Homework 7

Emma Spors and Cole Brown

11/17/2020

```r
library(readr)
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.0.3

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
#crime_data <- read_csv("C:/Users/espor/Dropbox/Time Series/crime_data.csv")
#viscocity <- read_csv("C:/Users/espor/Dropbox/Time Series/Table_B3.csv")
crime_data <- read_csv("C:/Users/cbbro/OneDrive/Documents/Fall 2020/STAT 560/crime_data.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Crime_Rate = col_double()
## )
```

```r
viscocity <- read_csv("C:/Users/cbbro/OneDrive/Documents/Fall 2020/STAT 560/Table_B3.csv")
```

```
## Parsed with column specification:
## cols(
##   `Time Period` = col_double(),
##   Reading = col_double()
## )
```

Question 5.12

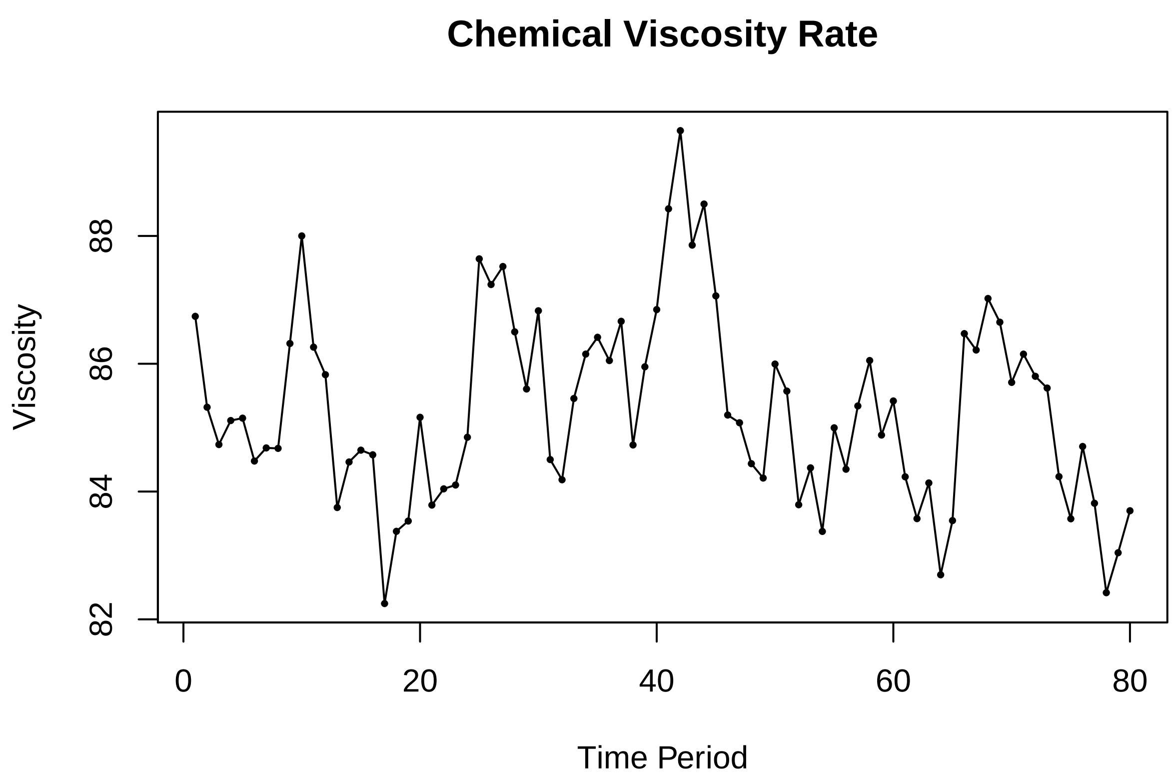
Table B.3 contains data on chemical viscosity.

a) Fit an ARIMA model to this time series, excluding the last 20 observations. Investigate the model adequacy. Explain how this model would be used for forecasting.

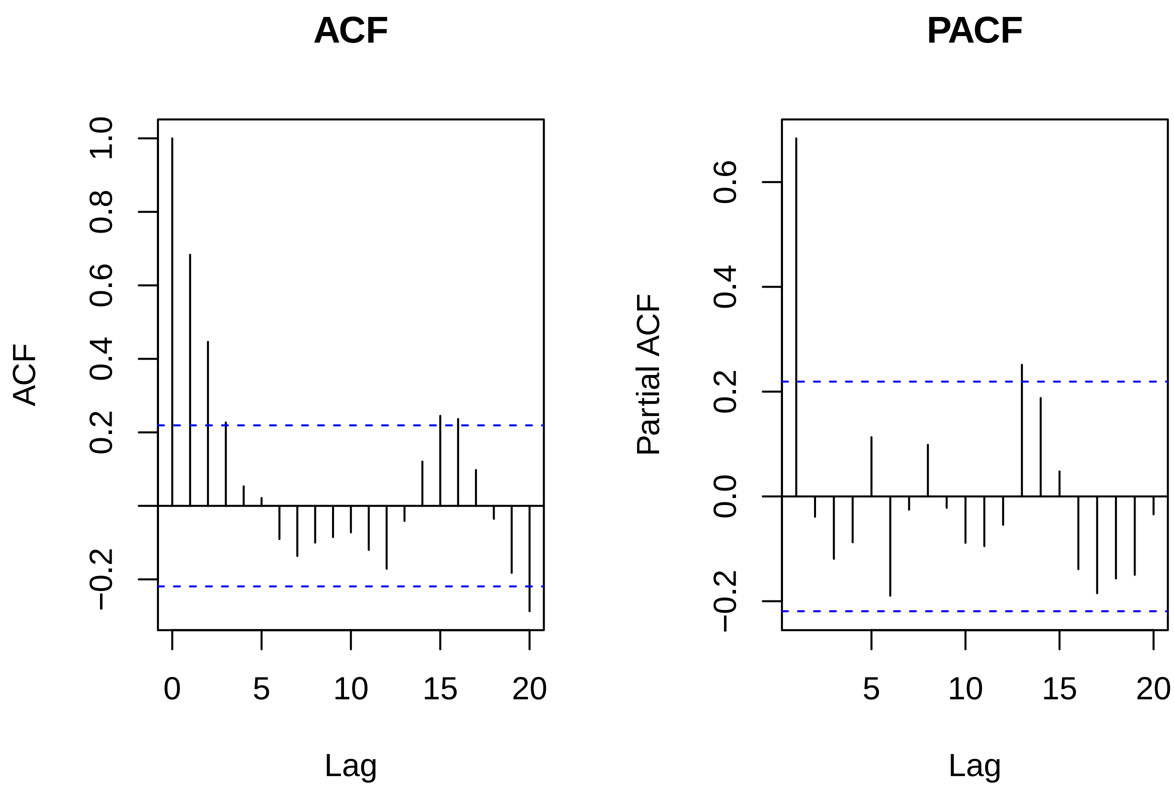
You must specify the model and give the estimates of the coefficients. Your answer should include the ACF and PACF of the data, then determine the "best" model. You can use the auto.arima() function to double check.

