MIS6357_Homework2_Keshan

Yash Keshan

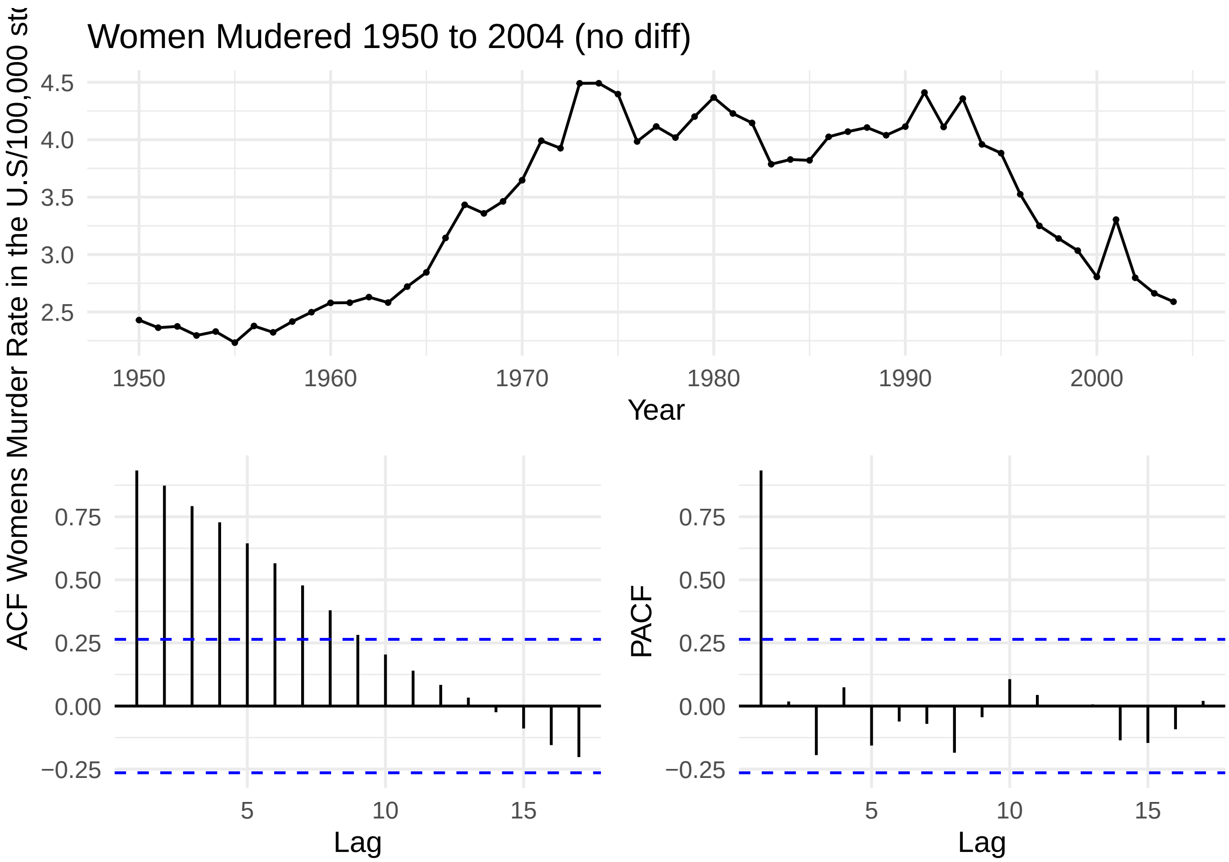
9/21/2020

```
#Importing Required Libraries and Data
#install.packages("pacman")
pacman::p_load(fpp3, fpp2, patchwork, purrr, tsbox, urca, gridExtra)
theme_set(theme_minimal())
```

```
#Storing the number of women murdered each year (per 100,000 standard population)
#in the U.S. into variable "no_women" and printing out head of the data
no_women <- wmurders
head(no_women)
```

```
## Time Series:
## Start = 1950
## End = 1955
## Frequency = 1
## [1] 2.429415 2.363364 2.374305 2.295520 2.329716 2.233017
```

```
#Generating a plot with dataset to view murder rate over the years
set.seed(42)
ggtsdisplay(no_women,xlab="Year",
            ylab="Womens Murder Rate in the U.S/100,000 std population)",
            main="Women Mudered 1950 to 2004 (no diff)")
```



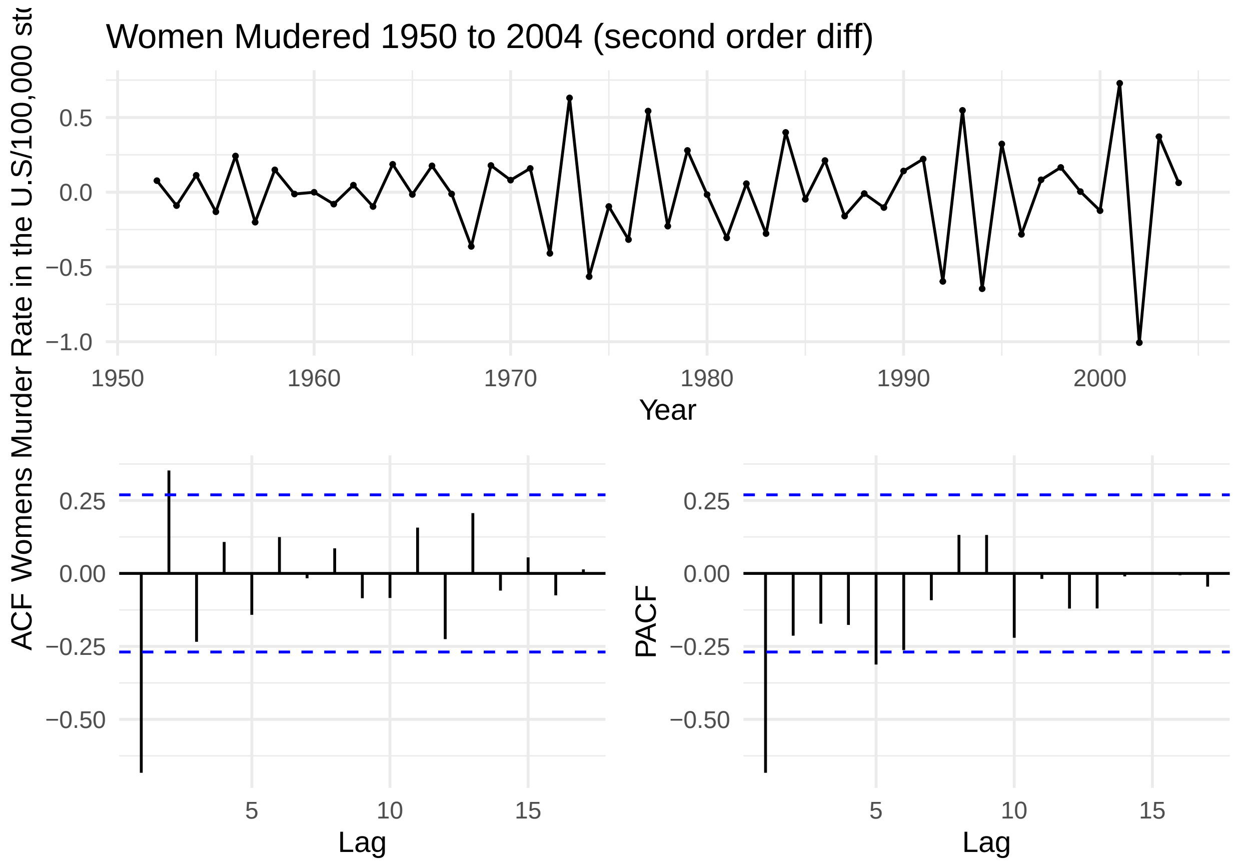
1> Through the plot we clearly observe an upwards trend in the data between 1957 - 1973. There also exist a downwards trend between 1994 - 2004. Hence, the dataset is not stationary and there is absence of white noise in the data. 2> In the above generated correlation graph we can easily identify that there is auto correlation between the first 9 lags. With the initial lags we see presence of high auto correlation. To implement ARIMA model our next steps would be to eliminate auto correlation and making mean and variance constant to make the series as stationary as possible / make it as close to white noise.

```r
#Finding number of differences required to stabilize the model using ndiffs()
ndiffs(no_women)
```

