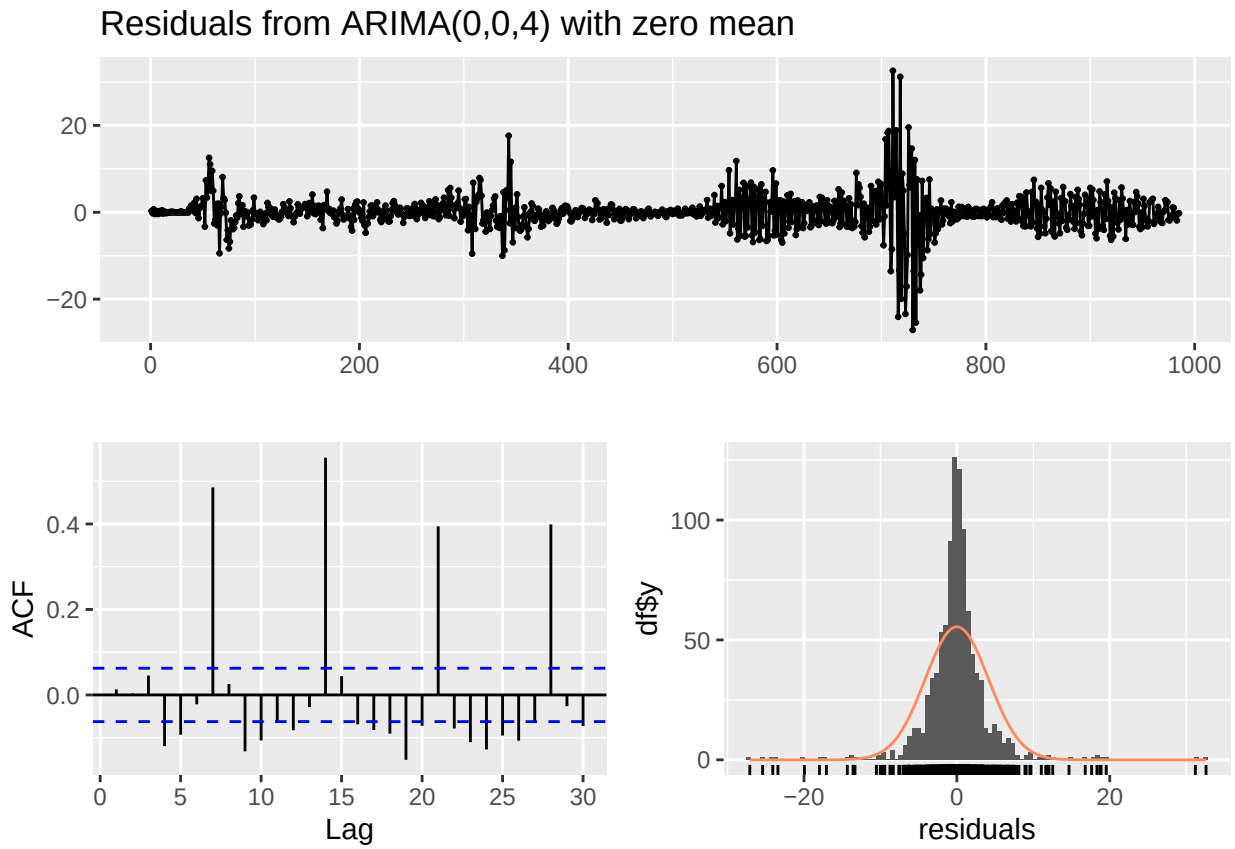
Comparing models:

MA(9) Check residuals:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,9) with zero mean
## Q* = 588.77, df = 21, p-value < 2.2e-16
##
## Model df: 9.   Total lags used: 30
```

MA(4) Check residuals:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,4) with zero mean
## Q* = 1043.3, df = 26, p-value < 2.2e-16
##
## Model df: 4.   Total lags used: 30
```

From the plots above, plots of MA(9) and MA(4) look similar, both showing zero mean and non-constant variance. The MA(4) ACF plot looks to be more stationary with more lags statistically 0 up to 30 lags.

MA(9) T-Test for Mean 0:

```
##
##  One Sample t-test
##
## data:  c
## t = 0.031885, df = 984, p-value = 0.9746
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2266527  0.2341396
## sample estimates:
##   mean of x
## 0.003743469
##
## [1] "T-Test: mean is zero, removed linear trend -> fail to reject H0"
```

```
## [1] TRUE
```

MA(4) T-Test for Mean 0:

```
##
##  One Sample t-test
##
## data:  c
## t = 0.014068, df = 984, p-value = 0.9888
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2572939  0.2610095
## sample estimates:
##   mean of x
## 0.001857826
##
## [1] "T-Test: mean is zero, removed linear trend -> fail to reject H0"
```

```
## [1] TRUE
```

Both MA(9) and MA(4) have mean zero since their respective 95% CI contains zero.

MA(9) Skewness and Kurtosis:

```
##     skew   lwr.ci   upr.ci
## 1.182903 1.240928 1.373842
```

```
## [1] FALSE
```

```
##     kurt   lwr.ci   upr.ci
## 23.35911 24.26973 24.92591
```

```
## [1] FALSE
```

MA(4) Skewness and Kurtosis:

```
##      skew    lwr.ci    upr.ci
## 0.2942145 0.2968289 0.3882439
```

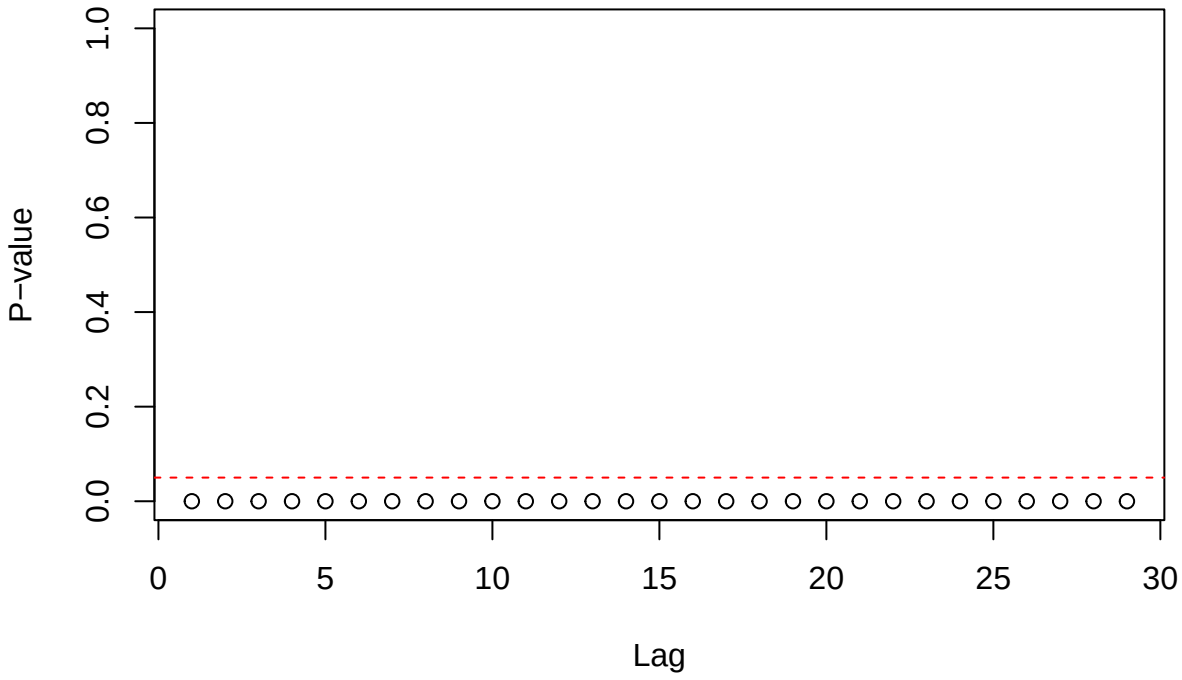
```
## [1] FALSE
```

```
##     kurt   lwr.ci   upr.ci
## 15.09061 15.40950 15.73647
```

```
## [1] FALSE
```

MA(9) tends to skew more right compared to MA(4) but since both models do not have 0 skewness or 0 (excess) Kurtosis, they both do not have normal distribution and thus both do not conform to a Gaussian PDF.

MA(9) Constant Variance:




```
## [1] FALSE

##
##  studentized Breusch-Pagan test
##
## data:  lm(c ~ seq(1, length(c)))
## BP = 7.6064, df = 1, p-value = 0.005816

##    BP
## FALSE
```

MA(4) Constant Variance:

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(c ~ seq(1, length(c)))
## BP = 13.262, df = 1, p-value = 0.0002708

##    BP
## FALSE
```

As we've seen in the respective plots, both MA(9) and MA(4) exhibit non-constant variance. Using the McLeod-Li and Breusch-Pagan tests formally confirms it.

Linear trend - MA(9):

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 7 lags.
##
## Value of test-statistic is: 0.035
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216

## [1] TRUE

##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -5.9687
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:46:55 2023 by user: Reed

##
## FALSE
```

Linear trend - MA(4):

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 7 lags.
##
## Value of test-statistic is: 0.0297
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216

## [1] TRUE

##
## Title:
```

```
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -5.9687
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:46:55 2023 by user: Reed

##
## FALSE
```

The KPSS and ADF tests for MA(9) and MA(4) show no unit roots and no linear trends for both models and therefore are respectively trend stationary and random walk stationary.

M(9) Lag independence:

```
##
##  Box-Ljung test
##
## data:  c
## X-squared = 588.77, df = 30, p-value < 2.2e-16

## [1] FALSE
```

M(4) Lag independence:

```
##
##  Box-Ljung test
##
## data:  c
## X-squared = 1043.3, df = 30, p-value < 2.2e-16

