STAT 560: Homework Assignment 7

Sakib Kabir

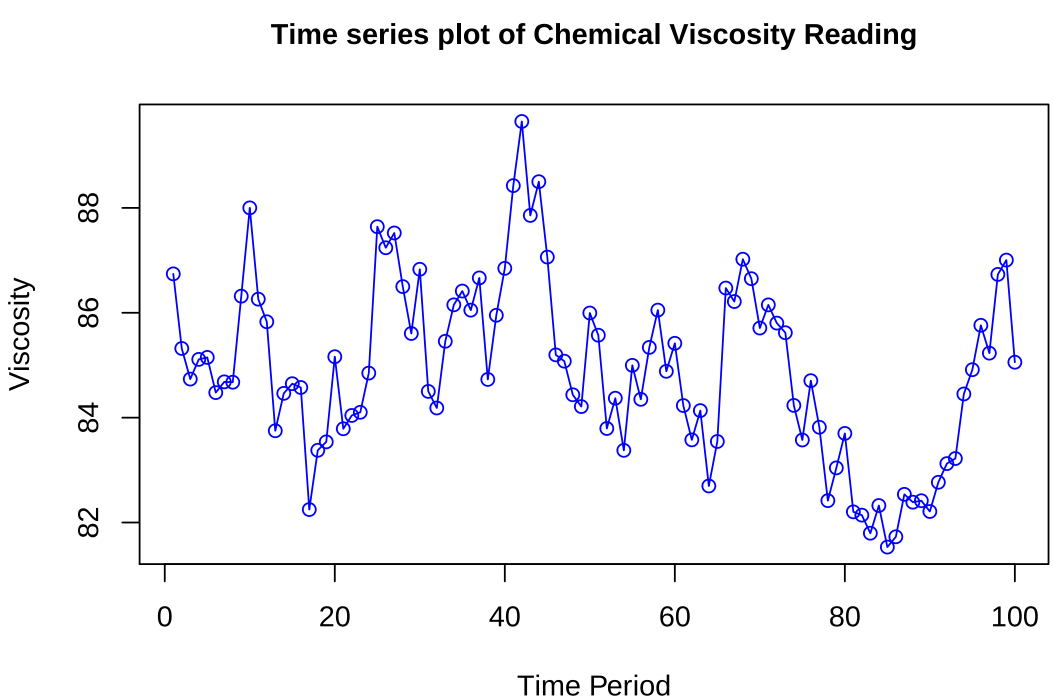
11/23/2020

Question 5.12

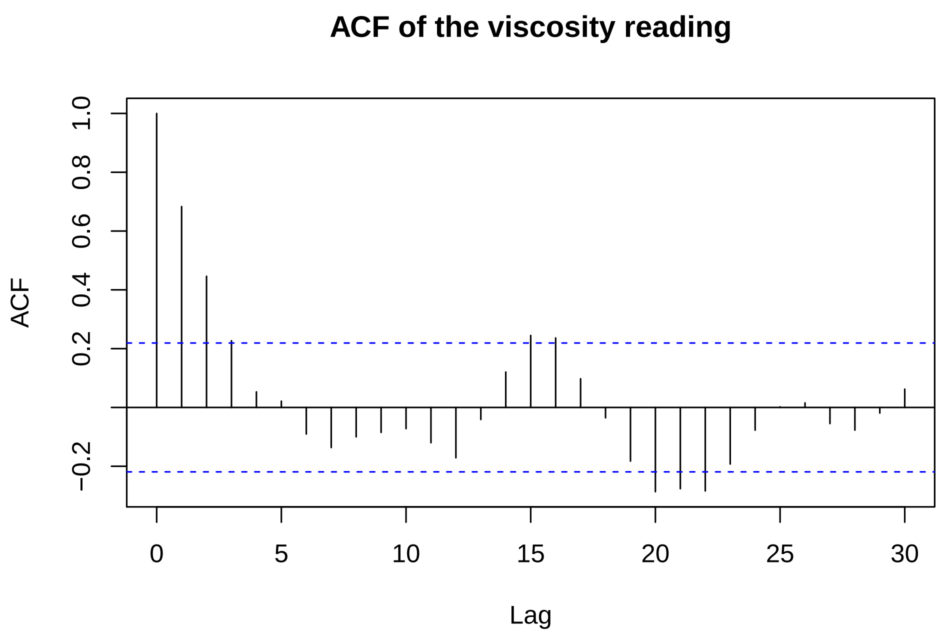
Table B.3 contains data on chemical process viscosity.
a. Fit an ARIMA model to this time series, excluding the last 20 observations. Investigate model adequacy. Explain how this model would be used for forecasting.
b. Forecast the last 20 observations.
c. Show how to obtain prediction intervals for the forecasts in part b above.

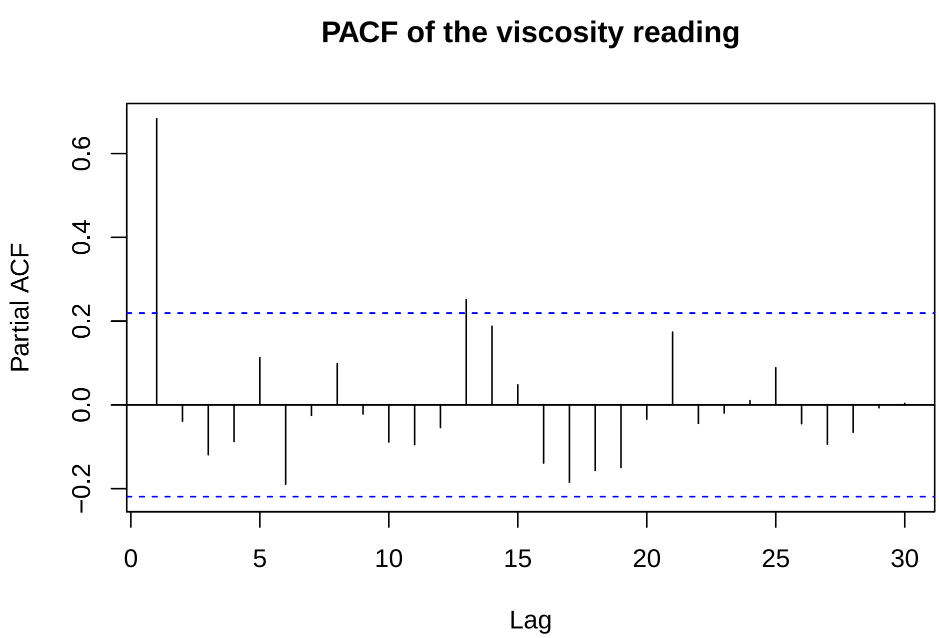
a. Fit an ARIMA model to this time series, excluding the last 20 observations. Investigate model adequacy. Explain how this model would be used for forecasting.