To investigate the model adequacy, show the 4-in-1 residual plots and ACF and PACF plots for the residuals. Then interpret your outputs/results. Show the forecast model. To specify the forecast model, you need to show how to get estimators of the Ψ_i's. Show at least the first 4 Ψ_i's. Similar to what we did for Example 5.3, 5.4, and 5.5 on page 381-382 in the class. The general forecast model is shown as the equation of (5.88) on page 379.

```
  #create the dataset without the last 20 observations
viscosity_short <- viscocity[1:80,]

  #plot the data for visual analysis
plot(viscosity_short, type = "o", pch = 16, cex = 0.5, xlab = "Time Period",
     ylab = "Viscosity", main = "Chemical Viscosity Rate")
```




```
  #plot the acf and pacf to confirm data stationarity and determine ARIMA model
par(mfrow=c(1,2), oma = c(0,0,0,0))
acf(viscosity_short[,2], lag.max = 20, type = "correlation", main = "ACF")
acf(viscosity_short[,2], lag.max = 20, type = "partial", main = "PACF")
```




```
  #confirm analysis of model
auto.arima(viscosity_short[,2])
```

```
## Series: viscosity_short[, 2]
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1     mean
##       0.6934  85.2721
## s.e.  0.0802   0.3756
##
## sigma^2 estimated as 1.15:  log likelihood=-118.42
## AIC=242.84   AICc=243.16   BIC=249.99
```

```
#create the arima model
viscosity.ar1 <- arima(viscosity_short[,2], order = c(1,0,0))
viscosity.ar1
```

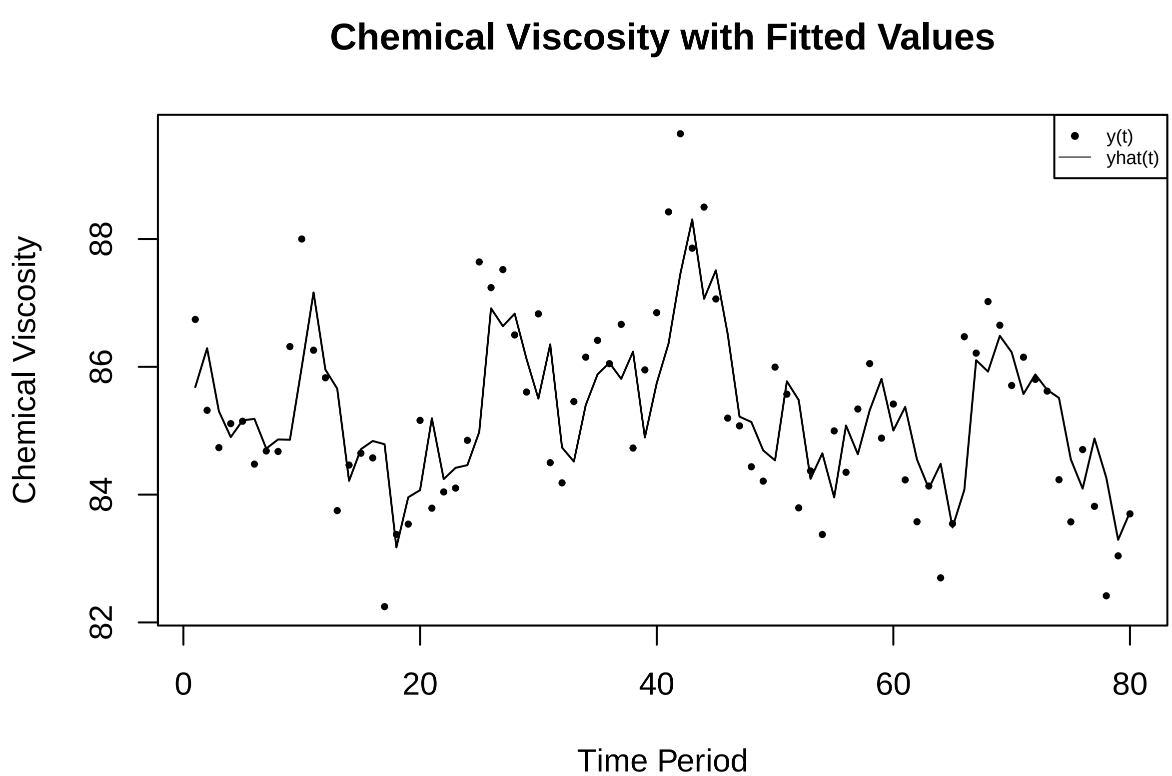
```
##
## Call:
## arima(x = viscosity_short[, 2], order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.6934    85.2721
```

```
## s.e.  0.0802    0.3756
##
## sigma^2 estimated as 1.121:  log likelihood = -118.42,  aic = 242.84
```

ANSWER: Looking at the time series plot, the data does not show significant non-stationary characteristics. The ACF has an initial decrease in values which is followed by a sinusoidal pattern about 0. This is consistent with stationary data, thus confirming visual analysis of the data (which is consistent with the results from the book). In addition to the dampened sinusoid pattern observed in the ACF, the PACF cuts off after lag 1. This is indicative of an AR(1) model. By running the auto.arima() function, we see that this is confirmed. In this case, we have that $\phi = 0.6834$, $\mu = 85.2721$, and $\delta = 26.1444$ Thus our fitted AR(1) model is

