BUAN6357_Homework2_Bhatia

Rishabh Bhatia

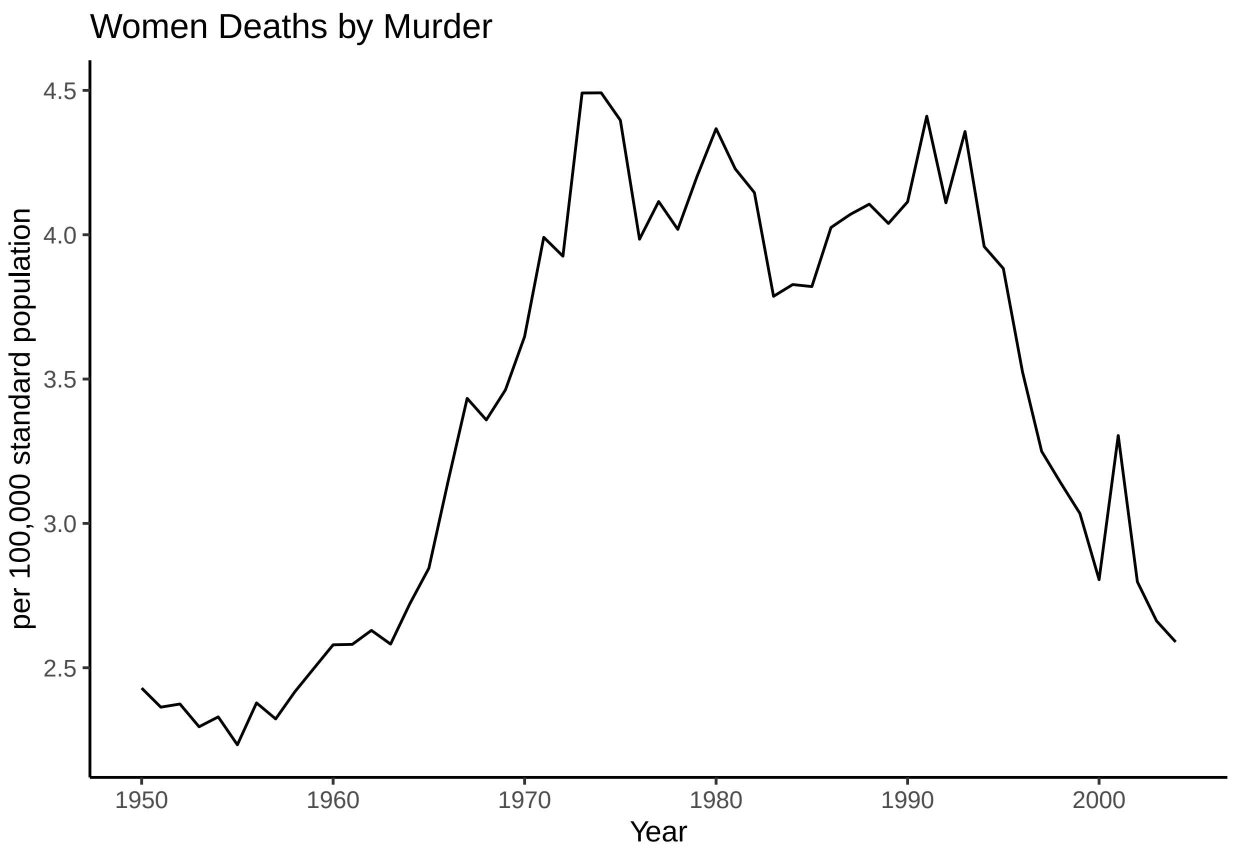
09/24/2020

Data Set: wmurder. This data set provides information on the number of women murdered each year (per 100,000 standard population) in the U.S. between 1950 and 2004.