The chemical process viscosity data table contains 100 viscosity readings, which have been plotted in the figure below. Looking at the first 80 observations (as last 20 observations will not be used for model fitting), this time series looks like stationary, though it is hard to be certain about stationarity from this plot only since there seems to be a downward trend from time period 44 to 80. To be certain about the stationarity, sample ACF has been plotted.



The sample ACF plot of viscosity reading data can be seen in the figure displayed below. The ACF shows exponential decay and approximately sinusoidal patter at higher lag. These properties of the time series confirms that the viscosity reading time series is stationary. Additionally, these properties also suggest either AR(p) or ARMA(p,q) process will be good fit for this data. To be certain about the model for this data, partial ACF has been plotted below.





The partial acf appears to cut off after lag 1, which confirms that AR(1) model will be appropriate for this data. Exponential decay and damped sinusoid in the ACF plot and PACF cut off after lag 1 is a strong evidence that AR(1) will be a good model for viscosity reading data.

We know that the first-order atuoregressive or AR(1) model is expressed in this form: $y_t = \delta + \phi y_{t-1} + \epsilon_t$, where $\delta = (1-\phi)\mu$ and ϵ_t is the white noise. Here $\phi = 0.693$, mean $\mu = 85.27$ (can be seen in the r-code output below). So, $\delta = (1-\phi)\mu = (1-0.693) \times 85.27 = 27.18$. Therefore, the AR(1) model can be expressed as:

$$y_t = 27.18 + 0.693 y_{t-1} + \epsilon_t$$

Fitting ARIMA(1,0,0) model

```
library(forecast)
```

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

```
arimaModel_AR1  <- arima(cp.viscosity[1:80,2], c(1,0,0))
arimaModel_AR1
```

```
##
## Call:
## arima(x = cp.viscosity[1:80, 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
```

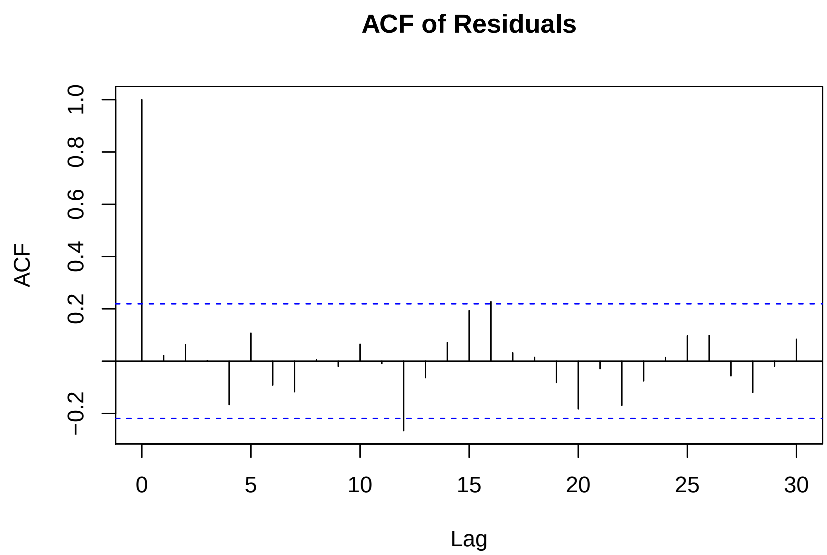
In order to check the appropriateness of the model above, an automatic function ("auto.arima()") has been used. The auto.arima() function output shows that the ARIMA(1,0,0) model is the "best" model for the viscosity data, which is essentially a AR(1) model. Thus, it can be said that the above shown model is appropriate.

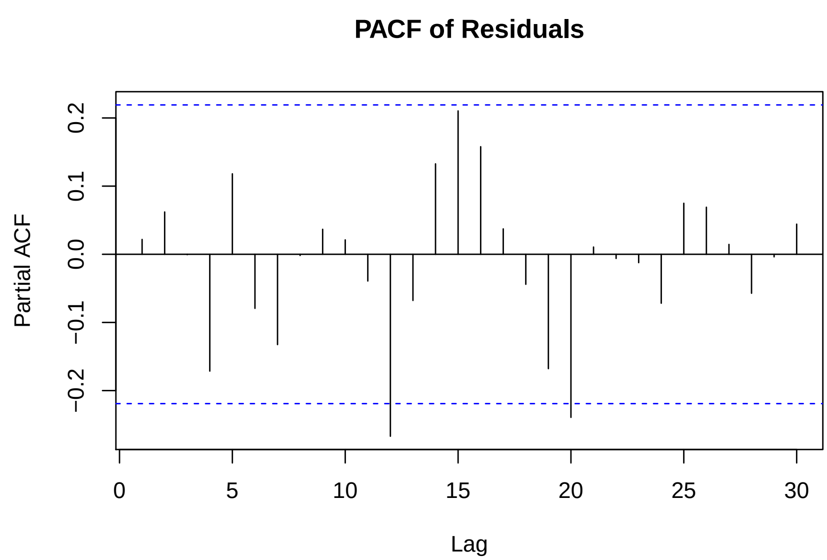
```
library(forecast)
arimaModel_1  <- auto.arima(cp.viscosity[1:80,2])
arimaModel_1
```

```
## Series: cp.viscosity[1:80, 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
```