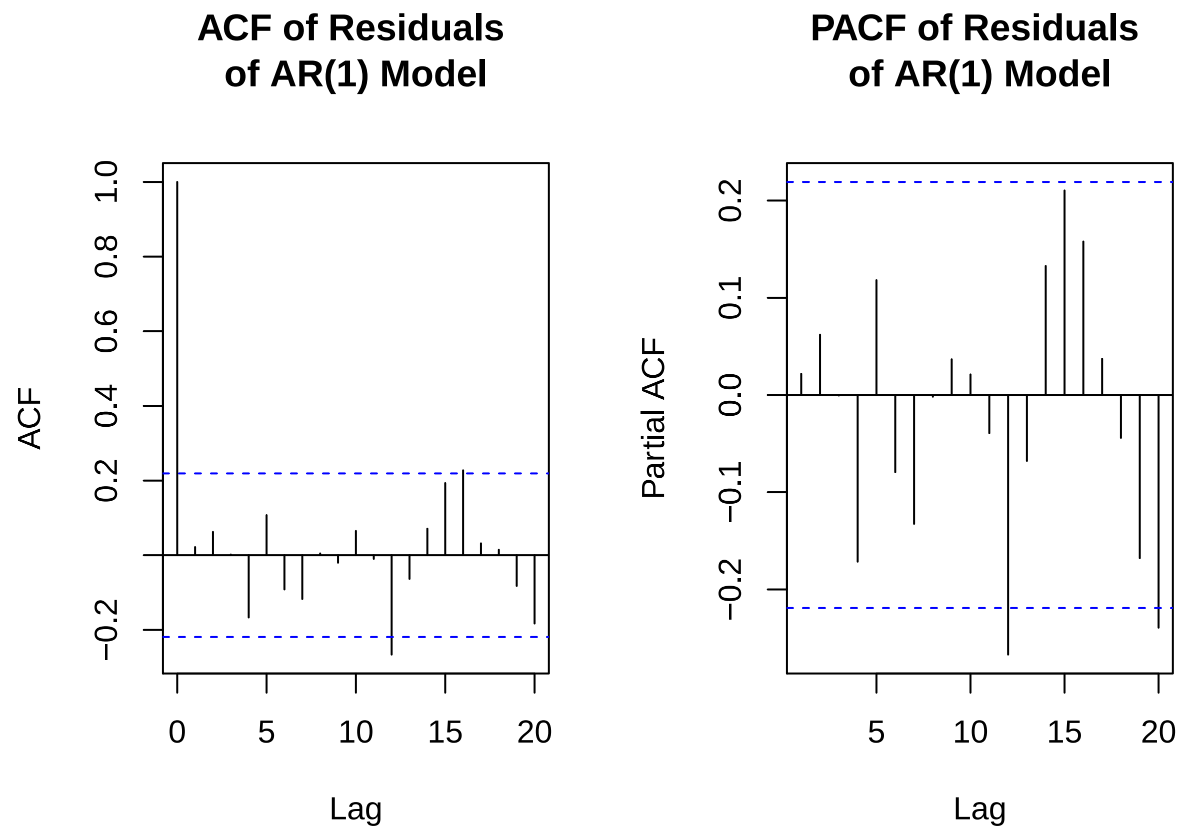
$$y_t = 26.1444 + 0.6934y_{t-1} + \epsilon_t.$$

```r
#fit the values and determine residuals
res.viscosity.ar1 <- as.vector(residuals(viscosity.ar1))
fit.viscosity.ar1 <- as.vector(fitted(viscosity.ar1))

#plot original values vs fitted values
plot(viscosity_short, type = "p", pch = 16, cex = .5, xlab = "Time Period",
     ylab = "Chemical Viscosity", main = "Chemical Viscosity with Fitted Values")
lines(fit.viscosity.ar1)
legend("topright", c("y(t)", "yhat(t)"), pch = c(16, NA), lwd = c(NA, .5), cex = .55)
```

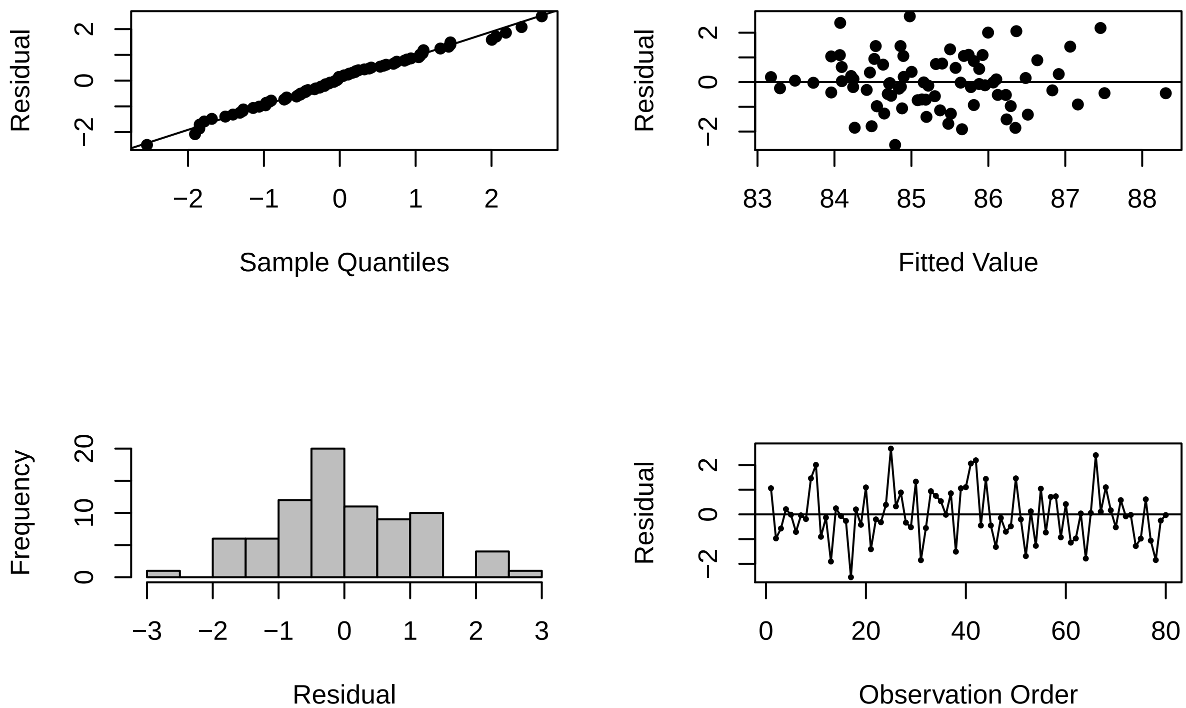



```r
#plot acf and pacf for analysis
par(mfrow=c(1,2), oma = c(0,0,0,0))
```

```r
acf(res.viscosity.ar1, lag.max = 20, type = "correlation",
    main = "ACF of Residuals \nof AR(1) Model")
acf(res.viscosity.ar1, lag.max = 20, type = "partial",
    main = "PACF of Residuals \nof AR(1) Model")
```




```r
  #plot 4 in 1 residual plots for analysis
par(mfrow=c(2,2),oma=c(0,0,0,0))
qqnorm(res.viscosity.ar1,datax=TRUE,pch=16,xlab='Residual',main='')
qqline(res.viscosity.ar1,datax=TRUE)
plot(fit.viscosity.ar1,res.viscosity.ar1,pch=16, xlab='Fitted Value',
ylab='Residual')
abline(h=0)
hist(res.viscosity.ar1,col="gray",xlab='Residual',main='')
plot(res.viscosity.ar1,type="l",xlab='Observation Order',
ylab='Residual')
points(res.viscosity.ar1,pch=16,cex=.5)
abline(h=0)
```



ANSWER: This time series plot shows the the original values with the fitted values from the AR(1) model. It appears to have smoothed out the highs and lows of the data. With the ACF and PACF, there are a few lags that raise a flag. On the ACF, Lag 12 and Lag 16 are on or outside the limits which indicates that their may be some autocorrelation in the residuals, but it is arguable. The 4 in 1 residual plots indicate that the fit is acceptable. The Fitted Value and Histogram of the Residuals appear to generally follow a normal distribution and the variance appears to be constant. We can use the values from the fitted model to get the forecasted values. In general, the forecast model is

$$\hat{y}_{T+\tau} = \mu + \sum_{i=\tau}^{\infty} \psi_i \epsilon_{T+\tau-i}$$

Subsequently, forecast error is defined as

$$e_T(\tau) = \sum_{i=0}^{\tau-1} \psi_i \epsilon_{T+\tau-i}$$

Then the variance of a model is defined as

$$Var[e_T(\tau)] = \sigma^2 \sum_{i=0}^{\tau-1} \psi_i^2$$

To estimate the ψ's, we used this equation for the AR(1) model:

$$(\psi_0 + \psi_1 B + \psi_2 B^2 + ...)(1 - \phi_1 B) = 1$$

By equation like powers of B, we find that

$$B^0 : \psi_0 = 1$$

$$B^1 : \psi_1 - \phi_1 = 0, \psi_1 = \phi_1$$
$$B^2 : \psi_2 - \phi_1\psi_1 = 0, \psi_2 = \phi_1\psi_1$$
$$B^3 : \psi_3 - \phi_1\psi_2 = 0, \psi_3 = \phi_1\psi_2$$

