

36-618 HW3

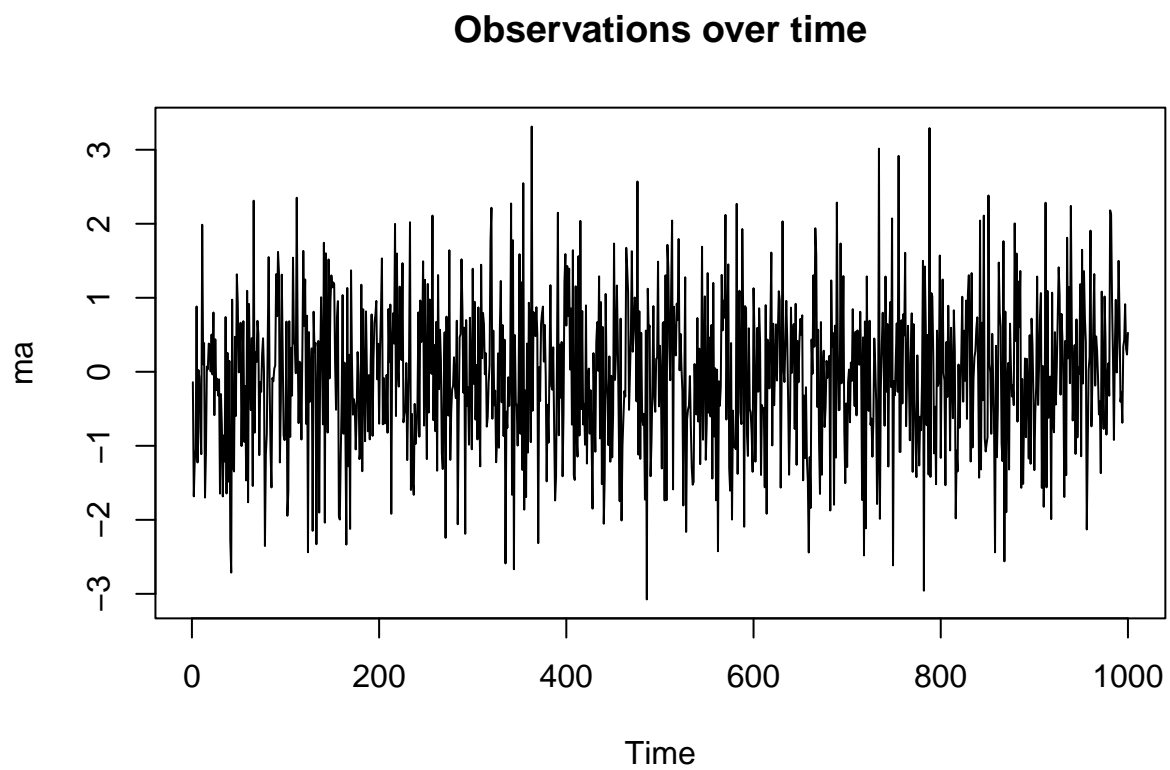
Daniel Nason

3/3/2022

Question 1

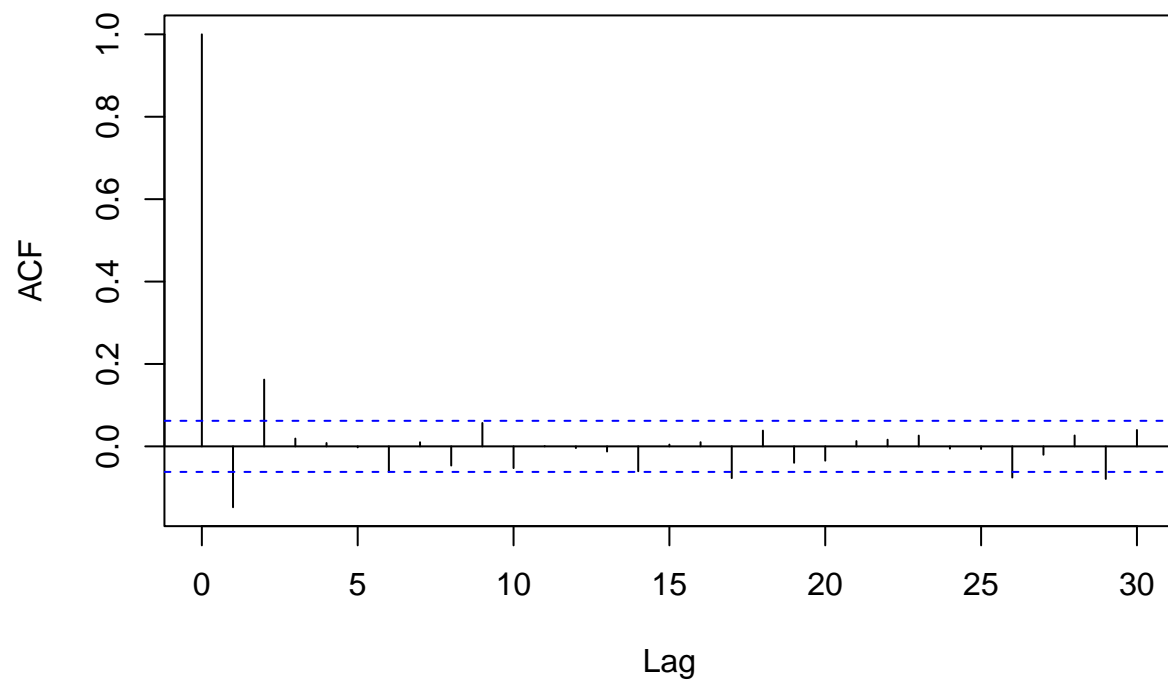
a)

```
ma <- ts(as.vector(t(read.table("ma_data.dat"))))  
plot(ma, main = "Observations over time")
```



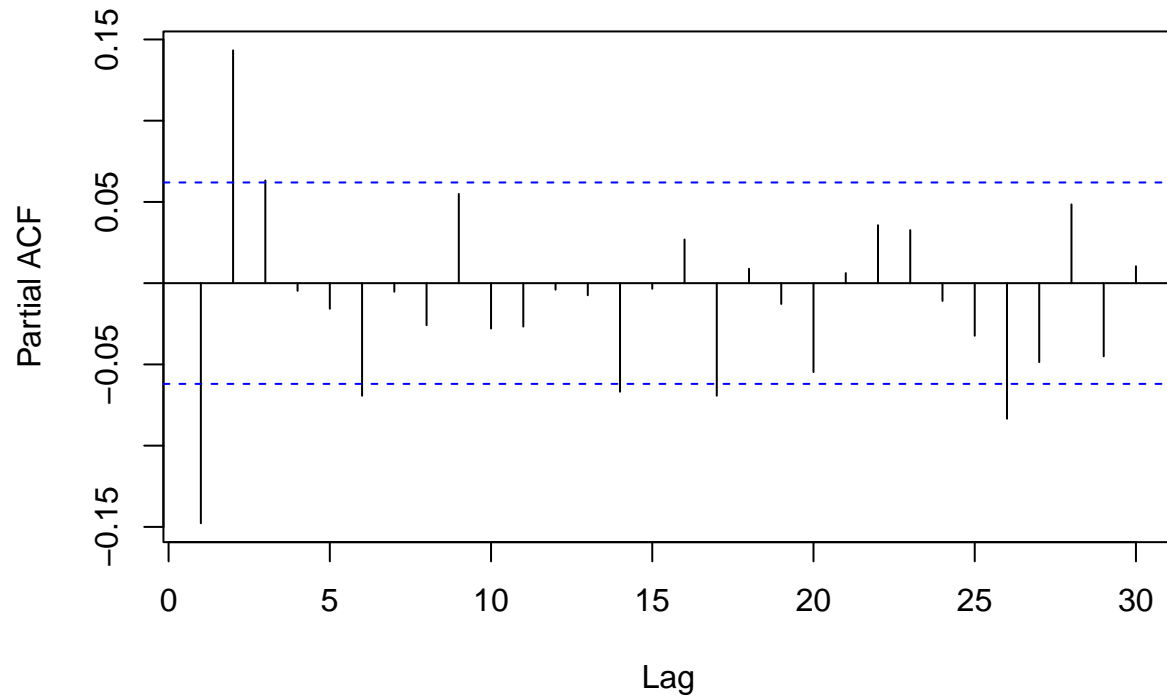
```
acf(ma, main = "ACF values of observations")
```

ACF values of observations



```
pacf(ma, main = "PACF values of observations")
```

