36-618 HW4

Daniel Nason

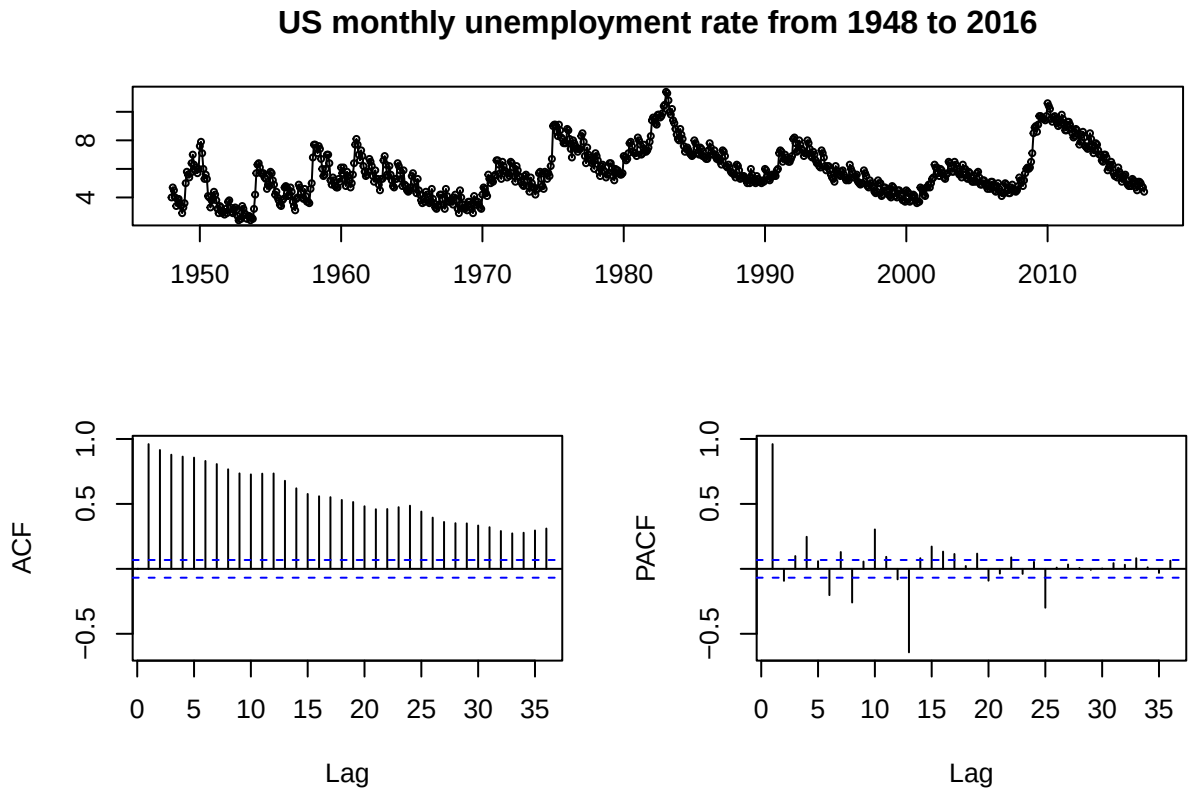
3/18/2022

```
setwd("C:/Users/Owner/CMU/Spring/36-618/HW/HW4")
library(forecast)
library(astsa)
library(tidyverse)
```

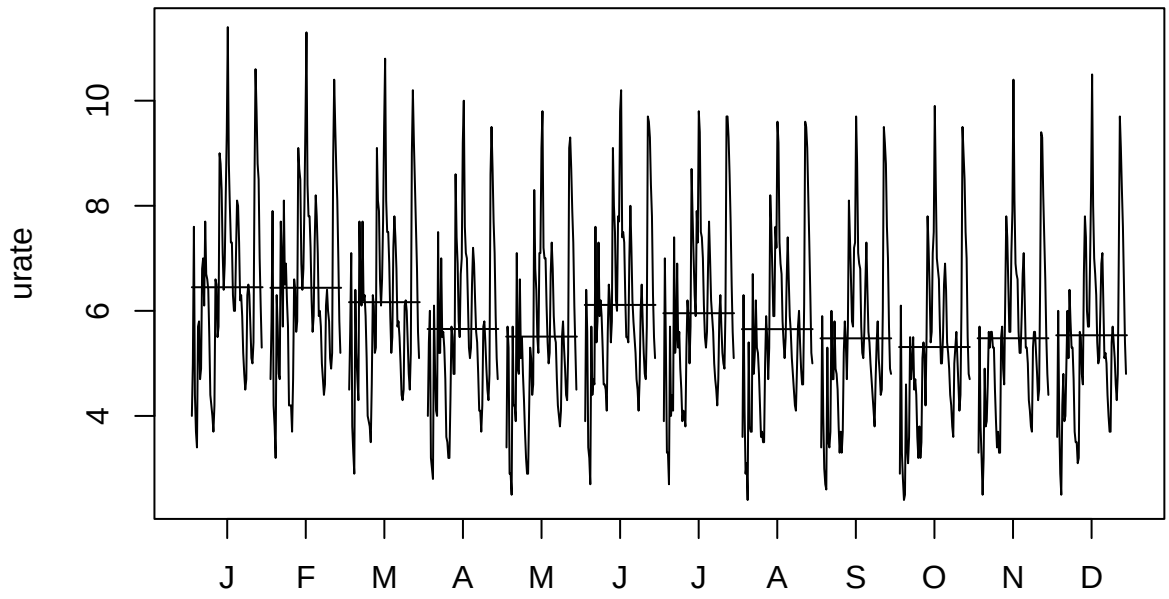
Question 1

a)

```
urate <- UnempRate
tsdisplay(urate, main = "US monthly unemployment rate from 1948 to 2016")
```

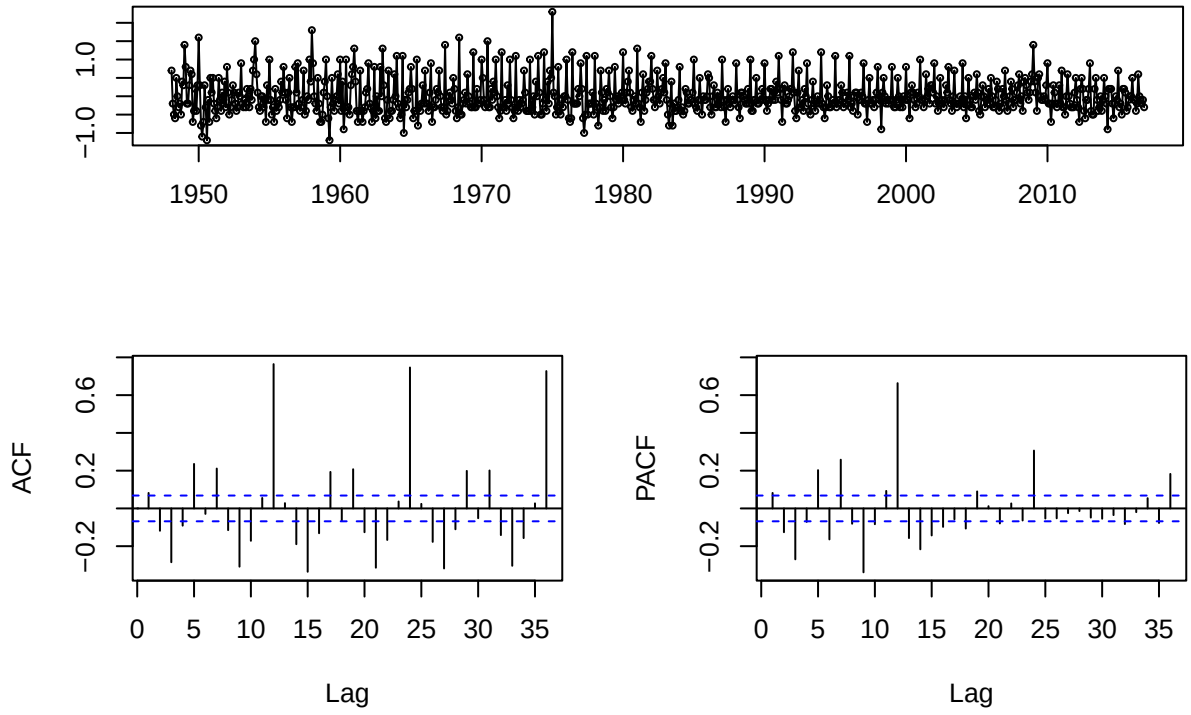


```
monthplot(urate)
```

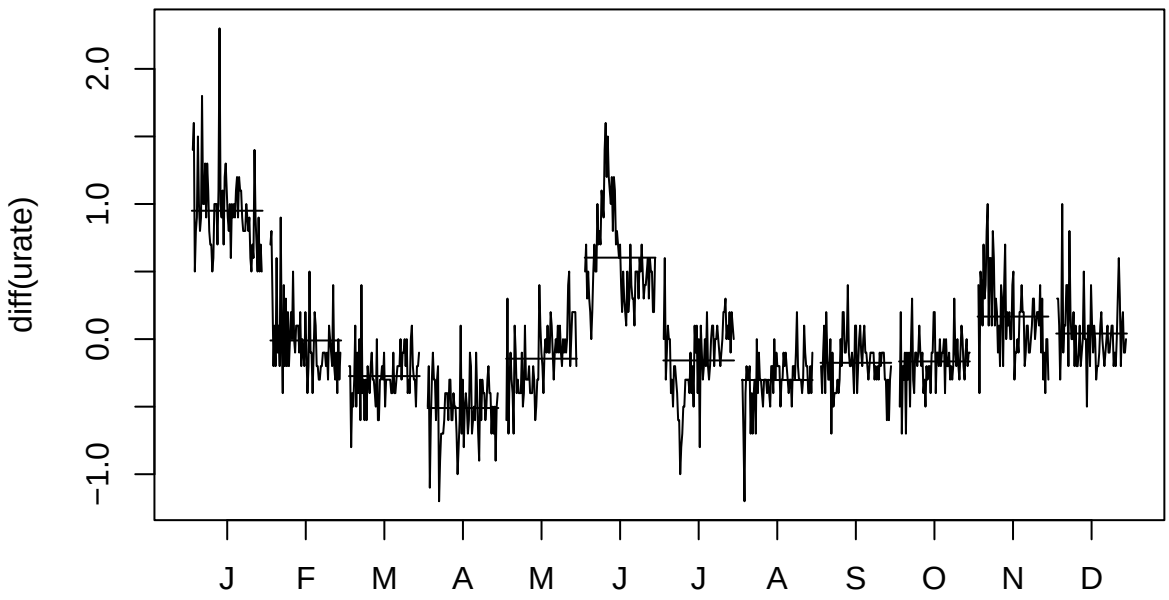


From the raw data we see that there is evidence seasonal cycles in the data. The monthplot shows that while there is constant variance across each of the months, there are cyclical fluctuations that suggest evidence of a seasonal cycle as well as a drift since the mean is not constant across each month. The ACF and PACF plots also show this since the ACF values slowly decay (suggesting a drift) and there are spikes in PACF values approximately around lags 12 and 22 (suggesting a seasonal cycle that is approximately annual).

```
tsdisplay(diff(urate), main ='Differenced US monthly unemployment rate from 1948 to 2016')
```