Thus for the AR(1) model,

$$\psi_j = \phi_1\psi_{j-1} = \phi^j$$

In our case, $\phi_1 = 0.69$, so $\psi_j = 0.69\psi_{j-1} = .69^j$ and $\mu = 85.2721$ We would then replace for these values in

$$\hat{y}_{T+\tau} = \mu + \sum_{i=\tau}^{\infty} \psi_i \epsilon_{T+\tau-i}$$

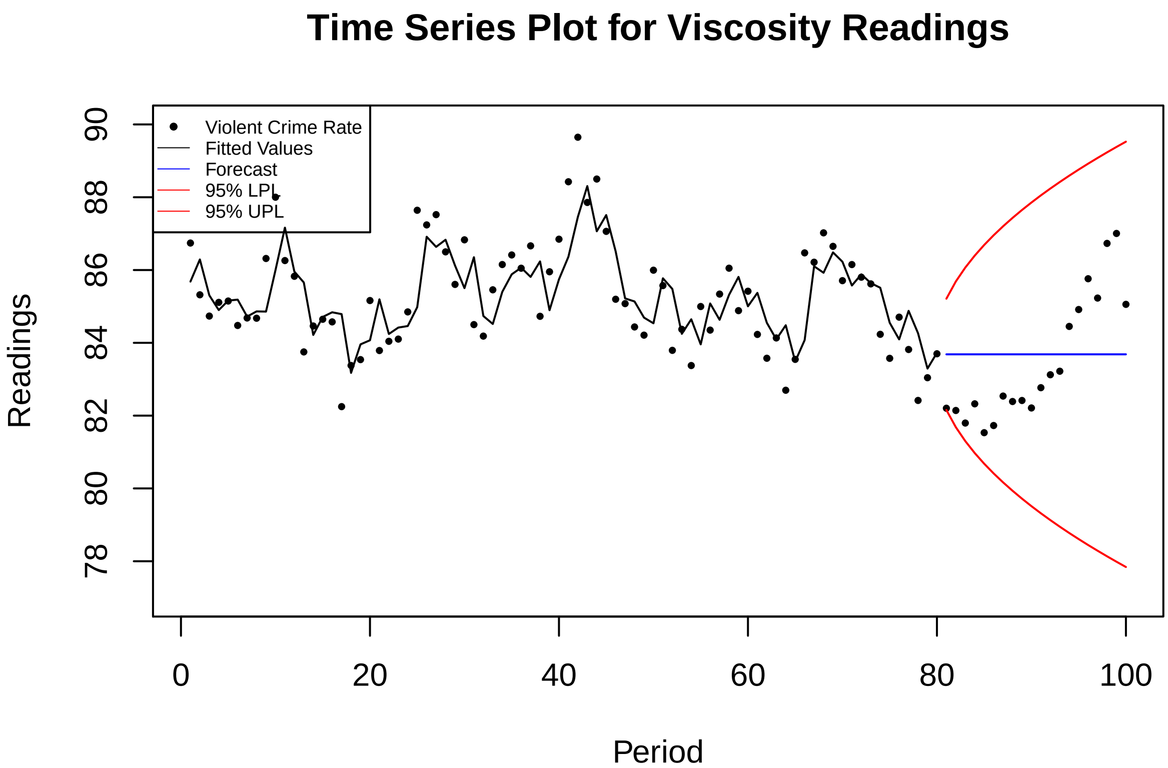
to get the forcasted values. The forcasted values can be seen in the dataset below.

b) Forecast the last 20 observations. No matter what model that you got in part a, to answer question b, please use auto.arima() to get the best model. (This is just for the grading purpose.) Then use the forecast() to get the forecast values. See the R code on page 408. This will give you a 1- to 20- step ahead forecasts.

```
forecast.viscosity <-as.array(forecast(fit.viscosity.ar1, h = 20))
forecast.viscosity
```

```
##     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
##  81       83.68393 82.68427 84.68359 82.15508 85.21278
##  82       83.68393 82.37455 84.99331 81.68140 85.68646
##  83       83.68393 82.12518 85.24268 81.30003 86.06783
##  84       83.68393 81.91053 85.45733 80.97175 86.39611
##  85       83.68393 81.71918 85.64868 80.67910 86.68876
##  86       83.68393 81.54487 85.82299 80.41251 86.95535
##  87       83.68393 81.38372 85.98414 80.16606 87.20180
##  88       83.68393 81.23313 86.13473 79.93575 87.43211
##  89       83.68393 81.09127 86.27659 79.71879 87.64907
##  90       83.68393 80.95677 86.41109 79.51309 87.85477
##  91       83.68393 80.82859 86.53927 79.31706 88.05080
##  92       83.68393 80.70591 86.66195 79.12944 88.23842
##  93       83.68393 80.58808 86.77978 78.94924 88.41862
##  94       83.68393 80.47457 86.89328 78.77564 88.59221
##  95       83.68393 80.36494 87.00292 78.60797 88.75989
##  96       83.68393 80.25880 87.10906 78.44565 88.92221
##  97       83.68393 80.15585 87.21201 78.28820 89.07966
##  98       83.68393 80.05582 87.31204 78.13521 89.23265
##  99       83.68393 79.95846 87.40940 77.98631 89.38155
## 100       83.68393 79.86358 87.50428 77.84120 89.52666
```

```
plot(viscocity, pch = 16, cex = .5, xlab = "Period", ylab = "Readings", ylim = c(77,90),
     main = "Time Series Plot for Viscosity Readings")
lines(1:80, fit.viscosity.ar1)
lines(81:100, forecast.viscosity$mean, col = "blue")
lines(81:100, forecast.viscosity$lower[,2], col = "red")
lines(81:100, forecast.viscosity$upper[,2], col = "red")
legend("topleft", c("Violent Crime Rate", "Fitted Values", "Forecast", "95% LPL", "95% UPL"),
       pch=c(16, NA, NA, NA, NA), lwd=c(NA, .5, .5, .5, .5), cex =.55,
       col = c("black", "black", "blue", "red", "red"))
```



ANSWER: The printed dataframe shows the forecasted values in the point in the "Point Forecast". The time series plot shows the forcasted values in blue with their 95% confidence interval in red. Because they are 1-20 step ahead forecast values, they are not as accurate if they would all be 1-step ahead values because the future can be hard to determine.

c) Show how to obtain prediction intervals for the forecasts in part b above.The 80% and 95% Prediction Intervals should have been obtained from part b if you use forecast() function. You don't have to calculate the Prediction intervals again. For this question, just show the prediction interval formula.