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

As we can observe we need to perform 2nd order difference to remove trends and present auto correlation and hence make the the it stable/white noise.

```r
#Performing 2nd order differences to the data
no_women_2 <- no_women %>% diff(lag = 1) %>% diff(lag = 1)
set.seed(42)
ggtsdisplay(no_women_2,xlab="Year",
            ylab="Womens Murder Rate in the U.S/100,000 std population)",
            main="Women Mudered 1950 to 2004 (second order diff)")
```

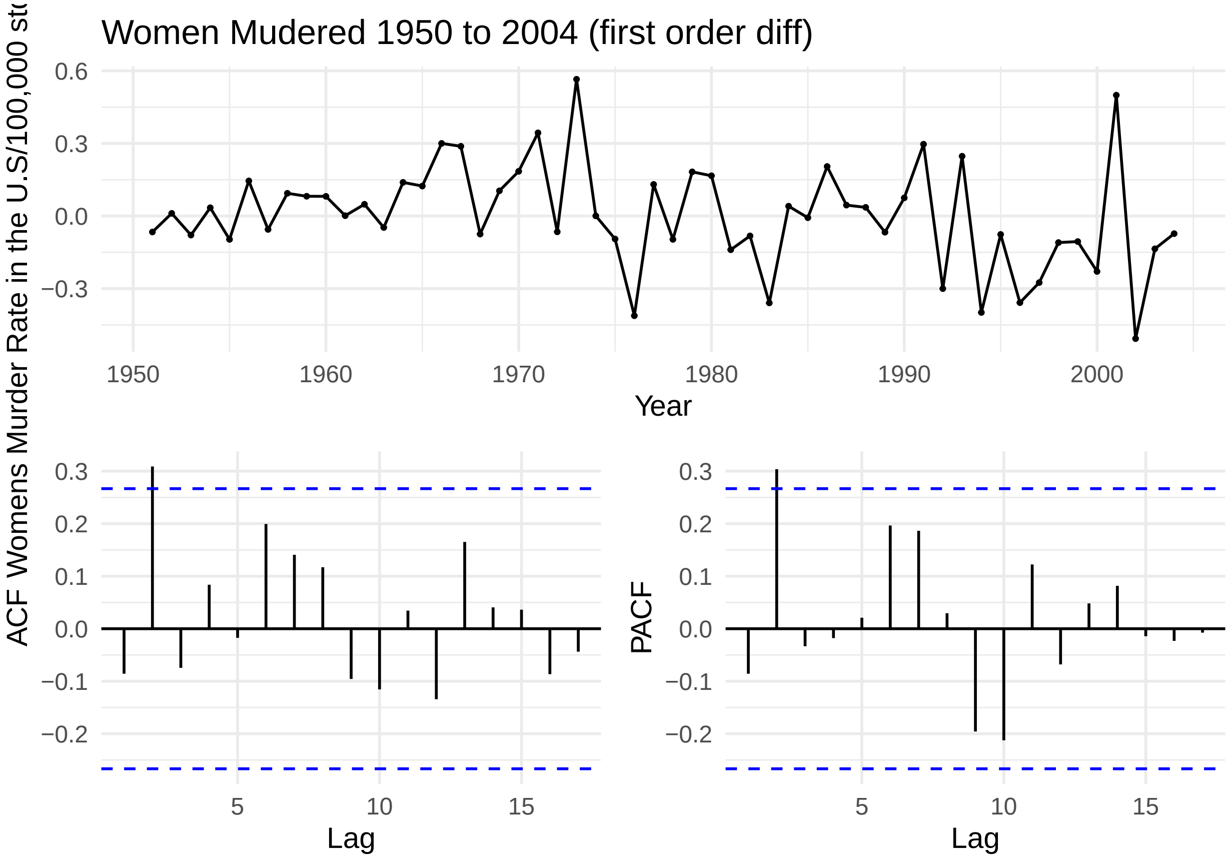


1> As we can see from the above generated graph after 2nd degree difference the generated graph is much more stationary and stable with a constant mean and constant variance. 2> With updated plot starting negative almost 90% of the lags within the range of significance level we can say that majority of auto correlation has been discarded.

```r
#We are required to perform first order difference
#Hence generating graphs for 1st order difference we get
no_women_1 <- diff(no_women) # <- Performing First Order Difference
no_women
```

```
## Time Series:
## Start = 1950
## End = 2004
## Frequency = 1
##  [1] 2.429415 2.363364 2.374305 2.295520 2.329716 2.233017 2.378179 2.322671
##  [9] 2.416556 2.498199 2.579453 2.580840 2.629293 2.581853 2.720940 2.844774
## [17] 3.144862 3.433044 3.358418 3.462620 3.647342 3.991080 3.925702 4.490962
## [25] 4.491541 4.396567 3.984491 4.115111 4.018538 4.201107 4.367459 4.228103
## [33] 4.145889 3.786691 3.827373 3.820376 4.025134 4.070130 4.105920 4.039027
## [41] 4.113978 4.410670 4.110586 4.357700 3.959040 3.882907 3.524803 3.249564
## [49] 3.139884 3.034263 2.805041 3.304467 2.797697 2.662227 2.589383
```

```r
# Generating plots for first order difference
set.seed(42)
ggtsdisplay(no_women_1,xlab="Year",
            ylab="Womens Murder Rate in the U.S/100,000 std population)",
            main="Women Mudered 1950 to 2004 (first order diff)")
```



```r
#To statistically check if the graph is stationary
#we will perform KPSS Unit Root Test
#where our null hypothesis is that the series is stationary.
#H0: Series is stationary
#H1: Series is not stationary
ur.kpss(no_women_1) %>%  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
```

t-statistics: 0.4679 < t-critical: 0.739 Since t-statistics is less than the test critical value we failed to reject the null hypothesis of test (H0: Series is stationary). Hence we conclude that series is stationary and we may incorporate first order difference in out ARIMA model.