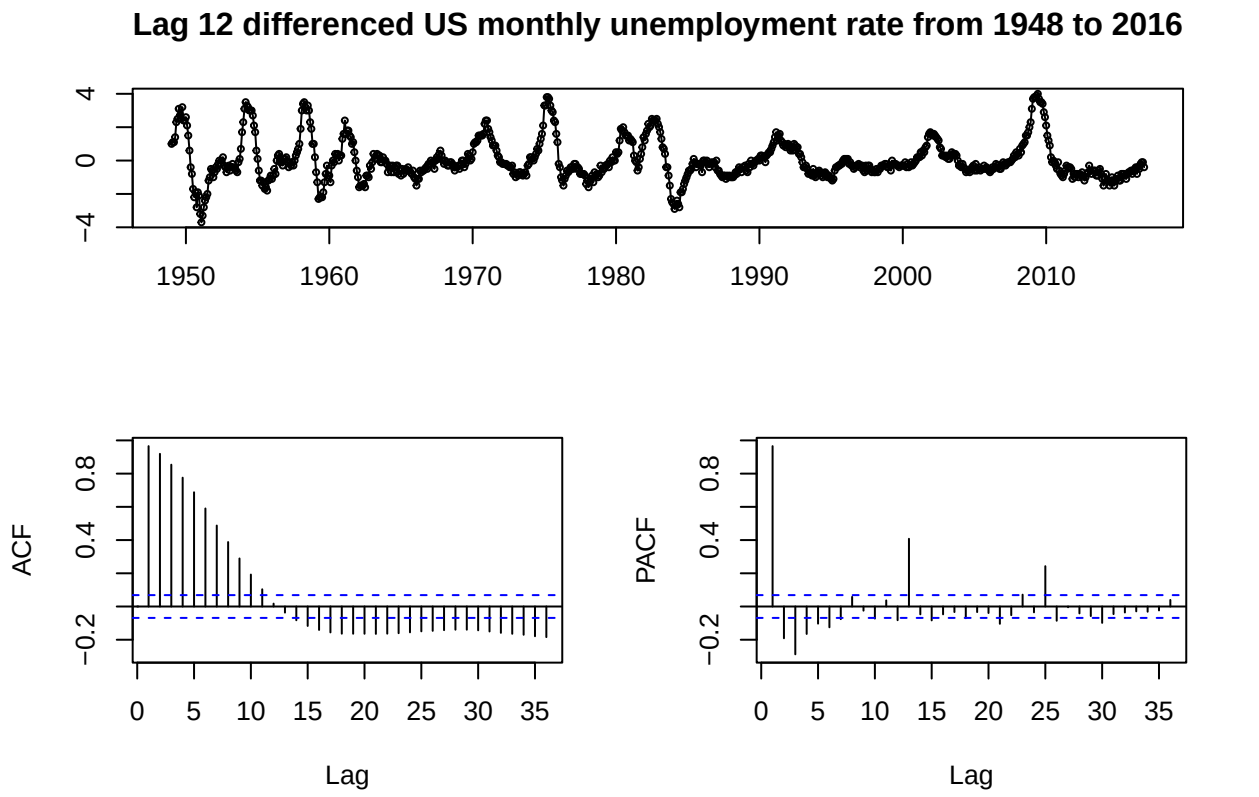



```
monthplot(diff(urate))
```

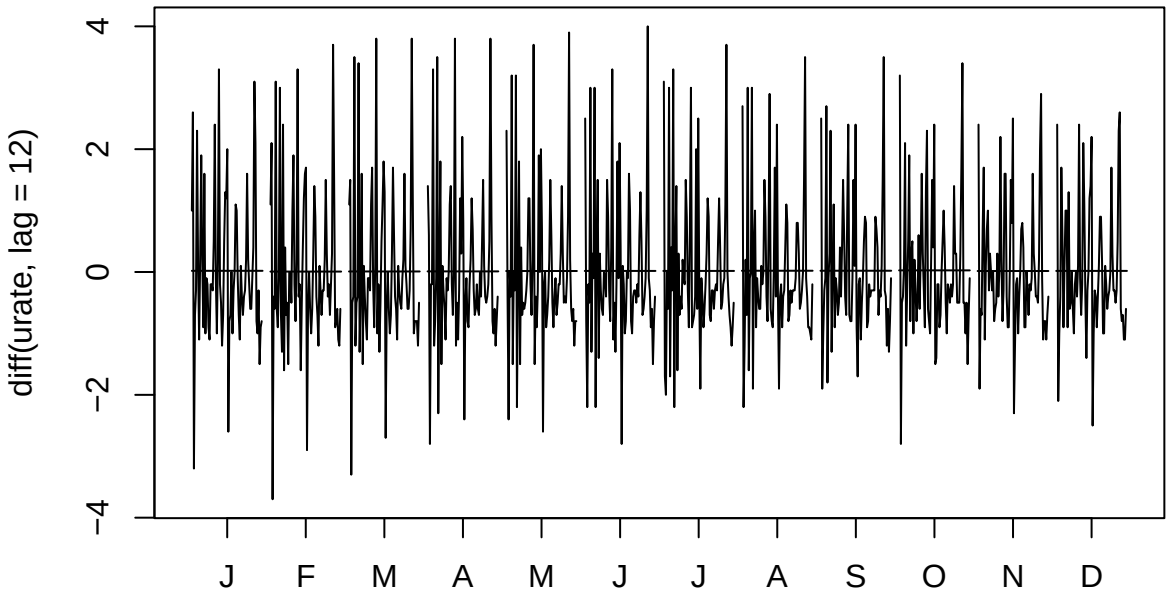


The plot of the differenced time series suggests that the differencing removed the trend, but the ACF and PACF plots show large values at around lags 10 and 20 which indicates that a seasonal cycle is still present in the data. The monthplot of the differenced time series shows evidence of a seasonal cycle based on the month of the year, and that the variance is not constant.

```
tsdisplay(diff(urate, lag = 12), main = "Lag 12 differenced US monthly unemployment rate from 1948 to 2
```




```
monthplot(diff(urate, lag = 12))
```



The plot of the differenced time series at lag 12 suggests that the differencing removed the trend but there is still evidence of a seasonal cycle. The seasonal cycle also does not appear to have constant variance over time, such as the difference in the size of the spikes between 1960 and 1970 versus 2000 and 2010. The ACF plot suggests that an AR process is present by the exponential decay in ACF values, and and PACF plot has seasonal spikes around lags 11 and 21 in addition to the additional spikes at the first few lags, which suggests a seasonal cycle is present. The monthplot of the differenced time series at lag 12 appears to have relatively constant variance and a mean of 0 across each of the months.

Given the behavior in the ACF, PACF and monthplots, it would not make sense to model the differenced time series at lag 1 with a stationary time series process since there is evidence of a seasonal cycle. However, it would make sense to model the differenced time series at lag 12 with a stationary process since this differenced series has constant variance and mean of 0.

b)

```
auto.arima(urate, d = 1, D = 0, trace=T, approximation = F, allowdrift = F)
```

i)

```
##
##  ARIMA(2,1,2)(1,0,1)[12]                    : Inf
##  ARIMA(0,1,0)                               : 1092.599
##  ARIMA(1,1,0)(1,0,0)[12]                    : 327.7329
##  ARIMA(0,1,1)(0,0,1)[12]                    : 726.7668
```

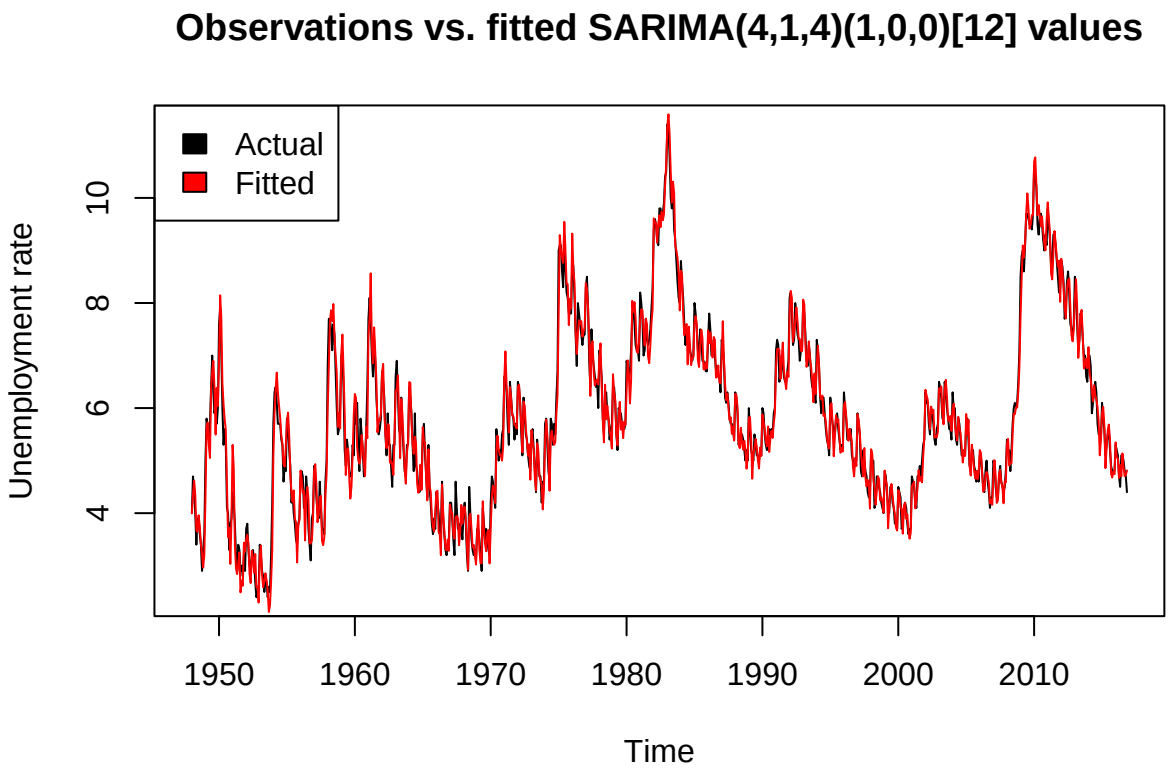
```
##  ARIMA(1,1,0)                    : 1089.04
##  ARIMA(1,1,0)(2,0,0)[12]         : Inf
##  ARIMA(1,1,0)(1,0,1)[12]         : Inf
##  ARIMA(1,1,0)(0,0,1)[12]         : 726.376
##  ARIMA(1,1,0)(2,0,1)[12]         : Inf
##  ARIMA(0,1,0)(1,0,0)[12]         : 348.19
##  ARIMA(2,1,0)(1,0,0)[12]         : 290.813
##  ARIMA(2,1,0)                    : 1077.817
##  ARIMA(2,1,0)(2,0,0)[12]         : Inf
##  ARIMA(2,1,0)(1,0,1)[12]         : Inf
##  ARIMA(2,1,0)(0,0,1)[12]         : 728.3918
##  ARIMA(2,1,0)(2,0,1)[12]         : Inf
##  ARIMA(3,1,0)(1,0,0)[12]         : 288.2691
##  ARIMA(3,1,0)                    : 1017.359
##  ARIMA(3,1,0)(2,0,0)[12]         : Inf
##  ARIMA(3,1,0)(1,0,1)[12]         : Inf
##  ARIMA(3,1,0)(0,0,1)[12]         : 713.7386
##  ARIMA(3,1,0)(2,0,1)[12]         : Inf
##  ARIMA(4,1,0)(1,0,0)[12]         : 289.4376
##  ARIMA(3,1,1)(1,0,0)[12]         : 288.1331
##  ARIMA(3,1,1)                    : 1017.625
##  ARIMA(3,1,1)(2,0,0)[12]         : Inf
##  ARIMA(3,1,1)(1,0,1)[12]         : Inf
##  ARIMA(3,1,1)(0,0,1)[12]         : 715.7188
##  ARIMA(3,1,1)(2,0,1)[12]         : Inf
##  ARIMA(2,1,1)(1,0,0)[12]         : 286.1726
##  ARIMA(2,1,1)                    : 1050.379
##  ARIMA(2,1,1)(2,0,0)[12]         : Inf
##  ARIMA(2,1,1)(1,0,1)[12]         : Inf
##  ARIMA(2,1,1)(0,0,1)[12]         : Inf
##  ARIMA(2,1,1)(2,0,1)[12]         : Inf
##  ARIMA(1,1,1)(1,0,0)[12]         : 295.3931
##  ARIMA(2,1,2)(1,0,0)[12]         : 288.106
##  ARIMA(1,1,2)(1,0,0)[12]         : 286.7999
##  ARIMA(3,1,2)(1,0,0)[12]         : 256.8508
##  ARIMA(3,1,2)                    : Inf
##  ARIMA(3,1,2)(2,0,0)[12]         : Inf
##  ARIMA(3,1,2)(1,0,1)[12]         : Inf
##  ARIMA(3,1,2)(0,0,1)[12]         : 686.664
##  ARIMA(3,1,2)(2,0,1)[12]         : Inf
##  ARIMA(4,1,2)(1,0,0)[12]         : 275.6778
##  ARIMA(3,1,3)(1,0,0)[12]         : 245.8101
##  ARIMA(3,1,3)                    : 936.4631
##  ARIMA(3,1,3)(2,0,0)[12]         : Inf
##  ARIMA(3,1,3)(1,0,1)[12]         : Inf
##  ARIMA(3,1,3)(0,0,1)[12]         : 687.5865
##  ARIMA(3,1,3)(2,0,1)[12]         : Inf
##  ARIMA(2,1,3)(1,0,0)[12]         : 290.1371
##  ARIMA(4,1,3)(1,0,0)[12]         : 246.8908
##  ARIMA(3,1,4)(1,0,0)[12]         : 246.524
##  ARIMA(2,1,4)(1,0,0)[12]         : 247.0463
##  ARIMA(4,1,4)(1,0,0)[12]         : 207.8515
##  ARIMA(4,1,4)                    : 939.0479
##  ARIMA(4,1,4)(2,0,0)[12]         : Inf
```

```
##  ARIMA(4,1,4)(1,0,1)[12]                    : Inf
##  ARIMA(4,1,4)(0,0,1)[12]                    : Inf
##  ARIMA(4,1,4)(2,0,1)[12]                    : Inf
##  ARIMA(5,1,4)(1,0,0)[12]                    : Inf
##  ARIMA(4,1,5)(1,0,0)[12]                    : 233.9243
##  ARIMA(3,1,5)(1,0,0)[12]                    : 243.7767
##  ARIMA(5,1,3)(1,0,0)[12]                    : Inf
##  ARIMA(5,1,5)(1,0,0)[12]                    : 233.8066
##
##  Best model: ARIMA(4,1,4)(1,0,0)[12]
```

```
## Series: urate
## ARIMA(4,1,4)(1,0,0)[12]
##
## Coefficients:
##          ar1     ar2     ar3      ar4      ma1     ma2      ma3     ma4   sar1
##       0.7324  0.1027  0.5936  -0.7669  -0.6993  0.0815  -0.7340  0.8859  0.862
## s.e.  0.0332  0.0408  0.0374   0.0309   0.0276  0.0232   0.0215  0.0286  0.019
##
## sigma^2 = 0.07297:  log likelihood = -93.79
## AIC=207.58   AICc=207.85   BIC=254.75
```