ANSWER: The prediction intervals can be seen in the dataframe from Part (b). To get the prediction interval, we note that the variance (from above) for an AR(1) model is:

$$Var[e_T(\tau)] = \sigma^2 \sum_{i=0}^{\tau-1} \psi_i^2 = \sigma^2 \frac{1-\phi^{2\tau}}{1-\phi^2}$$

Thus the $100(1-\alpha)$ prediction interval for $y_{T+\tau}$ is

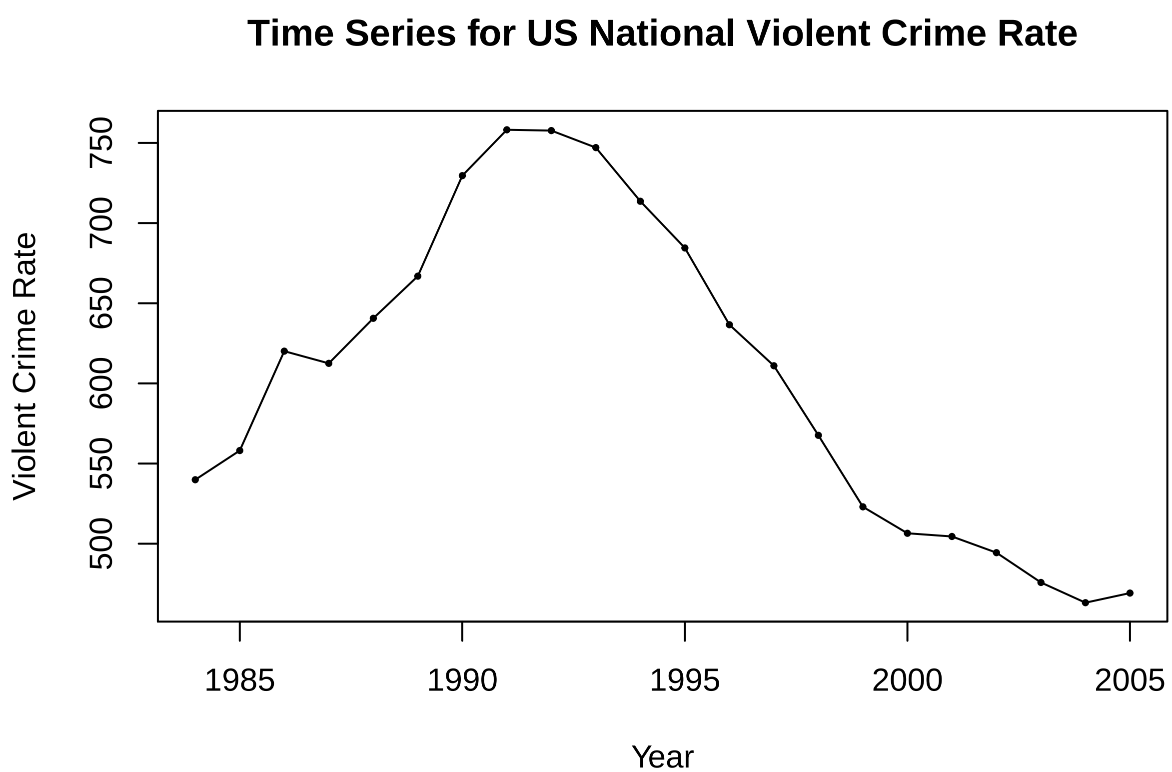
$$\hat{y}_{T+\tau}(T) \pm Z_{\frac{\alpha}{2}} * \sigma \sqrt{\frac{1-\phi^{2\tau}}{1-\phi^2}}$$

As is seen in the time series plot, the formula shows that the forecast error, and hence prediction interval, gets bigger with increasing forecast lead times. This is intuitive because there is more uncertainty in the future.

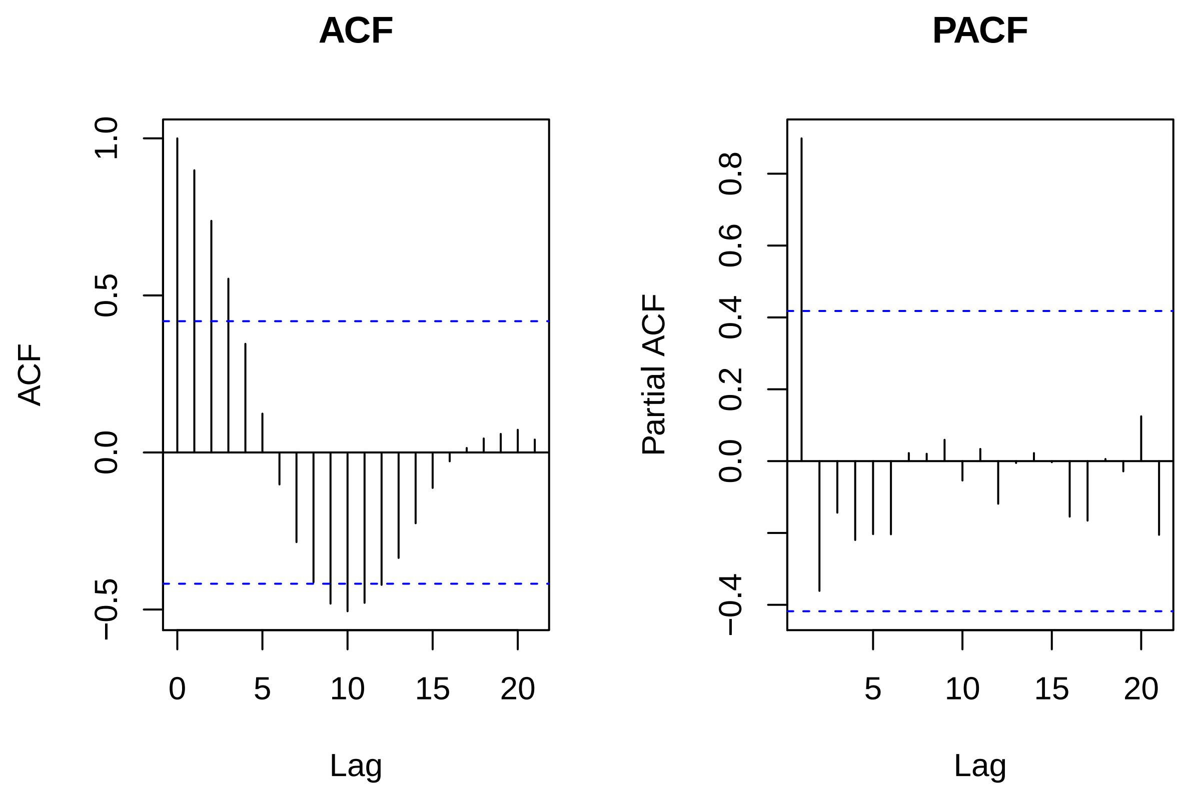
Question 5.33

Table B.15 presents data on the occurrence of violent crimes. Develop an appropriate ARIMA model and procedure for forecasting these data. Explain how prediction intervals would be computed. Show the model and compute the 1- to 10-step ahead forecasts. Similar to 5.12, but you can skip the calculation of Ψ_i's. show the prediction intervals formula. Compute the 95% prediction intervals for the 1- to 10-step ahead forecasts.

```
  #plot the time series for visual analysis
plot(crime_data, type = "o", pch = 16, cex = 0.5, xlab = "Year", ylab = "Violent Crime Rate",
     main = "Time Series for US National Violent Crime Rate")
```




```
  #plot the acf and pacf for analysis
par(mfrow=c(1,2), oma = c(0,0,0,0))
acf(crime_data[,2], lag.max = 25, type = "correlation", main = "ACF")
acf(crime_data[,2], lag.max = 25, type = "partial", main = "PACF")
```