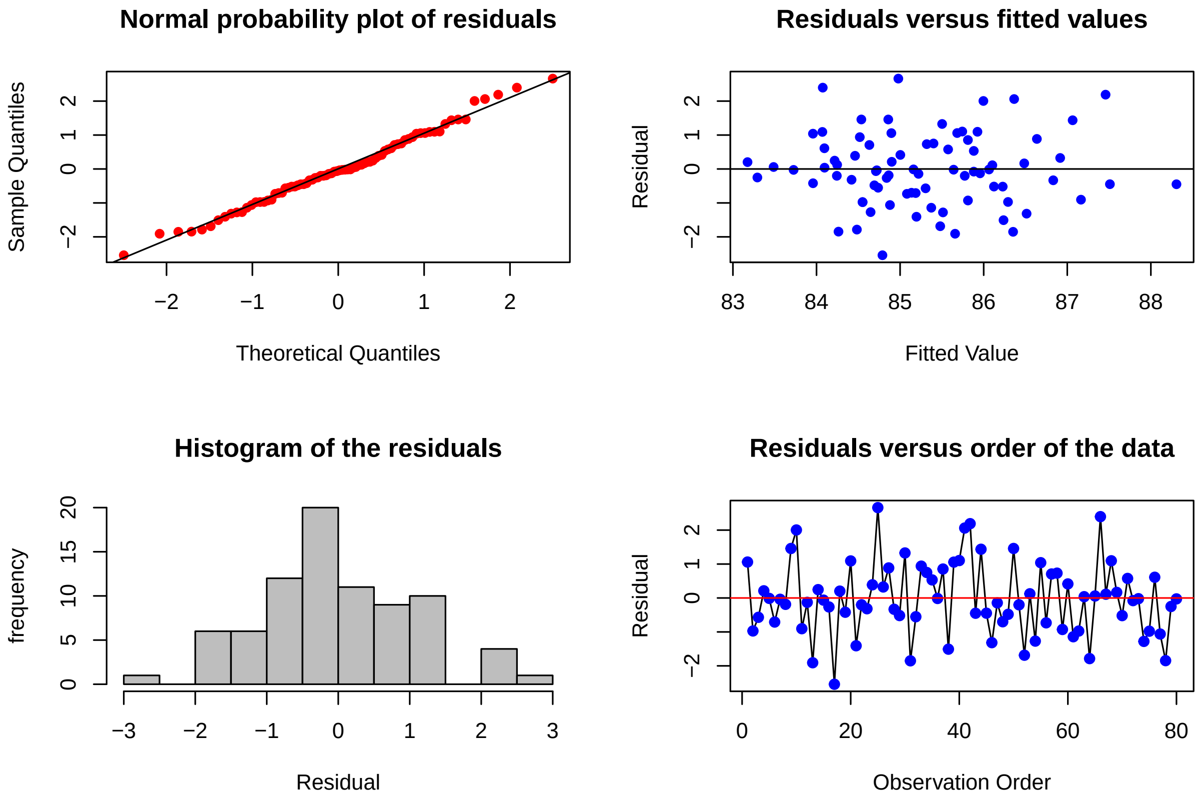
Model Adequacy Investigation:

The model adequacy can be investigated by looking at the acf and pacf plot of residuals and four diagnostic plots. The acf and pacf plot displayed below shows that all the acfs and pacfs are approximately close to zero (with few exceptions), meaning there is no significant autocorrelation left in the data.


ACF of Residuals
ACF
1.0
0.8
0.6
0.4
0.2
−0.2
0
5
10
15
20
25
30
Lag


PACF of Residuals
Partial ACF
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20
25
30
Lag

Residual Plots:



The Q-Q plot and histogram shown above suggest that the residuals are normally distributed. The time series plot of residuals and residual versus fitted value plot do not indicate any significant deviation from common variance assumption. Therefore, it can be said that the AR(1) model provides a decent fit to the viscosity data.

How this model would be used for forecasting?

Best forecast model in mean square sense is:

$$\hat{y}_{T+\tau}(T) = E[y_{T+\tau}|y_T, y_{T-1,\ldots}] = \mu + \sum_{i=\tau}^{\infty} \psi_i \epsilon_{T+\tau-i} \tag{1}$$

To calculate the forecast ψ weights should be obtained. The ψ weights for the general ARIMA(p,d,q) model may be obtained by equation like powers of B in the expansion of

$$(\psi_0 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 \ldots)(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)(1-B)^d = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) \tag{2}$$

For AR(1) model or ARIMA(1,0,0) model, equation 2 becomes:

$$\begin{aligned}
(\psi_0 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 \ldots)(1 - \phi_1 B) &= 1 \\
(\psi_0 - \psi_0\phi_1 B + \psi_1 B - \psi_1\phi_1 B^2 + \psi_2 B^2 - \psi_2\phi_1 B^3 - +\psi_3 B^3 - \psi_3\phi_1 B^4 \ldots) &= 1 \\
\psi_0 + B(\psi_1 - \psi_0\phi_1) + B^2(\psi_2 - \psi_1\phi_1) + B^3(\psi_3 - \psi_2\phi_1) &= 1
\end{aligned}$$