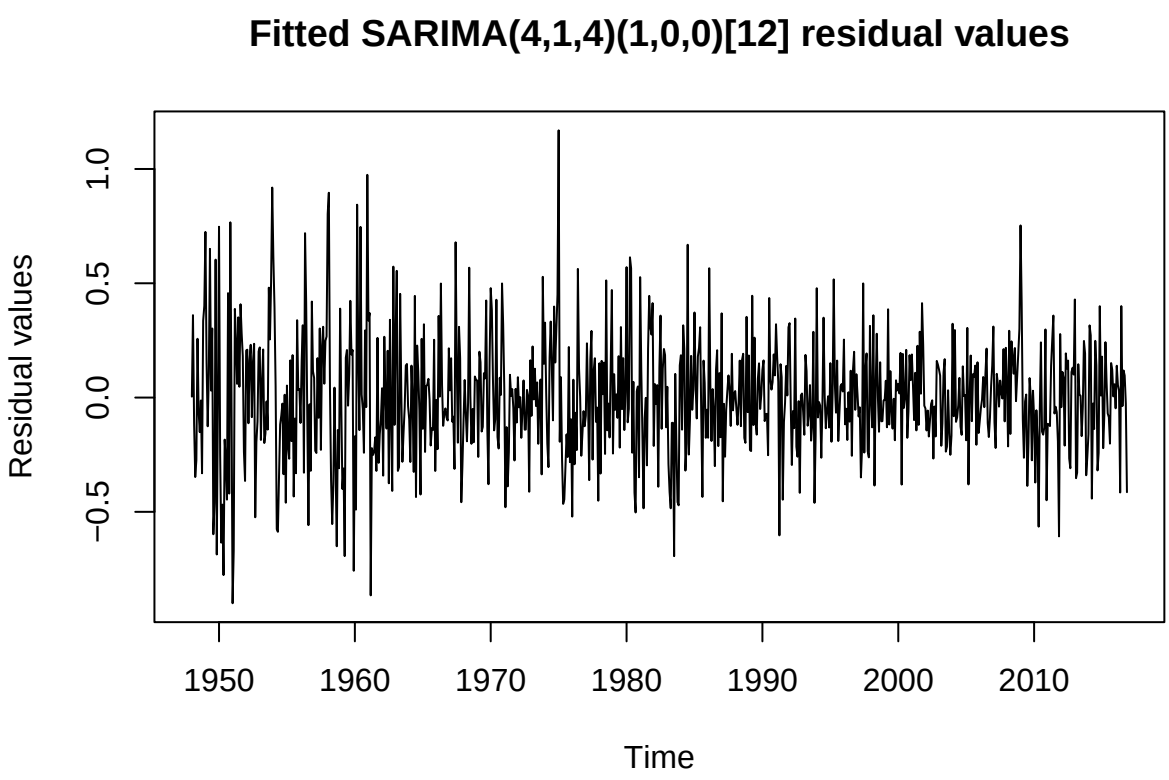
The order of the parameters identified is $\text{SARIMA}(4,1,4)(1,0,0)_{12}$.

```
q1bi_fit <- Arima(urate, order = c(4,1,4), seasonal = list(order=c(1,0,0), period = 12))
plot(urate,
     main = "Observations vs. fitted SARIMA(4,1,4)(1,0,0)[12] values",
     ylab = "Unemployment rate")
lines(urate-q1bi_fit$resid, col="red")
legend("topleft",legend = c("Actual", "Fitted"), fill = c("black", "red"))
```

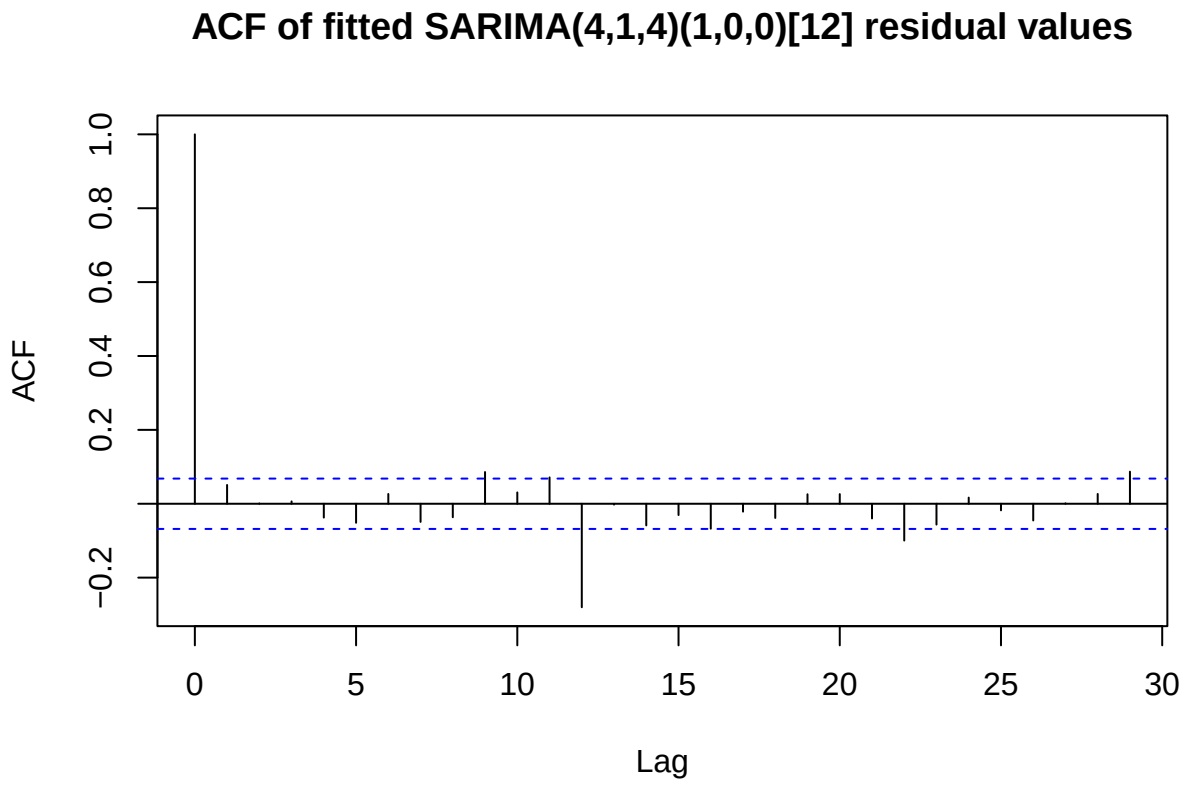


Comparing the observations to the fitted values, we see that the fitted values fit the data extremely well and capture almost all of the fluctuations in the sample. This suggests that the model is a good fit for the data.

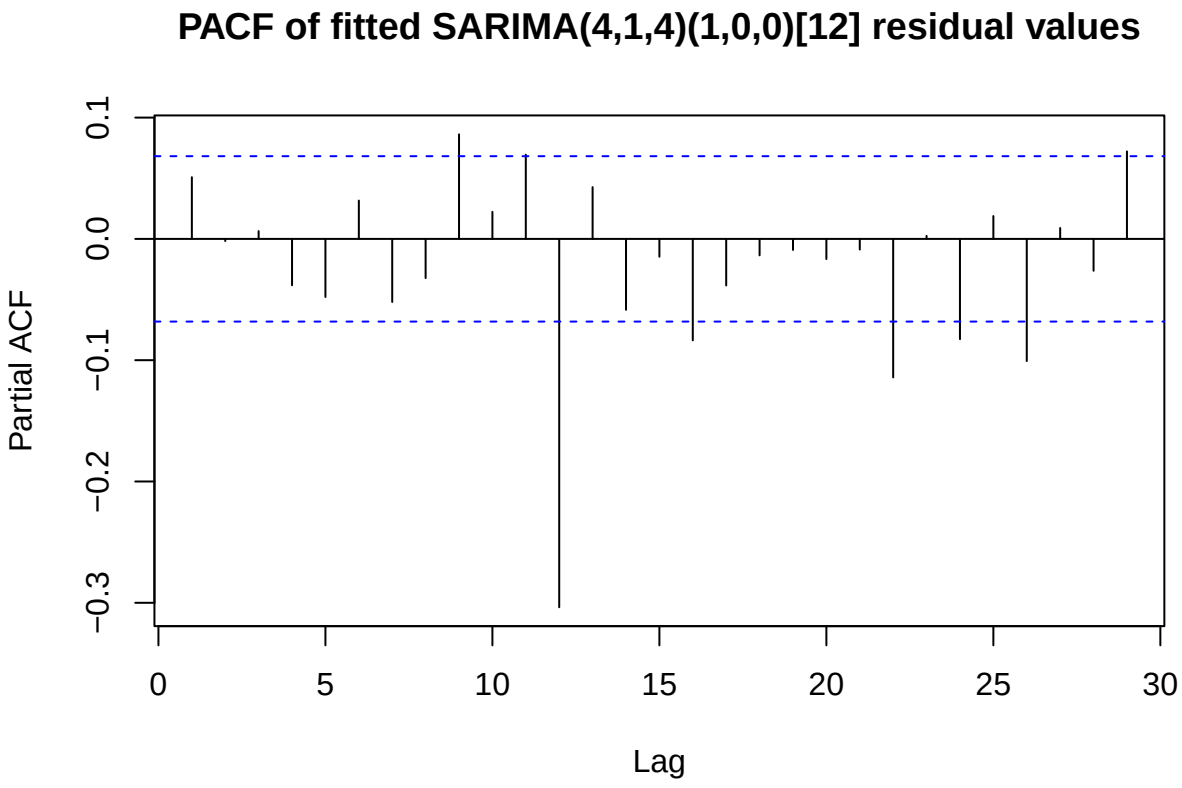
```
plot(q1bi_fit$resid, main = "Fitted SARIMA(4,1,4)(1,0,0)[12] residual values", ylab = "Residual values"
```



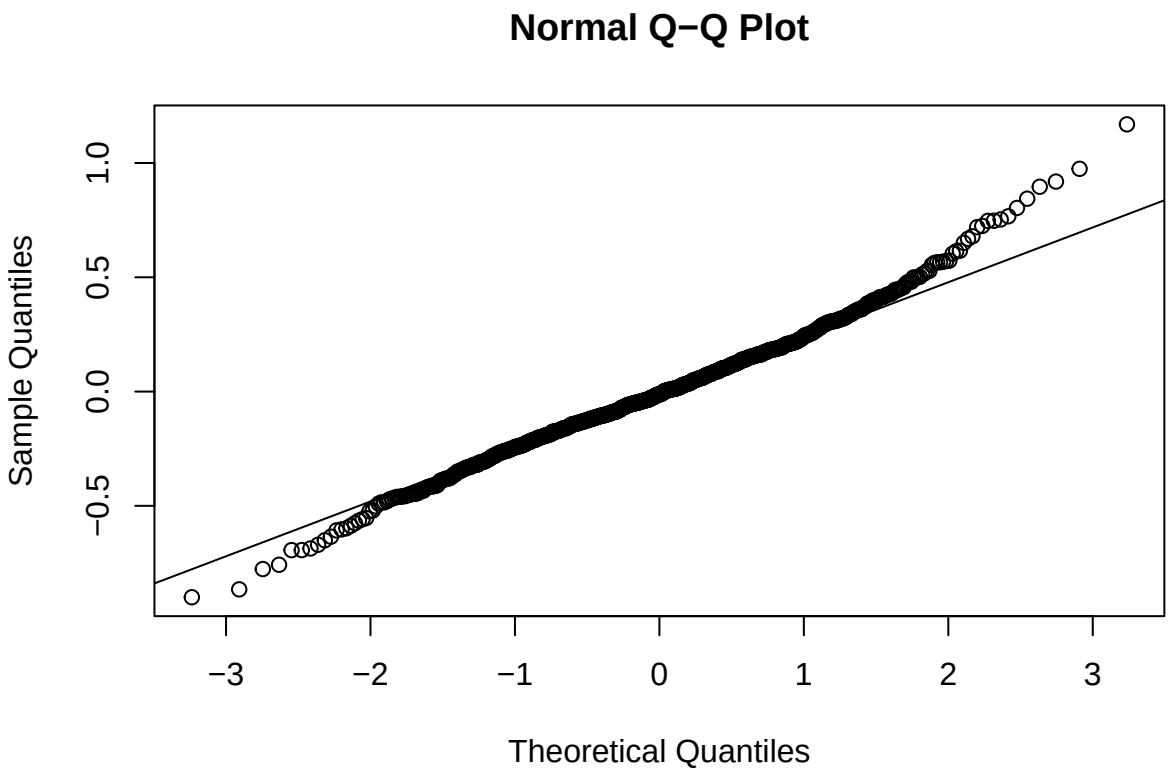

```
acf(q1bi_fit$resid[1:length(q1bi_fit$residuals)], main = "ACF of fitted SARIMA(4,1,4)(1,0,0)[12] residua
```




```
pacf(q1bi_fit$resid[1:length(q1bi_fit$residuals)], main = "PACF of fitted SARIMA(4,1,4)(1,0,0)[12] resi
```




```
qqnorm(q1bi_fit$resid)
qqline(q1bi_fit$resid)
```