PACF values of observations



The first two lag values of the ACF are outside of the error bars and the remaining values are within the error bars and are approximately 0. The first 2 lag values of the PACF are outside of the error bars but decay to within the error bars. Therefore, it is reasonable to fit an MA to these data based on the ACF and PACF and 2 is the appropriate order.

b)

```
q1b <- arima(ma, order=c(0,0,2), method = "CSS-ML")
q1b
```

```
##
## Call:
## arima(x = ma, order = c(0, 0, 2), method = "CSS-ML")
##
## Coefficients:
##      ma1      ma2  intercept
##    -0.1408  0.1704   -0.0555
## s.e.   0.0315  0.0304    0.0329
##
## sigma^2 estimated as 1.022:  log likelihood = -1429.67,  aic = 2867.33
```

Since we're using an MA(2) model based on Q1a, the model is of the form $x_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$, where μ , θ_1 , θ_2 and w_t are the estimated model parameters. The estimated coefficient values for θ_1 and θ_2 are approximately -0.1408446 and 0.1703555. Both estimates have standard errors that are less than half of

the magnitude of the coefficient estimates (0.0315209 and 0.030449, respectively), so these coefficients would be statistically significant from 0 at the 5% level of significance. The intercept estimate for μ is -0.0554754, but this is not statistically significantly different from 0 at the 5% level of significance since the standard error is more than half the size of the parameter estimate. Additionally, the estimated parameter value for w_t is 1.0107476.

c)

```
print("ACFs for fitted values")
```

```
## [1] "ACFs for fitted values"
```

```
acf(ma, plot = F, lag.max = 20)
```

```
##
## Autocorrelations of series 'ma', by lag
##
##      0      1      2      3      4      5      6      7      8      9     10
## 1.000 -0.148  0.162  0.019  0.008 -0.002 -0.062  0.010 -0.047  0.057 -0.053
##     11     12     13     14     15     16     17     18     19     20
## 0.001 -0.004 -0.012 -0.060  0.004  0.010 -0.077  0.038 -0.040 -0.035
```

```
cat("\n")
```

```
print("ACFs for theoretical values")
```

```
## [1] "ACFs for theoretical values"
```

```
ARMAacf(ma=c(q1b$coef[1],q1b$coef[2]), lag.max = 20) %>% round(3)
```

```
##      0      1      2      3      4      5      6      7      8      9     10
## 1.000 -0.157  0.162  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
##     11     12     13     14     15     16     17     18     19     20
## 0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
```

Comparing the fitted and theoretical ACF values, we see that both have an ACF value of 1 at lag 0, and similar magnitude and identical sign estimates for lags 1 and 2. While the theoretical ACF values are all 0 for values beyond lag 2 by definition, the ACF values for the fitted model slightly deviation from 0 beyond lag 2 likely due to randomness in the sample data. As seen from the previous ACF plot in Q1a, all of these values are within the 95% error bars. Therefore, it is reasonable to say that the theoretical and sample ACF values are consistent with one another.

```
print("PACFs for fitted values")
```

```
## [1] "PACFs for fitted values"
```

```
pacf(ma, plot = F, lag.max = 20)
```

```
##
## Partial autocorrelations of series 'ma', by lag
##
##      1      2      3      4      5      6      7      8      9     10     11
## -0.148  0.143  0.063 -0.005 -0.016 -0.069 -0.005 -0.026  0.055 -0.028 -0.027
##      12     13     14     15     16     17     18     19     20
## -0.004 -0.007 -0.067 -0.003  0.027 -0.069  0.009 -0.013 -0.055
```

```
cat("\n")
```

```
print("PACFs for theoretical values")
```

```
## [1] "PACFs for theoretical values"
```

```
ARMAacf(ma=c(q1b$coef[1],q1b$coef[2]), lag.max = 20, pacf = T) %>% round(3)
```

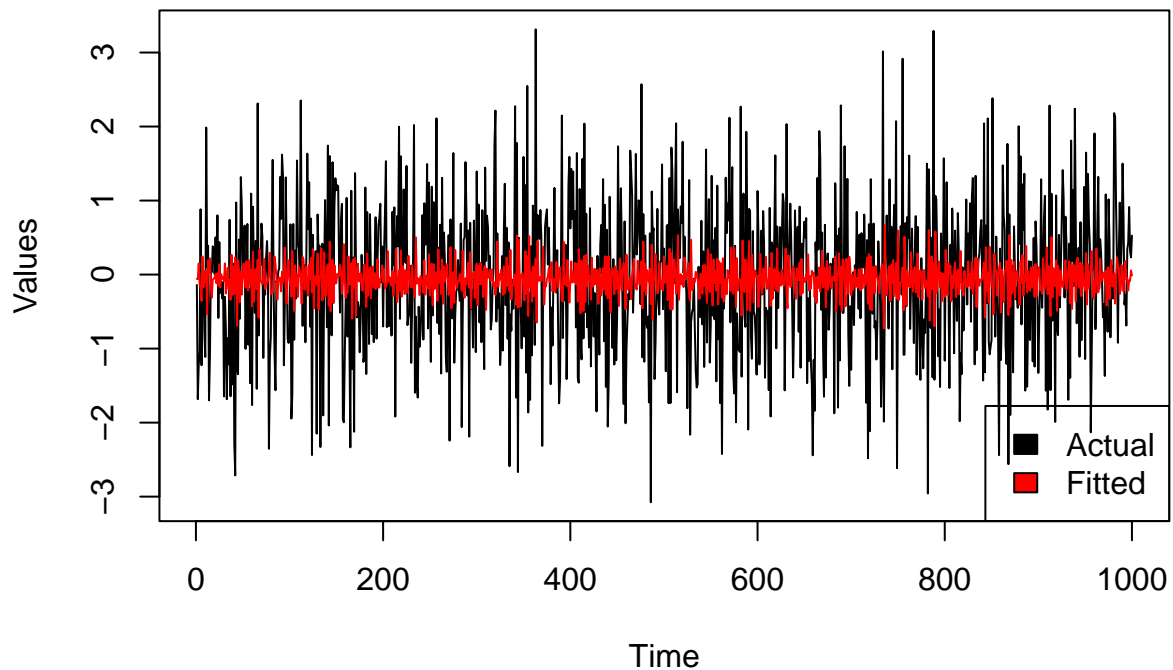
```
## [1] -0.157  0.141  0.046 -0.018 -0.010  0.002  0.002  0.000  0.000  0.000
## [11]  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
```

Comparing the fitted and theoretical PACF values, we see that both have an PACFs of similar magnitude and identical sign estimates for lags 1, 2 and 3. The theoretical PACF values approximately decay to 0 by lagged term 7, and the sample PACF values also decay to around 0 in later lags as well. However, there is some randomness in the values so there is slightly more deviation from 0 in the fitted model, but these are within the 95% error bars as seen in the PACF plot from Q1a. Therefore, it is reasonable to say that the theoretical and sample PACF values are consistent with one another.

d)

```
plot(ma, main = "Observations vs. fitted MA(2) values", ylab = "Values")
lines(ma-q1b$resid, col="red")
legend("bottomright", legend = c("Actual", "Fitted"), fill = c("black", "red"))
```