## [1] FALSE
```

Via the Box-Ljung test, both MA(9) and MA(4) exhibit lag dependency over their respective number of lags.

The tests performed above show that MA(9) and MA(4) are similar in their respects.

We will compare their business cycles in section 2.3 and forecasts in sections 2.4 and 2.5.

2.3. Construct a MA() model from either part 2.1. or part 2.2. Perform model checking to validate the fitted model. Interpret the diagnostics. What are the business cycles in the Covid data? What do they mean?

MA(9) Business Cycles:

```
## [1]  5.593  2.251  3.066 23.207
```

MA(4) Business Cycles:

```
## [1] 3.454
```

MA(9) has 4 business cycles: 5.5-, 2-, 3-, and 23-month cycles, while MA(4) only has a 3.5-month business cycle. The less business cycles favor MA(4) as it shows more stationarity than MA(9).

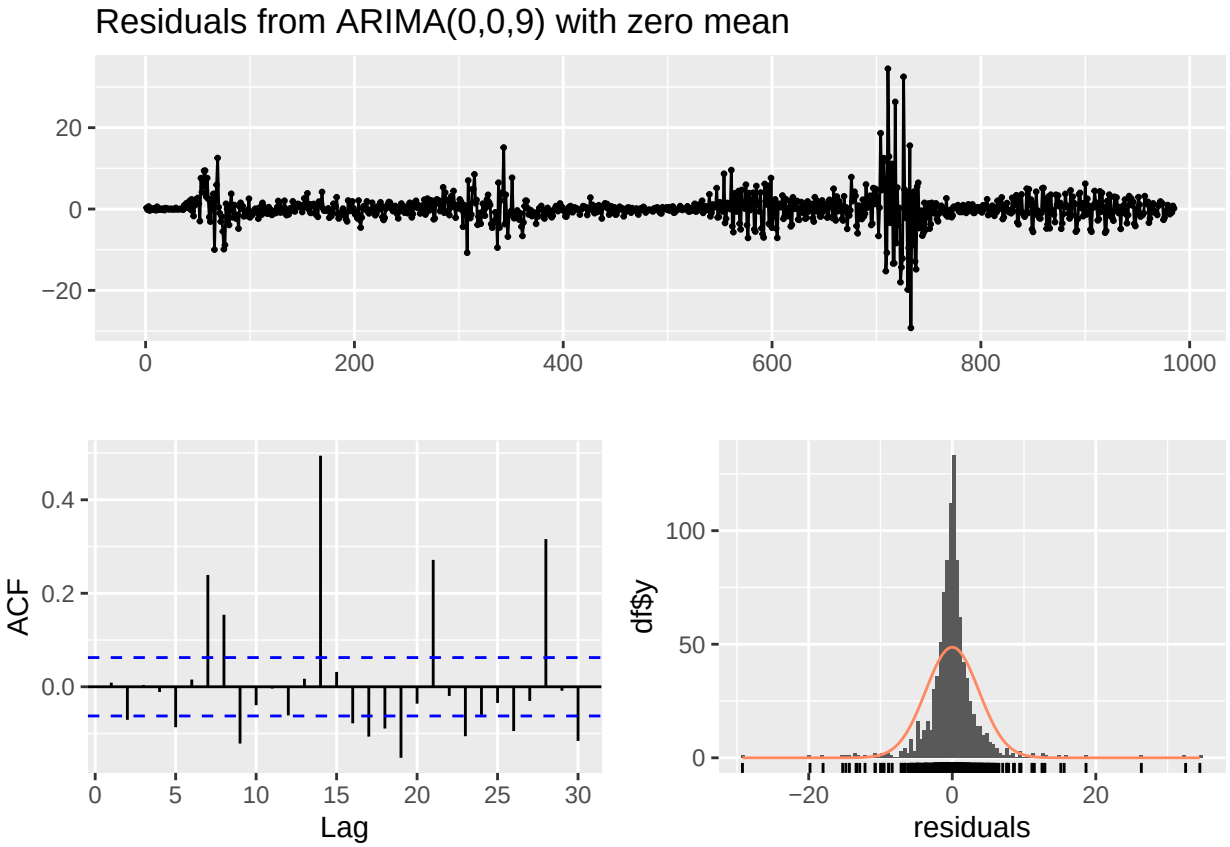
We can attempt to reduce the MA(9) model with the parameter test and identify statistically insignificant components:

```
##               t       pval_t       pval_z Pr(>|t|)
## ma1 -19.5910183 0.000000e+00 0.000000e+00      ***
## ma2  -2.1626508 3.081034e-02 3.056805e-02        *
## ma3   2.9406213 3.352846e-03 3.275547e-03       **
## ma4  -0.7717343 4.404586e-01 4.402718e-01
## ma5  -1.8530877 6.417146e-02 6.386979e-02        .
## ma6   5.8900168 5.309513e-09 3.861564e-09      ***
## ma7   9.2588816 0.000000e+00 0.000000e+00      ***
## ma8  -8.0819747 1.776357e-15 6.661338e-16      ***
## ma9  -0.6033663 5.464052e-01 5.462650e-01
```

From the parameter test, we have ma4, ma5,and ma9 that are statistically insignificant, and will remove them from the model, thus creating a MA() model reduced to 6 components.

```
## Series: y
## ARIMA(0,0,9) with zero mean
##
## Coefficients:
##           ma1      ma2     ma3  ma4  ma5     ma6     ma7      ma8  ma9
##       -0.6927  -0.0502  0.0705    0    0  0.2072  0.3195  -0.3854    0
## s.e.   0.0290   0.0373  0.0302    0    0  0.0286  0.0323   0.0280    0
##
## sigma^2 = 13.77:  log likelihood = -2687.64
## AIC=5389.28   AICc=5389.4   BIC=5423.53
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set 0.00382086 3.699156 2.056041 NaN  Inf 0.4460563 0.008793946
```

Reduced MA() Model: check residuals




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,9) with zero mean
## Q* = 625.16, df = 21, p-value < 2.2e-16
##
## Model df: 9.   Total lags used: 30
```

The residuals from the reduced model show a mean zero with non-constant variance. The distribution's high kurtosis and slight right skewness might not make it normal based on a Gaussian PDF. The ACF might show some stationarity but also shows a 6-lag peak-to-peak cycle. We will test the stationarity below.

Reduced MA() Model: t-test for mean 0

```
##
##  One Sample t-test
##
## data:  c
## t = 0.032401, df = 984, p-value = 0.9742
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2275920  0.2352337
## sample estimates:
##  mean of x
## 0.00382086
##
```

```
## [1] "T-Test: mean is zero, removed linear trend -> fail to reject H0"
```

```
## [1] TRUE
```

From the t-test, the reduced MA() model's 95% CI contain zero, therefore can say the mean of the reduced model's is statistically zero.

Reduced MA() Model: normalcy test via Skewness and Kurtosis

```
##     skew   lwr.ci   upr.ci
## 1.124993 1.145734 1.279481
```

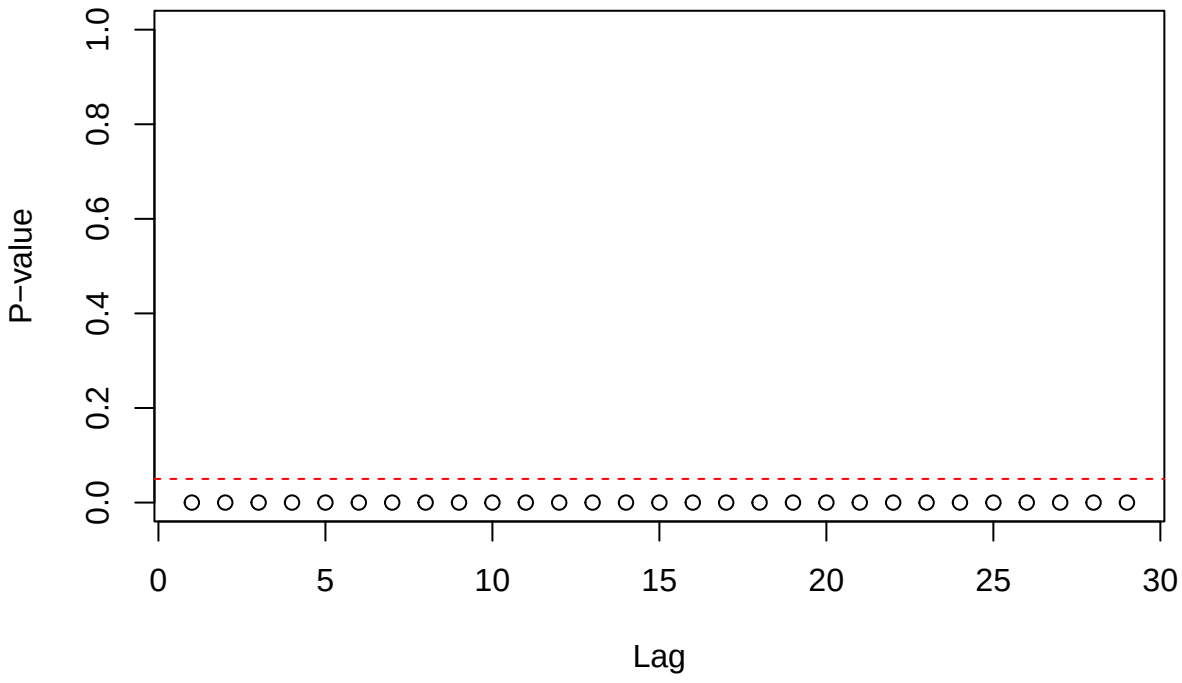
```
## [1] FALSE
```

```
##     kurt   lwr.ci   upr.ci
## 23.11977 24.04455 24.64115
```

```
## [1] FALSE
```

The skewness and kurtosis high 95% confidence intervals show that is they are not zero, therefore the distribution is not normal based on a Gaussian PDF.