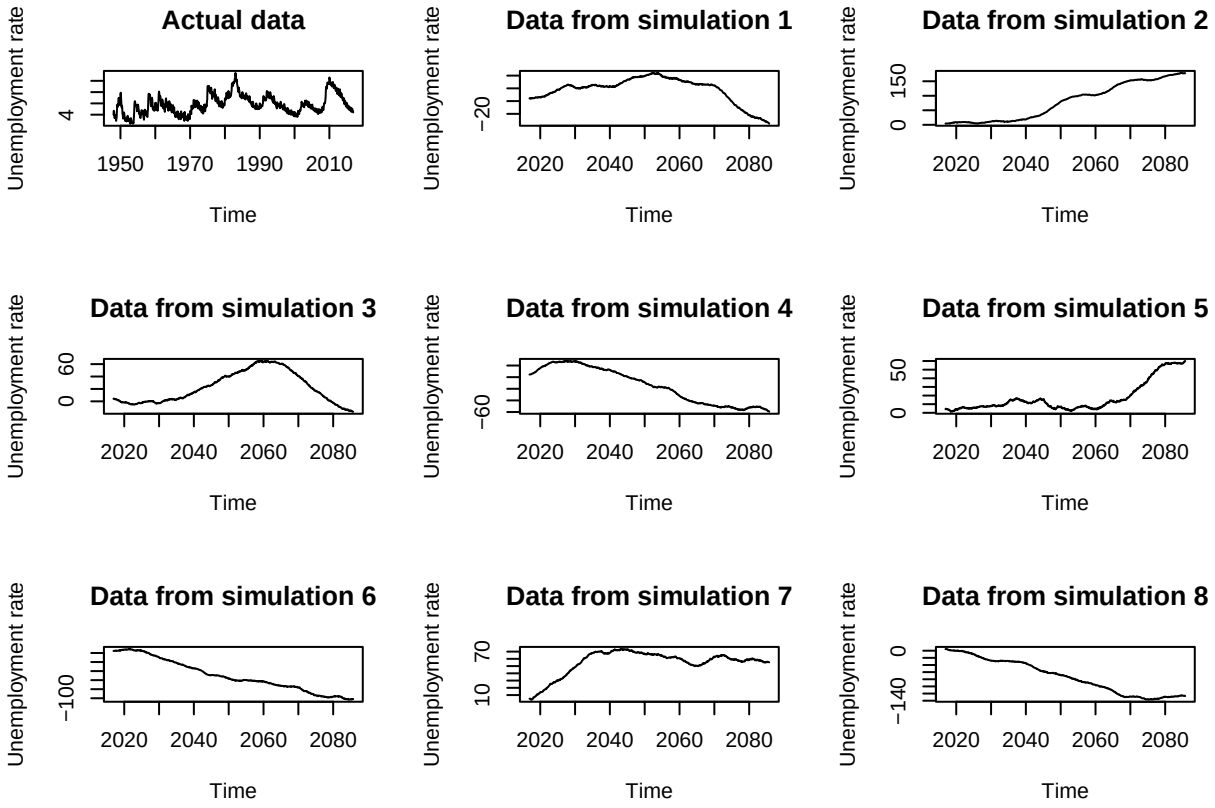
The residual values of SARIMA$(4,1,4)(1,0,0)_{12}$ model are centered at 0, have relatively constant variance. The variance is slightly larger between around 1950 through 1960 compared to the rest of the data and there are a few spikes, but they do not exhibit behaviors of autocorrelation since sequential terms do not usually have similar values based on the fluctuations.

The ACF and PACF plots show that at lag 10 in both the plots, there is a noticeable spike in respective values, suggesting that the model fails to capture the seasonal cycle around this time. However, the the majority of the other lagged values beyond lag 0 of the residuals are approximately within or around the 95% error bars.

The Q-Q plot illustrates that the residuals follow an approximately normal distribution since the theoretical and sample quantiles are linearly related for the majority of the data; however, there are some minor deviations from normality at the tails.

The spikes in the ACF and PACF plots at the seasonal lag suggest that the residuals are not Gaussian white noise.

```
par(mfrow=c(3,3))
plot(urate, ylab = 'Unemployment rate', main = 'Actual data')
for (i in 1:8){
  set.seed(i)
  plot(simulate(q1bi_fit),
     ylab = 'Unemployment rate',
     main = paste('Data from simulation', i, sep = ' '))
}
```




```
par(mfrow=c(1,1))
```

The behavior of the simulated values from the fitted model do not seem to mirror the sample data, since the simulated values do not capture the seasonal cycle present in the original unemployment rate data. Additionally, the simulated data does not appear to be sensible in the context of the problem since by definition the unemployment requires that the data be between 0 and 100 and is usually not greater than 10-20%. However, the simulations from this process show that the values of the data vary considerably by simulation and do not resemble the original data. For example, simulations 1, 3, 4, 6, and 8 are negative at some point during the simulation, while simulations 2, 5, and 7 have extremely percentages by the final year of the simulation. This does not strengthen the argument that the sample data could be from a SARIMA$(4,1,4)(1,0,0)_{12}$ process.

While the fitted values fit the sample observations well, the residuals do not display Gaussian white noise behavior and the simulated values are not similar to the sample observations, this suggests that the SARIMA$(4,1,4)(1,0,0)_{12}$ is not necessarily a good fit for the data.

```
auto.arima(urate, d =0, D = 1, trace = T, approximation = F, allowdrift = F)
```

ii)

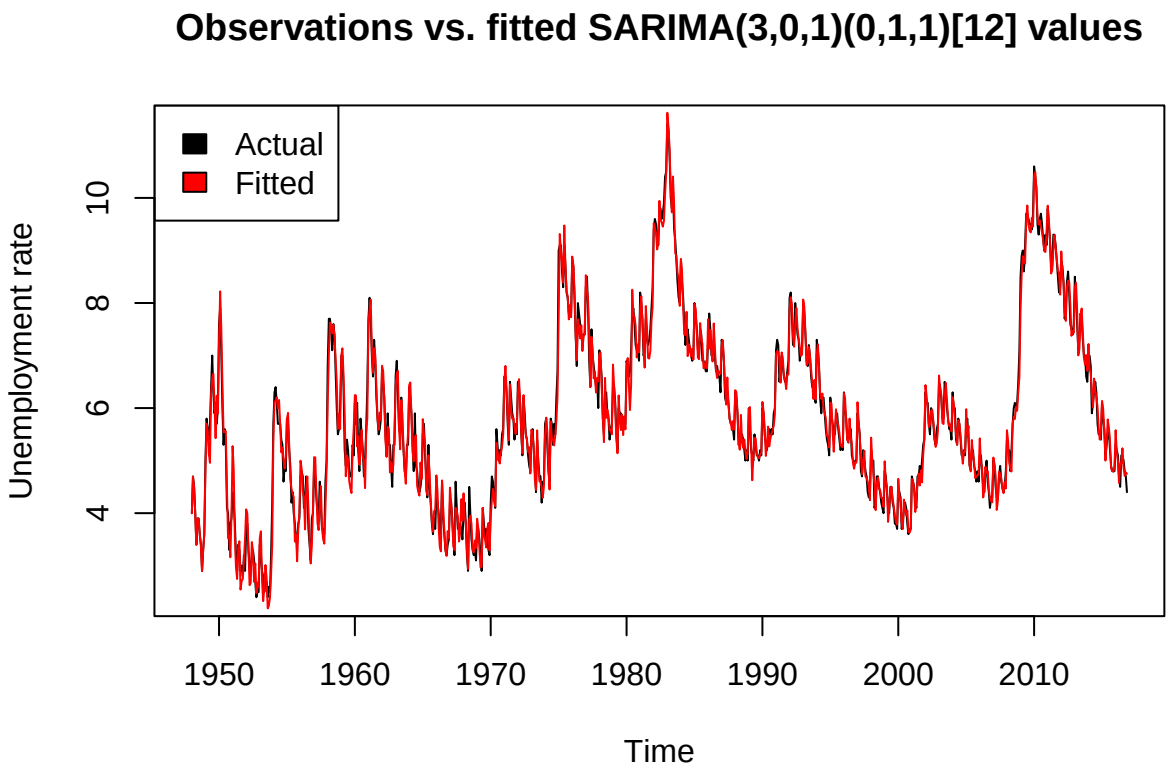
```
##
##  ARIMA(2,0,2)(1,1,1)[12]                    : -35.94212
##  ARIMA(0,0,0)(0,1,0)[12]                    : 2617.837
```

```
##  ARIMA(1,0,0)(1,1,0)[12]                    : 207.6273
##  ARIMA(0,0,1)(0,1,1)[12]                    : 1730.539
##  ARIMA(2,0,2)(0,1,1)[12]                    : -37.97278
##  ARIMA(2,0,2)(0,1,0)[12]                    : 283.019
##  ARIMA(2,0,2)(0,1,2)[12]                    : -35.94192
##  ARIMA(2,0,2)(1,1,0)[12]                    : 92.53043
##  ARIMA(2,0,2)(1,1,2)[12]                    : Inf
##  ARIMA(1,0,2)(0,1,1)[12]                    : -7.003126
##  ARIMA(2,0,1)(0,1,1)[12]                    : -34.91358
##  ARIMA(3,0,2)(0,1,1)[12]                    : -36.93028
##  ARIMA(2,0,3)(0,1,1)[12]                    : -36.94979
##  ARIMA(1,0,1)(0,1,1)[12]                    : 25.94754
##  ARIMA(1,0,3)(0,1,1)[12]                    : -11.95983
##  ARIMA(3,0,1)(0,1,1)[12]                    : -38.48627
##  ARIMA(3,0,1)(0,1,0)[12]                    : 282.4463
##  ARIMA(3,0,1)(1,1,1)[12]                    : -36.4594
##  ARIMA(3,0,1)(0,1,2)[12]                    : -36.45905
##  ARIMA(3,0,1)(1,1,0)[12]                    : 91.50597
##  ARIMA(3,0,1)(1,1,2)[12]                    : Inf
##  ARIMA(3,0,0)(0,1,1)[12]                    : -22.19614
##  ARIMA(4,0,1)(0,1,1)[12]                    : -37.03467
##  ARIMA(2,0,0)(0,1,1)[12]                    : 18.07318
##  ARIMA(4,0,0)(0,1,1)[12]                    : -29.50972
##  ARIMA(4,0,2)(0,1,1)[12]                    : -34.99152
##
##  Best model: ARIMA(3,0,1)(0,1,1)[12]
```

```
## Series: urate
## ARIMA(3,0,1)(0,1,1)[12]
##
## Coefficients:
##          ar1      ar2      ar3      ma1     sma1
##       1.6979  -0.5957  -0.1132  -0.6197  -0.7538
## s.e.  0.0908   0.1312   0.0469   0.0872   0.0268
##
## sigma^2 = 0.05448:  log likelihood = 25.3
## AIC=-38.59   AICc=-38.49   BIC=-10.37
```

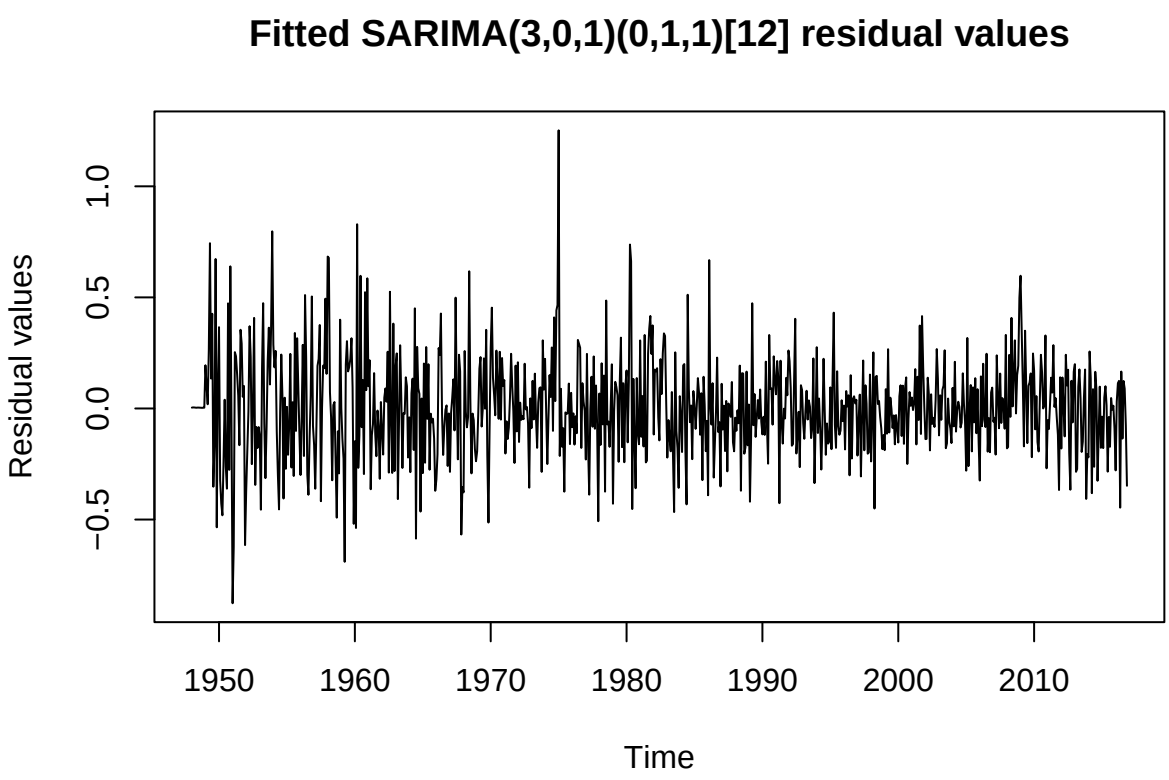
The order of the parameters identified is SARIMA$(3,0,1)(0,1,1)_{12}$.

```
q1bii_fit <- Arima(urate, order = c(3,0,1), seasonal = list(order=c(0,1,1), period = 12))
plot(urate,
     main = "Observations vs. fitted SARIMA(3,0,1)(0,1,1)[12] values",
     ylab = "Unemployment rate")
lines(urate-q1bii_fit$resid, col="red")
legend("topleft",legend = c("Actual", "Fitted"), fill = c("black", "red"))
```