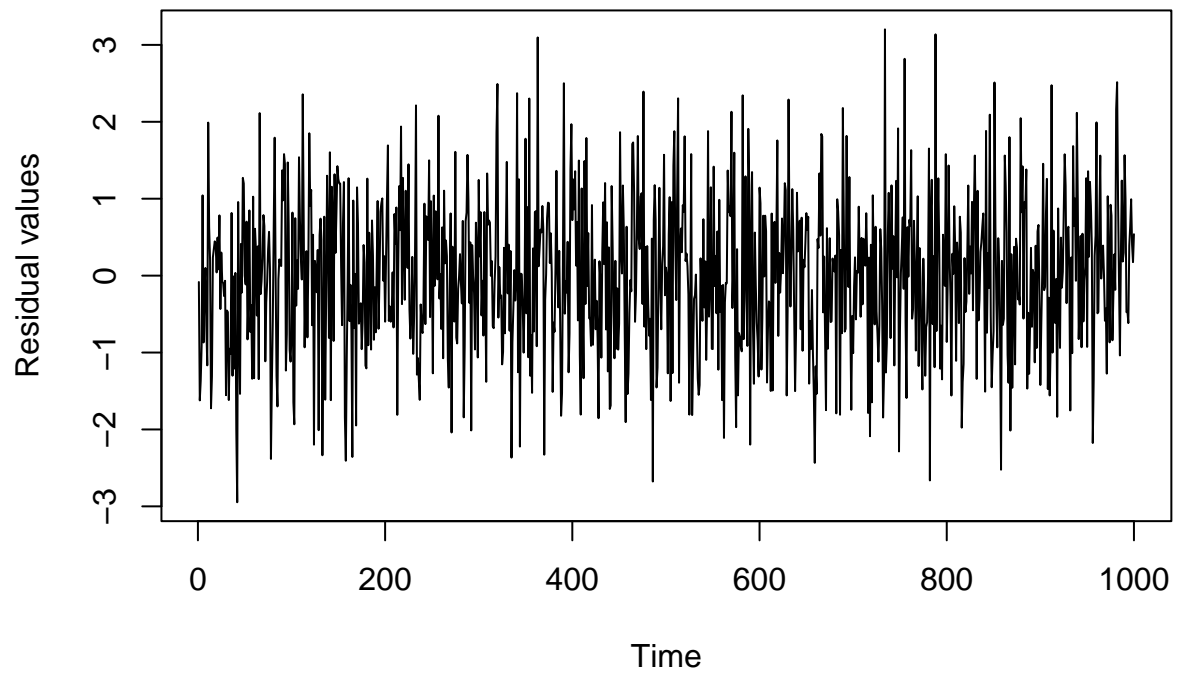
Observations vs. fitted MA(2) values



Comparing the observations to the fitted values, we see that the fitted values tend to model the center of the observations well but do not capture the magnitude of the noise fluctuations very well. However, since an $MA(q)$ is trying to model with a weighted combination of other noise terms and can only have a maximum ACF value of 0.5 the white noise terms are assumed to be uncorrelated. Therefore, this lack of fit of the noise terms is not unexpected from the reduced autocorrelation and suggests that the MA model is a good fit for the data.

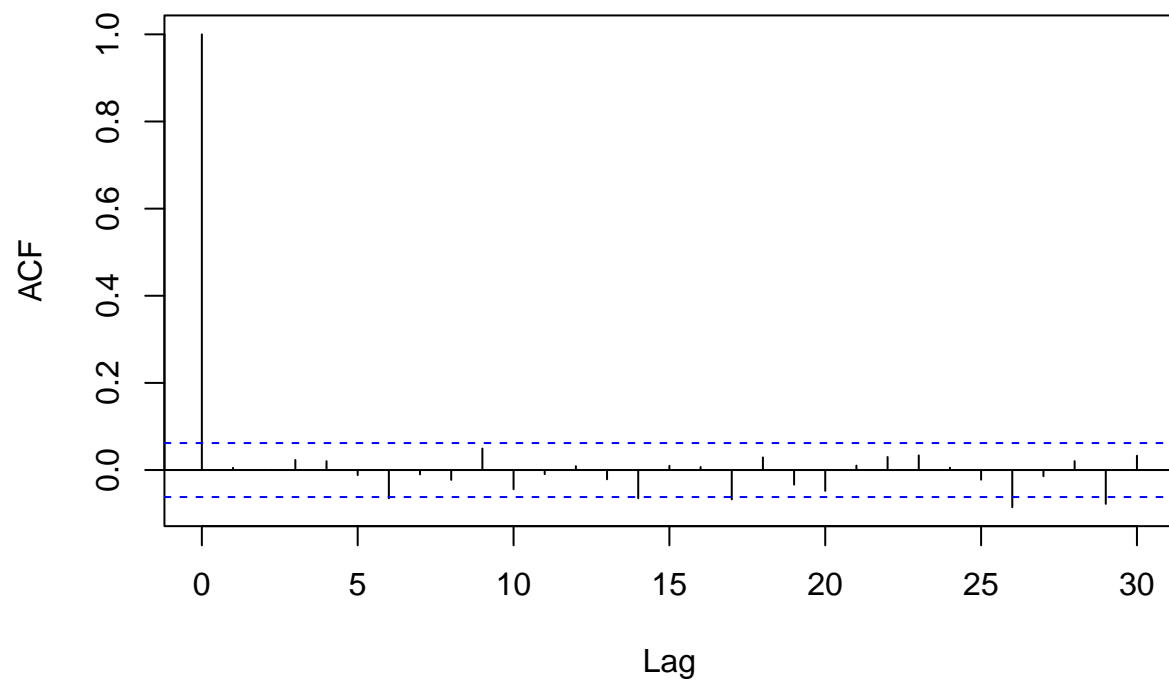
```
plot(q1b$resid, main = "Fitted MA(2) residual values", ylab = "Residual values")
```

Fitted MA(2) residual values



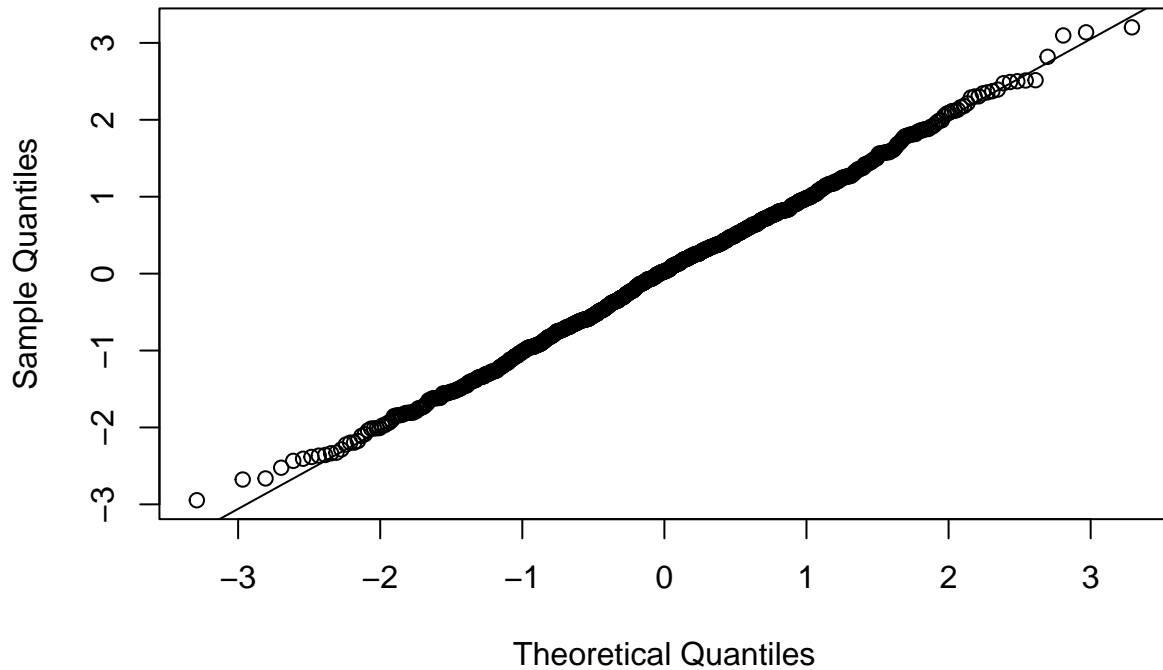
```
acf(q1b$resid, main = "ACF of fitted MA(2) residual values")
```

ACF of fitted MA(2) residual values



```
qqnorm(q1b$resid)  
qqline(q1b$resid)
```