Reduced MA() Model: Constant variance




```
## [1] FALSE
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(c ~ seq(1, length(c)))
## BP = 7.7679, df = 1, p-value = 0.005318

##    BP
## FALSE
```

The McLeod-Li and Breusch-Pagan tests show that the reduced MA() model's residuals has non-constant variance, which we've observed in the plot.

Reduced MA() Model: Linear Trend

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 7 lags.
##
## Value of test-statistic is: 0.0348
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216

## [1] TRUE

##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -5.9687
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:46:56 2023 by user: Reed

##
## FALSE
```

The KPSS and ADF tests show that the reduced MA() model is stationary.

Reduced MA() Model: Lag Independence

```
##
##  Box-Ljung test
##
## data:  c
## X-squared = 625.16, df = 30, p-value < 2.2e-16
```

```
## [1] FALSE
```

```
## [1]  5.481  2.248  3.076 22.815
```

The reduced MA() model has 4 business cycles: 5.5-, 2-, 3-, and 23-month cycles, similar to the full MA(9) model. In the spirit of parsimony, I would prefer this reduced model over the full MA(9) since it is a simpler model and the model diagnostics have similar results.

Comparing the reduced model to the auto.arima() MA(4) model, I would prefer the MA(4) model as the model diagnostics are similar to the reduced model but only has one business cycle and more stationarity in the residuals, thus favoring it via parsimony.

2.4. Obtain 1-step to 7-step ahead points with 95% interval forecasts for the total cases data using the model you chose in part 2.3.

```
##     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 988       96546145 96399408 96692998 96321776 96770784
## 989       96594783 96290547 96899514 96129694 97061030
## 990       96623769 96118625 97130279 95851771 97398963
## 991       96624120 95904584 97346431 95524809 97729923
## 992       96624120 95687886 97565059 95194178 98065067
## 993       96624120 95512915 97741959 94927364 98336394
## 994       96624120 95362158 97894645 94697581 98570689
```

Using the auto-arima MA(4) model:

- total_cases forecast for the next 7 days on the lower 95% CI:

96321776, 96129694, 95851771, 95524809, 95194178, 94927364, 94697581

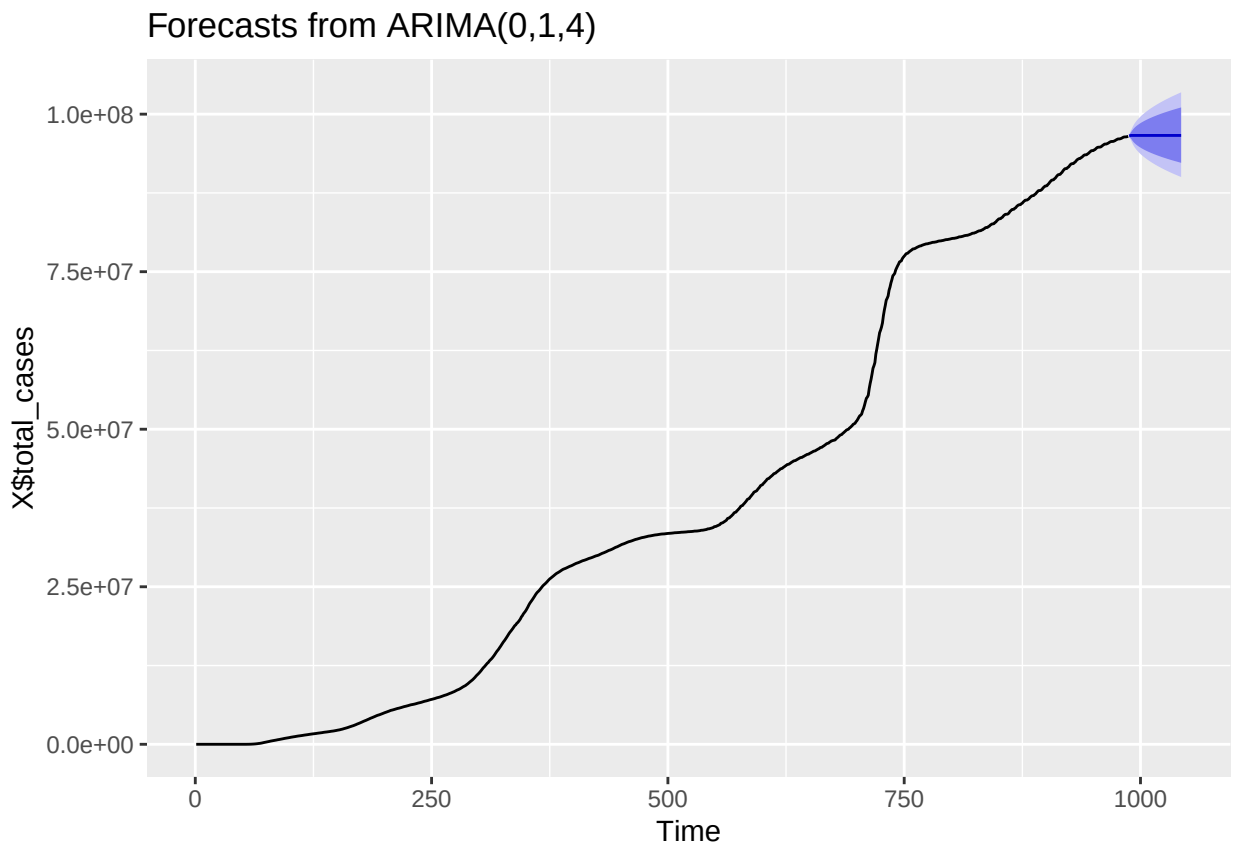
- total_cases forecast for the next 7 days on the upper 95% CI:

96770784, 97061030, 97398963, 97729923, 98065067, 98336394, 98570689

I would take the lower 95% CI forecast with a grain of salt to the number of total_cases decreasing. total_cases is a cumulative sum that is ever-increasing and would not go lower.

2.5. Forecast total cases with the forecast origin the last observed data point using the model you chose in part 2.3. Interpret.

Using the auto-arima MA(4) model, we will generate a forecast plot for the next eight weeks:



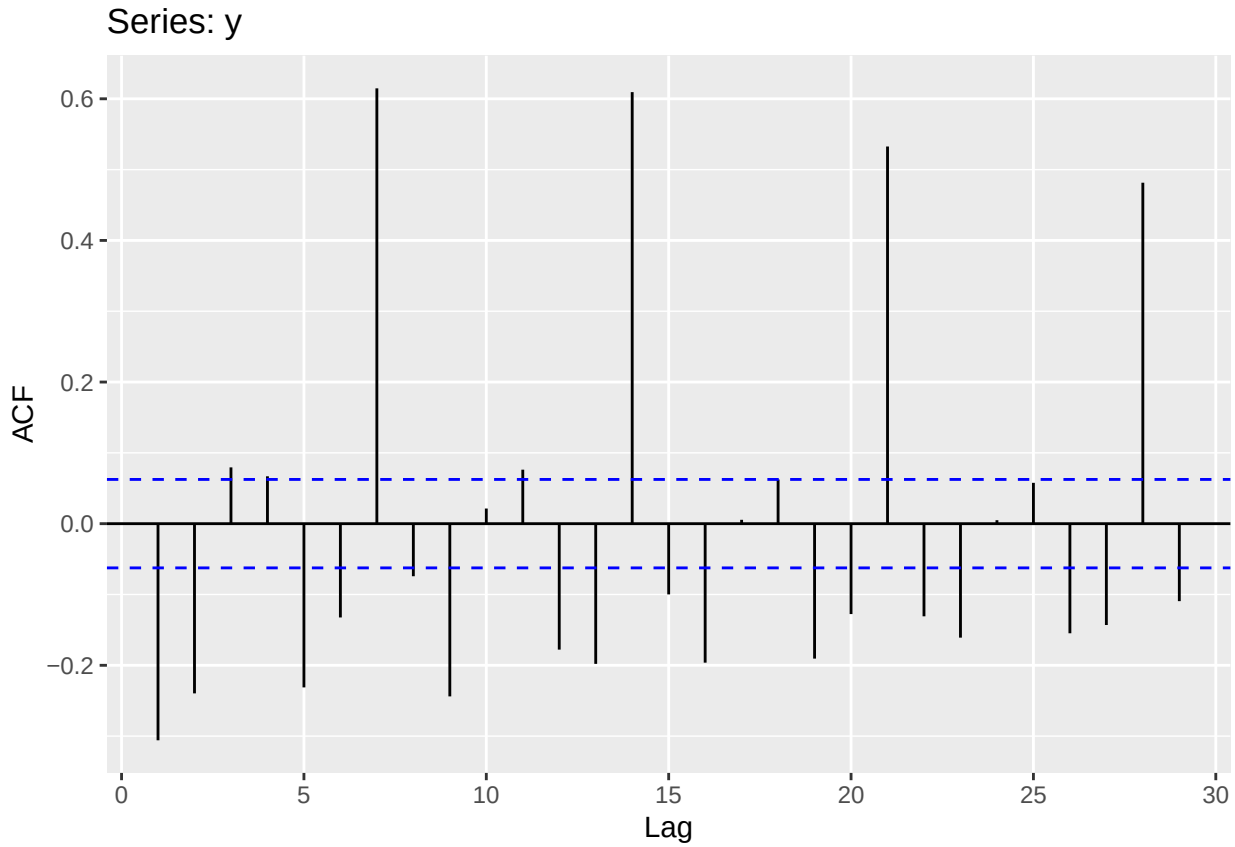
From the forecast plot above, the MA(4) model's dark blue prediction line does not show much of a change from the last data point in the time series, while the 80% CI range shown by the blue area spreads out over an eight week period. The 95% CI range shown by the light blue area has a wider spread.

As explained in section 2.4, I would take the lower CI section forecast with a grain of salt to the number of total_cases decreasing. total_cases is a cumulative sum that is ever-increasing and would not go lower.

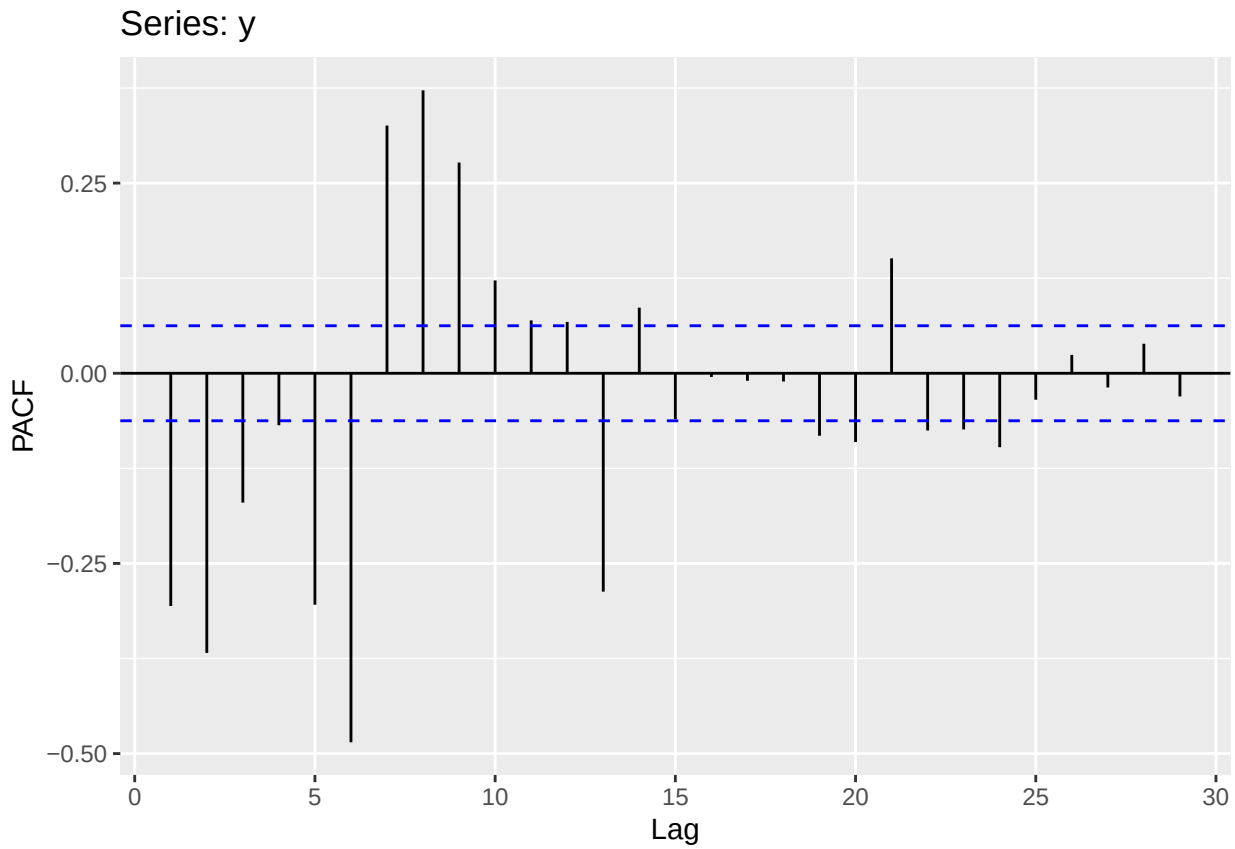
Overall, while this was a good exercise in understanding MA() models, I would not recommend using the MA(4), MA(9), or the reduced MA() model to predict the growth of total covid cases as the prediction line is a straight horizontal line. Hopefully ARMA() models in the next section can show some improvement.

3. Autoregressive Moving Average (ARMA) Models (20 points)

3.1. Use the ACF and the PACF to determine the order to fit an ARMA model to the Covid data. Justify your choice of order.



As shown in section 2.1, from the rule of thumb document, with an ACF plot 'the function drops off to 0 after lag q (i.e. D(q)).' Based on the ACF plot above, the dropoff to 0 is at lag 10; lag 10 is after lag 9 (with q=9, D(q)=D(9)), therefore we could propose q=9 in the ARMA() model.



From the rule of thumb document, with an PACF plot 'the function drops off to 0 after lag p (i.e. D(p)).' Based on the ACF plot above, the dropoff to 0 is at lag 15; lag 15 is after lag 14 (with p=14, D(q)=D(14)).

Therefore we could propose an ARMA(14,9) model.

A summary of ARMA(14,9) is as follows:

```
## Series: y
## ARIMA(14,1,9)
##
## Coefficients:
##           ar1      ar2      ar3     ar4     ar5     ar6     ar7     ar8     ar9
##       -0.5834  -0.2908  -0.0109  0.0367  0.1843  0.3743  0.9635  0.4234  0.1786
## s.e.   0.1230   0.1400   0.1331  0.1294  0.1224  0.1211  0.1010  0.1082  0.1092
##          ar10     ar11     ar12     ar13     ar14      ma1      ma2     ma3
##       -0.0929  -0.1326  -0.2420  -0.4112  -0.1420  -1.1273  -0.0352  0.0693
## s.e.   0.0960   0.0924   0.0909   0.0927   0.0793   0.1194   0.1305  0.1588
##          ma4      ma5     ma6      ma7     ma8      ma9
##       0.1292  -0.1883  0.0299  -0.1374  0.6378  -0.3779
## s.e.  0.1706   0.1705  0.1729   0.1979  0.1683   0.0751
##
## sigma^2 = 8.858:  log likelihood = -2465.41
## AIC=4978.82   AICc=4980.07   BIC=5096.22
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set -0.05874896 2.939736 1.516353 NaN  Inf 0.3289716 -0.00859798
```

3.2. Use the command auto.arima(y, $d = 0$, $max.p = 0$, $stationary = TRUE$), where y is your stationary *total_cases* time series, to find the AR and MA orders. Interpret and compare with your ACF and PACF choices. Give the model degrees of freedom (df).

auto.arima() proposes using an ARMA(2,2) model.

```
am2 <- auto.arima(y,d = 1,max.p=lags, max.q=lags,stationary=TRUE)
summary(am2)
```

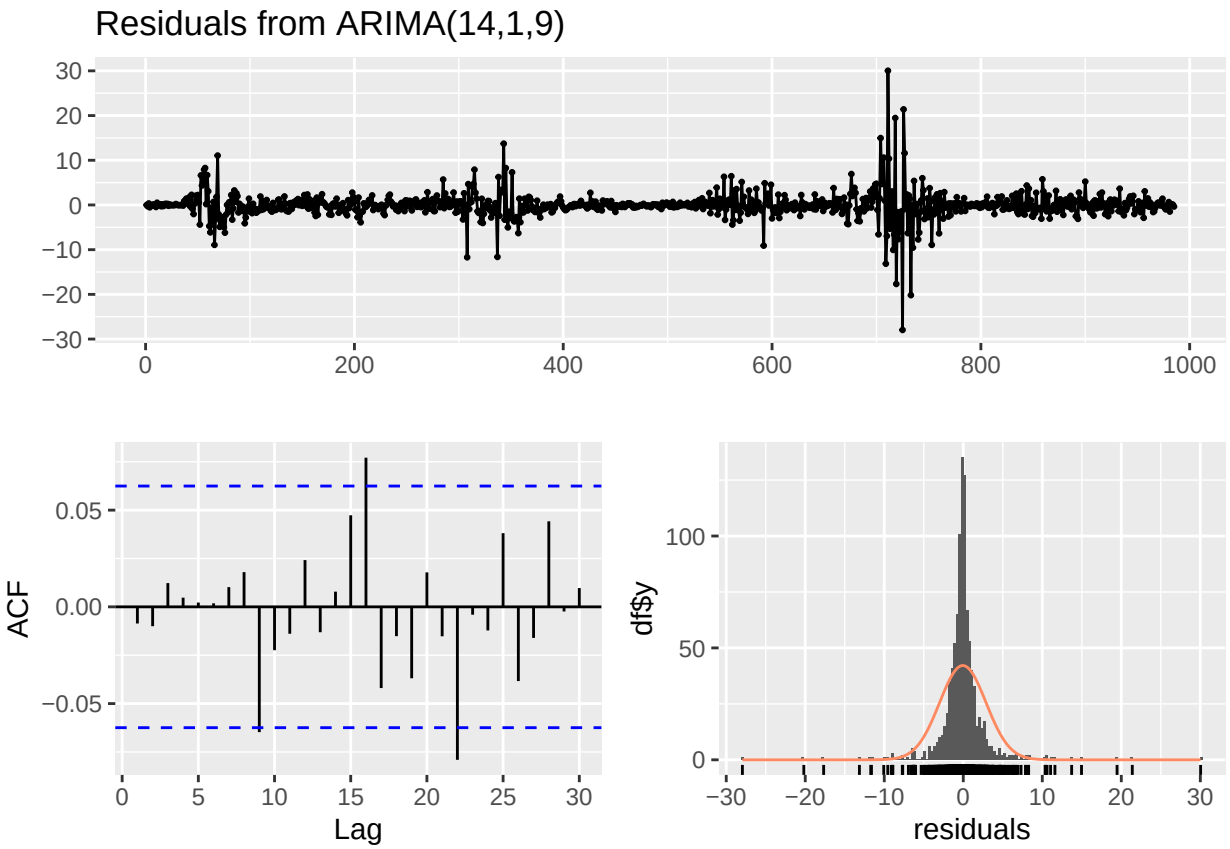
```
## Series: y
## ARIMA(2,0,2) with zero mean
##
## Coefficients:
##          ar1      ar2      ma1     ma2
##       0.9511  -0.5466  -1.6101  0.8736
## s.e.  0.0307   0.0351   0.0167  0.0171
##
## sigma^2 = 16.7:  log likelihood = -2783.39
## AIC=5576.79   AICc=5576.85   BIC=5601.25
##
## Training set error measures:
##                       ME    RMSE      MAE MPE MAPE      MASE       ACF1
## Training set 0.003354573 4.07816 2.342763 NaN  Inf 0.5082604 -0.0583303
```

ARMA(2,2) degrees of freedom: number of parameters p+d+q = 2 + 0 + 2 = 4

ARMA(14,1,9) degrees of freedom: number of parameters p+d+q = 14 + 1 + 9 = 24

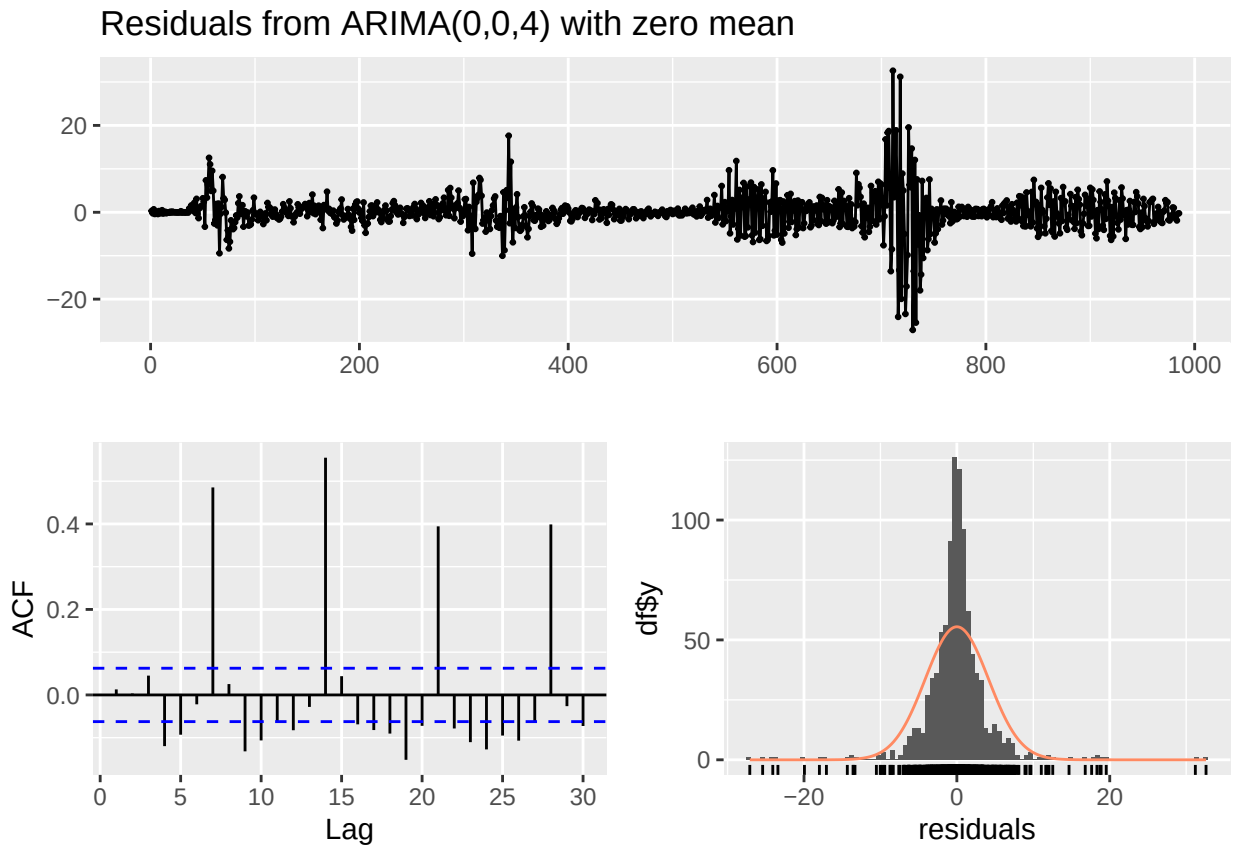
Comparing models:

ARMA(14,9) Check residuals:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(14,1,9)
## Q* = 30.377, df = 7, p-value = 8.096e-05
##
## Model df: 23.   Total lags used: 30
```