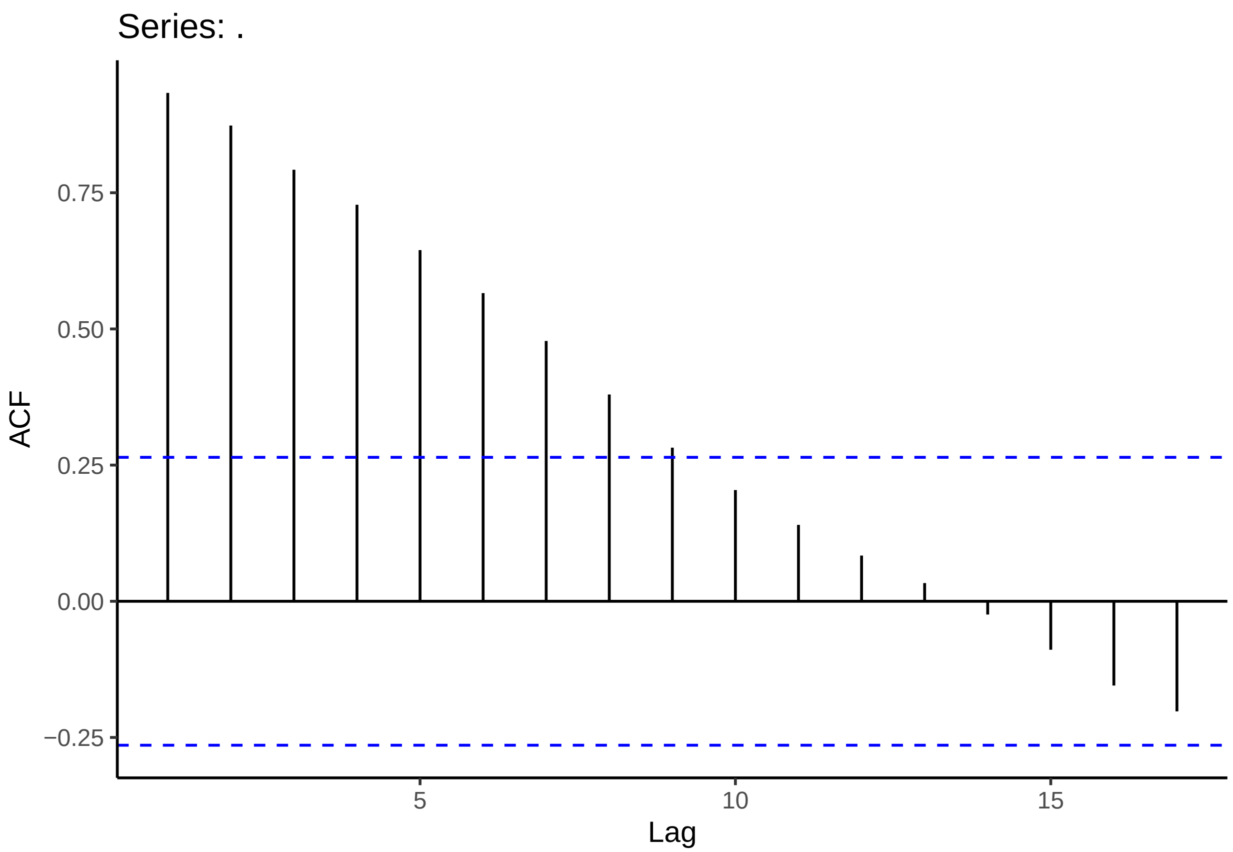
```
pacman::p_load(fpp2, dplyr, gridExtra, urca)
theme_set(theme_classic())

data("wmurders")
```

1) By studying appropriate graphs of the series in R, find an appropriate ARIMA(p,d,q) model using first difference. If there are two equally likely candidate models, then choose the one with a moving average process (MA).

```
set.seed(42)
wmurders %>% autoplot() +
  ggtitle("Women Deaths by Murder") +
  xlab("Year") + ylab("per 100,000 standard population")
```



```
wmurders %>% ggAcf()
```