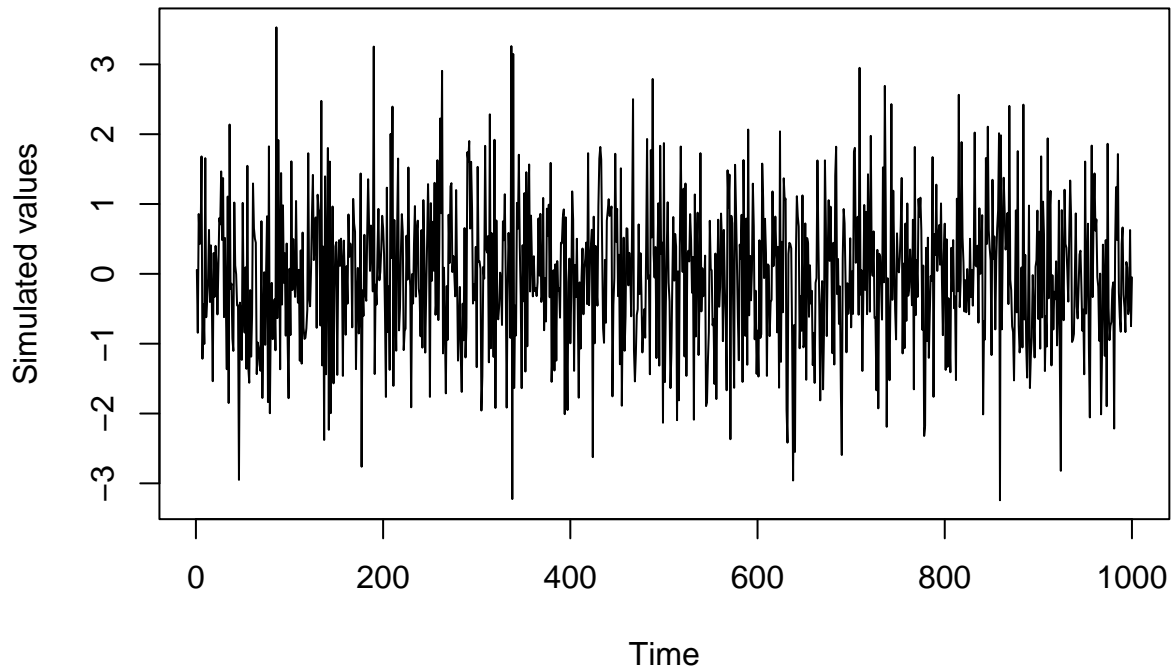

Normal Q-Q Plot



The residual values of MA(2) model are centered at 0, have relatively constant variance, and do not exhibit behaviors of autocorrelation since sequential terms do not usually have similar values based on the fluctuations in the residual values. The ACF plot shows that all the lagged values beyond lag 0 of the residuals are approximately within the 95% error bars, and Q-Q plot illustrates that the residuals follow a normal distribution since the theoretical and sample quantiles are linearly related with only minor deviations from normality at the tails. These suggest that the residuals are Gaussian white noise.

```
set.seed(36618)
plot(arima.sim(n = length(ma),
              list(ma = c(q1b$coef[1], q1b$coef[2]),
                   sd = sqrt(q1b$sigma2) + q1b$coef[3])),
     main = "Simulated MA(2) values using the fitted coefficients",
     ylab = "Simulated values")
```

Simulated MA(2) values using the fitted coefficients



The behavior of the simulated values using the coefficients from the fitted model mirror the sample data observations, which also strengthens the argument that the sample data are from an MA(2) process.

Given that the fitted values roughly approximate the sample observations (based on the definition of the MA model), the residuals display Gaussian white noise behavior, and the simulated values are similar to the sample observations, this implies that the MA(2) is appropriate and a good fit for the sample data.

e)

```
q1e <- arima(ma, order=c(2,0,0), method = "CSS-ML")  
AIC(q1e)
```

```
## [1] 2870.784
```

```
AIC(q1b)
```

```
## [1] 2867.331
```

If an AR(p) model were to be fit to the data, the appropriate order p would be 2 since the PACF values of the first 2 lags are outside of the 95% error bars as seen in Q1a. The AIC value for the AR(2) model is 2870.7844679, which is greater than the AIC for MA(2) of 2867.3310266. Therefore, the MA(2) model is preferred since it has the lower AIC value.

Question 2

a)

The model can be rewritten as $(1 - B + 0.5B^2)x_t = (1 - B)w_t$ where B is the backshift operator.

```
# AR roots
abs(polyroot(c(1,-1,0.5)))
```

```
## [1] 1.414214 1.414214
```

```
# MA roots
abs(polyroot(c(1,-1)))
```

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

Common factors: there are no common factors between the AR polynomial $\phi(B)$ and the MA polynomial $\theta(B)$, so the order of the model is ARMA(2, 1).

AR: For the AR polynomial $1 - B + 0.5B^2$, the roots are strictly greater than 1 in absolute value so the ARMA model is both stationary and causal.

MA: Since the For the MA polynomial $1 - B$, the root is not strictly greater than 1 in absolute value so the model is not invertible.

b)

The model can be rewritten as $(1 - 1.5B + 0.5B^2)x_t = (1 - B)w_t$ where B is the backshift operator.

```
# AR roots
abs(polyroot(c(1,-1.5,0.5)))
```

```
## [1] 1 2
```

```
# MA roots
abs(polyroot(c(1,-1)))
```

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

Common factors: since there is a common factor $1 - B$ in both the AR polynomial $\phi(B)$ and the MA polynomial $\theta(B)$, this can be removed and the model can be rewritten as $(B - 2)x_t = w_t$. The order of the model is ARMA(1, 0) which could be rewritten as AR(1).

AR: For the AR polynomial $(B - 2)$, the roots are strictly greater than 1 in absolute value so the ARMA model is both stationary and causal.

MA: Since the polynomial can be written as an AR(1) model and the AR models are always invertible by definition since they can be written as an MA(∞) model, the model is invertible.

c)

The model can be rewritten as $(1 - 0.8B + 0.15B^2)x_t = (1 - 0.3B)w_t$ where B is the backshift operator.

```
# AR roots
abs(polyroot(c(1,-.8,0.15)))
```

```
## [1] 2.000000 3.333333
```

```
# MA roots
abs(polyroot(c(1,-.3)))
```

```
## [1] 3.333333
```

Common factors: since there is a common factor $1 - \frac{3}{10}B$ in both the AR polynomial $\phi(B)$ and the MA polynomial $\theta(B)$, this can be removed and the model can be rewritten as $(B - 2)x_t = w_t$. The order of the model is ARMA(1, 0) which could be rewritten as AR(1).

AR: For the AR polynomial $(B - 2)$, the roots are strictly greater than 1 in absolute value so the ARMA model is both stationary and causal.

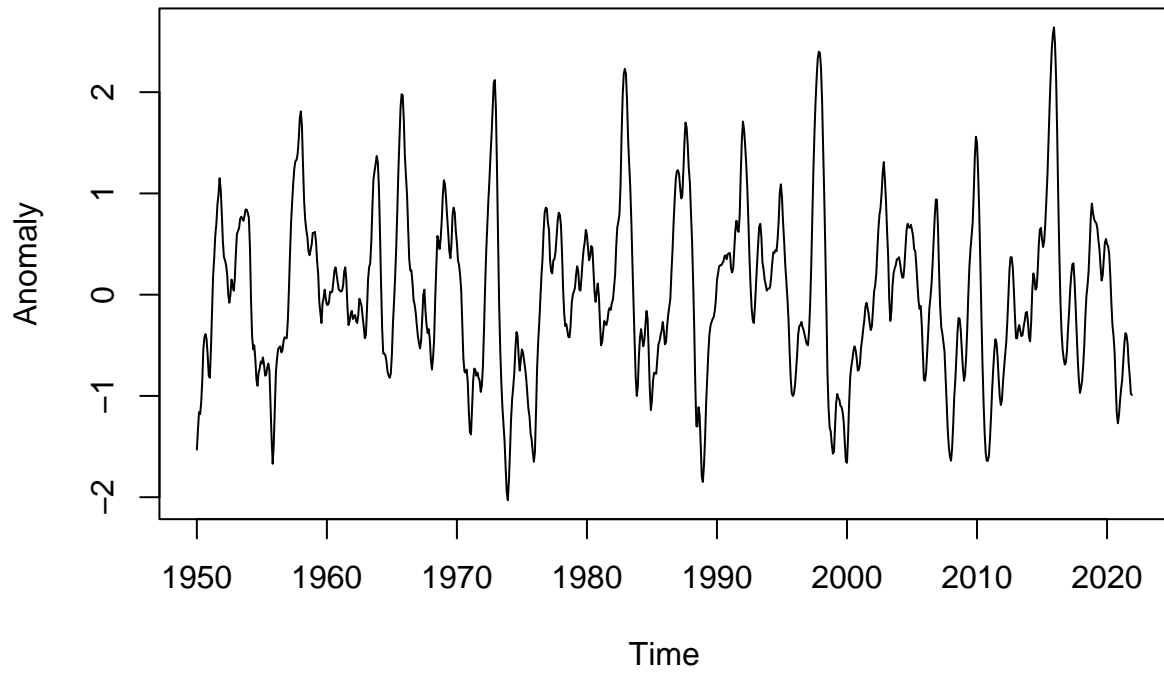
MA: Since the polynomial can be written as an AR(1) model and the AR models are always invertible by definition since they can be written as an MA(∞) model, the model is invertible.