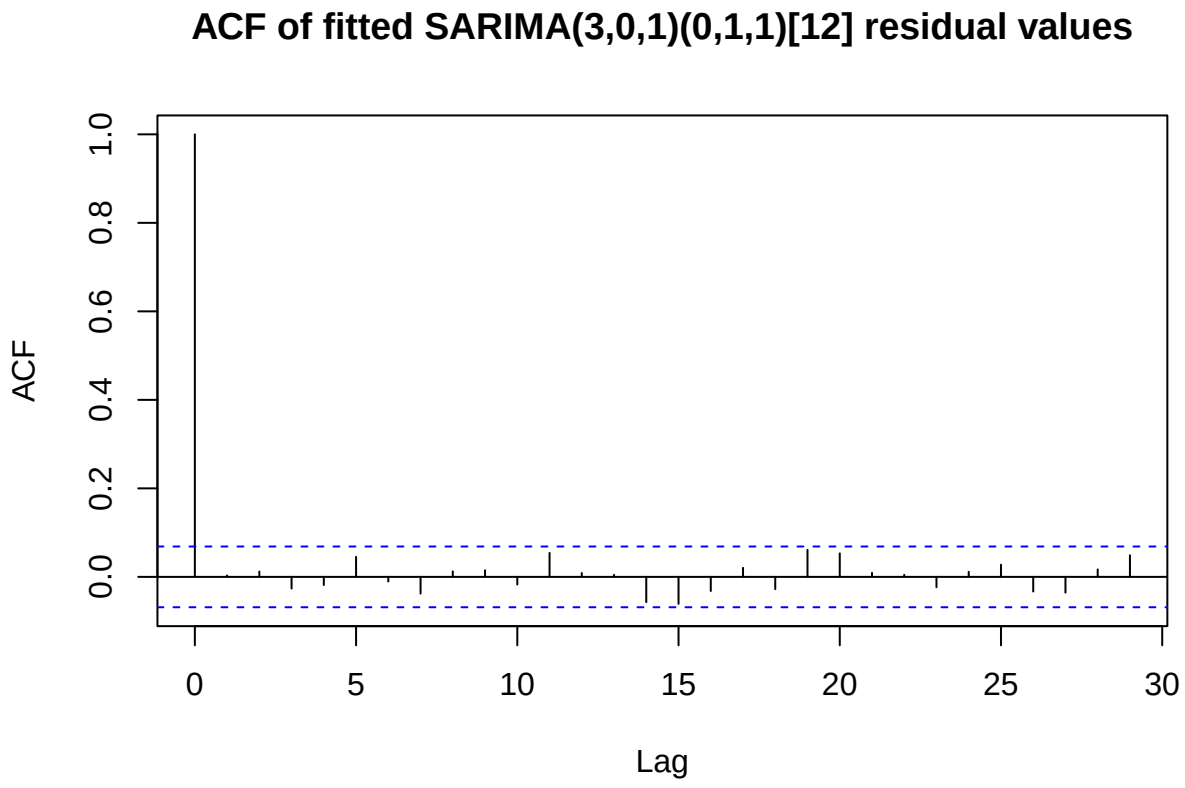


Comparing the observations to the fitted values, we see that the fitted values fit the data extremely well and capture almost all of the fluctuations in the sample. This suggests that the model is a good fit for the data.

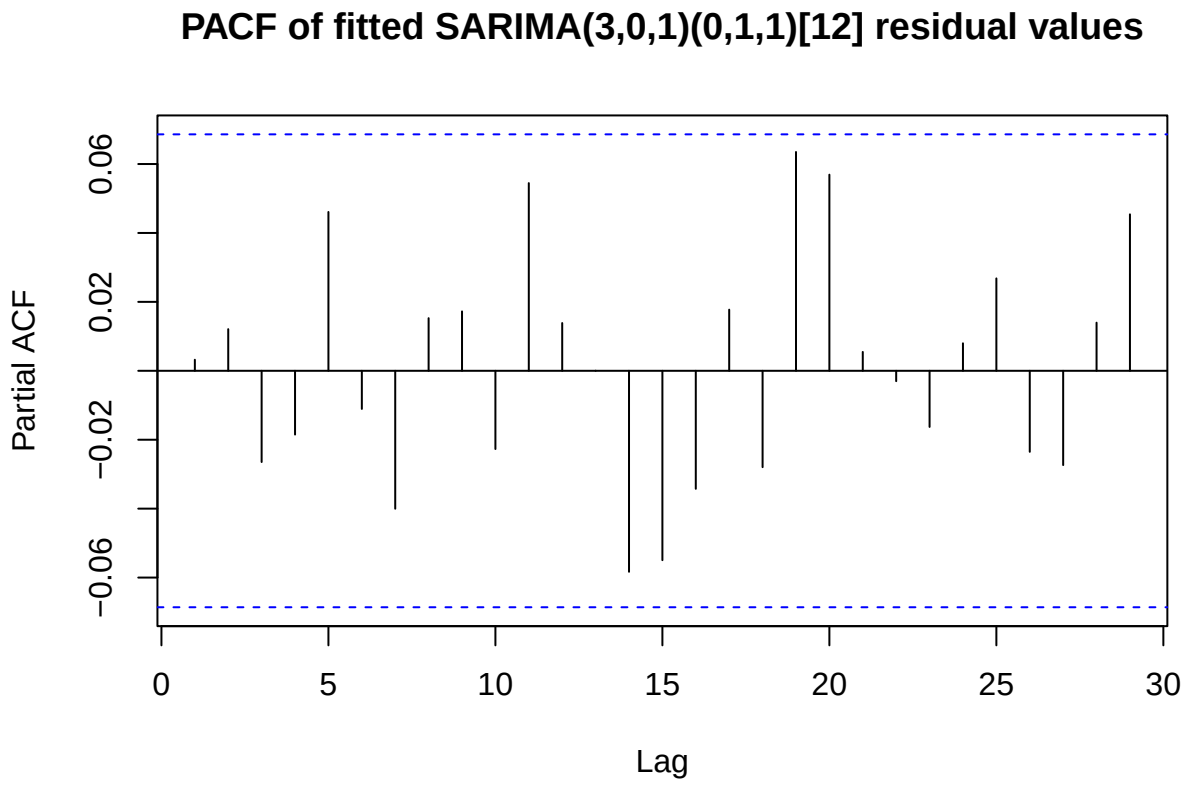
```
plot(q1bii_fit$resid, main = "Fitted SARIMA(3,0,1)(0,1,1)[12] residual values", ylab = "Residual values
```



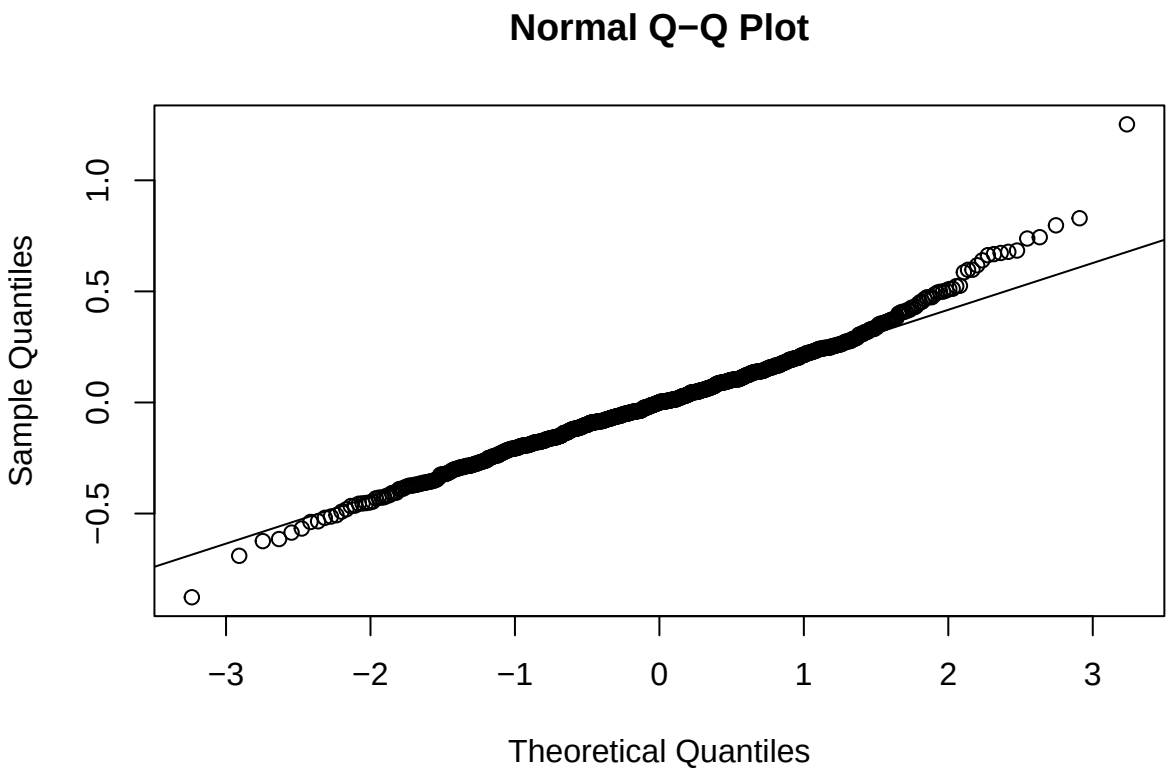

```
acf(q1bii_fit$resid[12:length(q1bii_fit$residuals)], main = "ACF of fitted SARIMA(3,0,1)(0,1,1)[12] resi
```




```
pacf(q1bii_fit$resid[12:length(q1bii_fit$residuals)], main = "PACF of fitted SARIMA(3,0,1)(0,1,1)[12] re
```




```
qqnorm(q1bii_fit$resid)
qqline(q1bii_fit$resid)
```



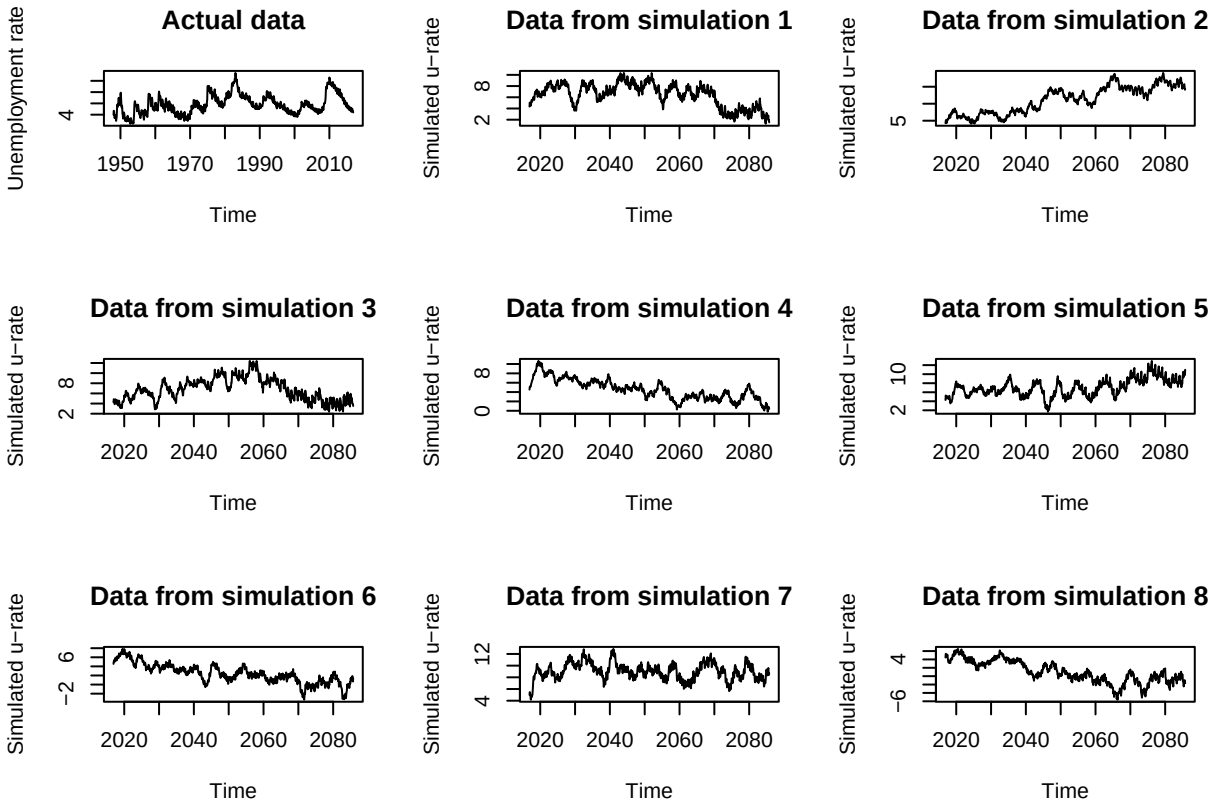
The residual values of SARIMA$(3,0,1)(0,1,1)_{12}$ model are centered at 0, have relatively constant variance. The variance is slightly larger between around 1950 through 1960 compared to the rest of the data and there are a few spikes, but they do not exhibit behaviors of autocorrelation since sequential terms do not usually have similar values based on the fluctuations.

The ACF and PACF plots show that all the lagged values beyond lag 0 of the residuals are approximately within the 95% error bars.

The Q-Q plot illustrates that the residuals follow a normal distribution since the theoretical and sample quantiles are linearly related for the majority of the data; however, there are some minor deviations from normality at the tails.