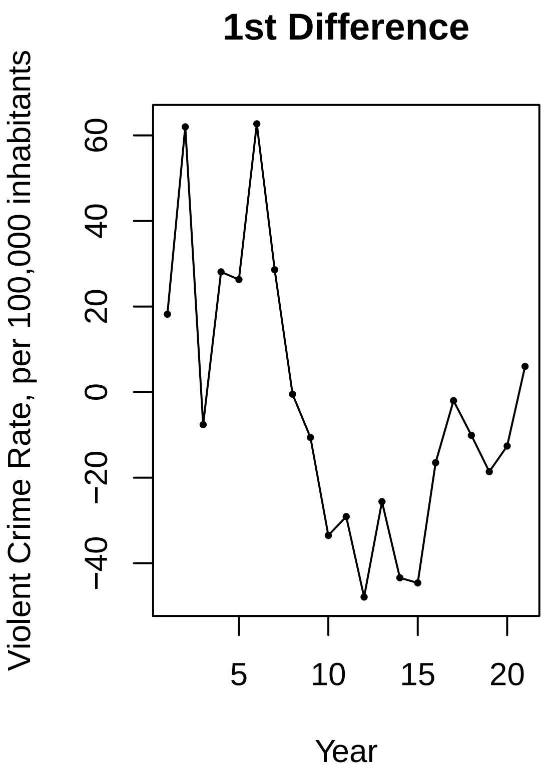

```
  #see suggested arima model
auto.arima(crime_data[,2], stepwise = FALSE, approximation = FALSE)
```

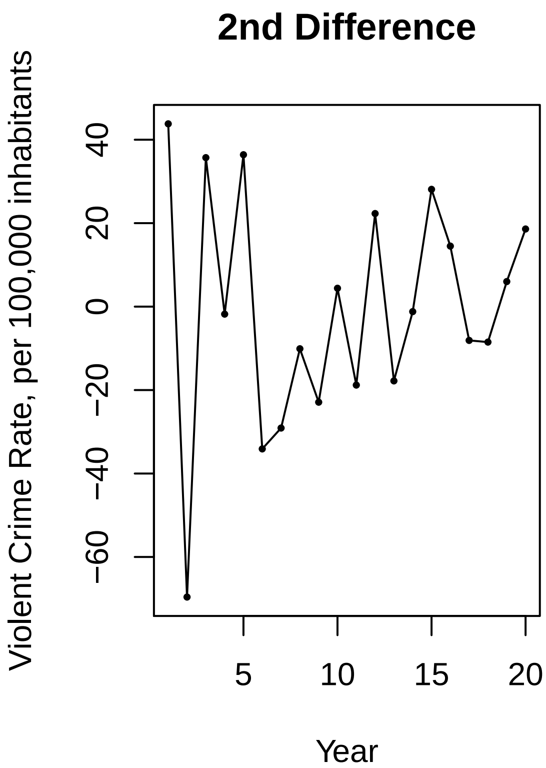
```
## Series: crime_data[, 2]
## ARIMA(1,2,0)
##
## Coefficients:
##           ar1
##       -0.4533
## s.e.   0.2091
##
## sigma^2 estimated as 623.5:  log likelihood=-92.33
## AIC=188.67   AICc=189.37   BIC=190.66
```

ANSWER: From the ACF and visual analysis of the data, the data does not appear to stationary. Additional work will be needed to understand this. The auto.arima() function suggests that differences are needed.

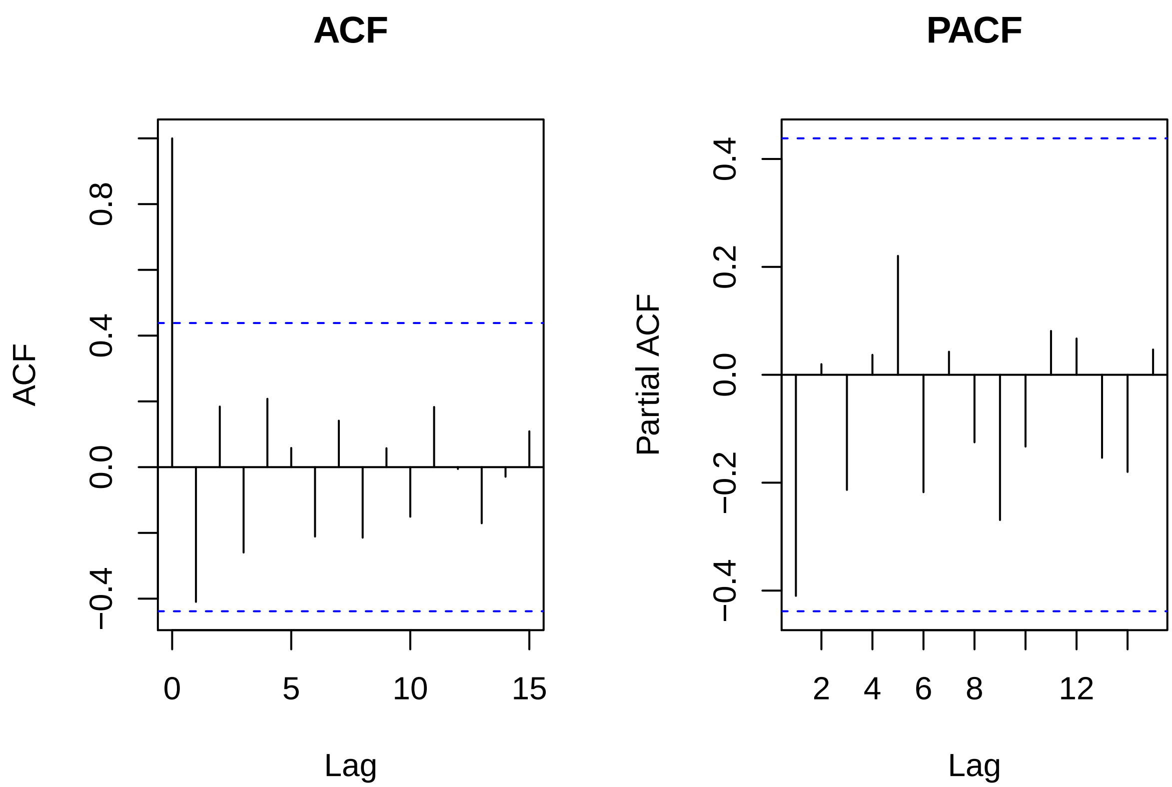
```
par(mfrow=c(1,2), oma = c(0,0,0,0))

  #Take the first and second differece the data
crime <- as.matrix(crime_data)
crime1 <- diff(crime[,2], differences = 1)
crime2 <- diff(crime[,2], differences = 2)

  #Plot the new data
```

```r
plot(crime1, type = "o", pch = 16, cex = 0.5, xlab = "Year",
     ylab = "Violent Crime Rate, per 100,000 inhabitants", main = "1st Difference")
plot(crime2, type = "o", pch = 16, cex = 0.5, xlab = "Year",
     ylab = "Violent Crime Rate, per 100,000 inhabitants", main = "2nd Difference")
```






```r
  #ACF and PACF for the second difference data
acf(crime2, lag.max = 15, type = "correlation", main = "ACF")
acf(crime2, lag.max = 15, type = "partial", main = "PACF")
```