ARMA(2,2) Check residuals:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,4) with zero mean
## Q* = 1043.3, df = 26, p-value < 2.2e-16
##
## Model df: 4.   Total lags used: 30
```

From the plots above, plots of ARMA(14,9) and ARMA(2,2) look similar, both showing zero mean and non-constant variance. The ARMA(14,9) ACF plot looks to be more stationary with more lags statistically 0 up to 30 lags.

ARMA(14,9) T-Test for Mean 0:

```
##
##  One Sample t-test
##
## data:  c
## t = -0.62701, df = 984, p-value = 0.5308
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2426171  0.1251192
## sample estimates:
##   mean of x
## -0.05874896
##
```

```
## [1] "T-Test: mean is zero, removed linear trend -> fail to reject H0"
```

```
## [1] TRUE
```

ARMA(2,2) T-Test for Mean 0:

```
##
##  One Sample t-test
##
## data:  c
## t = 0.025803, df = 984, p-value = 0.9794
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2517681  0.2584773
## sample estimates:
##   mean of x
## 0.003354573
##
## [1] "T-Test: mean is zero, removed linear trend -> fail to reject H0"
```

```
## [1] TRUE
```

Both ARMA(14,9) and ARMA(2,2) have mean zero since their respective 95% CI contains zero.

ARMA(14,9) Skewness and Kurtosis:

```
##      skew    lwr.ci    upr.ci
## 0.5738201 0.5177015 0.6987736
```

```
## [1] FALSE
```

```
##     kurt   lwr.ci   upr.ci
## 29.38400 30.64198 31.43417
```

```
## [1] FALSE
```

ARMA(2,2) Skewness and Kurtosis:

```
##      skew    lwr.ci    upr.ci
## 0.3050378 0.2600036 0.3697134
```

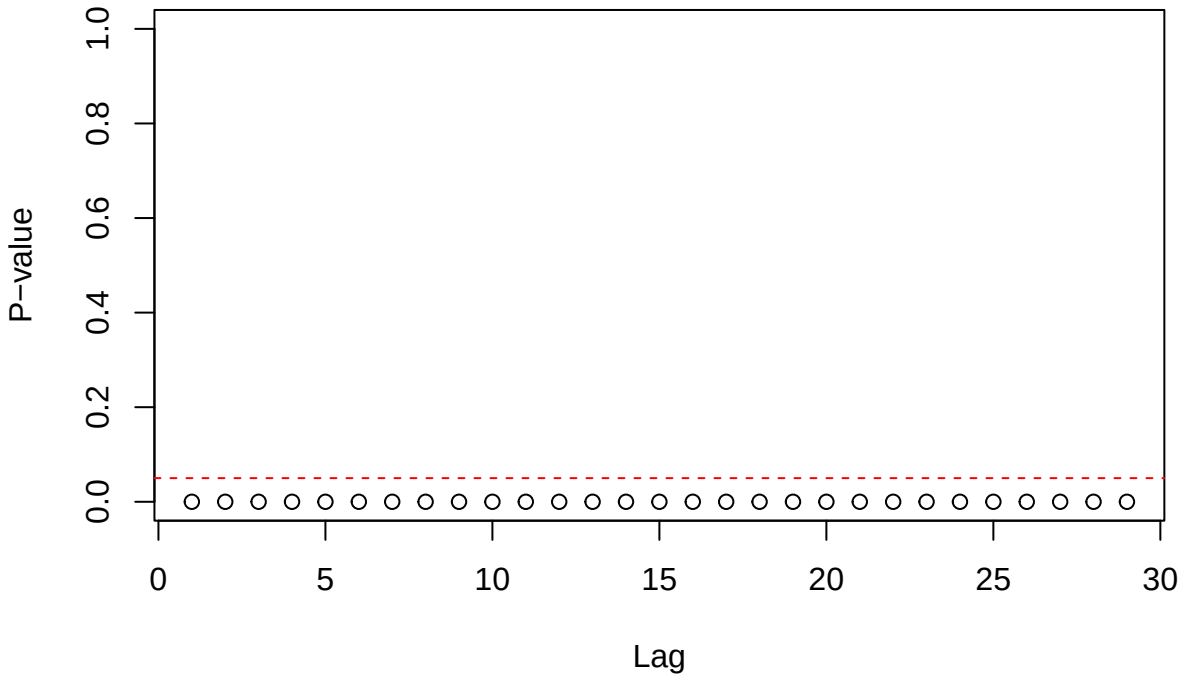
```
## [1] FALSE
```

```
##     kurt   lwr.ci   upr.ci
## 16.77910 17.14482 17.51257
```

```
## [1] FALSE
```

ARMA(14,9) tends to skew more right compared to ARMA(2,2) but since both models do not have 0 skewness or 0 (excess) Kurtosis, they both do not have normal distribution and thus both do not conform to a Gaussian PDF.

ARMA(14,9) Constant Variance:




```
## [1] FALSE
```

ARMA(2,2) Constant Variance:

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(c ~ seq(1, length(c)))
## BP = 11.714, df = 1, p-value = 0.0006202
```

```
##    BP
## FALSE
```

As we've seen in the respective plots, both ARMA(14,9) and ARMA(2,2) exhibit non-constant variance. Using the McLeod-Li method for ARMA(14,9) and Breusch-Pagan for ARMA(2,2) formally confirms it.