Question 1: Answers-> Comparing the plots above with first order difference and one without difference we come to a conclusion that we should select ARIMA(0,d,q) model since we observe ACF & PCAF with similar significant number. Selecting MA model. Selecting lag ‘q = 2’ due to resulting observation noting a spike at position lag ‘q’ and no other spikes beyond lag ‘q = 2’ in the ACF model. After performing KPSS test we observed that first order difference provides with stationary series. Due to all these reasons we select ARIMA(0,1,2) where; AR = 0, D(d) = 1, MA(q) = 2

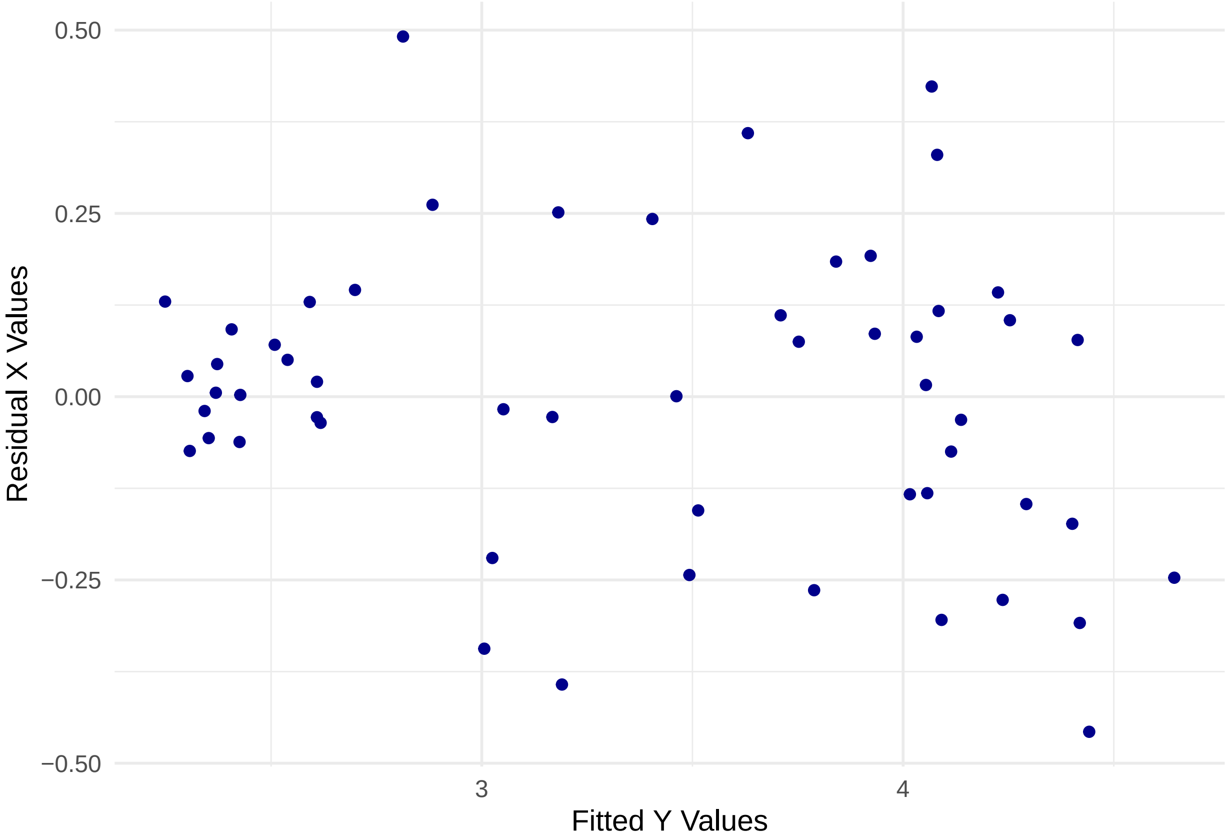
Question 2: Answers -> Our model above has first order difference a constant mean of 0, which clearly relates to the model having no significant trend or drift. Hence, since there is no drift after taking difference we do not have to include a constant.

```
set.seed(42)
#Fitting ARIMA(0,1,2) Model and generating residual plot for the same
fitmodel_1 <- Arima(no_women,order = c(0,1,2))
summary(fitmodel_1)
```