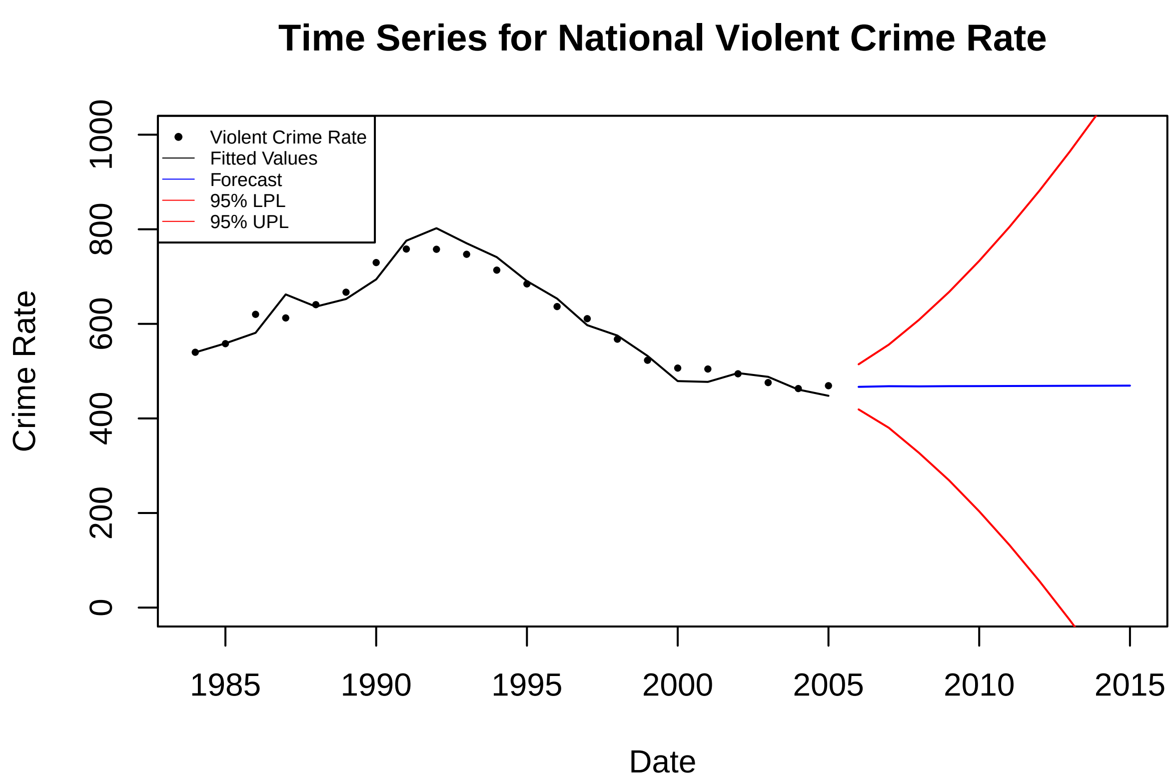

```r
#fit the data to the arima model
crime_fit_mod <- arima(crime_data[,2], order = c(1,2,0))
crime_fit_mod
```

```
##
## Call:
## arima(x = crime_data[, 2], order = c(1, 2, 0))
##
## Coefficients:
##           ar1
##       -0.4533
## s.e.   0.2091
##
## sigma^2 estimated as 592.3:  log likelihood = -92.33,  aic = 188.67
```

ANSWER: The time series plots showed the new data after the first and second difference respectively. The ACF from the second difference demonstrates stationarity. The fact that the 2nd difference was needed is consistent with the fact that the data that has a quadratic trend. While not perfect,the ACF and PACF do show characteristics consistent with AR(1). Thus, we will go with an ARIMA(1,2,0) model, as suggested by the auto.arima() function with $\phi_1 = -0.4533$.

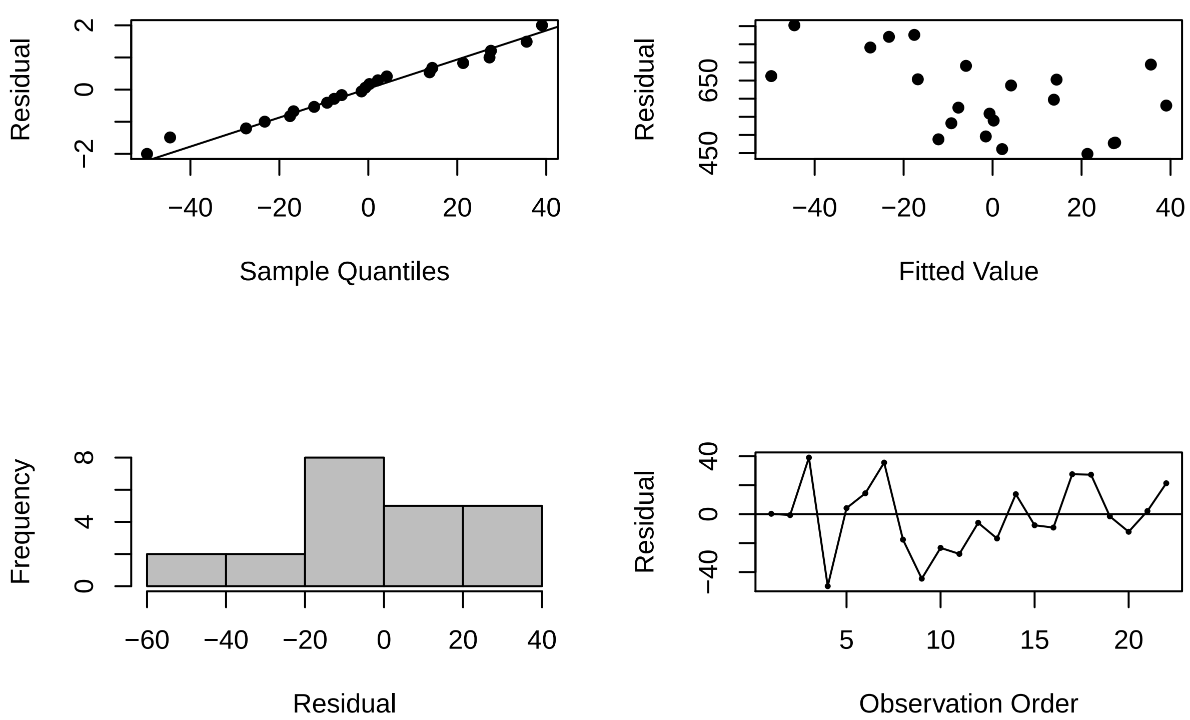
```r
  #fit the current values
fitted_crime <- as.array(fitted(crime_fit_mod))

  #forecast the next ten valeues
```

```
crime_fit <- as.array(forecast(crime_fit_mod, h = 10))

  #plot the fitted data with the forecasted data
plot(crime_data, pch = 16, cex = .5, xlab = "Date", ylab = "Crime Rate", ylim = c(0, 1000),
     xlim = c(1984, 2015), main = "Time Series for National Violent Crime Rate")
lines(1984:2005, fitted_crime)
lines(2006:2015, crime_fit$mean, col = "blue")
lines(2006:2015, crime_fit$lower[,2], col = "red")
lines(2006:2015, crime_fit$upper[,2], col = "red")
legend("topleft", c("Violent Crime Rate", "Fitted Values", "Forecast", "95% LPL", "95% UPL"),
       pch=c(16, NA, NA, NA, NA), lwd=c(NA, .5, .5, .5, .5), cex =.55,
       col = c("black", "black", "blue", "red", "red"))
```