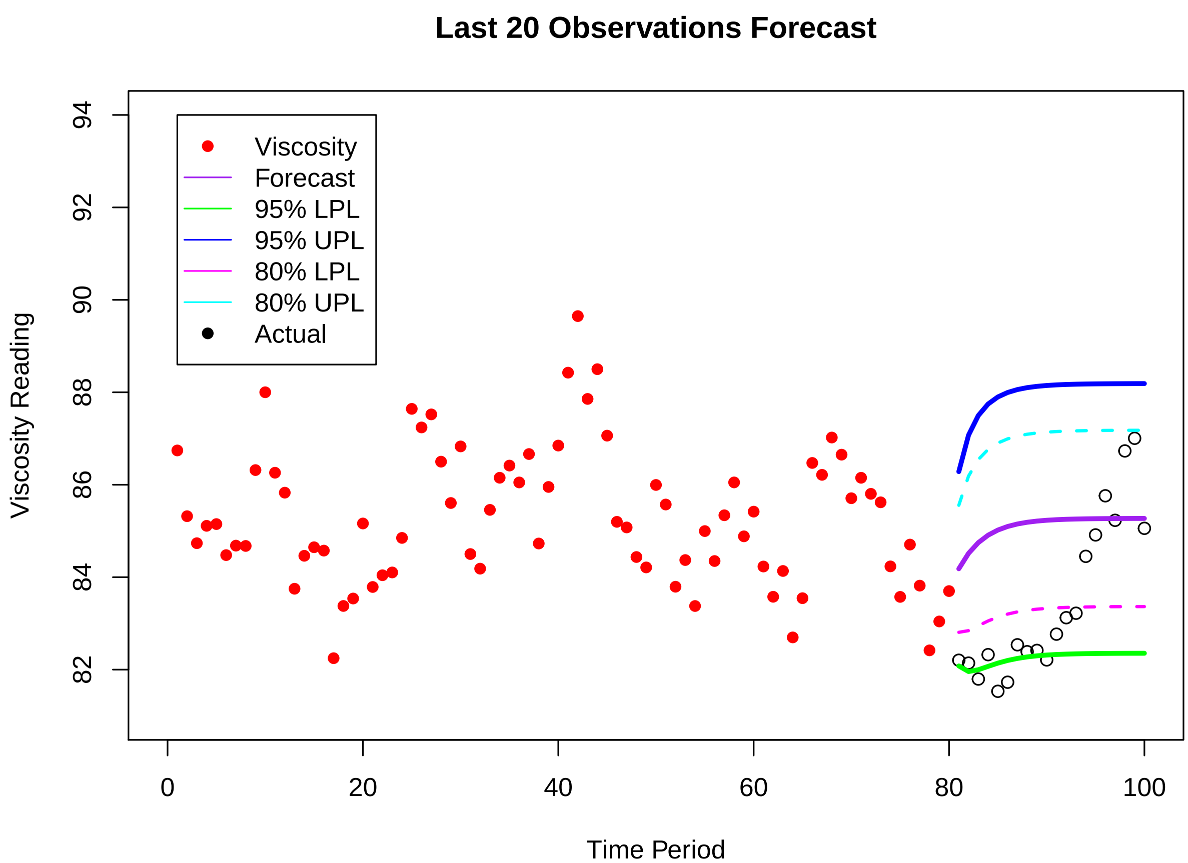
Equating like power of B, we find

$$B^0 : \psi_0 = 1$$
$$B^1 : \psi_1 - \psi_0\phi_1 = 0;\ or,\ \psi_1 = \phi_1\psi_0 = \phi_1 = 0.693$$
$$B^2 : \psi_2 - \psi_1\phi_1 = 0;\ or,\ \psi_2 = \phi_1\psi_1 = \phi_1^2 = 0.480$$
$$B^3 : \psi_3 - \psi_2\phi_1 = 0;\ or,\ \psi_3 = \phi_1\psi_2 = \phi_1^3 = 0.333$$

In general, it is evident that the weight is in this form: $\psi_i = \phi_1\psi_{i-1}$ or $\psi_i = \phi_1^i$. This weight equation should be used in the equation 1 to calculate the forecast.

b. Forecast the last 20 observations.

The last 20 observations forecast has been presented in the figure below. The purple line is the forecast, and green and blue line represent 95% lower and upper prediction levels.



c. Show how to obtain prediction intervals for the forecasts in part b above.

For obtaining prediction intervals, we need to know the variance of the forecast error, which can be obtained by:

$$Var[e_T(\tau)] = \sigma^2 \sum_{i=0}^{\tau-1} \psi_i^2 = \sigma^2 \sum_{i=0}^{\tau-1} \phi_1^{2i} = \sigma^2 \frac{1-\phi^{2\tau}}{1-\phi^2}$$

The $100(1-\alpha)$ prediction interval for $\hat{y}_{T+\tau}(T)$ can be expressed as:

$$\hat{y}_{T+\tau}(T) \pm Z_{\alpha/2}\sqrt{Var[e_T(\tau)]}$$
$$\hat{y}_{T+\tau}(T) \pm Z_{\alpha/2}\sqrt{\sigma^2\frac{1-\phi^{2\tau}}{1-\phi^2}}$$
$$\hat{y}_{T+\tau}(T) \pm Z_{\alpha/2} \times \sigma\sqrt{\frac{1-\phi^{2\tau}}{1-\phi^2}}$$

Here, $\sigma^2 = 1.15$ and $\phi = 0.691$, which can be seen in the r-output in part a. Using this parameters in the above equation, prediction intervals can be calculated for each $\tau -$ *step ahead* and time period T. The forecast function (used in part b) gives 95% and 80% prediction interval. The figure in part b shows the prediction intervals, the dotted line represent 80% prediction interval.The table below presents the 95% and 80% prediction intervals.

```
##     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
##  81       84.18154 82.80723 85.55585 82.07972 86.28337
##  82       84.51592 82.84356 86.18828 81.95827 87.07357
##  83       84.74777 82.94962 86.54592 81.99774 87.49780
##  84       84.90853 83.05294 86.76412 82.07064 87.74642
##  85       85.02000 83.13741 86.90258 82.14083 87.89917
##  86       85.09729 83.20186 86.99271 82.19848 87.99609
##  87       85.15088 83.24930 87.05245 82.24267 88.05908
##  88       85.18803 83.28352 87.09255 82.27533 88.10074
##  89       85.21380 83.30787 87.11973 82.29893 88.12867
##  90       85.23166 83.32505 87.13828 82.31575 88.14758
##  91       85.24405 83.33711 87.15099 82.32764 88.16046
##  92       85.25264 83.34554 87.15974 82.33599 88.16929
##  93       85.25860 83.35142 87.16577 82.34183 88.17536
##  94       85.26272 83.35552 87.16993 82.34590 88.17955
##  95       85.26559 83.35836 87.17281 82.34874 88.18244
##  96       85.26757 83.36034 87.17481 82.35071 88.18444
##  97       85.26895 83.36171 87.17619 82.35208 88.18582
##  98       85.26990 83.36266 87.17714 82.35303 88.18678
##  99       85.27057 83.36332 87.17781 82.35369 88.18744
## 100       85.27102 83.36378 87.17827 82.35415 88.18790
```