These suggest that the residuals are roughly Gaussian white noise.

```
par(mfrow=c(3,3))
plot(urate, ylab = 'Unemployment rate', main = 'Actual data')
for (i in 1:8){
  set.seed(i)
  plot(simulate(q1bii_fit),
     ylab = 'Simulated u-rate',
     main = paste('Data from simulation', i, sep = ' '))
}
```



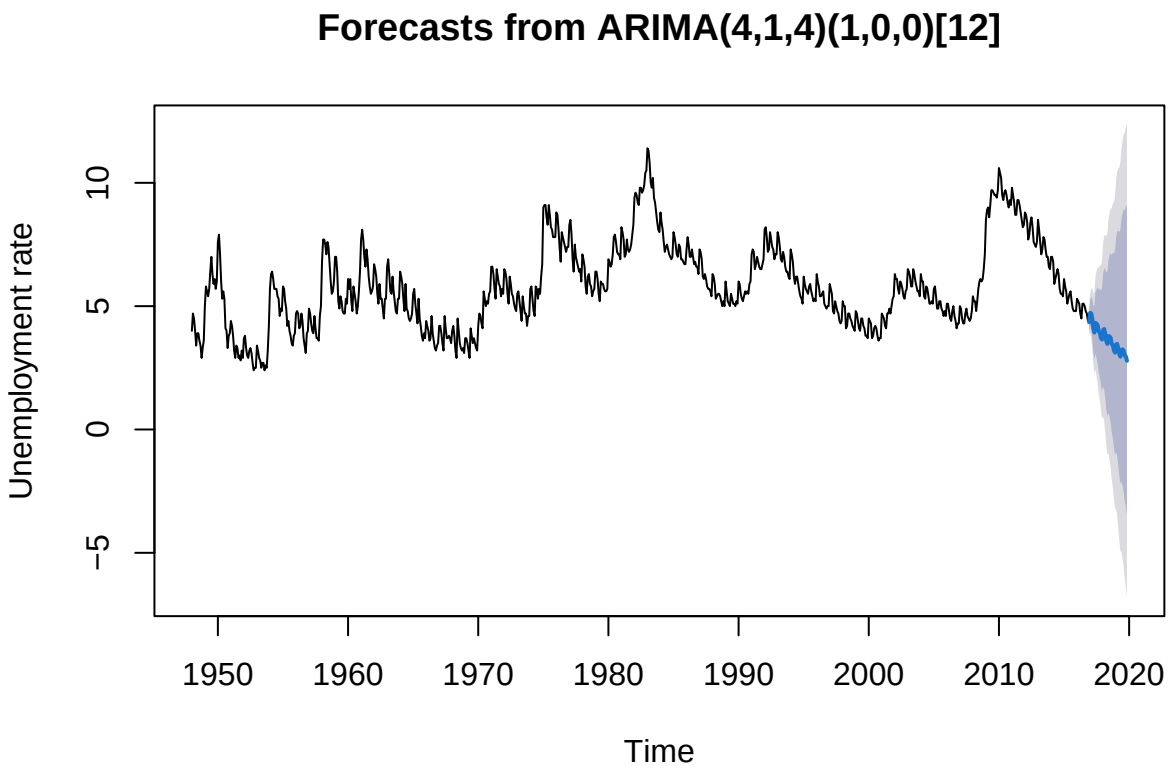

```
par(mfrow=c(1,1))
```

The behavior of the simulated values using the fitted model mirror the sample data in that they capture a seasonal cycle with a drift over time. Specifically, almost all the simulations have values that are sensible within the context of the data in that they are non-negative and fluctuate within historical patterns (i.e. do not exceed 10-20%). This strengthens the argument that the sample data could be from a $\text{SARIMA}(3,0,1)(0,1,1)_{12}$ process.

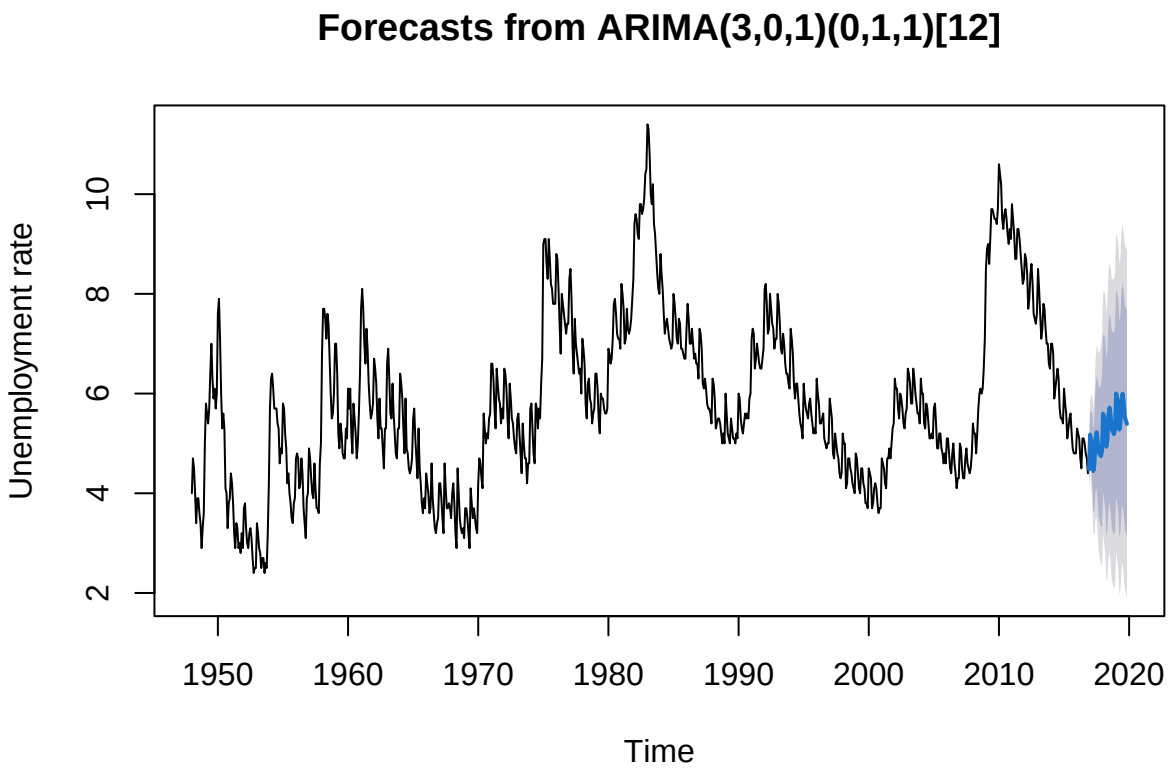
Given that the fitted values fit the sample observations reasonably well, the residuals display Gaussian white noise behavior, and the simulated values are similar to the sample observations, this implies that the $\text{SARIMA}(3,0,1)(0,1,1)_{12}$ is appropriate and a good fit for the data. Therefore, the $\text{SARIMA}(3,0,1)(0,1,1)_{12}$ model is a better fit for the data than the $\text{SARIMA}(4,1,4)(1,0,0)_{12}$ model.

c)

```
plot(forecast(q1bi_fit,h=36), ylab = "Unemployment rate", xlab = 'Time')
```




```
plot(forecast(q1bii_fit,h=36), ylab = "Unemployment rate", xlab = 'Time')
```



The mean forecast from the SARIMA$(4,1,4)(1,0,0)_{12}$ model is reasonable, but the lower and upper bounds of the forecast are very wide. The lower bound is less than 0, which is not sensible in the context of the problem. Additionally, based on the seasonal pattern in the data, it would probably not make sense for the data to continue trending downward and an increase (i.e. recession) could be expected.

The forecast from the SARIMA$(3,0,1)(0,1,1)_{12}$ model is reasonable in that the mean forecast is trending upwards and similar to historical data and the lower and upper bounds of the forecast are similar to historical patterns i.e. greater than 0 and less than 10. The model's prediction interval is also smaller implying less uncertainty. This better aligns with the context of the problem.

Based on the model validation from part b) and the forecasted values for the next 36 months, the SARIMA$(3,0,1)(0,1,1)_{12}$ with the seasonal difference provides a more trustworthy forecast.

Question 2

```
oni <- read.delim("oni.ascii_Dec_2021.txt", sep = "")
anom <- ts(oni$ANOM, start = 1950, frequency = 12)
```

a)

```
q2a_train <- window(anom, end = c(2015,12))
q2a_test <- window(anom, start = c(2016,1))
auto.arima(q2a_train, max.p = 5, max.q = 5, max.order = 10, stationary = T, seasonal = F, trace = T, st
```

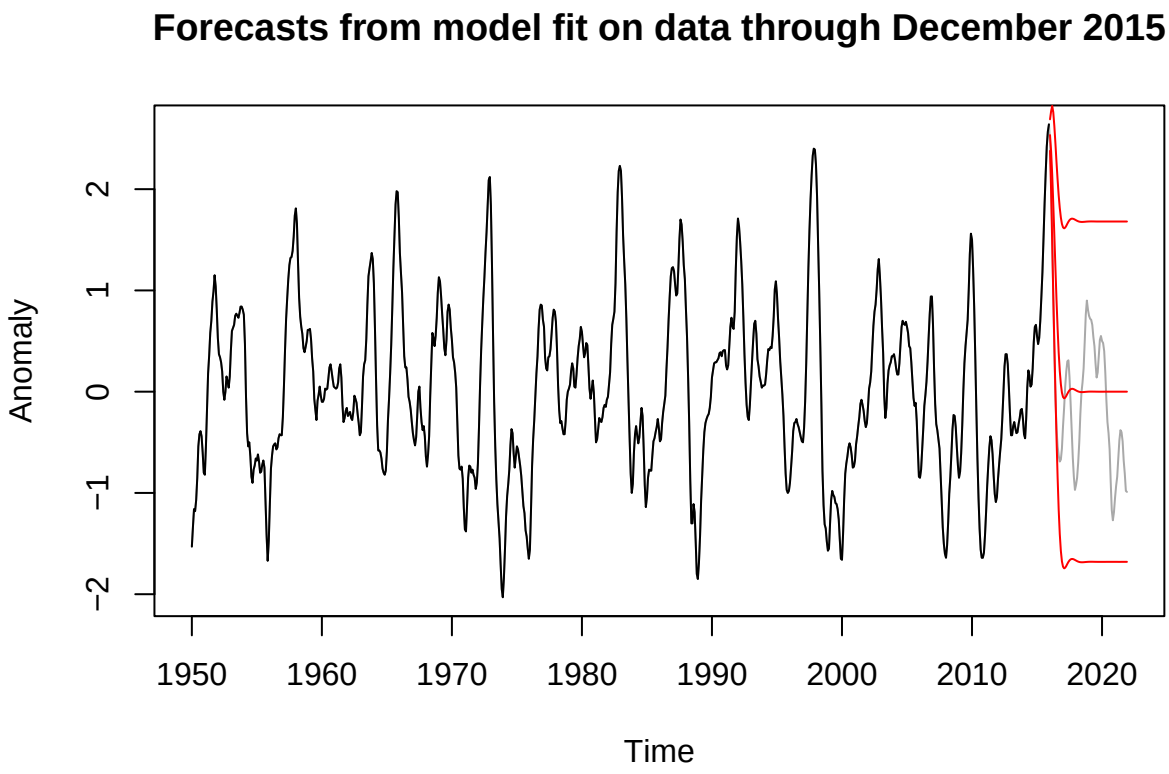
```
##
##  ARIMA(0,0,0)           with zero mean     : 1964.704
##  ARIMA(0,0,0)           with non-zero mean : 1966.303
##  ARIMA(0,0,1)           with zero mean     : 948.6383
##  ARIMA(0,0,1)           with non-zero mean : 950.2418
##  ARIMA(0,0,2)           with zero mean     : 95.57068
##  ARIMA(0,0,2)           with non-zero mean : 97.18147
##  ARIMA(0,0,3)           with zero mean     : Inf
##  ARIMA(0,0,3)           with non-zero mean : Inf
##  ARIMA(0,0,4)           with zero mean     : -1074.168
##  ARIMA(0,0,4)           with non-zero mean : -1072.454
##  ARIMA(0,0,5)           with zero mean     : -1332.198
##  ARIMA(0,0,5)           with non-zero mean : -1330.416
##  ARIMA(1,0,0)           with zero mean     : -353.1174
##  ARIMA(1,0,0)           with non-zero mean : Inf
##  ARIMA(1,0,1)           with zero mean     : -980.2484
##  ARIMA(1,0,1)           with non-zero mean : -978.2728
##  ARIMA(1,0,2)           with zero mean     : -1558.772
##  ARIMA(1,0,2)           with non-zero mean : -1556.776
##  ARIMA(1,0,3)           with zero mean     : -1694.562
##  ARIMA(1,0,3)           with non-zero mean : -1692.57
##  ARIMA(1,0,4)           with zero mean     : -1699.966
##  ARIMA(1,0,4)           with non-zero mean : -1697.974
##  ARIMA(1,0,5)           with zero mean     : -1705.324
##  ARIMA(1,0,5)           with non-zero mean : -1703.323
##  ARIMA(2,0,0)           with zero mean     : -1324.764
##  ARIMA(2,0,0)           with non-zero mean : -1322.808
##  ARIMA(2,0,1)           with zero mean     : -1443.622
##  ARIMA(2,0,1)           with non-zero mean : -1441.654
##  ARIMA(2,0,2)           with zero mean     : -1710.756
##  ARIMA(2,0,2)           with non-zero mean : -1708.771
##  ARIMA(2,0,3)           with zero mean     : -1708.991
##  ARIMA(2,0,3)           with non-zero mean : -1707.001
##  ARIMA(2,0,4)           with zero mean     : -1732.681
##  ARIMA(2,0,4)           with non-zero mean : -1730.71
##  ARIMA(2,0,5)           with zero mean     : -1732.528
##  ARIMA(2,0,5)           with non-zero mean : -1730.545
##  ARIMA(3,0,0)           with zero mean     : -1467.937
##  ARIMA(3,0,0)           with non-zero mean : -1465.956
##  ARIMA(3,0,1)           with zero mean     : -1472.776
##  ARIMA(3,0,1)           with non-zero mean : -1470.794
##  ARIMA(3,0,2)           with zero mean     : -1708.775
##  ARIMA(3,0,2)           with non-zero mean : -1706.783
##  ARIMA(3,0,3)           with zero mean     : -1709.304
##  ARIMA(3,0,3)           with non-zero mean : -1707.309
##  ARIMA(3,0,4)           with zero mean     : -1731.966
##  ARIMA(3,0,4)           with non-zero mean : Inf
##  ARIMA(3,0,5)           with zero mean     : Inf
##  ARIMA(3,0,5)           with non-zero mean : -1733.729
##  ARIMA(4,0,0)           with zero mean     : -1484.184
##  ARIMA(4,0,0)           with non-zero mean : -1482.208
##  ARIMA(4,0,1)           with zero mean     : -1475.638
##  ARIMA(4,0,1)           with non-zero mean : -1473.644
##  ARIMA(4,0,2)           with zero mean     : -1726.435
```

```
##  ARIMA(4,0,2)            with non-zero mean : -1724.444
##  ARIMA(4,0,3)            with zero mean     : -1733.458
##  ARIMA(4,0,3)            with non-zero mean : -1731.474
##  ARIMA(4,0,4)            with zero mean     : -1734.447
##  ARIMA(4,0,4)            with non-zero mean : -1732.451
##  ARIMA(4,0,5)            with zero mean     : -1735.405
##  ARIMA(4,0,5)            with non-zero mean : Inf
##  ARIMA(5,0,0)            with zero mean     : -1560.482
##  ARIMA(5,0,0)            with non-zero mean : -1558.523
##  ARIMA(5,0,1)            with zero mean     : -1591.904
##  ARIMA(5,0,1)            with non-zero mean : -1589.934
##  ARIMA(5,0,2)            with zero mean     : -1728.982
##  ARIMA(5,0,2)            with non-zero mean : -1726.99
##  ARIMA(5,0,3)            with zero mean     : -1731.412
##  ARIMA(5,0,3)            with non-zero mean : -1729.422
##  ARIMA(5,0,4)            with zero mean     : Inf
##  ARIMA(5,0,4)            with non-zero mean : Inf
##  ARIMA(5,0,5)            with zero mean     : Inf
##  ARIMA(5,0,5)            with non-zero mean : Inf
##
##
##
##  Best model: ARIMA(4,0,5)            with zero mean

## Series: q2a_train
## ARIMA(4,0,5) with zero mean
##
## Coefficients:
##          ar1      ar2     ar3      ar4      ma1     ma2      ma3     ma4
##       2.5879  -2.5262  1.0705  -0.1581  -0.2706  0.0813  -0.5625  0.5092
## s.e.  0.1518   0.3659  0.3382   0.1225   0.1514  0.1226   0.0502  0.1134
##          ma5
##       0.2003
## s.e.  0.0993
##
## sigma^2 = 0.006376:  log likelihood = 877.84
## AIC=-1735.69   AICc=-1735.41   BIC=-1688.94
```