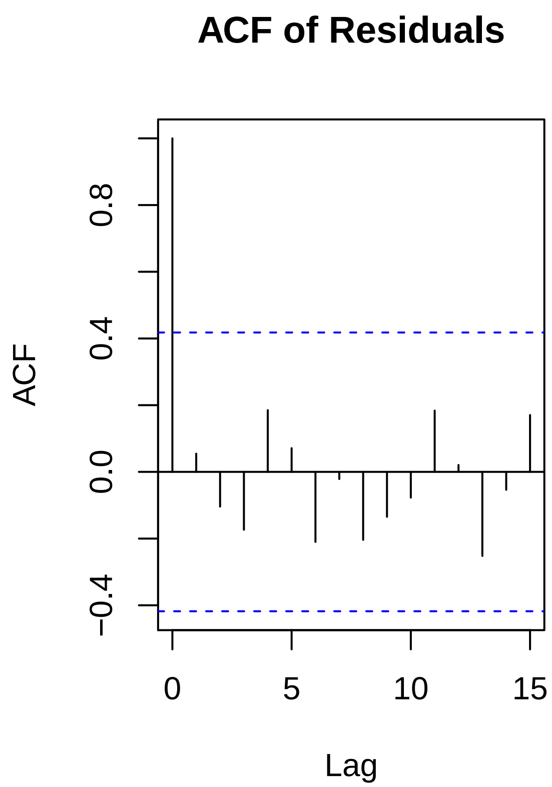


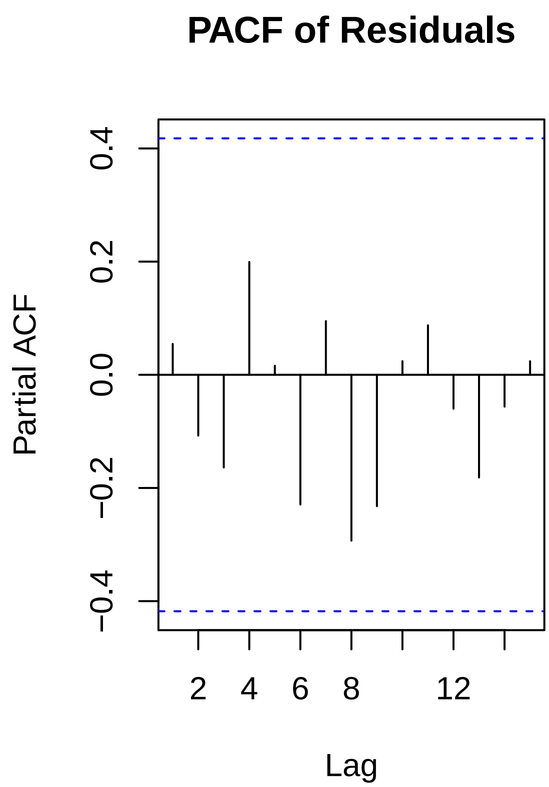

```
res.crime <- as.vector(residuals(crime_fit))

#plot 4 in 1 residual plots for analysis
par(mfrow=c(2,2),oma=c(0,0,0,0))
qqnorm(res.crime ,datax=TRUE,pch=16,xlab='Residual',main='')
qqline(res.crime,datax=TRUE)
plot(res.crime,fitted_crime,pch=16, xlab='Fitted Value',
ylab='Residual')
abline(h=0)
hist(res.crime,col="gray",xlab='Residual',main='')
plot(res.crime,type="l",xlab='Observation Order',
ylab='Residual')
```

```r
points(res.crime,pch=16,cex=.5)
abline(h=0)
```




```r
par(mfrow=c(1,2), oma = c(0,0,0,0))
acf(res.crime, lag.max = 15, type = "correlation", main = "ACF of Residuals")
acf(res.crime, lag.max = 15, type = "partial", main = "PACF of Residuals")
```





ANSWER: The time series plot shows the fitted values (black line) and the forcasted values (blue line). The residual plots for the fitted model show that the residuals generally follow a normal distribution although there may be a violation in the constant variance assumption. This could be investigated further. The ACF and PACF shows that is not correlation left in the residuals, which is good.

```
crime_fit
```

```
##    Point Forecast     Lo 80    Hi 80      Lo 95     Hi 95
## 23       466.7685 435.57949 497.9575  419.06902  514.4680
## 24       468.1591 410.71483 525.6033  380.30568  556.0124
## 25       467.8171 375.94389 559.6902  327.30918  608.3249
## 26       468.2604 338.06244 598.4584  269.13979  667.3811
## 27       468.3478 295.21144 641.4842  203.55859  733.1370
## 28       468.5965 248.80248 688.3906  132.45054  804.7425
## 29       468.7721 198.69568 738.8486   55.72585  881.8184
## 30       468.9809 145.31996 792.6418  -26.01578  963.9775
## 31       469.1746  88.78690 849.5623 -112.57818 1050.9274
## 32       469.3751  29.29789 909.4524 -203.66493 1142.4152
```

ANSWER: The general formula for forecasting values is

$$\hat{y}_{T+\tau} = \mu + \sum_{i=\tau}^{\infty} \psi_i \epsilon_{T+\tau-i}$$

The ψ can be calculated by equation powers of B in

$$(\psi_0 + \psi_1 B + \psi_2 B^2 + ...)(1 - 2B + B^2)(1 - \phi_1 B) = 1$$

The forcasted values can be found in the "Point Forecast" column with the 80% and 95% prediction intervals in subsequent columns. Again, the variance is defined as

$$Var[e_T(\tau)] = \sigma^2 \sum_{i=0}^{\tau-1} \psi_i^2$$

So the prediction interval is

$$\hat{y}_{T+\tau}(T) \pm Z_{\frac{\alpha}{2}} * \sigma \sum_{i=0}^{\tau-1} \psi_i^2$$