Question 5.33

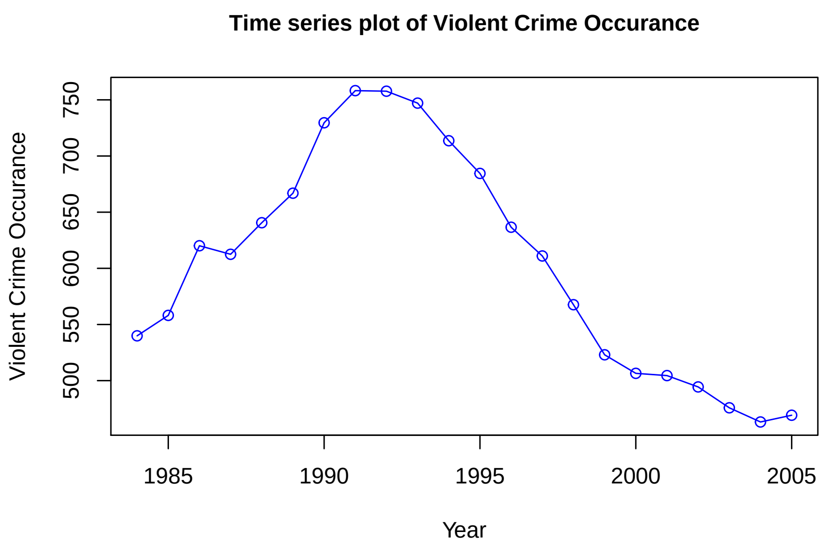
Table B.15 presents data on the occurrence of violent crimes. Develop an appropriate ARIMA model and a procedure for forecasting for these data. Explain how prediction intervals would be computed.

Answer

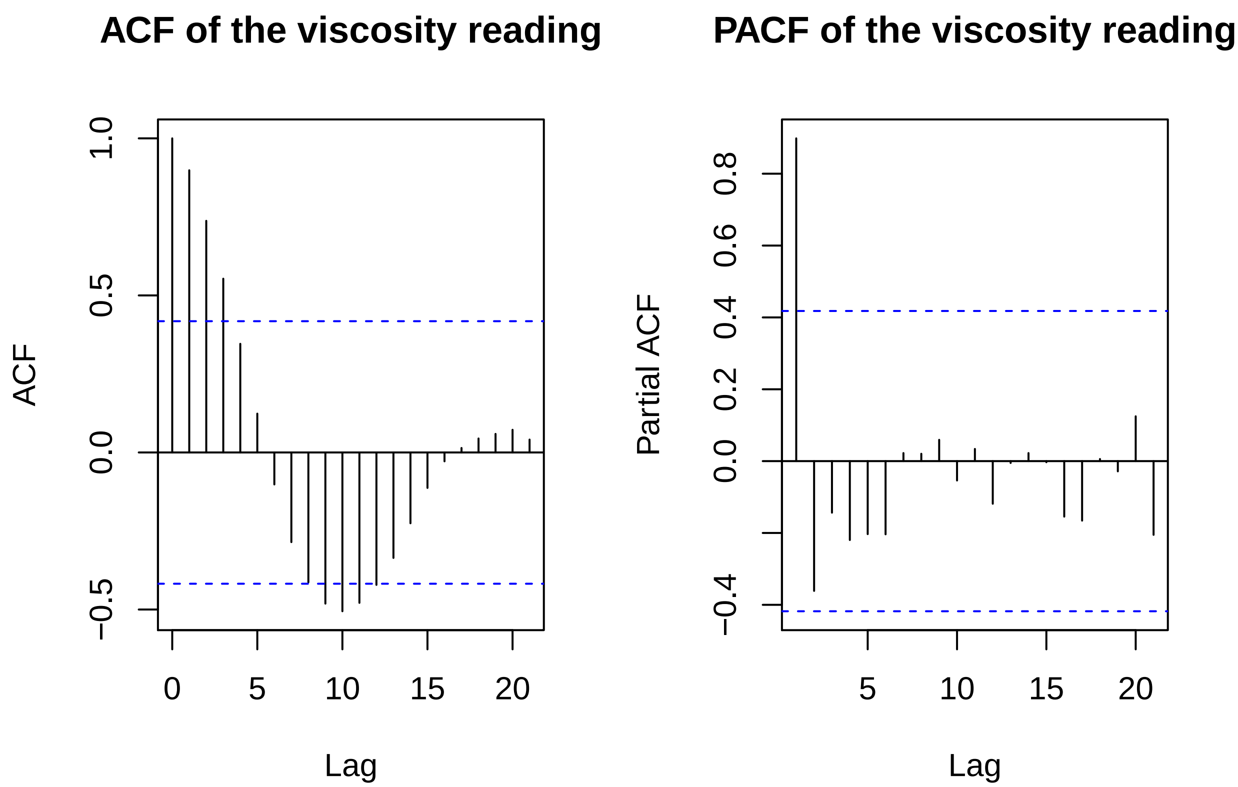
A three-step iterative procedure is used to build an ARIMA model. In step 1, a tentative model of the ARIMA class is identified through analysis of historical data. Unknown parameters of the model are estimated in step 2. In step 3, through residual analysis, diagnostic checks are performed to determine the adequacy of the model, or to indicate potential improvements. These three steps have been presented in the forthcoming text.