For the full data, the model selected by auto.arima is the ARIMA(3,0,5). Looking at the data ending at December 2015, the model selected by auto.arima is ARIMA(4,0,5), so the same model is not identified compared to before.

```
q2a_fit <- arima(q2a_train, order = c(4,0,5), include.mean = F)
q2a_pred <- predict(q2a_fit, n.ahead = length(floor(window(time(q2a_test)))))
plot(anom, col = 'darkgray',ylab = 'Anomaly', main = 'Forecasts from model fit on data through December
lines(q2a_train)
lines(q2a_pred$pred, col = 'red')
lines(q2a_pred$pred - 1.96*q2a_pred$se, col = 'red')
lines(q2a_pred$pred + 1.96*q2a_pred$se, col = 'red')
```



The prediction interval for the ARIMA(4,0,5) model is sufficiently large to capture the true values of the anomalies, and the predicted expected value is roughly at the center of the data left out of training the model. However, we also see that after approximately 1 year (December 2016), the forecasts are constant for every time period going forward and therefore not very useful in forecasting future anomalies.

b)

```
q2b_train <- window(anom, end = c(2014,12))
q2b_test <- window(anom, start = c(2015,1))
auto.arima(q2b_train, max.p = 5, max.q = 5, max.order = 10, stationary = T, seasonal = F, trace = T, st
```

```
##
##  ARIMA(0,0,0)            with zero mean     : 1898.229
##  ARIMA(0,0,0)            with non-zero mean : 1900.228
##  ARIMA(0,0,1)            with zero mean     : 895.7
##  ARIMA(0,0,1)            with non-zero mean : 897.6999
##  ARIMA(0,0,2)            with zero mean     : 54.91988
##  ARIMA(0,0,2)            with non-zero mean : 56.92402
##  ARIMA(0,0,3)            with zero mean     : Inf
##  ARIMA(0,0,3)            with non-zero mean : Inf
##  ARIMA(0,0,4)            with zero mean     : -1090.296
##  ARIMA(0,0,4)            with non-zero mean : -1088.289
##  ARIMA(0,0,5)            with zero mean     : -1335.657
##  ARIMA(0,0,5)            with non-zero mean : -1333.641
```

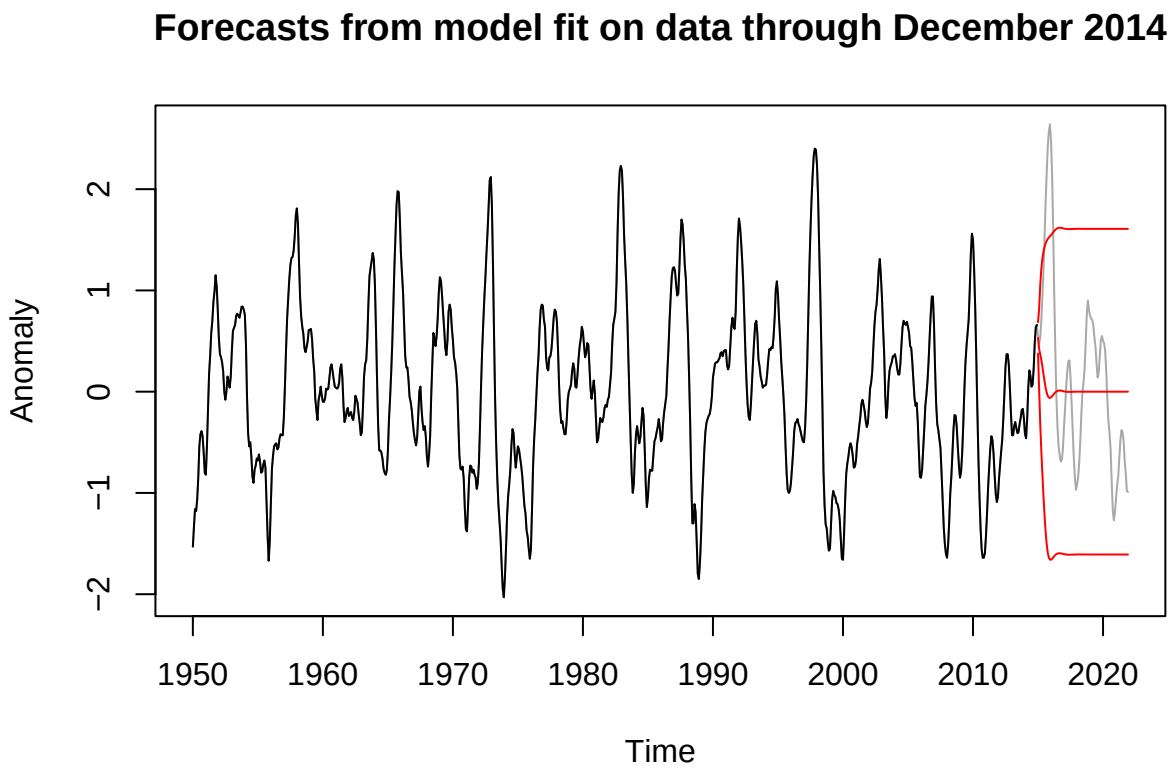
```
##  ARIMA(1,0,0)           with zero mean     : -355.5391
##  ARIMA(1,0,0)           with non-zero mean : -353.5502
##  ARIMA(1,0,1)           with zero mean     : -970.2124
##  ARIMA(1,0,1)           with non-zero mean : -968.2074
##  ARIMA(1,0,2)           with zero mean     : -1539.895
##  ARIMA(1,0,2)           with non-zero mean : -1537.89
##  ARIMA(1,0,3)           with zero mean     : -1672.211
##  ARIMA(1,0,3)           with non-zero mean : -1670.198
##  ARIMA(1,0,4)           with zero mean     : -1677.044
##  ARIMA(1,0,4)           with non-zero mean : -1675.028
##  ARIMA(1,0,5)           with zero mean     : -1682.357
##  ARIMA(1,0,5)           with non-zero mean : -1680.336
##  ARIMA(2,0,0)           with zero mean     : -1307.587
##  ARIMA(2,0,0)           with non-zero mean : -1305.582
##  ARIMA(2,0,1)           with zero mean     : -1422.788
##  ARIMA(2,0,1)           with non-zero mean : -1420.777
##  ARIMA(2,0,2)           with zero mean     : -1687.457
##  ARIMA(2,0,2)           with non-zero mean : -1685.446
##  ARIMA(2,0,3)           with zero mean     : -1685.694
##  ARIMA(2,0,3)           with non-zero mean : -1683.679
##  ARIMA(2,0,4)           with zero mean     : -1710.407
##  ARIMA(2,0,4)           with non-zero mean : -1708.377
##  ARIMA(2,0,5)           with zero mean     : -1710.197
##  ARIMA(2,0,5)           with non-zero mean : -1708.164
##  ARIMA(3,0,0)           with zero mean     : -1445.669
##  ARIMA(3,0,0)           with non-zero mean : -1443.663
##  ARIMA(3,0,1)           with zero mean     : -1450.472
##  ARIMA(3,0,1)           with non-zero mean : -1448.458
##  ARIMA(3,0,2)           with zero mean     : -1685.474
##  ARIMA(3,0,2)           with non-zero mean : -1683.458
##  ARIMA(3,0,3)           with zero mean     : -1685.998
##  ARIMA(3,0,3)           with non-zero mean : -1683.976
##  ARIMA(3,0,4)           with zero mean     : -1709.693
##  ARIMA(3,0,4)           with non-zero mean : -1707.659
##  ARIMA(3,0,5)           with zero mean     : Inf
##  ARIMA(3,0,5)           with non-zero mean : Inf
##  ARIMA(4,0,0)           with zero mean     : -1462.06
##  ARIMA(4,0,0)           with non-zero mean : -1460.042
##  ARIMA(4,0,1)           with zero mean     : -1453.258
##  ARIMA(4,0,1)           with non-zero mean : -1451.244
##  ARIMA(4,0,2)           with zero mean     : -1703.086
##  ARIMA(4,0,2)           with non-zero mean : -1701.062
##  ARIMA(4,0,3)           with zero mean     : -1710.916
##  ARIMA(4,0,3)           with non-zero mean : -1708.883
##  ARIMA(4,0,4)           with zero mean     : -1711.604
##  ARIMA(4,0,4)           with non-zero mean : -1709.566
##  ARIMA(4,0,5)           with zero mean     : -1712.279
##  ARIMA(4,0,5)           with non-zero mean : -1710.237
##  ARIMA(5,0,0)           with zero mean     : -1539.442
##  ARIMA(5,0,0)           with non-zero mean : -1537.417
##  ARIMA(5,0,1)           with zero mean     : -1570.036
##  ARIMA(5,0,1)           with non-zero mean : -1568.006
##  ARIMA(5,0,2)           with zero mean     : -1705.652
##  ARIMA(5,0,2)           with non-zero mean : -1703.622
```

```
##  ARIMA(5,0,3)            with zero mean     : -1708.869
##  ARIMA(5,0,3)            with non-zero mean : -1706.83
##  ARIMA(5,0,4)            with zero mean     : Inf
##  ARIMA(5,0,4)            with non-zero mean : Inf
##  ARIMA(5,0,5)            with zero mean     : -1704.801
##  ARIMA(5,0,5)            with non-zero mean : Inf
##
##
##
##  Best model: ARIMA(4,0,5)            with zero mean
```

```
## Series: q2b_train
## ARIMA(4,0,5) with zero mean
##
## Coefficients:
##          ar1      ar2     ar3      ar4      ma1     ma2      ma3     ma4
##       2.5716  -2.4917  1.0467  -0.1542  -0.2639  0.0764  -0.5730  0.5062
## s.e.  0.1554   0.3734  0.3454   0.1255   0.1550  0.1269   0.0505  0.1173
##          ma5
##       0.1994
## s.e.  0.1033
##
## sigma^2 = 0.006348:  log likelihood = 866.28
## AIC=-1712.56   AICc=-1712.28   BIC=-1665.97
```