Question 3

a)

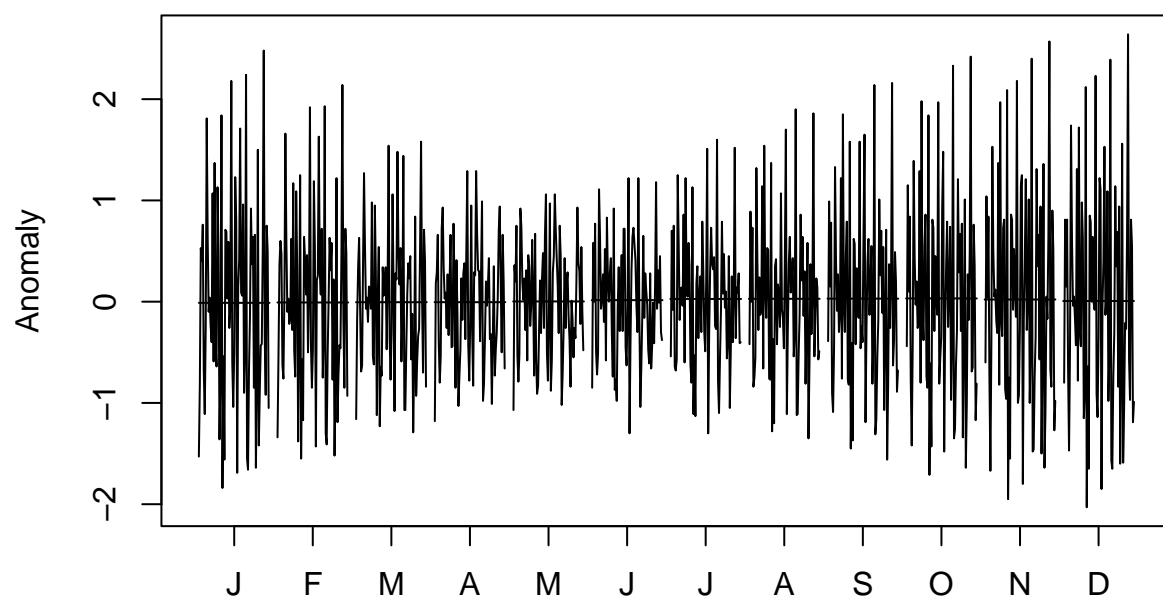
```
oni <- read.delim("oni.ascii_Dec_2021.txt", sep = "")
anom <- ts(oni$ANOM, start = 1950, frequency = 12)
plot(anom, main = "Anomalies over time", ylab = "Anomaly")
```

Anomalies over time



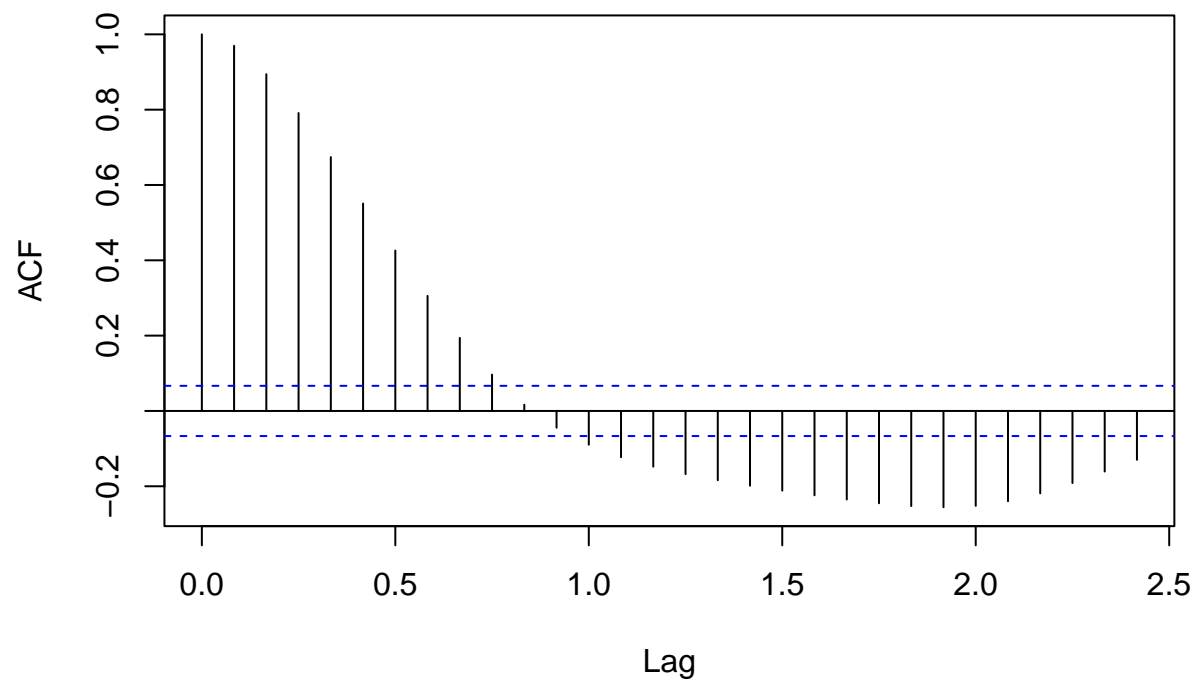
```
monthplot(anom, main = "Anomalies over time by month", ylab = "Anomaly")
```

Anomalies over time by month

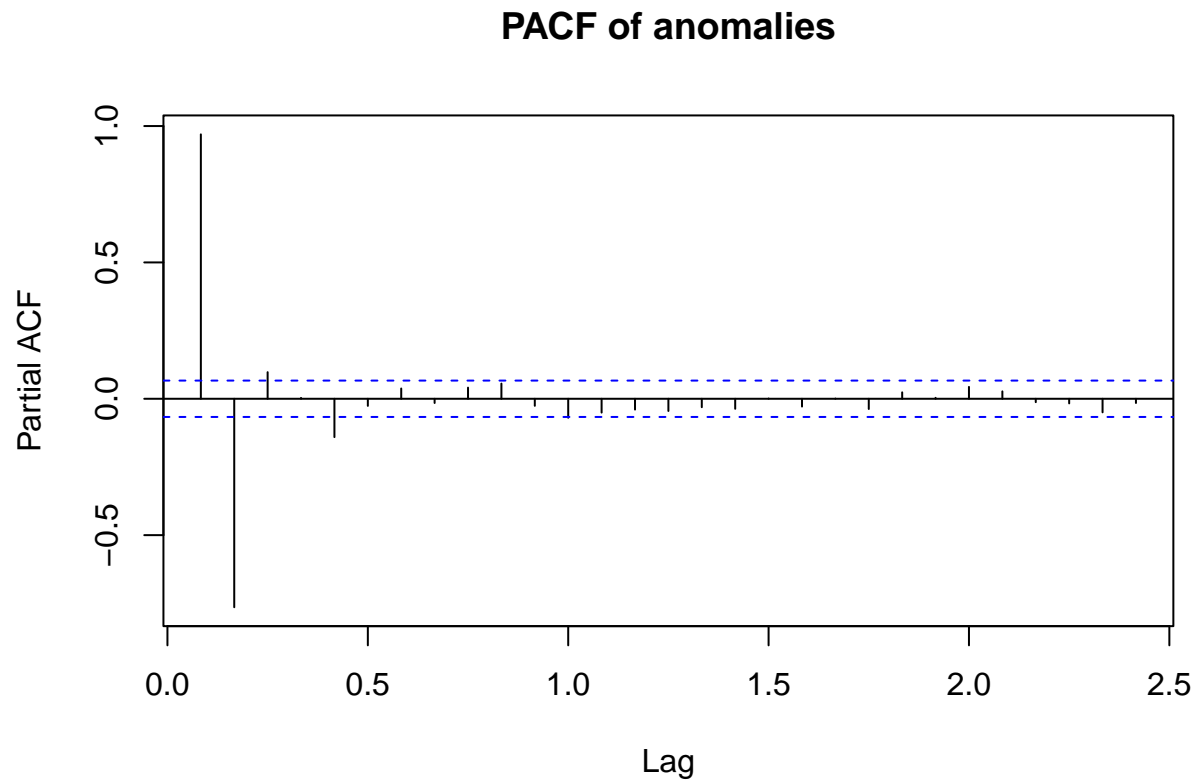


```
acf(anom, main = "ACF of anomalies")
```

ACF of anomalies



```
pacf(anom, main = "PACF of anomalies")
```



Based on the plot and monthly plot of the anomalies, there is no evidence of a mean trend since the mean stays relatively constant around 0 and the variance stays approximately constant over time suggesting the data are stationary. However, there is a evidence of a seasonal cycle as indicated by the monthly plot, where the anomalies decrease from January to April and/or May, stay relatively constant through June, and then continue to increase starting from July until December. The ACF and PACF plots suggest that this could be an AR(2) process because of the exponential decay in the lagged ACF values and the cut off in PACF values after lag 2 since most of the values are within or close to the 95% error bars. While there are some lags (3 and 5) that are slightly outside of the error bars for the PACF plot, lag 2 was selected since it is likely that the PACF values are outside due to randomness in the data and therefore the more parsimonious model is selected.