Step 1: ARIMA model building

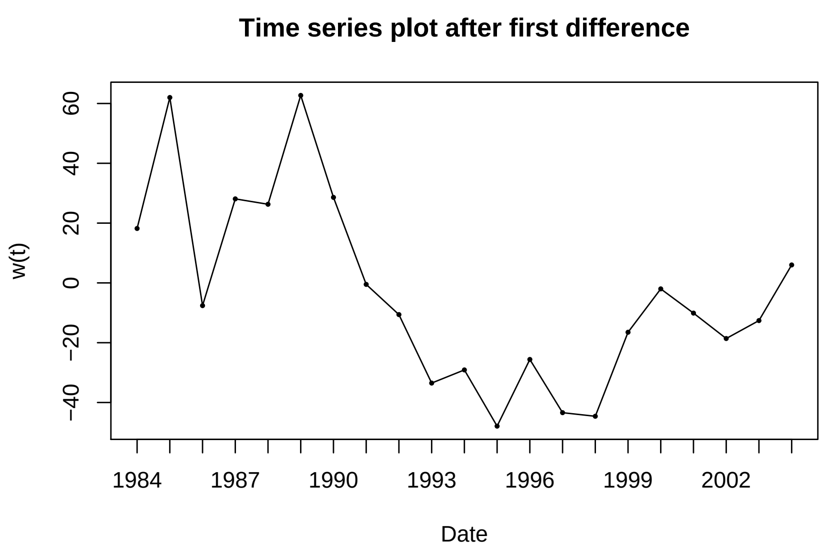
The violent crime occurrence from 1984 to 2005 can be seen in the figure below. There is a clear downward trend from 1992 to 2005 present in the data, which is a indication that this may be a nonstationary time series. To be certain, sample acf has been plotted below.



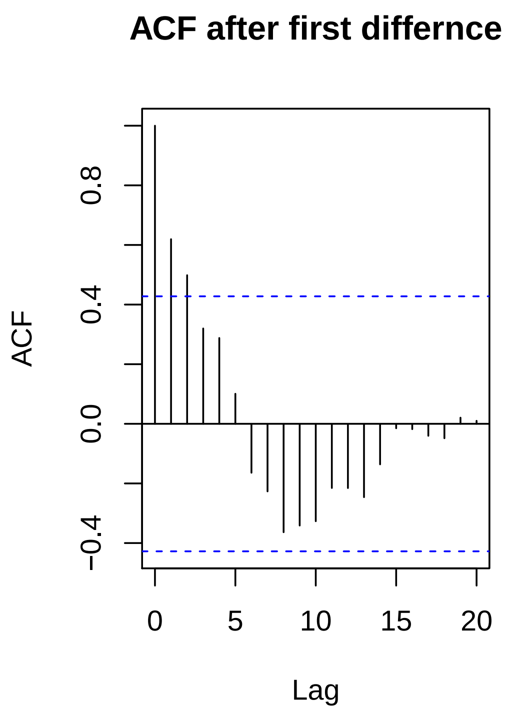
The sample ACF plot of the data can be seen in the figure displayed below. The ACF shows exponential decay and approximately sinusoidal patter at higher lag. These are the properties of stationary time series. But, the ACF and PACF is close to 1 at lag 1, which confirms that this is a nonstationary time series. Consequently, ARIMA model will be good fit for this data.

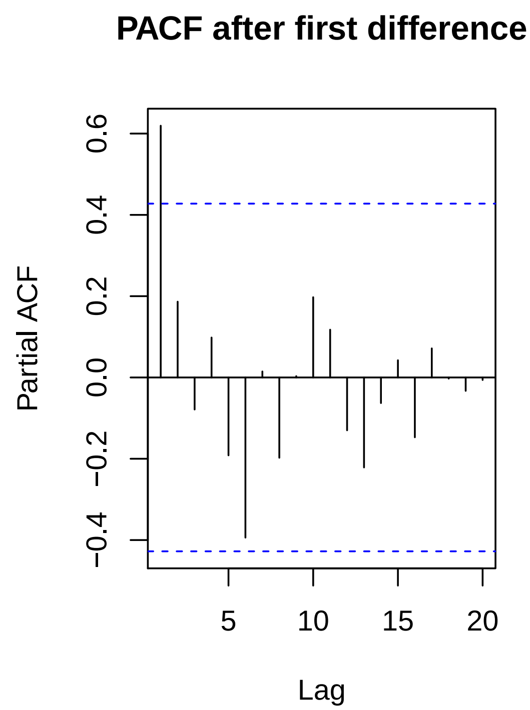


In order to remove the trend from the data in other words making this time series stationary, differencing technique should be applied. The first difference w(t) of the crime data has been plotted in the figure below. The time series after first difference appears to be still nonstationary, since the data shows overall downward trend.

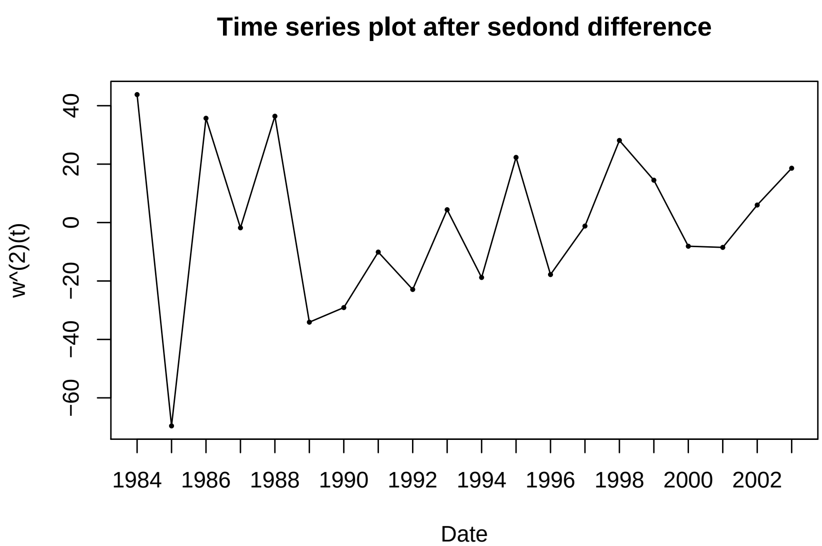


The ACF and PACF plot can be seen below. The ACF and PACF at lag 1 is still significant. Consequently, second difference should be applied to this data.