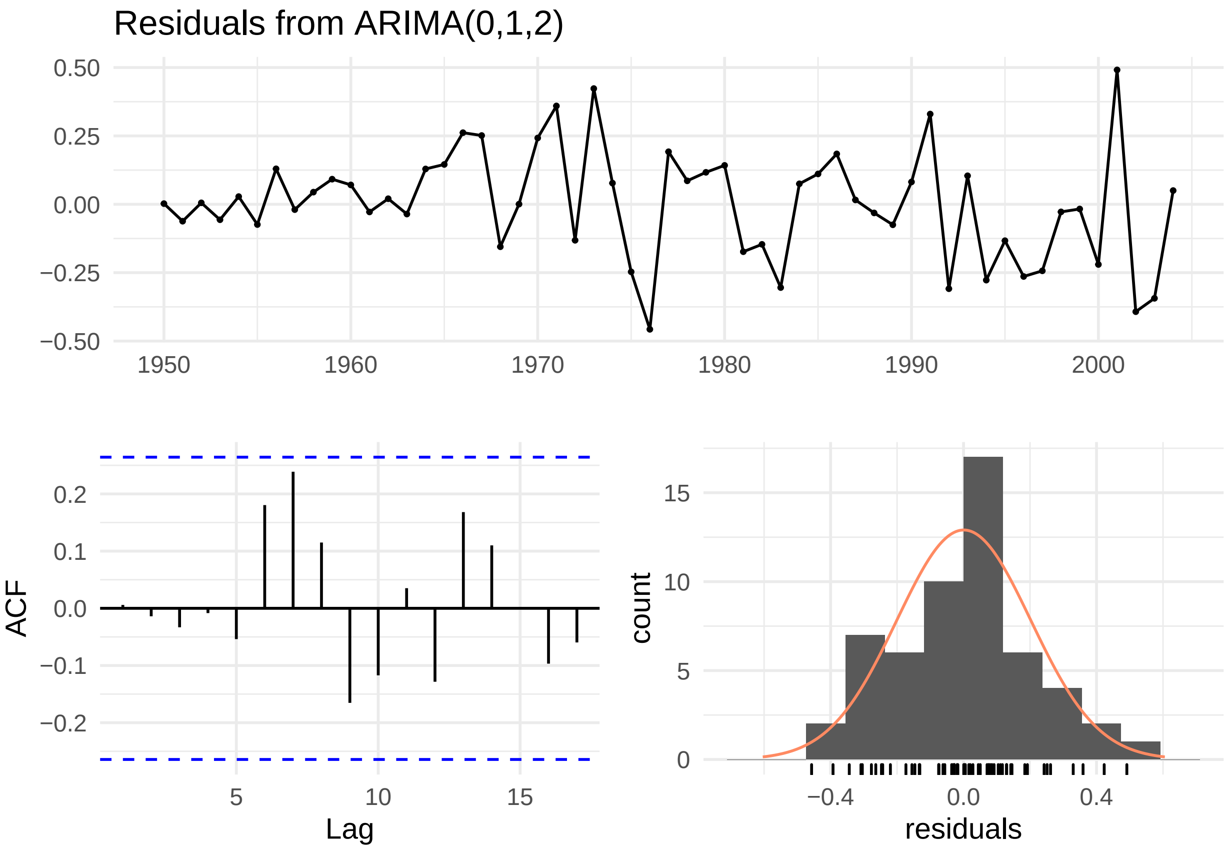
```
## Series: no_women
## ARIMA(0,1,2)
##
## Coefficients:
##           ma1     ma2
##       -0.0660  0.3712
## s.e.   0.1263  0.1640
##
## sigma^2 estimated as 0.0422:  log likelihood=9.71
## AIC=-13.43   AICc=-12.95   BIC=-7.46
##
## Training set error measures:
##                        ME      RMSE       MAE         MPE     MAPE      MASE
## Training set 0.0007242355 0.1997392 0.1543531 -0.08224024 4.434684 0.9491994
##                     ACF1
## Training set 0.005880608
```

```
ggplot(data = no_women) +
  geom_point(mapping = aes(fitted(fitmodel_1), resid(fitmodel_1)),
           col="darkblue")+ ylab("Residual X Values") +
  xlab("Fitted Y Values")
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



```
#Checking results from ARIMA(0,1,2) model
checkresiduals(fitmodel_1)
```

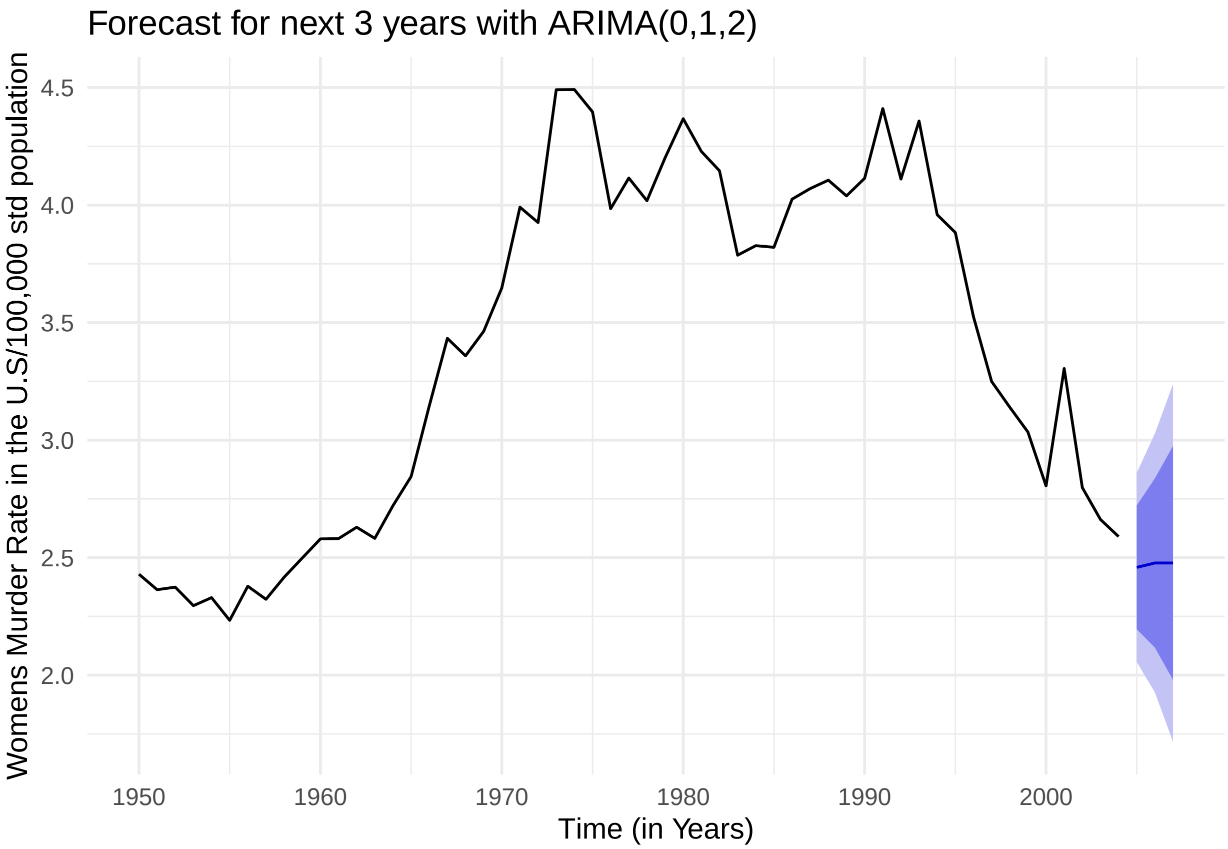


```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,2)
## Q* = 9.7748, df = 8, p-value = 0.2812
##
## Model df: 2.   Total lags used: 10
```

```
#Generating forecast for next 3 years (i.e. 2005,2006 and 2007)
#and generating plot for the same using ARIMA(0,1,2)
forecastmodel1 <- forecast(fitmodel_1, h = 3)
forecastmodel1
```