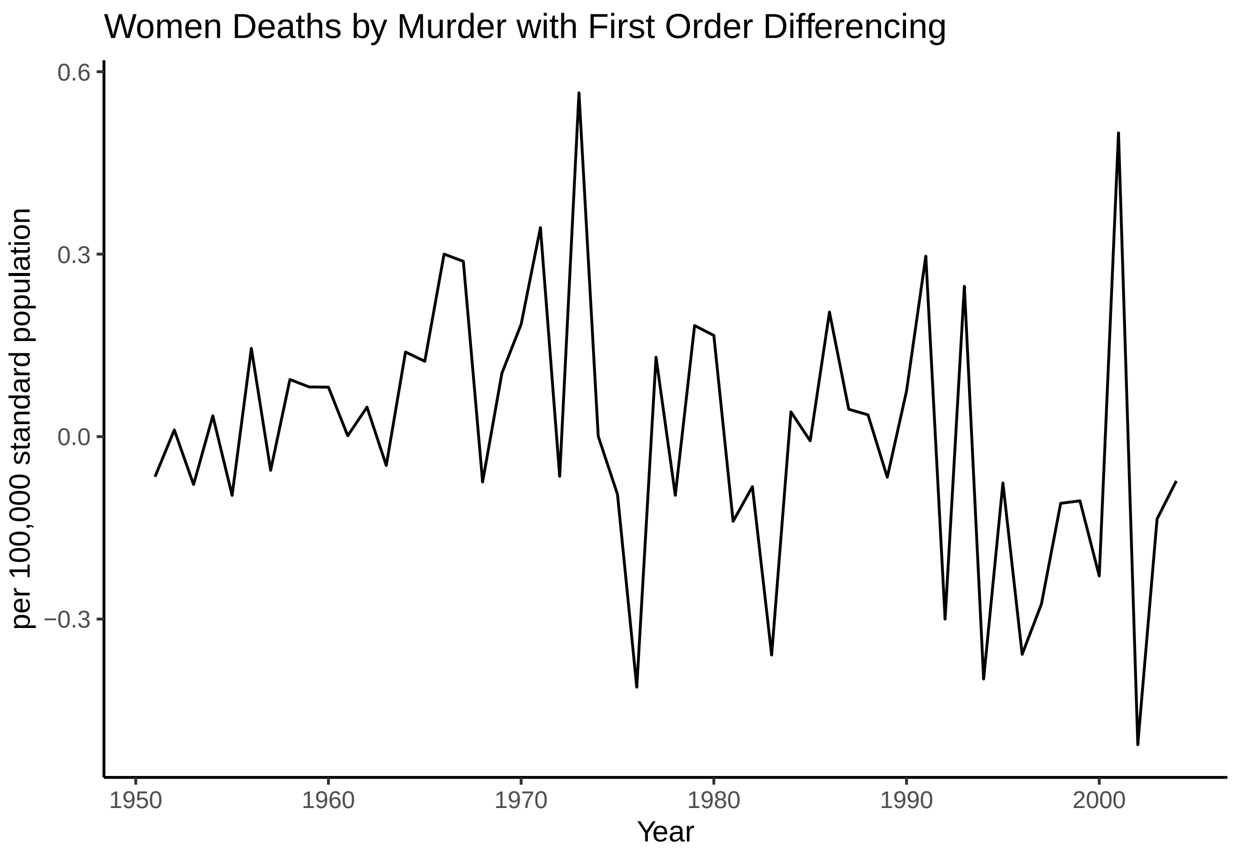


The line plot shows a positive trend from mid 1950s to mid 1970s. There seems to be a stationary trend from mid 1970s to mid 1990s. After that, there is a downward trend till 2004 with a spike in 2001. There is no seasonality at all. The auto-correlation plot also shows that there is a strong trend.

```
wmurders %>% ndiffs()
```

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

```
wm.ts <- wmurders %>% diff()
wm.ts %>% autoplot() +
  ggtitle("Women Deaths by Murder with First Order Differencing") +
  xlab("Year") + ylab("per 100,000 standard population")
```