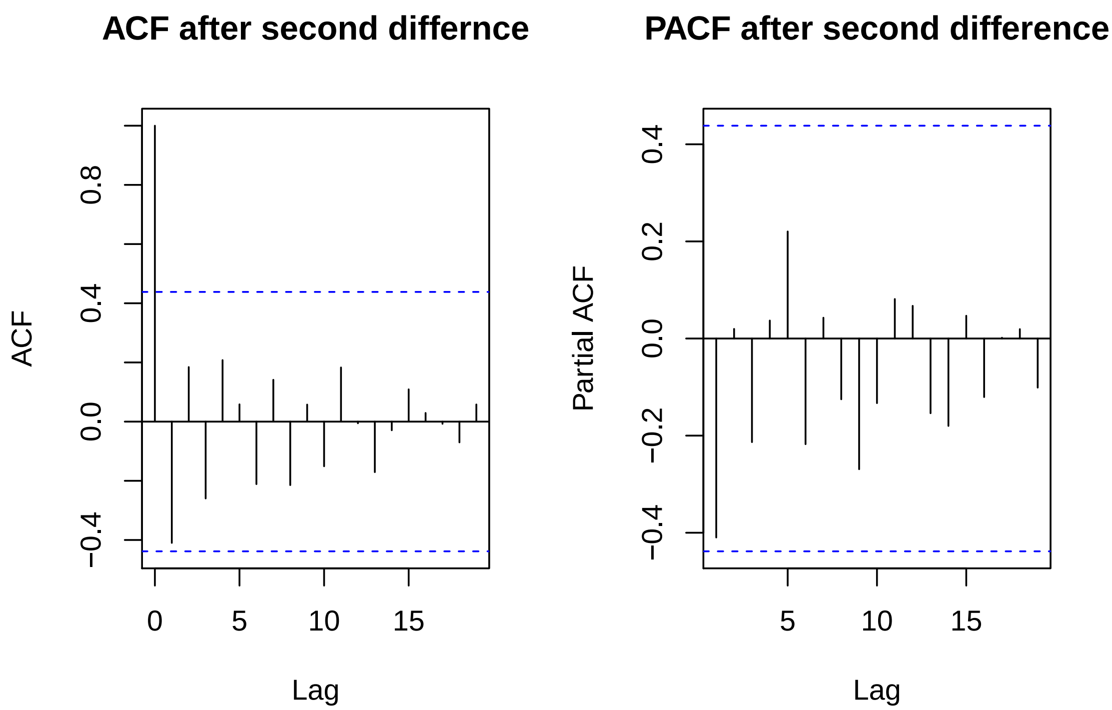




The second difference w(t) of the crime data has been plotted in the figure below. The time series after second difference appears to be stationary. Changing variances can be noticed in data.



After second differencing the ACF and PACF becomes insignificant, suggesting ARIMA(0,2,0). But at lag 1, ACF and PACF are close to the 95% confidence line. Therefore, maybe ARIMA(1,2,0) will be better fit for this data.



The r-code below shows the ARIMA(1,2,0) model fit to the data.

```
# fitting ARIMA(1,2,0) model
vcrime.ARIMA<-arima(v.crime[,2], order=c(1, 2, 0))
vcrime.ARIMA
```

```
##
## Call:
## arima(x = v.crime[, 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
```

To check the validity of the model fit above, auto.arima() function is used. The r-code output below shows that the model fit above is appropriate for the violent crime data.

```
library(forecast)
vcrime.ARIMAauto<- auto.arima(v.crime[,2])
vcrime.ARIMAauto
```

```
## Series: v.crime[, 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
```

Step 2: Parameter Estimation

The ARIMA(1,2,0) can be written as:

$$
\begin{aligned}
(1-\phi B)(1-B)^2 y_t &= \delta + \epsilon_t \\
(1-\phi B)(1-2B+B^2) y_t &= \delta + \epsilon_t \\
(1-2B+B^2-\phi B+2\phi B^2-\phi B^3) y_t &= \delta + \epsilon_t \\
y_t - 2y_{t-1} + y_{t-2} - \phi y_{t-1} + 2\phi y_{t-2} - \phi y_{t-3} &= \delta + \epsilon_t \\
y_t = \delta + \phi y_{t-1} - 2\phi y_{t-2} + \phi y_{t-3} + 2y_{t-1} - y_{t-2} + \epsilon_t &
\end{aligned}
$$

ϕ has been estimated to be -0.4533 (shown in above r-code). Plugging in -0.45 to the last equation:

$$y_t = \delta - 0.45y_{t-1} + 0.9y_{t-2} - 0.45y_{t-3} + 2y_{t-1} - y_{t-2} + \epsilon_t$$

Since after the second difference mean of the time series becomes close to zero, δ can be approximated close to zero. So the ARIMA(1,2,0) model will be:

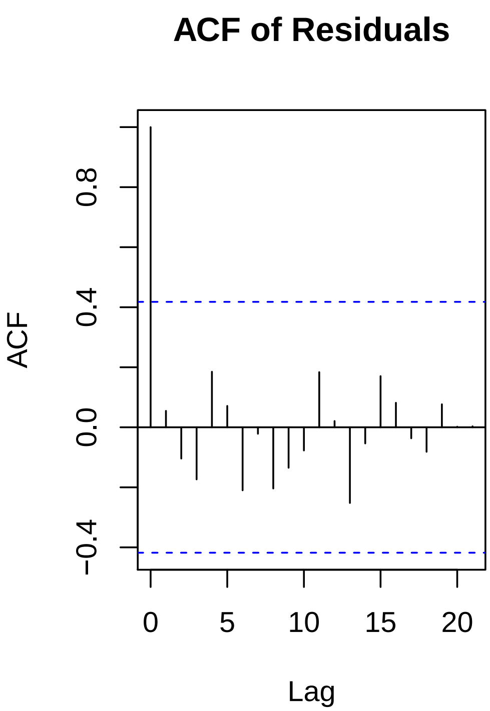
$$y_t = -0.45y_{t-1} + 0.9y_{t-2} - 0.45y_{t-3} + 2y_{t-1} - y_{t-2} + \epsilon_t$$