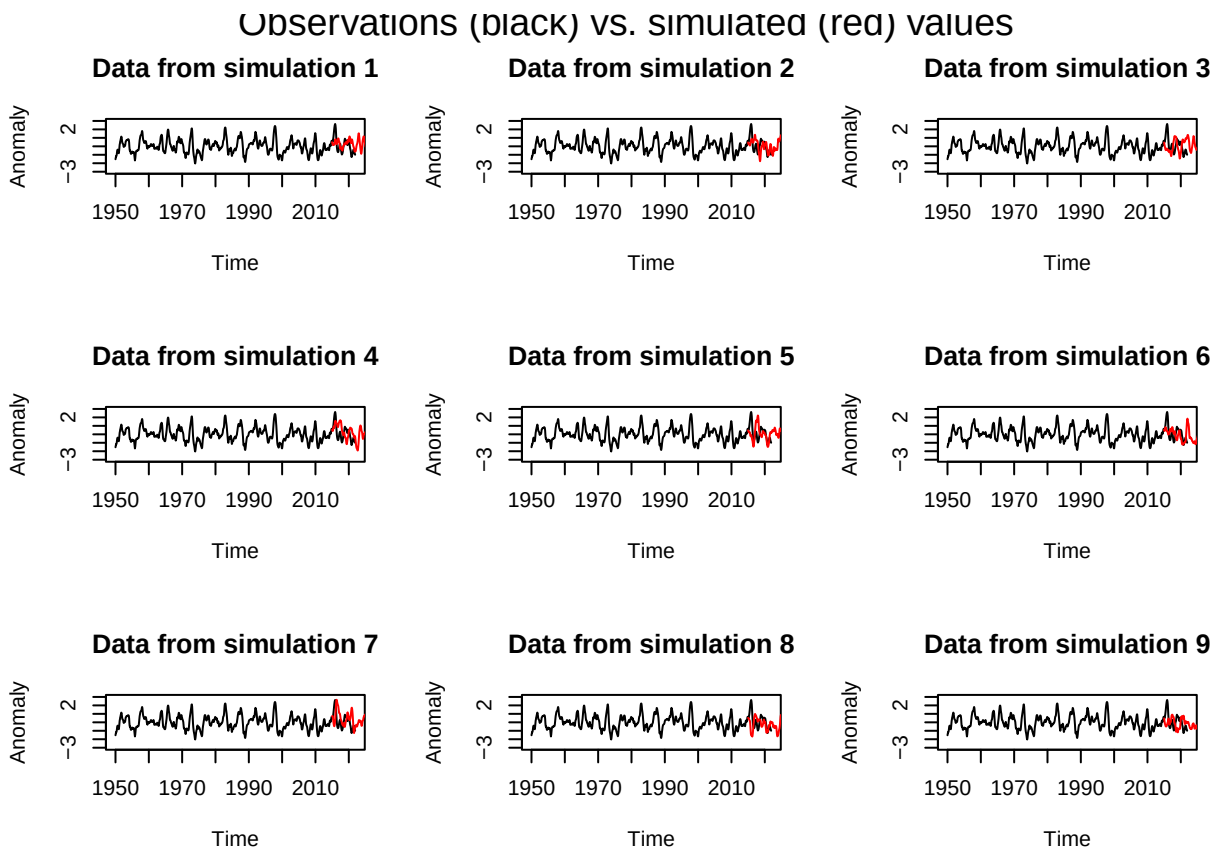
For the full data, the model selected by auto.arima is the ARIMA(3,0,5). Looking at the data ending at December 2014, the model selected by auto.arima is ARIMA(4,0,5), so the same model is not identified compared to before.

```
q2b_fit <- arima(q2b_train, order = c(4,0,5), include.mean = F)
q2b_pred <- predict(q2b_fit, n.ahead = length(floor(window(time(q2b_test)))))
plot(anom, col = 'darkgray', ylab = 'Anomaly',main = 'Forecasts from model fit on data through December
lines(q2b_train)
lines(q2b_pred$pred, col = 'red')
lines(q2b_pred$pred - 1.96*q2b_pred$se, col = 'red')
lines(q2b_pred$pred + 1.96*q2b_pred$se, col = 'red')
```



The prediction interval for the ARIMA(4,0,5) model fails to capture the large spike predicted in 2015. However, for the remainder of the data, the interval is large enough to capture true values of the anomalies, and the predicted expected value is roughly at the center of the data left out of training the model. We also see that after approximately 1 year (December 2015), the forecasts are constant for every time period going forward and therefore not very useful in forecasting future anomalies.

```
n <- 3
par(mfrow=c(n,n))
for (i in 1:n^2){
  plot(anom, col = 'black',
     ylab = 'Anomaly',
     ylim = c(-3,3),
     main = paste('Data from simulation', i, sep = ' '))
  set.seed(i)
  lines(simulate(q2b_fit, future = T),
      xlim = range(floor(window(time(q2b_test)))),
      col = 'red')
}
mtext('Observations (black) vs. simulated (red) values',
      side = 3, line = -1, cex = 1.2,
      outer = T)
```

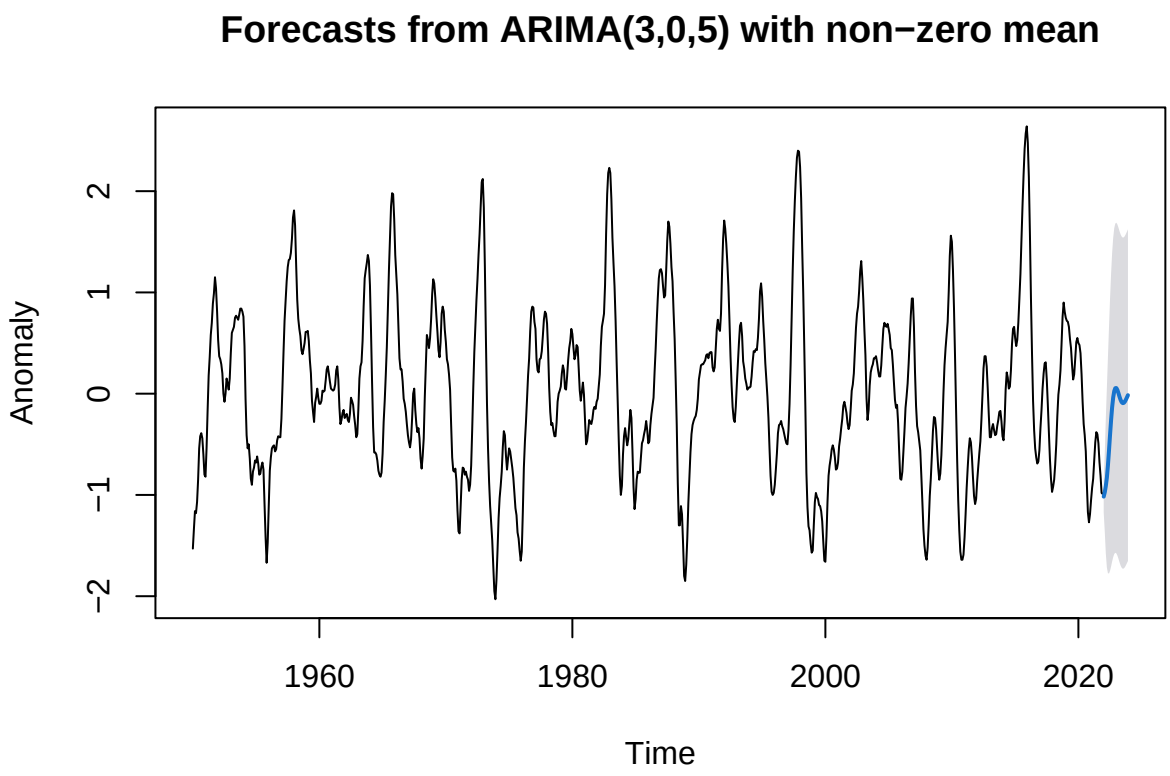



```
par(mfrow=c(1,1))
```

Based on the plot of the simulated data compared to the actual data, the model does not appear to successfully predict the anomaly in late 2015 for the majority of the simulations (only simulation 7 comes close to approximating this anomaly). This suggests that the model could not capture the anomaly and that it is likely an outlier. The model may also not be an appropriate fit for the data.

c)

```
q2c <- arima(anom, order=c(3,0,5))
plot(forecast(q2c,h=24, level = 95), ylab = 'Anomaly', xlab = 'Time')
```



The best prediction of the ONI value at March 2022 is -0.9110904, with a 95% predictive interval whose lower and upper bounds are -1.5756847 and -0.246496, respectively.

The best prediction of the ONI value at March 2023 is 0.0043949, with a 95% predictive interval whose lower and upper bounds are -1.6305078 and 1.6392977, respectively.

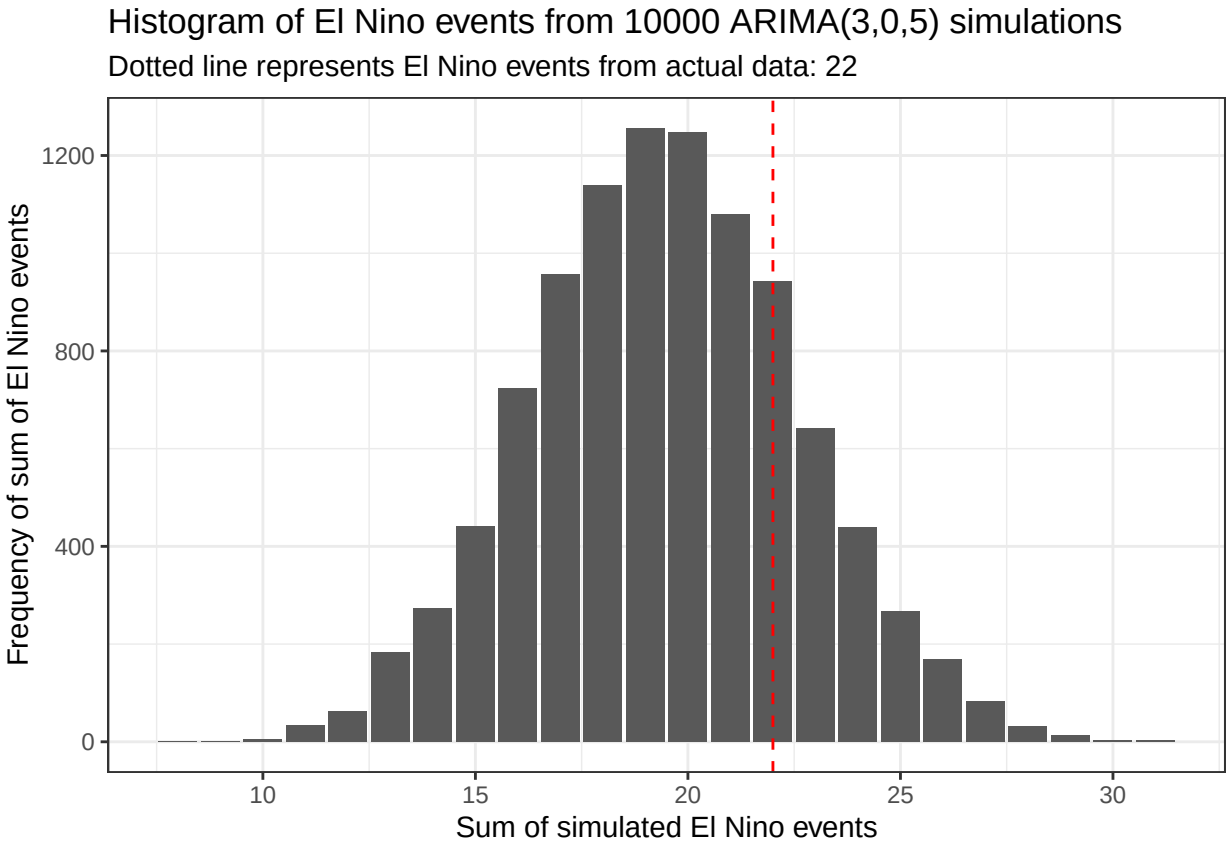
The mean prediction for the time period is roughly consistent with the historical data and around 0, while the prediction interval upper and lower bounds are with the historical data values as well. Therefore, the forecast is reasonable based on the data.

Question 3

```
q3_sim_func <- function(use_anom = T){
  ifelse(use_anom, dat <- anom, dat <- simulate(q2c, future = F))
  x <- ((dat %>% as.vector()) >= 0.5) %>% rle()
  total <- (x$lengths[x$values] >= 5) %>% sum()
  total
}
anoms_oni <- q3_sim_func()
set.seed(1)
sim_vec <- replicate(10000, q3_sim_func(use_anom = F))

sim_vec %>%
  as.data.frame() %>%
  ggplot(aes(x = sim_vec)) +
```

```
  #geom_histogram(binwidth = 1) +
  geom_bar() +
  geom_vline(xintercept = sum(anoms_oni), linetype = 'dashed', color = 'red') +
  labs(x = 'Sum of simulated El Nino events', y = 'Frequency of sum of El Nino events',
       title = 'Histogram of El Nino events from 10000 ARIMA(3,0,5) simulations',
       subtitle = paste('Dotted line represents El Nino events from actual data:', sum(anoms_oni))) +
  theme_bw()
```



Using the 10000 simulations from the ARIMA(3,0,5) model, 19.4688 El Nino event events would occur on average from January 1950 to December 2021. 22 El Nino events were observed in the ONI time series, and this is roughly consistent with the model predictions since it shows up roughly in 1000 of the simulations and is close to the mean of the simulations.