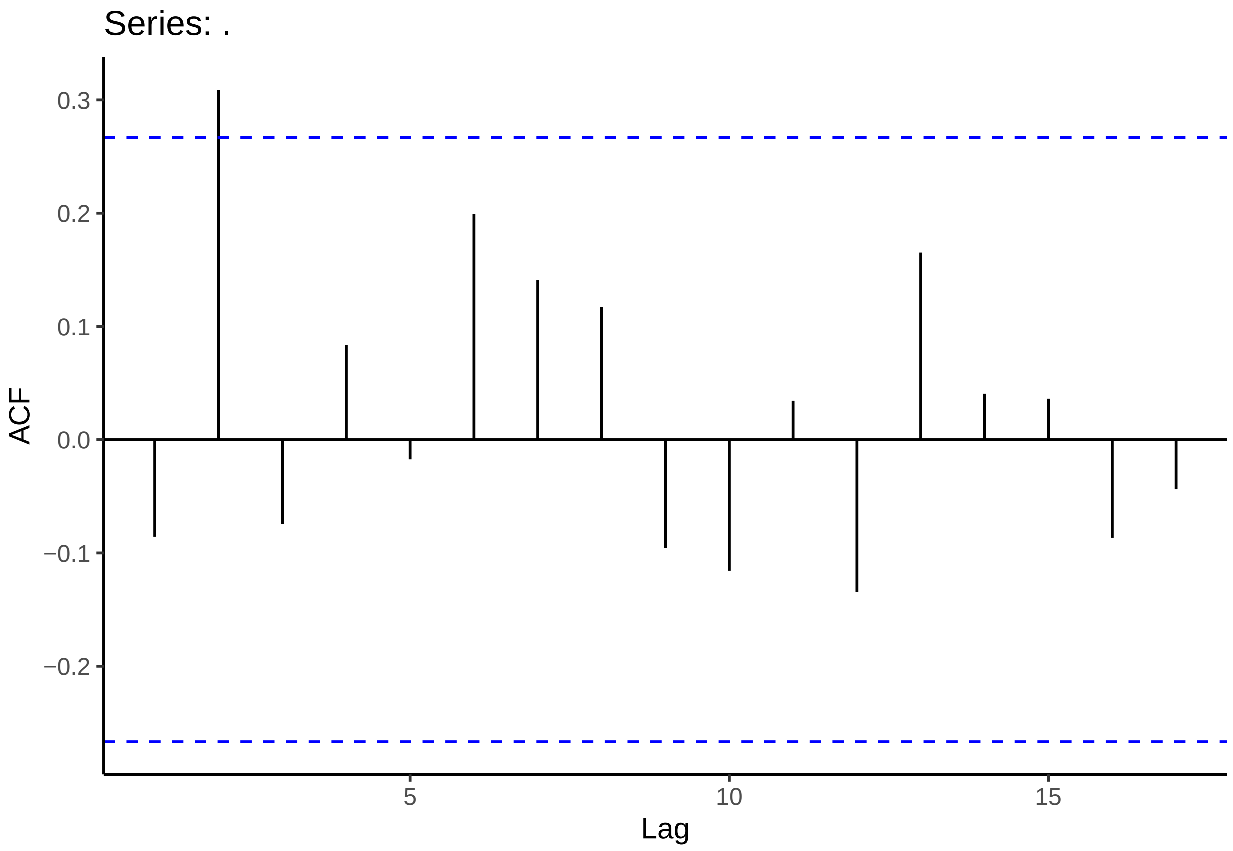


```
wm.ts %>% ur.kpss() %>% summary()
```

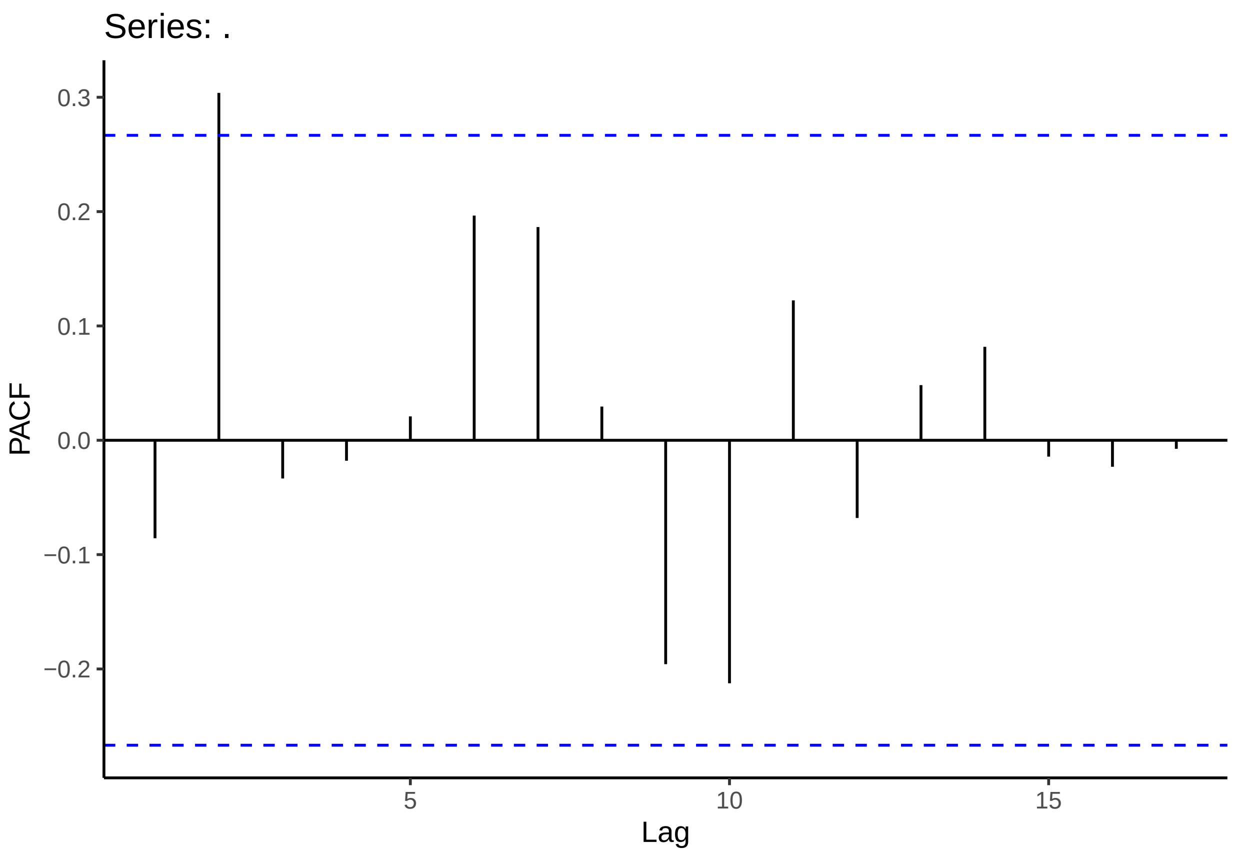
```
##
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 3 lags.
##
## Value of test-statistic is: 0.4697
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

From the time plot we can see that the first order differencing leads to a similar graph as the line graph above. The time series plot shows a positive trend from mid 1950s to mid 1970s. There seems to be a stationary trend from mid 1970s to mid 1990s. After that, there is a downward trend till 2004 with a spike in 2001. There is no seasonality at all. The auto-correlation plot also shows that there is a strong trend.The Kwiatkowski et al. Unit Root Test rejects the null hypothesis (i.e., the time series is stationary) providing evidence to our graphical assumption.