```
##      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
```

```
autoplot(forecastmodel1) + xlab("Time (in Years)") +
  ylab("Womens Murder Rate in the U.S/100,000 std population") +
  ggtitle("Forecast for next 3 years with ARIMA(0,1,2)")
```



Question 4: Answer -> Checking the result generated using Ljung-Box with null hypothesis as: H0: The model is fine H1: The model is not fine With this test we get a result where generated p-value = 0.2812 which related to not rejecting the null hypothesis. (i.e. The model is fine). Hence, model is satisfactory and consists no auto correlation in the residuals and is observed to follow a normal distribution. Finally concluding that model is Satisfactory and using ARIMA(0,1,2) we can move to the next step of forecasting using this model. Generated forecast plot shows a downwards trend to a constant and hence, decrease in the rate of murders and then reaching a constant state

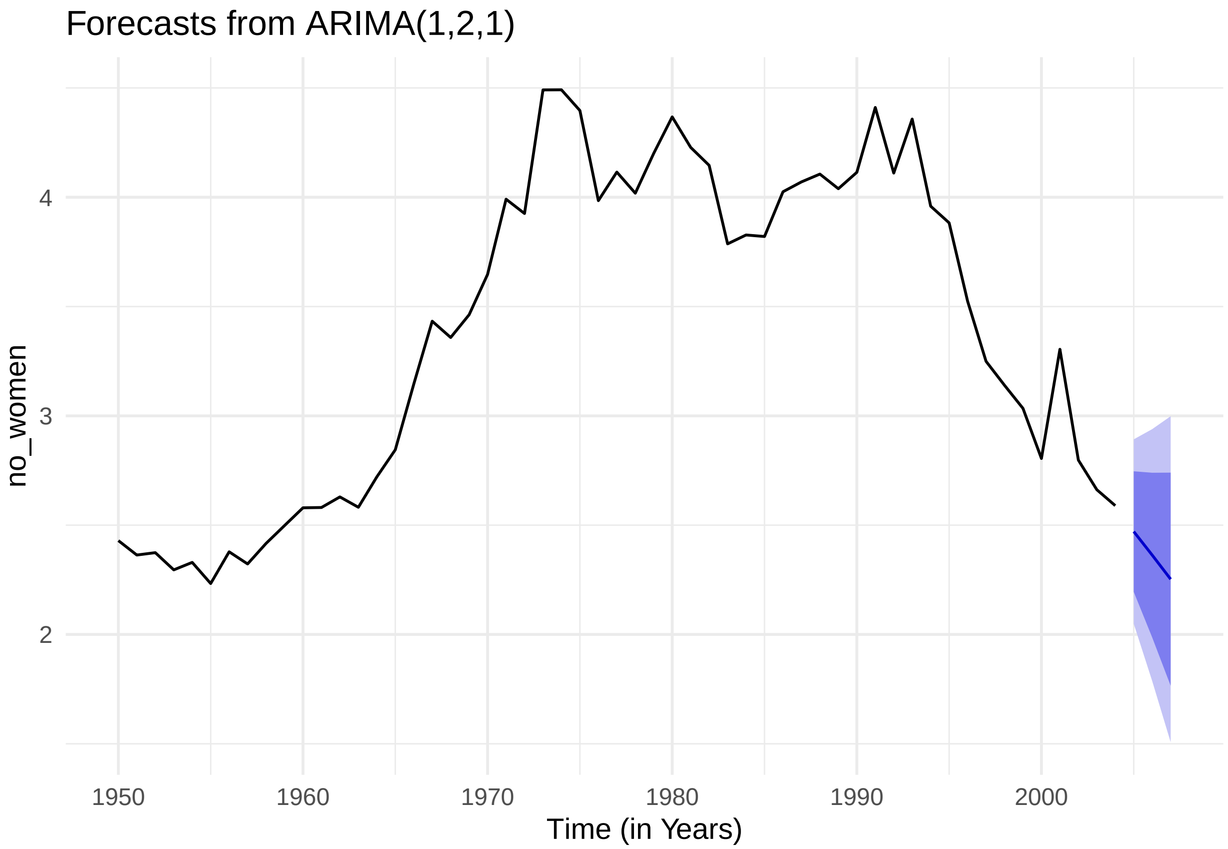
```
#Finding best ARIMA model and comparing it with the one generated above
#Results for model with approximation and stepwise
set.seed(42)
bestfitmodel_1 <- auto.arima(no_women,stepwise = TRUE, approximation = TRUE,
                             seasonal = FALSE)
summary(bestfitmodel_1)
```

```
## Series: no_women
## ARIMA(1,2,1)
##
## Coefficients:
##           ar1      ma1
##       -0.2434  -0.8261
## s.e.   0.1553   0.1143
##
## sigma^2 estimated as 0.04632:  log likelihood=6.44
## AIC=-6.88   AICc=-6.39   BIC=-0.97
##
## Training set error measures:
##                        ME      RMSE        MAE         MPE      MAPE       MASE
## Training set -0.01065956 0.2072523 0.1528734 -0.2149476 4.335214 0.9400996
##                    ACF1
## Training set 0.02176343
```

```
forecast(bestfitmodel_1,h=3)
```

```
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2005       2.470660 2.194836 2.746484 2.048824 2.892496
## 2006       2.363106 1.986351 2.739862 1.786908 2.939304
## 2007       2.252833 1.765391 2.740276 1.507354 2.998313
```