Linear trend - ARMA(14,9):

```
##
## ######################
```

```
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 7 lags.
##
## Value of test-statistic is: 0.0225
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216

## [1] TRUE

##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -5.9687
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:47:02 2023 by user: Reed

##
## FALSE
```

Linear trend - ARMA(2,2):

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 7 lags.
##
## Value of test-statistic is: 0.0324
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216

## [1] TRUE

##
## Title:
##  Augmented Dickey-Fuller Test
##
```

```
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -5.9687
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:47:02 2023 by user: Reed

##
## FALSE
```

The KPSS and ADF tests for ARMA(14,9) and ARMA(2,2) show no unit roots and no linear trends for both models and therefore are respectively trend stationary and random walk stationary.

ARMA(14,9) Lag independence:

```
##
##  Box-Ljung test
##
## data:  c
## X-squared = 30.377, df = 30, p-value = 0.4465

## [1] TRUE
```

ARMA(2,2) Lag independence:

```
##
##  Box-Ljung test
##
## data:  c
## X-squared = 1111, df = 30, p-value < 2.2e-16

## [1] FALSE
```

Via the Box-Ljung test, both ARMA(14,9) and ARMA(2,2) exhibit lag dependency over their respective number of lags.

The tests performed show that ARMA(14,9) and ARMA(2,2) are similar in those respects. We will compare their business cycles in section 3.3 and their forecasts in sections 3.4 and 3.5.

3.3. Construct a ARMA() model from either part 3.1. or part 3.2. Perform model checking to validate the fitted model. Interpret the diagnostics. What are the business cycles in the Covid data? What do they mean?

ARMA(14,9) Business Cycles:

```
##  [1]  5.780  2.726 40.727  4.172  2.278  3.384  3.241  8.455  6.740  2.362
## [11] 27.691
```

ARMA(2,2) Business Cycles:

```
## [1] 6.329
```

ARMA(14,9) has 8 business cycles: 2-, 3-, 4-, 6-, 7-, 8-, 28- and 41-month cycles.

ARMA(2,2) only has a single 6-month business cycle, which makes its residuals more stationary than ARMA(14,9).

We can attempt to reduce the ARMA(14,9) model with the parameter test and identify statistically insignificant components:

```
##                t       pval_t       pval_z Pr(>|t|)
## ar1  -4.74436171 2.408776e-06 2.091648e-06      ***
## ar2  -2.07792513 3.798124e-02 3.771626e-02        *
## ar3  -0.08212891 9.345613e-01 9.345442e-01
## ar4   0.28326949 7.770313e-01 7.769703e-01
## ar5   1.50615154 1.323569e-01 1.320283e-01
## ar6   3.09117176 2.051129e-03 1.993683e-03       **
## ar7   9.53585918 0.000000e+00 0.000000e+00      ***
## ar8   3.91219878 9.789830e-05 9.145958e-05      ***
## ar9   1.63634969 1.020938e-01 1.017664e-01
## ar10 -0.96757683 3.334992e-01 3.332558e-01
## ar11 -1.43405269 1.518825e-01 1.515572e-01
## ar12 -2.66153015 7.908432e-03 7.778637e-03       **
## ar13 -4.43640918 1.020438e-05 9.147190e-06      ***
## ar14 -1.79070024 7.365607e-02 7.334141e-02        .
## ma1  -9.44522057 0.000000e+00 0.000000e+00      ***
## ma2  -0.26942614 7.876596e-01 7.876018e-01
## ma3   0.43634306 6.626859e-01 6.625878e-01
## ma4   0.75714400 4.491492e-01 4.489636e-01
## ma5  -1.10457534 2.696201e-01 2.693436e-01
## ma6   0.17276408 8.628733e-01 8.628369e-01
## ma7  -0.69453188 4.875166e-01 4.873487e-01
## ma8   3.79011055 1.599566e-04 1.505803e-04      ***
## ma9  -5.03036860 5.840402e-07 4.895378e-07      ***
```

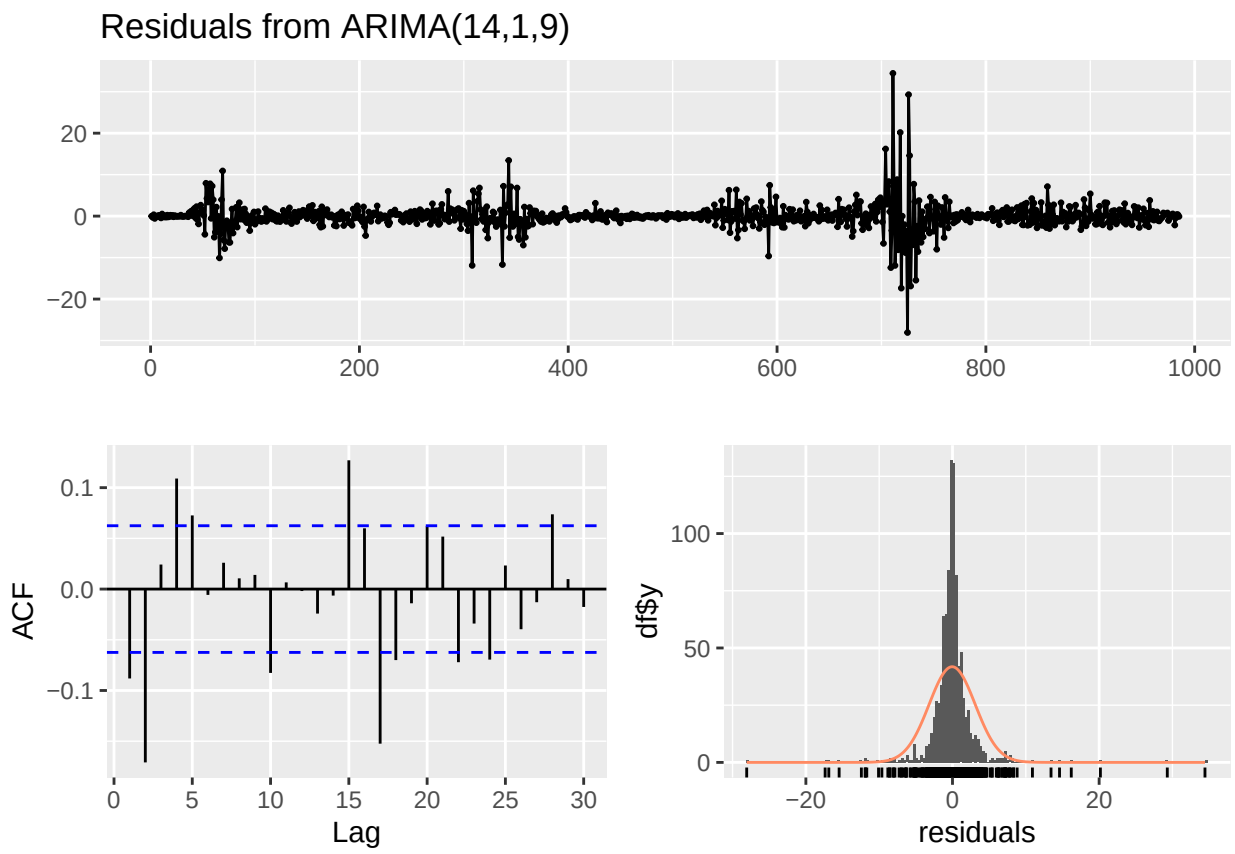
From the parameter test, we have ar3, ar4, ar5, ar6, ar9, ar10, ar11, ar14, ma2, ma3, ma4, ma5, ma6, and ma7 that are statistically insignificant, and will remove them from the model, thus creating an ARMA() model reduced from 23 components to 10 components.

Here is the summary of the reduced ARMA() model:

```
## Series: y
## ARIMA(14,1,9)
##
## Coefficients:
##           ar1      ar2  ar3  ar4  ar5     ar6     ar7     ar8  ar9  ar10  ar11
##       -0.4740  -0.1491    0    0    0  0.2367  0.7084  0.2539    0     0     0
## s.e.   0.0299   0.0220    0    0    0  0.0307  0.0265  0.0379    0     0     0
##          ar12     ar13  ar14      ma1  ma2  ma3  ma4  ma5  ma6  ma7     ma8
##       -0.0581  -0.3063     0  -1.0988    0    0    0    0    0    0  0.4356
## s.e.   0.0205   0.0292     0   0.0107    0    0    0    0    0    0  0.0480
##           ma9
##       -0.3365
## s.e.   0.0442
##
```

```
## sigma^2 = 9.978:  log likelihood = -2529.65
## AIC=5081.29   AICc=5081.57   BIC=5135.1
##
## Training set error measures:
##                       ME    RMSE     MAE MPE MAPE     MASE        ACF1
## Training set -0.04646548 3.14113 1.60474 NaN  Inf 0.348147 -0.08814631
```

Reduced ARMA() Check Residuals:




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(14,1,9)
## Q* = 136.97, df = 7, p-value < 2.2e-16
##
## Model df: 23.   Total lags used: 30
```

The reduced ARMA() model have independent lags from the Box-Ljung test and show no autocorrelation, similar to the full ARMA(14,9) model's diagnostics in section 3.2. The ACF for the full model shows more stationarity compared to the reduced model but we can further test it. The plot of the residuals for the reduced model shows a mean 0 with non-constant variance. The distribution plot shows a highly-peaked plot indicating very high Kurtosis and the positive outliers indicating some right skewness, therefore the reduced model does not meet normalcy under a Gaussian PDF.

Reduced ARMA() t-test mean of 0:

```
##
##  One Sample t-test
##
## data:  c
## t = -0.46408, df = 984, p-value = 0.6427
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2429478  0.1500168
## sample estimates:
##   mean of x
## -0.04646548
##
## [1] "T-Test: mean is zero, removed linear trend -> fail to reject H0"
```

```
## [1] TRUE
```

The 95% CI of the t-test contains zero, showing that the mean of the residuals of the reduced ARMA() model is statistically 0. With the mean zero, we've also removed the linear trend.