```
wm.ts %>% ggAcf()
```



```
wm.ts %>% ggPacf()
```



For Moving Averages (MA) series, ACF plot dies out gradually after a few lags and PACF plot dies out similarly. In general, differencing reduces positive autocorrelation and may even cause a switch from positive to negative autocorrelation and therefore MA series usually is negatively autcorrelated at lag 1 (evident from our correlation plots).

From the above assumptions, we can choose the p,d,q for ARIMA as follows:

p = 0, because there is no auto-regression part

d = 1, as mentioned in the question to take first order differencing

q -> vary q from 1 to 5

```
Arima(wmurders, order = c(0, 1, 1), include.constant = TRUE, include.drift = TRUE)
```

```
## Series: wmurders
## ARIMA(0,1,1) with drift
##
## Coefficients:
##           ma1   drift
##        -0.053  0.0031
## s.e.   0.107  0.0275
##
## sigma^2 estimated as 0.04716:  log likelihood=6.86
## AIC=-7.71   AICc=-7.23   BIC=-1.75
```

```r
Arima(wmurders, order = c(0, 1, 2), include.constant = TRUE, include.drift = TRUE)
```

```
## Series: wmurders
## ARIMA(0,1,2) with drift
##
## Coefficients:
##           ma1     ma2   drift
##       -0.0659  0.3711  0.0007
## s.e.   0.1263  0.1641  0.0355
##
## sigma^2 estimated as 0.04302:  log likelihood=9.71
## AIC=-11.43   AICc=-10.61   BIC=-3.47
```

```r
Arima(wmurders, order = c(0, 1, 3), include.constant = TRUE, include.drift = TRUE)
```

```
## Series: wmurders
## ARIMA(0,1,3) with drift
##
## Coefficients:
##           ma1     ma2     ma3   drift
##       -0.0557  0.3880  0.0273  0.0002
## s.e.   0.1401  0.2005  0.1726  0.0370
##
## sigma^2 estimated as 0.04383:  log likelihood=9.73
## AIC=-9.45   AICc=-8.2   BIC=0.49
```

```r
Arima(wmurders, order = c(0, 1, 4), include.constant = TRUE, include.drift = TRUE)
```

```
## Series: wmurders
## ARIMA(0,1,4) with drift
##
## Coefficients:
##           ma1     ma2     ma3      ma4    drift
##       -0.0951  0.4283  0.0894  -0.1151  -0.0004
## s.e.   0.1407  0.1561  0.1561   0.1406   0.0354
##
## sigma^2 estimated as 0.04401:  log likelihood=10.01
## AIC=-8.03   AICc=-6.24   BIC=3.91
```

```r
Arima(wmurders, order = c(0, 1, 5), include.constant = TRUE, include.drift = TRUE)
```