b)

```
auto.arima(anom, max.p = 5, max.q = 5, max.order = 10, stationary = T, seasonal = F, trace = T, stepwis
```

```
##
## ARIMA(0,0,0)          with zero mean      : 2128.866
## ARIMA(0,0,0)          with non-zero mean   : 2130.721
## ARIMA(0,0,1)          with zero mean      : 1013.938
## ARIMA(0,0,1)          with non-zero mean   : 1015.841
## ARIMA(0,0,2)          with zero mean      : 78.48925
## ARIMA(0,0,2)          with non-zero mean   : 80.43007
## ARIMA(0,0,3)          with zero mean      : -675.1306
## ARIMA(0,0,3)          with non-zero mean   : -673.1535
```


## ARIMA(0,0,4)	with zero mean	: -1199.996
## ARIMA(0,0,4)	with non-zero mean	: -1197.994
## ARIMA(0,0,5)	with zero mean	: -1481.07
## ARIMA(0,0,5)	with non-zero mean	: -1479.049
## ARIMA(1,0,0)	with zero mean	: -379.2739
## ARIMA(1,0,0)	with non-zero mean	: -377.3922
## ARIMA(1,0,1)	with zero mean	: -1075.641
## ARIMA(1,0,1)	with non-zero mean	: -1073.709
## ARIMA(1,0,2)	with zero mean	: -1707.165
## ARIMA(1,0,2)	with non-zero mean	: -1705.199
## ARIMA(1,0,3)	with zero mean	: -1866.701
## ARIMA(1,0,3)	with non-zero mean	: -1864.705
## ARIMA(1,0,4)	with zero mean	: -1873.705
## ARIMA(1,0,4)	with non-zero mean	: -1871.698
## ARIMA(1,0,5)	with zero mean	: -1882.325
## ARIMA(1,0,5)	with non-zero mean	: -1880.305
## ARIMA(2,0,0)	with zero mean	: -1474.635
## ARIMA(2,0,0)	with non-zero mean	: -1472.617
## ARIMA(2,0,1)	with zero mean	: -1605.674
## ARIMA(2,0,1)	with non-zero mean	: -1603.651
## ARIMA(2,0,2)	with zero mean	: -1888.339
## ARIMA(2,0,2)	with non-zero mean	: -1886.325
## ARIMA(2,0,3)	with zero mean	: -1887.596
## ARIMA(2,0,3)	with non-zero mean	: -1885.571
## ARIMA(2,0,4)	with zero mean	: -1912.779
## ARIMA(2,0,4)	with non-zero mean	: -1910.742
## ARIMA(2,0,5)	with zero mean	: -1913.989
## ARIMA(2,0,5)	with non-zero mean	: -1911.946
## ARIMA(3,0,0)	with zero mean	: -1631.241
## ARIMA(3,0,0)	with non-zero mean	: -1629.225
## ARIMA(3,0,1)	with zero mean	: -1636.841
## ARIMA(3,0,1)	with non-zero mean	: -1634.818
## ARIMA(3,0,2)	with zero mean	: -1886.519
## ARIMA(3,0,2)	with non-zero mean	: -1884.499
## ARIMA(3,0,3)	with zero mean	: -1887.328
## ARIMA(3,0,3)	with non-zero mean	: -1885.306
## ARIMA(3,0,4)	with zero mean	: -1912.954
## 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	: -1919.304
## ARIMA(4,0,0)	with zero mean	: -1649.031
## ARIMA(4,0,0)	with non-zero mean	: -1647.004
## ARIMA(4,0,1)	with zero mean	: -1639.257
## ARIMA(4,0,1)	with non-zero mean	: -1637.235
## ARIMA(4,0,2)	with zero mean	: -1908.726
## ARIMA(4,0,2)	with non-zero mean	: -1906.692
## ARIMA(4,0,3)	with zero mean	: -1914.499
## ARIMA(4,0,3)	with non-zero mean	: -1912.457
## ARIMA(4,0,4)	with zero mean	: Inf
## ARIMA(4,0,4)	with non-zero mean	: Inf
## ARIMA(4,0,5)	with zero mean	: Inf
## ARIMA(4,0,5)	with non-zero mean	: Inf
## ARIMA(5,0,0)	with zero mean	: -1723.995
## ARIMA(5,0,0)	with non-zero mean	: -1721.964

```
## ARIMA(5,0,1)          with zero mean      : -1759.155
## ARIMA(5,0,1)          with non-zero mean   : -1757.118
## ARIMA(5,0,2)          with zero mean      : -1910.828
## ARIMA(5,0,2)          with non-zero mean   : -1908.787
## ARIMA(5,0,3)          with zero mean      : -1912.457
## ARIMA(5,0,3)          with non-zero mean   : -1910.409
## ARIMA(5,0,4)          with zero mean      : -1918.229
## ARIMA(5,0,4)          with non-zero mean   : Inf
## ARIMA(5,0,5)          with zero mean      : -1909.113
## ARIMA(5,0,5)          with non-zero mean   : Inf
##
##
##
## Best model: ARIMA(3,0,5)          with non-zero mean

## Series: anom
## ARIMA(3,0,5) with non-zero mean
##
## Coefficients:
##          ar1          ar2          ar3          ma1          ma2          ma3          ma4          ma5
##          2.5050      -2.2111      0.6816      -0.1963      -0.0492      -0.5772      0.4450      0.3196
## s.e.      0.0865      0.1619      0.0799      0.0881      0.0447      0.0376      0.0654      0.0389
##          mean
##          0.0057
## s.e.      0.1018
##
## sigma^2 = 0.006199: log likelihood = 969.78
## AIC=-1919.56   AICc=-1919.3   BIC=-1871.95
```

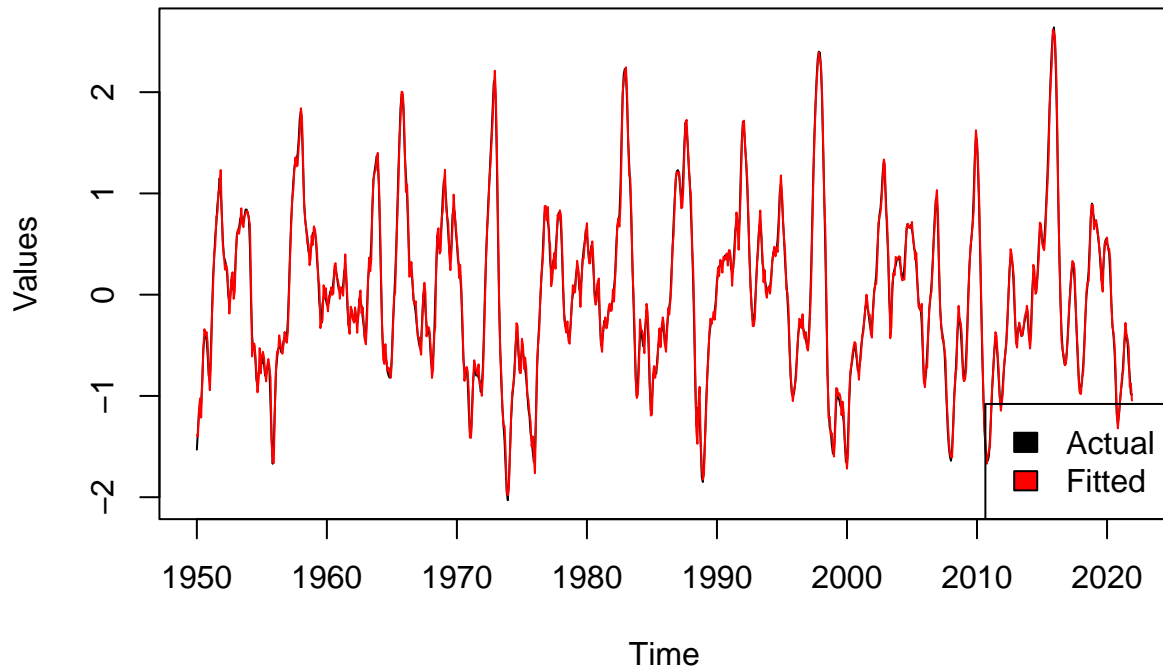
The `auto.arima` function, based on the values of the parameters inputted into the function), searches across the possible combinations of orders for the ARIMA (p, d, q) model where p is the order of lags for an AR process, d is the order of differencing, and q is the order of lags for a MA process. For the argument parameters, `max.p` and `max.q` limit the the AR and MA orders to the parameter values, respectively, and `max.order` limits the max summation of `max.p` and `max.q` if the model selection is not stepwise. The true stationary argument restricts the search to stationary models, the false seasonal argument restricts the search to non-seasonal models; trace set to T displays the ARIMA models considered, stepwise set to F searches over all possible model combinations, and approximation set to F uses MLE for each model to estimate their AIC/AICc/BIC values and chooses the selected model from these criteria.