Reduced ARMA() normal tests: skewness, Kurtosis

```
##     skew   lwr.ci   upr.ci
## 1.412731 1.544372 1.737357
```

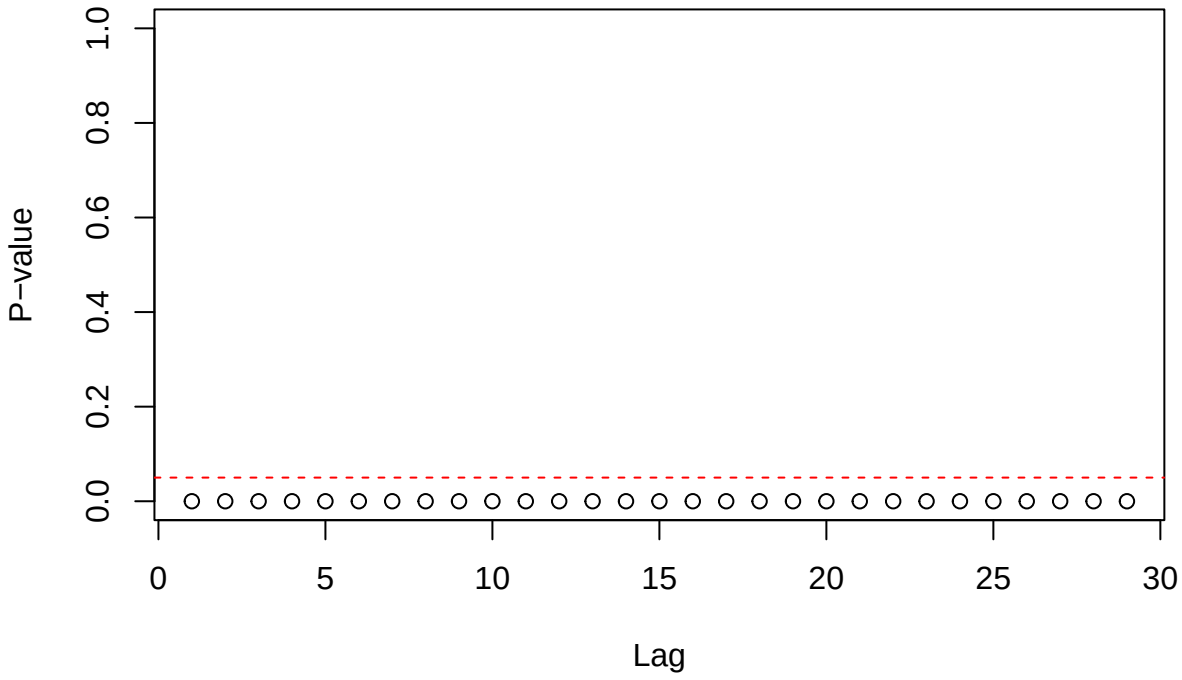
```
## [1] FALSE
```

```
##     kurt   lwr.ci   upr.ci
## 34.01539 35.89428 36.92409
```

```
## [1] FALSE
```

While the distribution plot showed Skewness and Kurtosis, we formally calculate its presence, thus the model is not normal to a Gausian PDF.

Reduced ARMA() constant variance:




```
## [1] FALSE
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(c ~ seq(1, length(c)))
## BP = 3.8678, df = 1, p-value = 0.04922
```

```
##    BP
## FALSE
```

We've observed that the plot of the reduced ARMA() model shows non-constant variance, and the McLeod-Li and Breusch-Pagan tests formally confirm it.

Reduced ARMA() linear trend test:

```
##
## ########################
## # KPSS Unit Root Test #
## ########################
##
## Test is of type: tau with 7 lags.
##
## Value of test-statistic is: 0.0202
##
```

```
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.119 0.146  0.176 0.216
```

```
## [1] TRUE
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 30
##   STATISTIC:
##     Dickey-Fuller: -5.9687
##   P VALUE:
##     0.01
##
## Description:
##  Tue Mar 28 19:47:03 2023 by user: Reed
```

```
##
## FALSE
```

While the full ARMA(14,9) model's ACF looks to be more stationary than the reduced ARMA() model, the KPSS and ADF tests show confirm the reduced model is stationary as well.

Reduced ARMA() business cycles

```
##  [1]  5.961  2.318  2.717  8.992  3.365 36.943  4.171  2.289  6.115  3.263
## [11] 25.143
```

The reduced ARMA() model has 7 business cycles: 2-,3-,4-, 6-, 9-, 25-, and 37-month cycles, one less than the full ARMA(14,9) model. In the spirit of parsimony, I would prefer this reduced model over the full ARMA(14,9) since it is a much simpler model and the model diagnostics have similar results, and one less business cycle. The less business cycles the more stationary the model's residuals are.

Comparing the ARMA(14,9) model with the reduced ARMA() model: AIC/BIC/RMSE/MAE

```
## Series: y
## ARIMA(14,1,9)
##
## Coefficients:
##           ar1      ar2      ar3     ar4     ar5     ar6     ar7     ar8     ar9
##       -0.5834  -0.2908  -0.0109  0.0367  0.1843  0.3743  0.9635  0.4234  0.1786
## s.e.   0.1230   0.1400   0.1331  0.1294  0.1224  0.1211  0.1010  0.1082  0.1092
##          ar10     ar11     ar12     ar13     ar14      ma1      ma2     ma3
##       -0.0929  -0.1326  -0.2420  -0.4112  -0.1420  -1.1273  -0.0352  0.0693
## s.e.   0.0960   0.0924   0.0909   0.0927   0.0793   0.1194   0.1305  0.1588
##          ma4      ma5     ma6      ma7     ma8      ma9
##       0.1292  -0.1883  0.0299  -0.1374  0.6378  -0.3779
## s.e.  0.1706   0.1705  0.1729   0.1979  0.1683   0.0751
##
```

```
## sigma^2 = 8.858:  log likelihood = -2465.41
## AIC=4978.82   AICc=4980.07   BIC=5096.22
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set -0.05874896 2.939736 1.516353 NaN  Inf 0.3289716 -0.00859798
```

```
## Series: y
## ARIMA(14,1,9)
##
## Coefficients:
##           ar1      ar2  ar3  ar4  ar5     ar6     ar7     ar8  ar9  ar10  ar11
##       -0.4740  -0.1491    0    0    0  0.2367  0.7084  0.2539    0     0     0
## s.e.   0.0299   0.0220    0    0    0  0.0307  0.0265  0.0379    0     0     0
##          ar12     ar13  ar14      ma1  ma2  ma3  ma4  ma5  ma6  ma7     ma8
##       -0.0581  -0.3063     0  -1.0988    0    0    0    0    0    0  0.4356
## s.e.   0.0205   0.0292     0   0.0107    0    0    0    0    0    0  0.0480
##           ma9
##       -0.3365
## s.e.   0.0442
##
## sigma^2 = 9.978:  log likelihood = -2529.65
## AIC=5081.29   AICc=5081.57   BIC=5135.1
##
## Training set error measures:
##                       ME    RMSE     MAE MPE MAPE     MASE        ACF1
## Training set -0.04646548 3.14113 1.60474 NaN  Inf 0.348147 -0.08814631
```

Despite the lower AIC/BIC/RMSE/MAE for the full ARMA(14,9) model over the reduced ARMA() model, as we've seen in the other model diagnostics the models are very similar. More components in a model will always favor a larger model, especially a model that has 13 more components. That being said, the improvements in AIC/BIC/RMSE/MAE for the full model seem to be marginal, especially if it has 13 more components. I would still prefer the reduced model over the full model in the spirit of parsimony as the full model only has marginal gains at the cost of much higher complexity.

Comparing the reduced ARMA() model with the auto-arima configured ARMA(2,2) model:

```
## Series: y
## ARIMA(14,1,9)
##
## Coefficients:
##           ar1      ar2  ar3  ar4  ar5     ar6     ar7     ar8  ar9  ar10  ar11
##       -0.4740  -0.1491    0    0    0  0.2367  0.7084  0.2539    0     0     0
## s.e.   0.0299   0.0220    0    0    0  0.0307  0.0265  0.0379    0     0     0
##          ar12     ar13  ar14      ma1  ma2  ma3  ma4  ma5  ma6  ma7     ma8
##       -0.0581  -0.3063     0  -1.0988    0    0    0    0    0    0  0.4356
## s.e.   0.0205   0.0292     0   0.0107    0    0    0    0    0    0  0.0480
##           ma9
##       -0.3365
## s.e.   0.0442
##
## sigma^2 = 9.978:  log likelihood = -2529.65
## AIC=5081.29   AICc=5081.57   BIC=5135.1
##
```

```
## Training set error measures:
##                        ME    RMSE     MAE MPE MAPE     MASE        ACF1
## Training set -0.04646548 3.14113 1.60474 NaN  Inf 0.348147 -0.08814631
```

```
## Series: y
## ARIMA(2,0,2) with zero mean
##
## Coefficients:
##          ar1      ar2      ma1     ma2
##       0.9511  -0.5466  -1.6101  0.8736
## s.e.  0.0307   0.0351   0.0167  0.0171
##
## sigma^2 = 16.7:  log likelihood = -2783.39
## AIC=5576.79   AICc=5576.85   BIC=5601.25
##
## Training set error measures:
##                        ME    RMSE      MAE MPE MAPE      MASE       ACF1
## Training set 0.003354573 4.07816 2.342763 NaN  Inf 0.5082604 -0.0583303
```

Comparing the reduced ARMA model to the auto-arima configured ARMA(2,2) model, we see that the reduced model's AIC, BIC, RMSE, and MAE are lower than the ARMA(2,2) model, indicating worse performance in the ARMA(2,2) model. We see these differences are not marginal, such as the difference between the AIC of both models is almost 500, almost a 10% difference. While the reduced ARMA() model is more complex with 10 components compared to the ARMA(2,2) model, the AIC, BIC, RMSE, and MAE significantly favor the reduced ARMA() model over the ARMA(2,2) model determined by auto-arima.