```
## Series: wmurders
## ARIMA(0,1,5) with drift
##
## Coefficients:
##           ma1     ma2     ma3      ma4      ma5   drift
##       -0.0491  0.5794  0.1857  -0.1216  -0.3079  0.0036
## s.e.   0.1345  0.1820  0.1487   0.1238   0.1302  0.0331
##
## sigma^2 estimated as 0.03943:  log likelihood=11.13
## AIC=-8.26   AICc=-5.83   BIC=5.66
```

As per the R documentation for including a constant in Arima, it is mentioned that if include.constant = TRUE, then include.mean is set to be TRUE for undifferenced series and include.drift is set to be TRUE for differenced series. Here, since the series is differenced, we set include.drift = true.

```
Arima(wmurders, order = c(0, 1, 1))
```

```
## Series: wmurders
## ARIMA(0,1,1)
##
## Coefficients:
##           ma1
##       -0.0527
## s.e.   0.1070
##
## sigma^2 estimated as 0.04628:  log likelihood=6.85
## AIC=-9.7   AICc=-9.47   BIC=-5.72
```

```
Arima(wmurders, order = c(0, 1, 2))
```

```
## Series: wmurders
## ARIMA(0,1,2)
##
## Coefficients:
##           ma1     ma2
##       -0.0660  0.3712
## s.e.   0.1263  0.1640
##
## sigma^2 estimated as 0.0422:  log likelihood=9.71
## AIC=-13.43   AICc=-12.95   BIC=-7.46
```

```
Arima(wmurders, order = c(0, 1, 3))
```

```
## Series: wmurders
## ARIMA(0,1,3)
##
## Coefficients:
##           ma1     ma2     ma3
##       -0.0557  0.3881  0.0274
## s.e.   0.1401  0.2000  0.1720
##
## sigma^2 estimated as 0.04298:  log likelihood=9.73
## AIC=-11.45   AICc=-10.64   BIC=-3.5
```

```
Arima(wmurders, order = c(0, 1, 4))
```

```
## Series: wmurders
## ARIMA(0,1,4)
##
## Coefficients:
##           ma1     ma2     ma3      ma4
##       -0.0951  0.4282  0.0893  -0.1151
## s.e.   0.1407  0.1559  0.1559   0.1407
##
## sigma^2 estimated as 0.04313:  log likelihood=10.01
## AIC=-10.03   AICc=-8.78   BIC=-0.08
```