```
bestfitmodel_1 %>% forecast(h=3) %>% autoplot() + xlab("Time (in Years)")
```

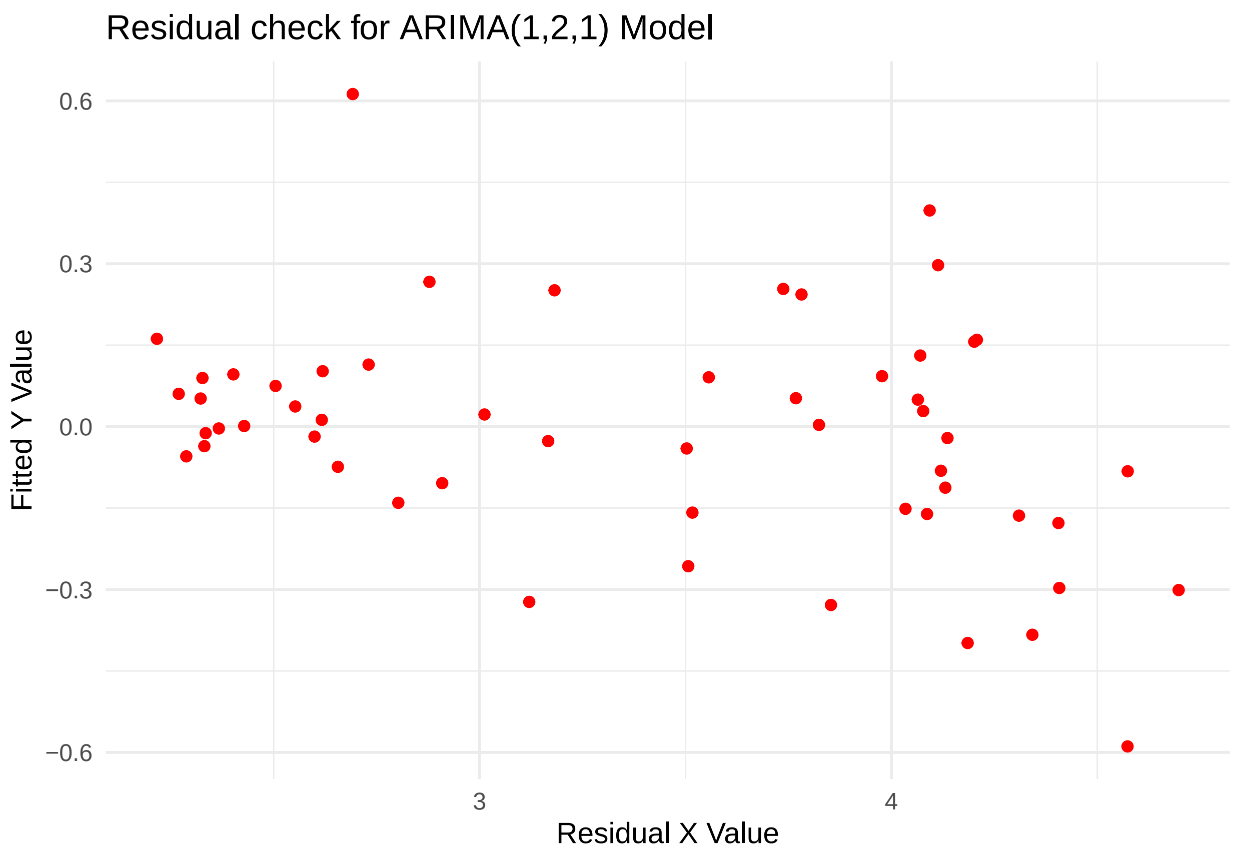


```
  ylab("Womens Murder Rate in the U.S/100,000 std population")
```

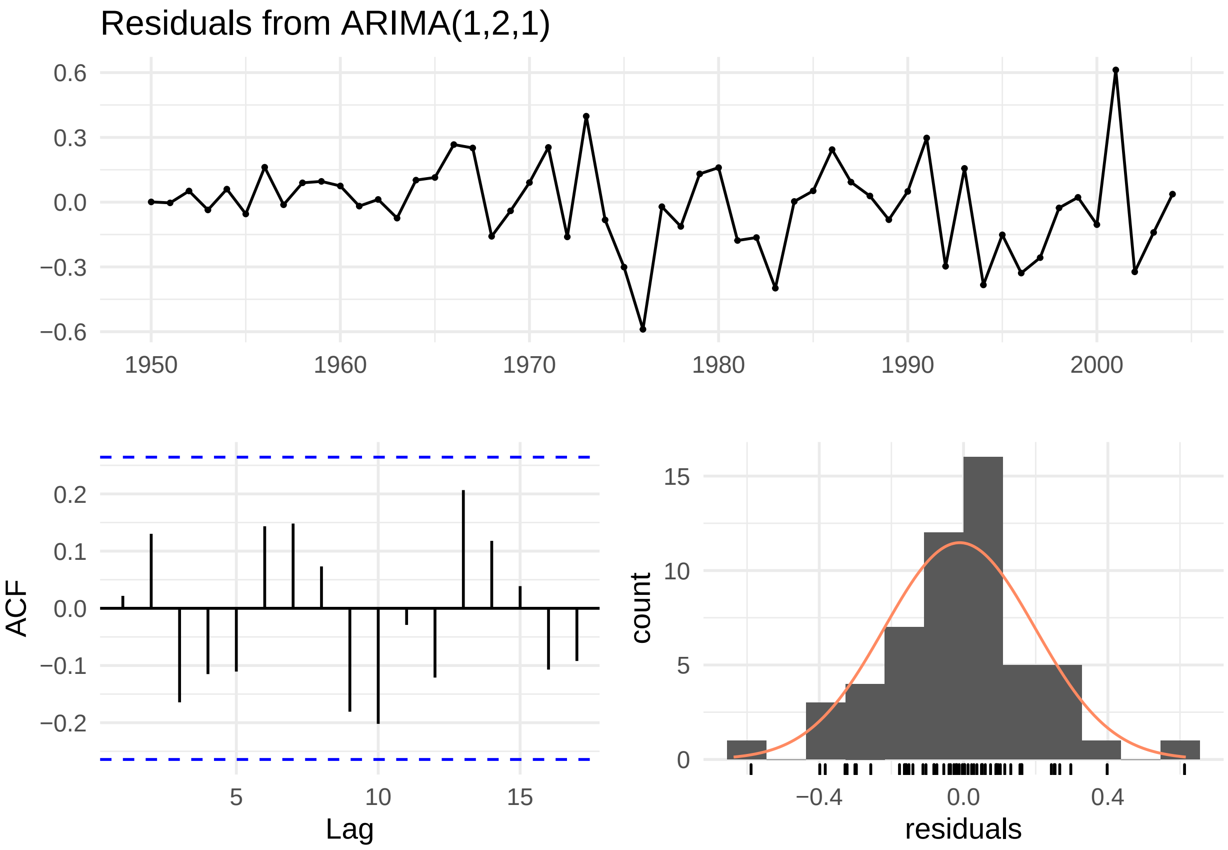
```
## $y
## [1] "Womens Murder Rate in the U.S/100,000 std population"
##
## attr(,"class")
## [1] "labels"
```

```
ggplot(data = no_women) +
  geom_point(mapping = aes(x=fitted(bestfitmodel_1), y=resid(bestfitmodel_1)),
             col="red") + xlab("Residual X Value") + ylab("Fitted Y Value") +
  ggtitle("Residual check for ARIMA(1,2,1) Model")
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



```
checkresiduals(bestfitmodel_1)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,2,1)
## Q* = 12.419, df = 8, p-value = 0.1335
##
## Model df: 2.   Total lags used: 10
```

For model with approximation and stepwise -> ARIMA(1,2,1): Checking the result generated using Ljung-Box with null hypothesis as: H0: The model is fine H1: The model is not fine With this test we get a result where generated p-value = 0.1335 which related to not rejecting the null hypothesis. (i.e. The model is fine). Hence, model is satisfactory and consists no auto correlation in the residuals and is observed to follow a normal distribution.