The model selected by `auto.arima` is the ARIMA(3,0,5) model, so $p = 3$ and $q = 5$. The parameter estimates are provided in the output. With the exception of the intercept and parameter estimate for the MA 2 term, all of the estimated parameter coefficients are at least twice the size of their standard error estimates, and so are statistically significant at the 5% level.

c)

```
q3c <- arima(anom, order=c(3,0,5))
plot(anom, main = "Observations vs. fitted ARIMA(3,0,5) values", ylab = "Values")
lines(anom-q3c$resid, col="red")
legend("bottomright", legend = c("Actual", "Fitted"), fill = c("black", "red"))
```

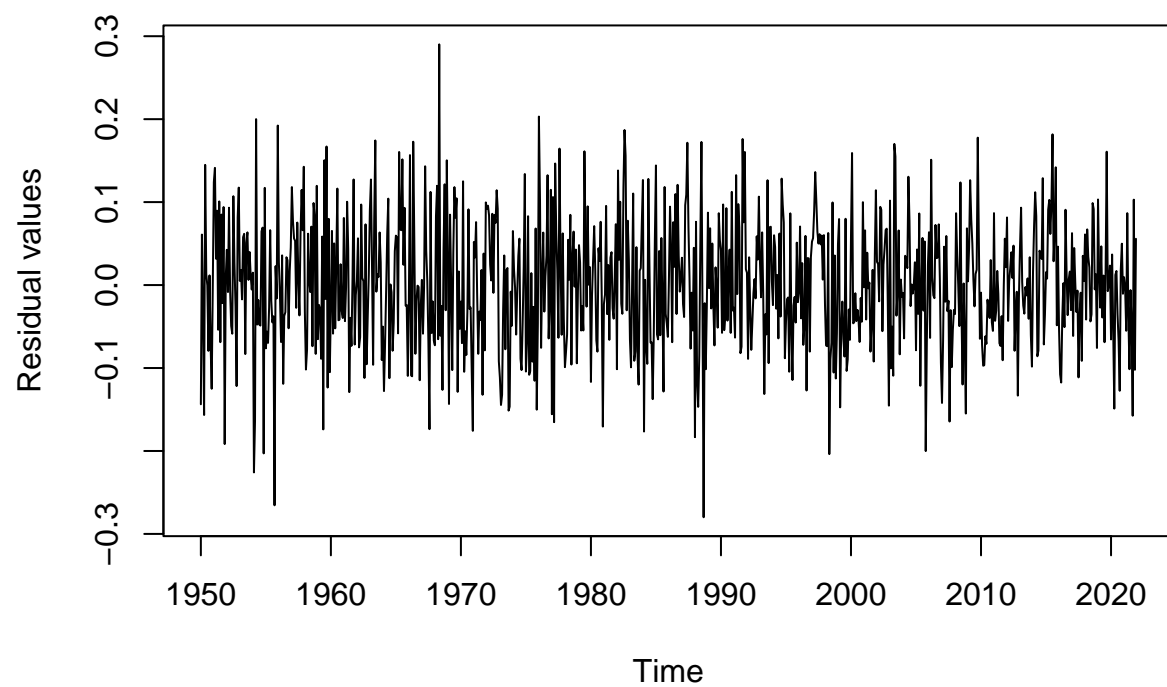
Observations vs. fitted ARIMA(3,0,5) values



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 ARIMA model is a good fit for the data.

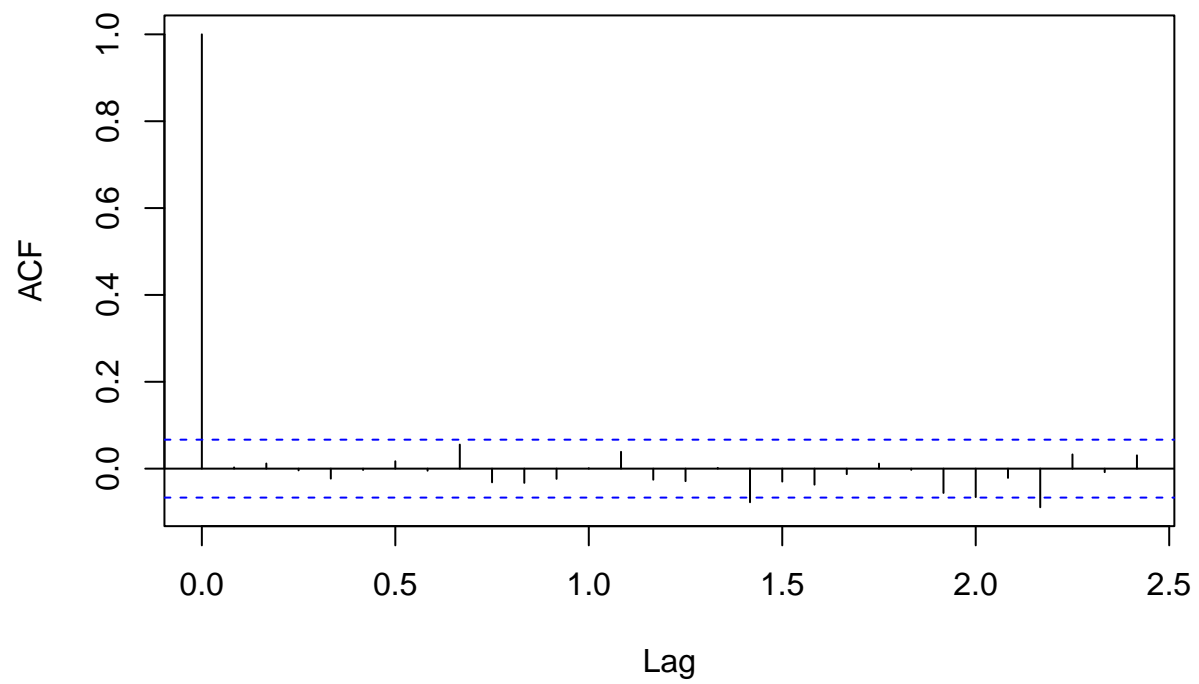
```
plot(q3c$resid, main = "Fitted ARIMA(3,0,5) residual values", ylab = "Residual values")
```

Fitted ARIMA(3,0,5) residual values



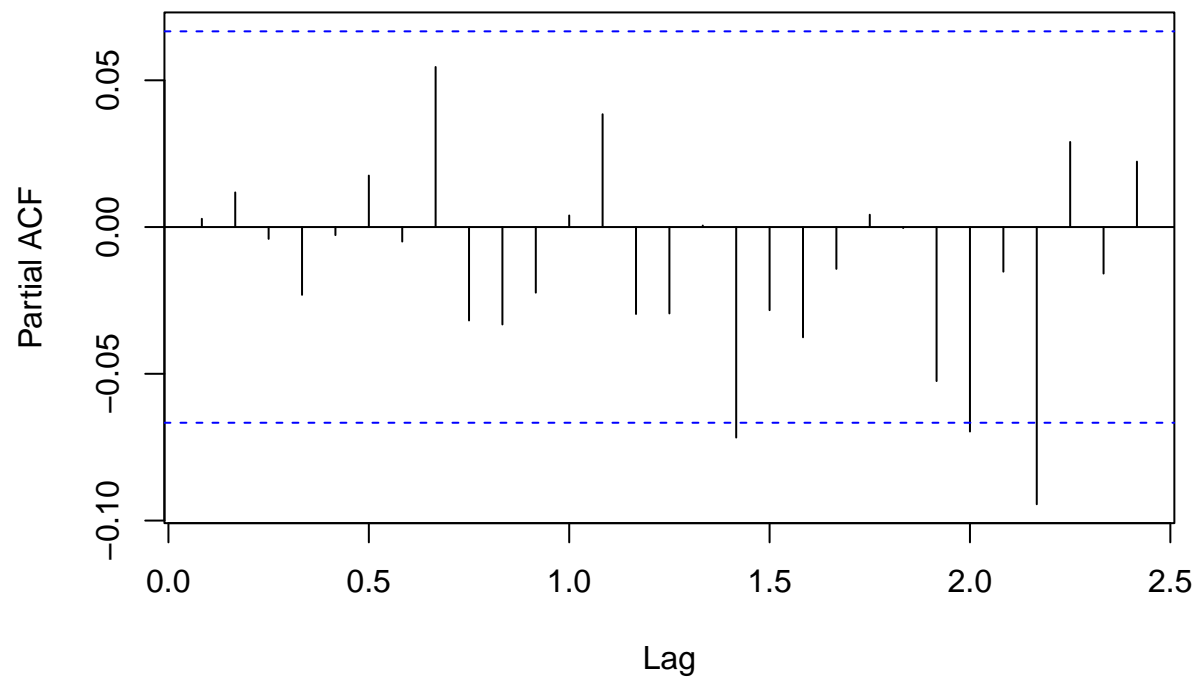
```
acf(q3c$resid, main = "ACF of fitted ARIMA(3,0,5) residual values")
```