3.4. Obtain 1-step to 7-step ahead points with 95% interval forecasts for the *total_cases* data using the model you chose in part 3.3.

```
##     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 988       96551997 96466687 96637346 96421543 96682543
## 989       96610833 96463080 96758702 96384912 96837026
## 990       96651419 96448155 96854903 96340644 96962711
## 991       96653709 96402456 96905299 96269587 97038620
## 992       96659371 96365768 96953435 96210531 97109289
## 993       96682519 96351059 97014566 96175832 97190580
## 994       96710966 96332331 97090367 96132204 97291520
```

Using the reduced ARMA() model:

- total_cases forecast for the next 7 days on the lower 95% CI:

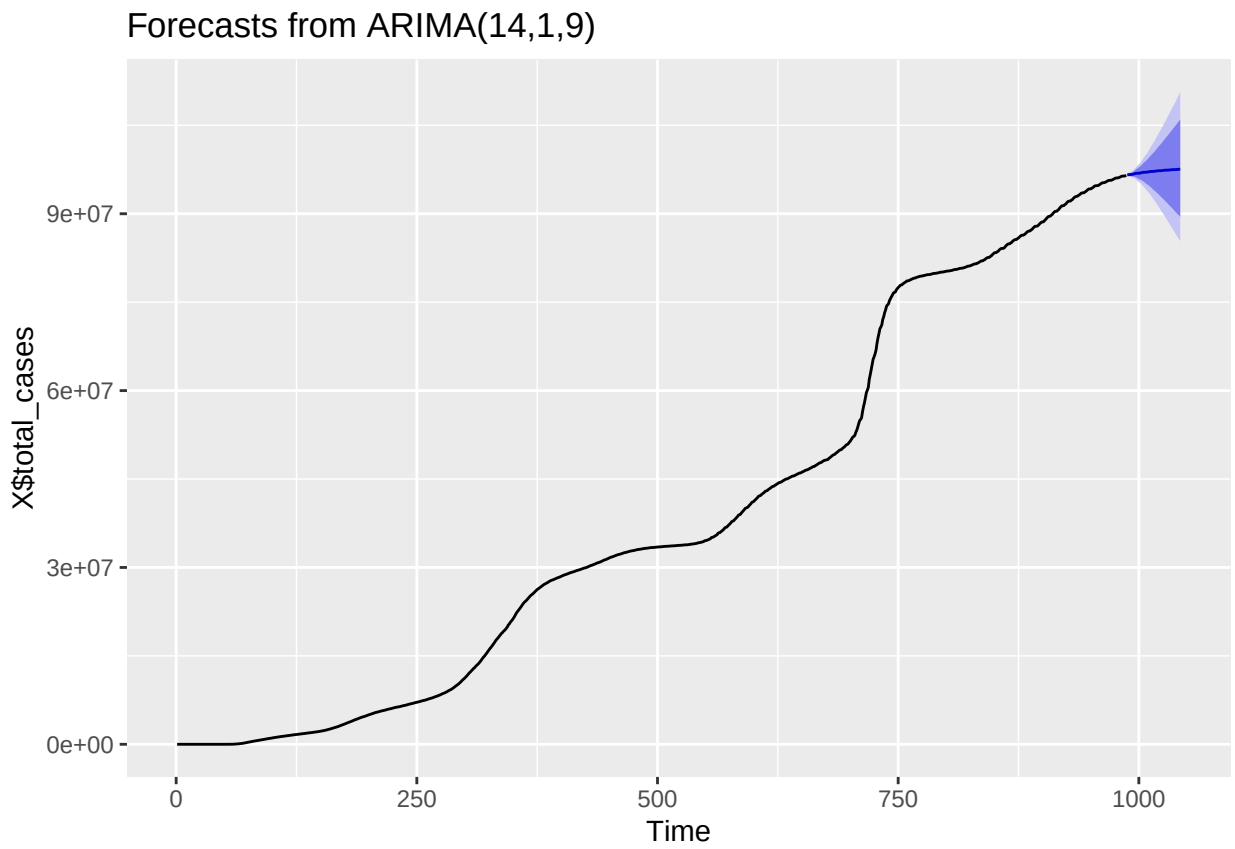
96421543, 96384912, 96340644, 96269587, 96210531, 96175832, 96132204

- total_cases forecast for the next 7 days on the upper 95% CI:

96682543, 96837026, 96962711, 97038620, 97109289, 97190580, 97291520

I would take the lower 95% CI forecast with a grain of salt to the number of total_cases decreasing. total_cases is a cumulative sum that is ever-increasing and would not go lower.

3.5. Forecast total cases with the forecast origin the last observed data point using the model you chose in part 2.3. Interpret.



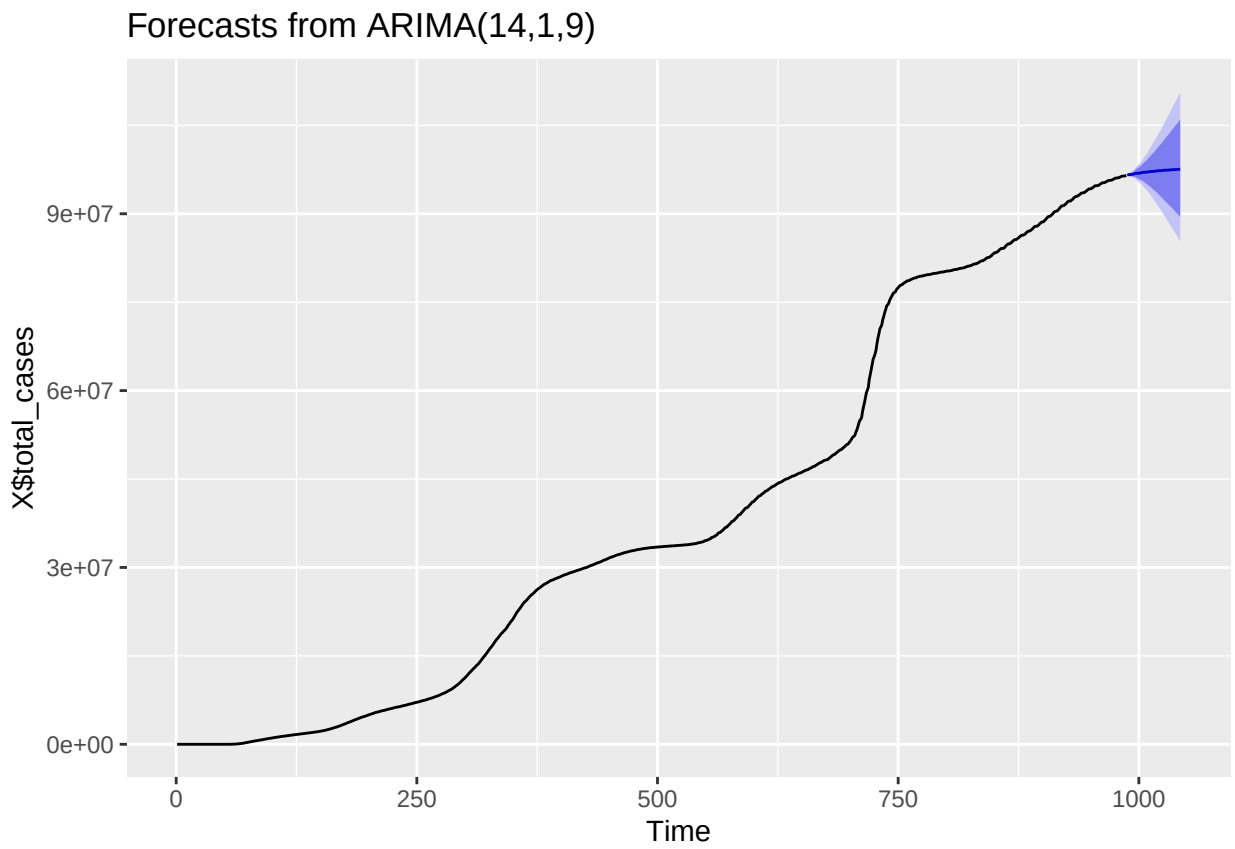
From the plot above, we observe the dark blue prediction line gently curving into a plateau of total_cases but still gradually increasing. The 80% CI shown by the blue area show a 'coning out' spread, while the 95% CI shown by the light blue area show a wider 'coning out' spread.

As explained in section 3.4, I would take the lower CI section forecast with a grain of salt to the number of total_cases decreasing. total_cases is a cumulative sum that is ever-increasing and would not go lower.

4. Report (20 points)

Choose the **"best"** model outcomes from the parts above. Write an executive report with information from the analysis such as forecasts from which decisions or actions can be made or taken.

(Based on the analysis, modelling, testing, and forecasting performed in sections 2 and 3, we will use the reduced ARMA(14,9) model for our forecasting executive report.)



This forecasting model attempts to predict the total number of COVID-19 cases in the United States over the next eight weeks. The data is sourced from the Our World in Data GitHub page, with daily US COVID-19 data starting from the first detected case in the nation on January 20, 2022. The model presented here is based on a model originally consisting of twenty-three parameters fitted by the time series data and has been optimized and simplified down to ten.

The forecast model shows the dark blue point forecast line gently curving into a plateau from the last data point of 96481081 total cases on October 4, 2022, but still shows a gradual increase in COVID-19 cases. Given the current pandemic's situation there doesn't seem to be major signs which could spark a sudden surge like what we've seen during the winter season of last year. The forecast looks to take a more conservative approach in future growth. In addition to the point forecast line we also have an 80% confidence interval (CI) shown by the blue area of possible values that could occur, as well as a 95% CI expanded in the lighter blue area. The range of these CIs expand quickly over the 8-week prediction period, showing the model to be less accurate in predicting total cases in much later dates. That being said, we should disregard the area below the forecast line as number of total cases is an ever-growing cumulative sum and cannot go down.

As for the model itself, given the current tools and methodologies we have at our disposal this is the most accurate model we've created so far in predicting future total cases of COVID-19. The range of possible future total cases eight weeks ahead seem to be fairly large and do not recommend to use this model for long-term planning. Short-term forecasts from the model, such as the next three or four days, can give us a modest idea of what to anticipate during that time to plan staffing for testing and vaccination at health centers, as well as manage supplies as needed.

As more data is collected for future dates, and as our own toolset and knowledge base expands, we will keep improving upon this model in an iterative fashion to hopefully provide a more accurate long-term forecast.