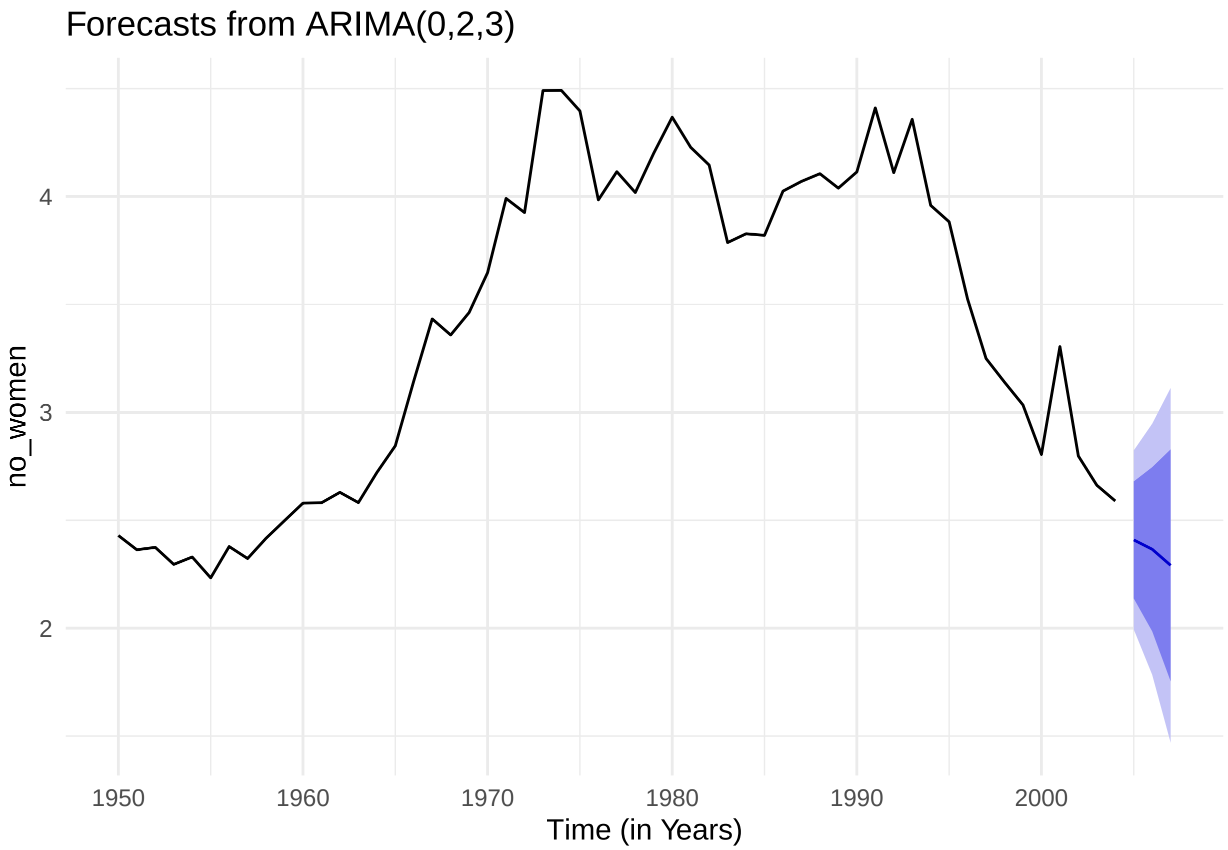
```
#Now trying ARIMA without approximation and without stepwise
#Finding best ARIMA model and comparing it with the one generated above
#Results for model with approximation and stepwise
set.seed(42)
bestfitmodel_2 <- auto.arima(no_women,stepwise = FALSE, approximation = FALSE,
                             seasonal = FALSE)
summary(bestfitmodel_2)
```

```
## Series: no_women
## 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
##
## Training set error measures:
##                       ME      RMSE       MAE        MPE     MAPE      MASE
## Training set -0.01336585 0.2016929 0.1531053 -0.3332051 4.387024 0.9415259
##                     ACF1
## Training set -0.03193856
```

```
forecast(bestfitmodel_2,h=3)
```

```
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2005       2.408817 2.137718 2.679916 1.994206 2.823428
## 2006       2.365555 1.985092 2.746018 1.783687 2.947423
## 2007       2.290976 1.753245 2.828706 1.468588 3.113363
```

```
bestfitmodel_2 %>% forecast(h=3) %>% autoplot() + xlab("Time (in Years)")
```

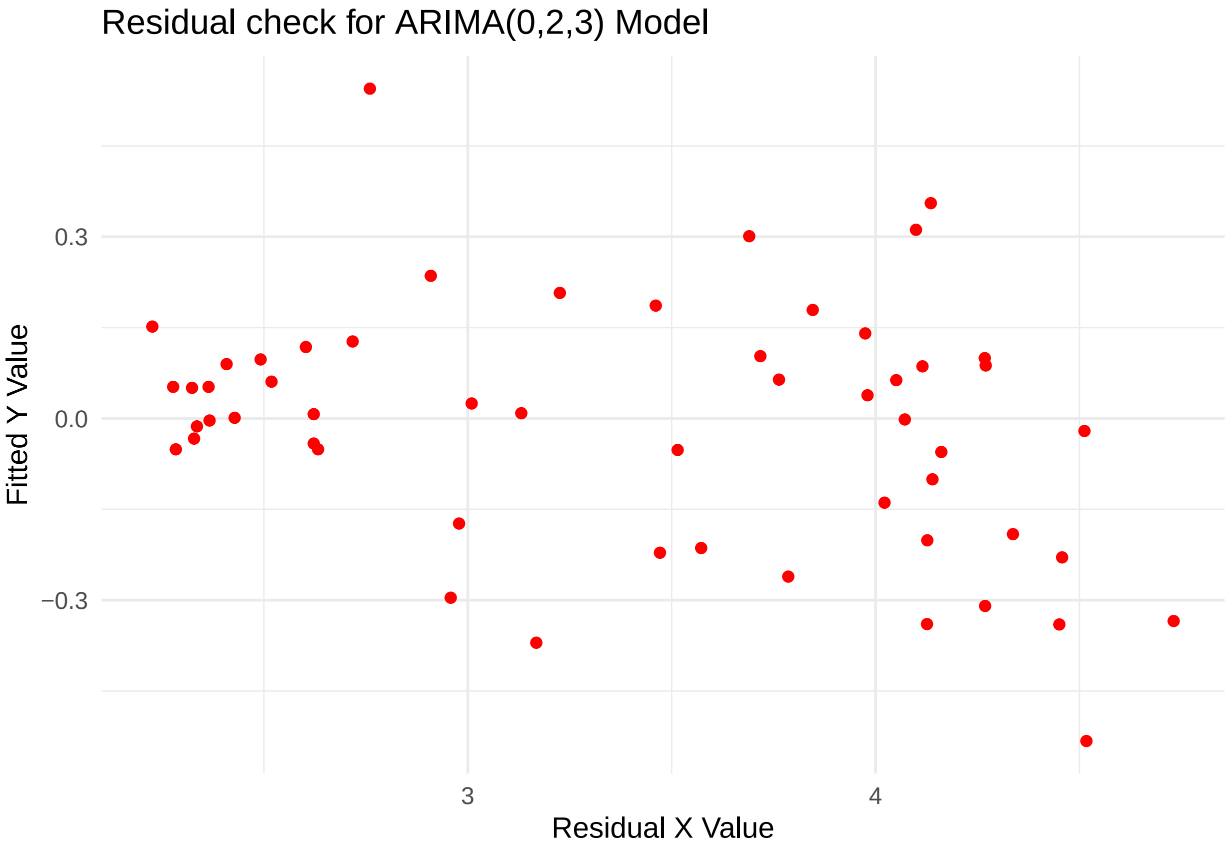


```
  ylab("Womens Murder Rate in the U.S/100,000 std population")
```