```
Arima(wmurders, order = c(0, 1, 5))
```

```
## Series: wmurders
## ARIMA(0,1,5)
##
## Coefficients:
##           ma1     ma2     ma3      ma4      ma5
##       -0.0490  0.5802  0.1867  -0.1214  -0.3067
## s.e.   0.1346  0.1821  0.1484   0.1238   0.1296
##
## sigma^2 estimated as 0.03864:  log likelihood=11.12
## AIC=-10.25   AICc=-8.46   BIC=1.69
```

As a criteria for model selection, AICc is chosen to compare models. I ran both models with contant at one run and without constant model at the other run. From the above results, ARIMA(p = 0, d = 1, q = 2) is the best model with the lowest AICc values of -12.95 and when Arima() was run without a constant.

2) Should you include a constant term in the model? Explain your answer.

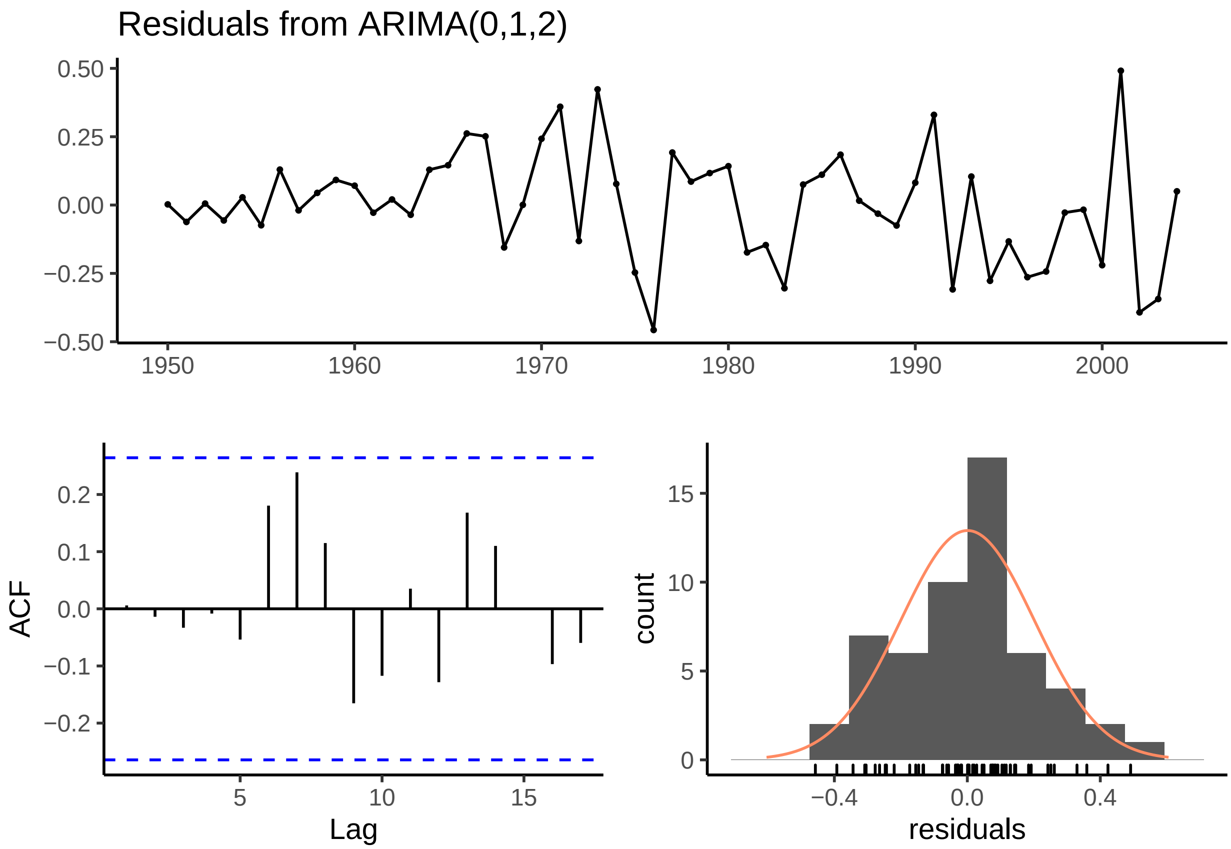
ARIMA model of the data includes first order differencing. If there is a constant in the model, twice integrated constant will yield quadratic trend, which is dangerous for forecasting. Therefore the constant is not included the model. Also, one can see above that the model with a constant does not do better compared to when without the constant.

3) Fit the model using R and examine the residuals. Is the model satisfactory?

```
(wm.fit <- Arima(wmurders,order = c(0, 1, 2)))
```

```
## Series: wmurders
## ARIMA(0,1,2)
##
## Coefficients:
##           ma1     ma2
##        -0.0660  0.3712
## s.e.    0.1263  0.1640
##
## sigma^2 estimated as 0.0422:  log likelihood=9.71
## AIC=-13.43   AICc=-12.95   BIC=-7.46
```