ACF of fitted ARIMA(3,0,5) residual values

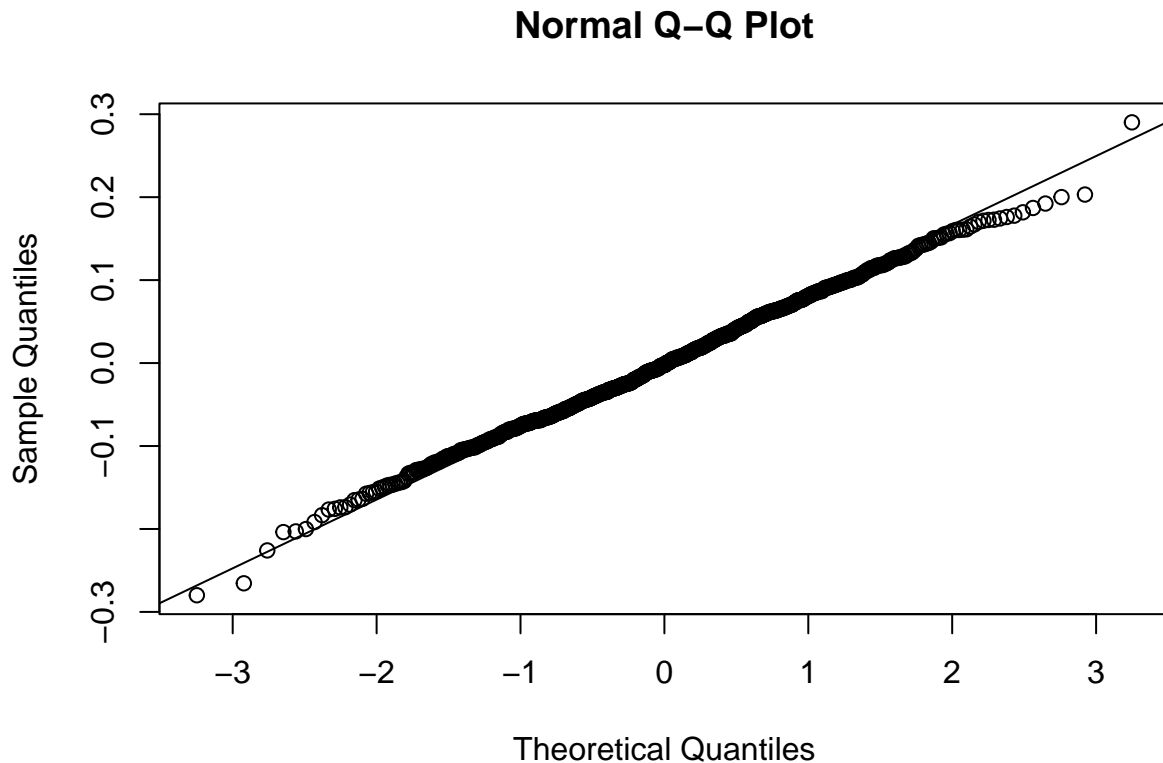


```
pacf(q3c$resid, main = "PACF of fitted ARIMA(3,0,5) residual values")
```

PACF of fitted ARIMA(3,0,5) residual values



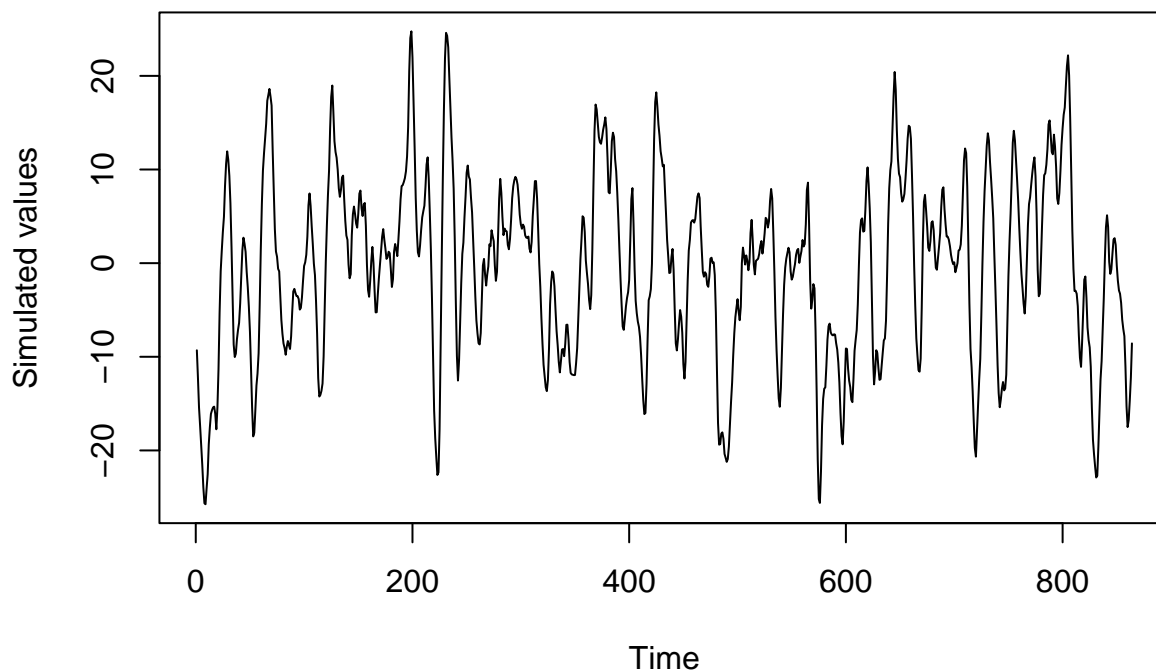
```
qqnorm(q3c$resid)  
qqline(q3c$resid)
```



The residual values of ARIMA(3,0,5) model are centered at 0, have relatively constant variance even with a few spikes and 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; there are some lags that slightly lie outside the error bars although these could correspond to the outliers seen in the plot of the data over time. 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 Gaussian white noise.

```
set.seed(36618)
plot(arima.sim(n = length(anom), list(ma = q3c$coef[4:8], ar = q3c$coef[1:3], sd = sqrt(q3c$sigma2) + q
      main = "Simulated ARIMA(3,0,5) values using the fitted coefficients",
      ylab = "Simulated values")
```

Simulated ARIMA(3,0,5) values using the fitted coefficients



The behavior of the simulated values using the coefficients from the fitted model mirror the sample data, which also strengthens the argument that the sample data could be from a $ARIMA(3,0,5)$ process.

Given that the fitted values fit the sample observations very well, the residuals display Gaussian white noise behavior, and the simulated values are similar to the sample observations, this implies that the $ARIMA(3,0,5)$ is appropriate and a good fit for the data.

d)

The issue with the monthly subseries plot is that there is seasonality in the data which varies by month. There is a cyclical trend based on the time of year where the anomalies are lowest during the summer months (May-July) and highest during the winter months (November-February). This behavior suggests that there is evidence of non-stationarity since the autocovariance in the data is dependent on time and not just the size of the lag between times. Some potential ways to correct for this non-constant variance would be to apply transformations to the data such as a logarithm or square root to try to stabilize the variance. Different modeling forms could also be considered, such as applying differencing and/or seasonal differencing on the data using SARIMA models. More parametric approaches could also be considered, such as GLS if the functional form of the variance is known or seasonal trend decomposition with loess.