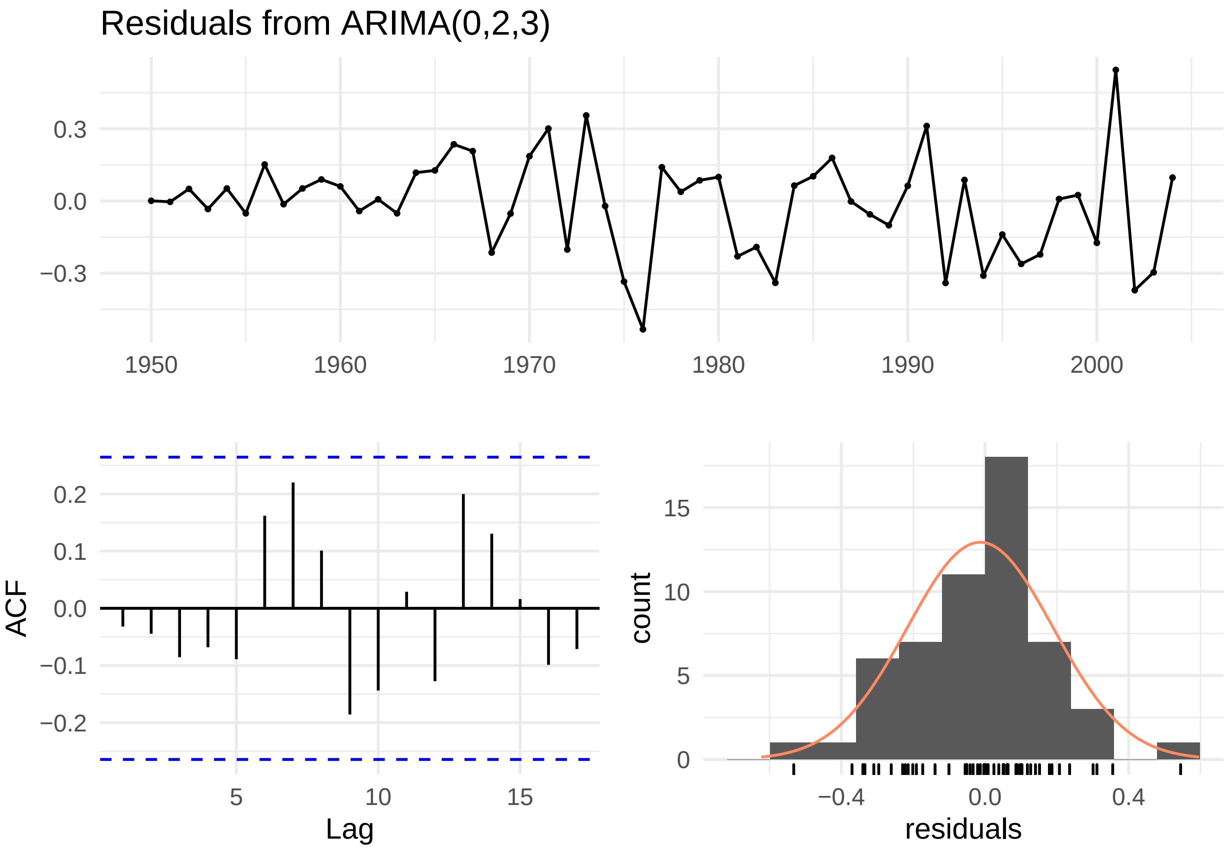
```
## $y
## [1] "Womens Murder Rate in the U.S/100,000 std population"
##
## attr(,"class")
## [1] "labels"
```

```
ggplot(data = no_women) +
  geom_point(mapping = aes(x=fitted(bestfitmodel_2), y=resid(bestfitmodel_2)),
             col="red") + xlab("Residual X Value") + ylab("Fitted Y Value") +
  ggtitle("Residual check for ARIMA(0,2,3) Model")
```

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



```
checkresiduals(bestfitmodel_2)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,2,3)
## Q* = 10.706, df = 7, p-value = 0.152
##
## Model df: 3.   Total lags used: 10
```

For model without approximation and stepwise -> ARIMA(0,2,3): Checking the result generated using Ljung-Box with null hypothesis as: H0: The model is fine H1: The model is not fine With this test we get a result where generated p-value = 0.152 which related to not rejecting the null hypothesis. (i.e. The model is fine). Hence, model is satisfactory and consists no auto correlation in the residuals and is observed to follow a normal distribution.

Question 5: Answer -> The results provided by auto ARIMA [ARIMA(1,2,1) and ARIMA(0,2,3)] are different than the ones predicted above.

For ARIMA(1,2,1) -> one with approximation and stepwise generates an AICc value of -6.39

For ARIMA(0,2,3) -> one without approximation and stepwise generates an AICc value of -6.7

Our ARIMA(0,1,2) -> one generated by us generated an AICc value of -12.95

Comparing all these models we can clearly see that: ARIMA(0,1,2) > ARIMA(0,2,3) > ARIMA(1,2,1) = -12.95 > -6.7 > -6.39

Based on AICc value we conclude that ARIMA(1,2,1) generated by Auto Arima is the best model of the three with a better fit and small criterion