```
checkresiduals(wm.fit)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,2)
## Q* = 9.7748, df = 8, p-value = 0.2812
##
## Model df: 2.   Total lags used: 10
```

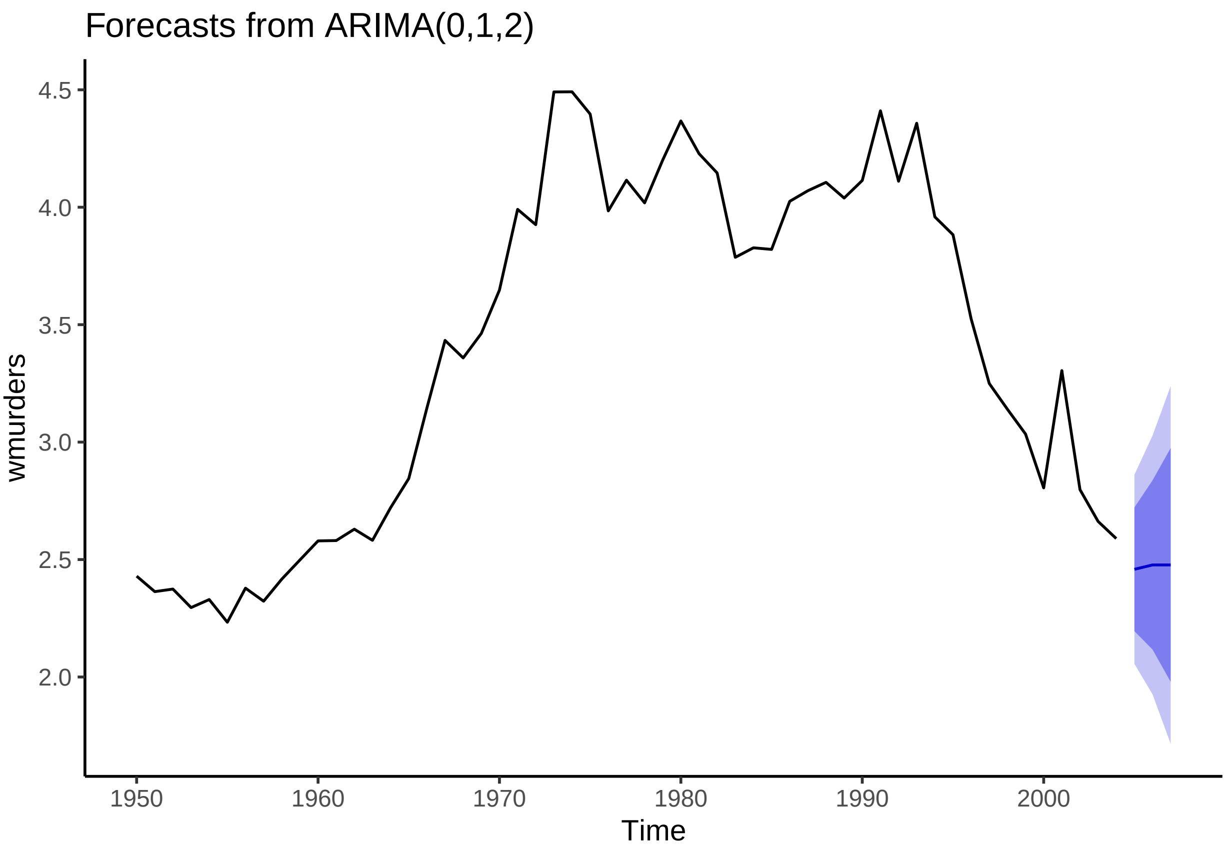
The model has the lowest AICc value compared to other models. In addition to that, the ACF plot shows that residuals are independent and identically distributed (iid) i.e., no correlation between the terms. The frequency plot of residuals shows that the distribution is normal. Hence we can presume that the model is satisfactory.

4) Forecast three times ahead and include the results in a table. Also, create a plot of the series with forecasts and prediction intervals for the next three periods shown.

```
(wm.forecast <- wm.fit %>% forecast(h = 3))
```

```
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2005       2.458450 2.195194 2.721707 2.055834 2.861066
## 2006       2.477101 2.116875 2.837327 1.926183 3.028018
## 2007       2.477101 1.979272 2.974929 1.715738 3.238464
```

```
wm.forecast %>% autoplot()
```



5) Does ARIMA() give the same model you have chosen? If not, which model do you think is better?

```r
auto.arima(wmurders, stepwise = FALSE, approximation = FALSE, seasonal = FALSE, max.d=1)
```

```
## Series: wmurders
## ARIMA(0,1,2)
##
## Coefficients:
##           ma1     ma2
##       -0.0660  0.3712
## s.e.   0.1263  0.1640
##
## sigma^2 estimated as 0.0422:  log likelihood=9.71
## AIC=-13.43   AICc=-12.95   BIC=-7.46
```

If differencing is set to 1 as we did for question 1, then auto.arima() gives the same model as the one chosen above. The reason is due to the detection of first order differencing and the subsequent negative auto-correlation factor at lag 1 to determine the time series has no auto-regression parts. If d was not set, in that case we may get different results.

```r
auto.arima(wmurders, stepwise = FALSE, approximation = FALSE, seasonal = FALSE)
```

```
## Series: wmurders
## ARIMA(0,2,3)
##
## Coefficients:
##           ma1     ma2      ma3
##       -1.0154  0.4324  -0.3217
## s.e.   0.1282  0.2278   0.1737
##
## sigma^2 estimated as 0.04475:  log likelihood=7.77
## AIC=-7.54   AICc=-6.7   BIC=0.35
```

auto.arima() gives a different model than the one chosen above. The reason is due to the detection of second order differencing overall and the subsequent negative auto-correlation factor at lag 1 to determine the time series has no auto-regression parts. Since AICc for the first model is lower, we will stick with that model.