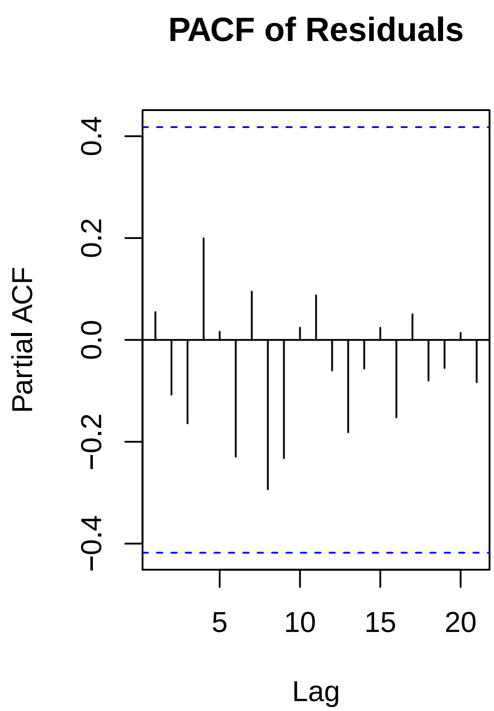
Step 3: Diagnostic Checking

The appropriateness of the model has been analysed through dianostic checking. The ACF and PACF of the residuals and four diagnostic plots have been presented below.

ACF and PACF of residuals:

The ACF and PACF of residuals appears to be insignificant, suggesting no autocorrelation left in the data.





Residual Plots:

The residual plots can be seen in the figure below. The q-q plot suggest that the residuals are approximately normally distributed. Though the histogram appears to be left skewed. Shaprio-wilk test can be performed to be sure about the normality of residuals.

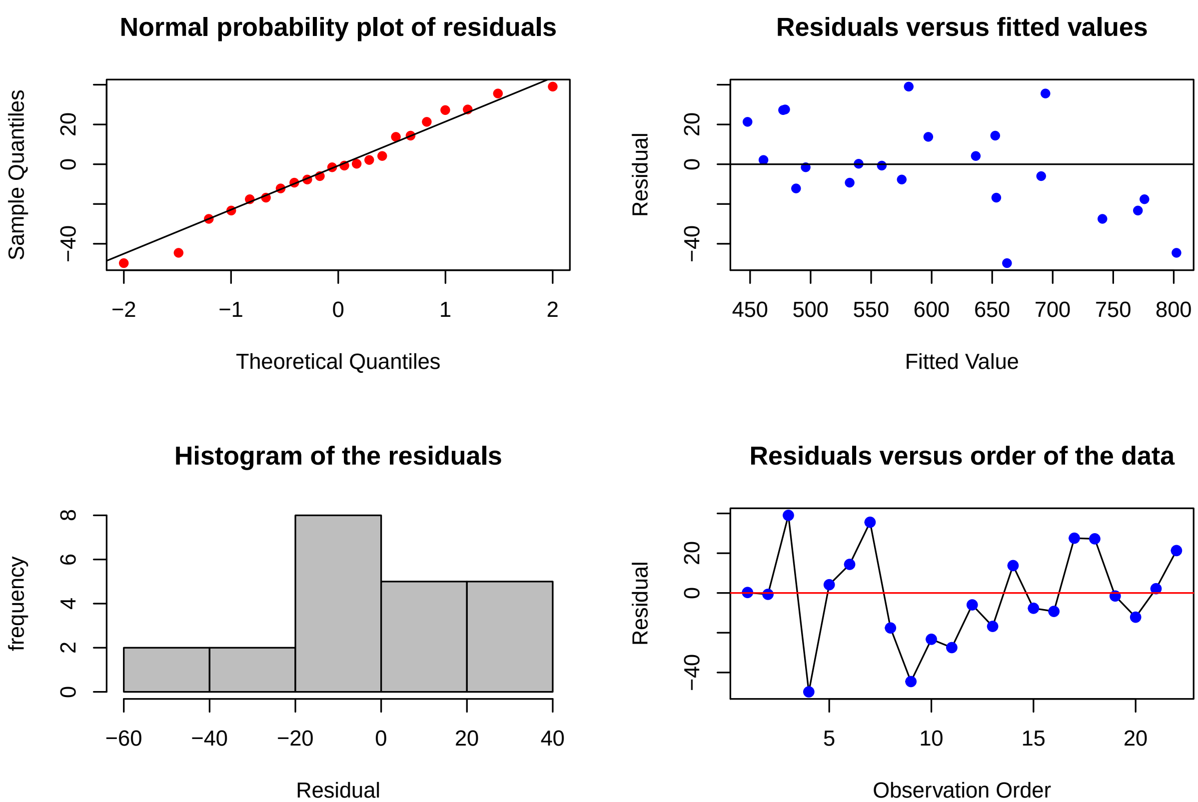
Shapiro-wilk test for normality:

$$H_0 : \textit{Residuals are normally distributed}$$
$$H_a : \textit{Residuals are not normally distributed}$$

```r
shapiro.test(vcrime.ARIMAauto$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  vcrime.ARIMAauto$residuals
## W = 0.97647, p-value = 0.853
```

Since p-value is higher than 0.01(assuming $\alpha = 0.01$), we fail to reject the null hypothesis and conclude that the residuals are normally distributed.
The variance appears to be changing in the residual vs order of the data plot. Residuals vs fitted value plot also suggest that there might be mild violation of common variance of residuals assumption.



How this model would be used for forecasting?

Best forecast model in mean square sense is:

$$\hat{y}_{T+\tau}(T) = E[y_{T+\tau}|y_T, y_{T-1,...}] = \mu + \sum_{i=\tau}^{\infty} \psi_i \epsilon_{T+\tau-i} \tag{3}$$

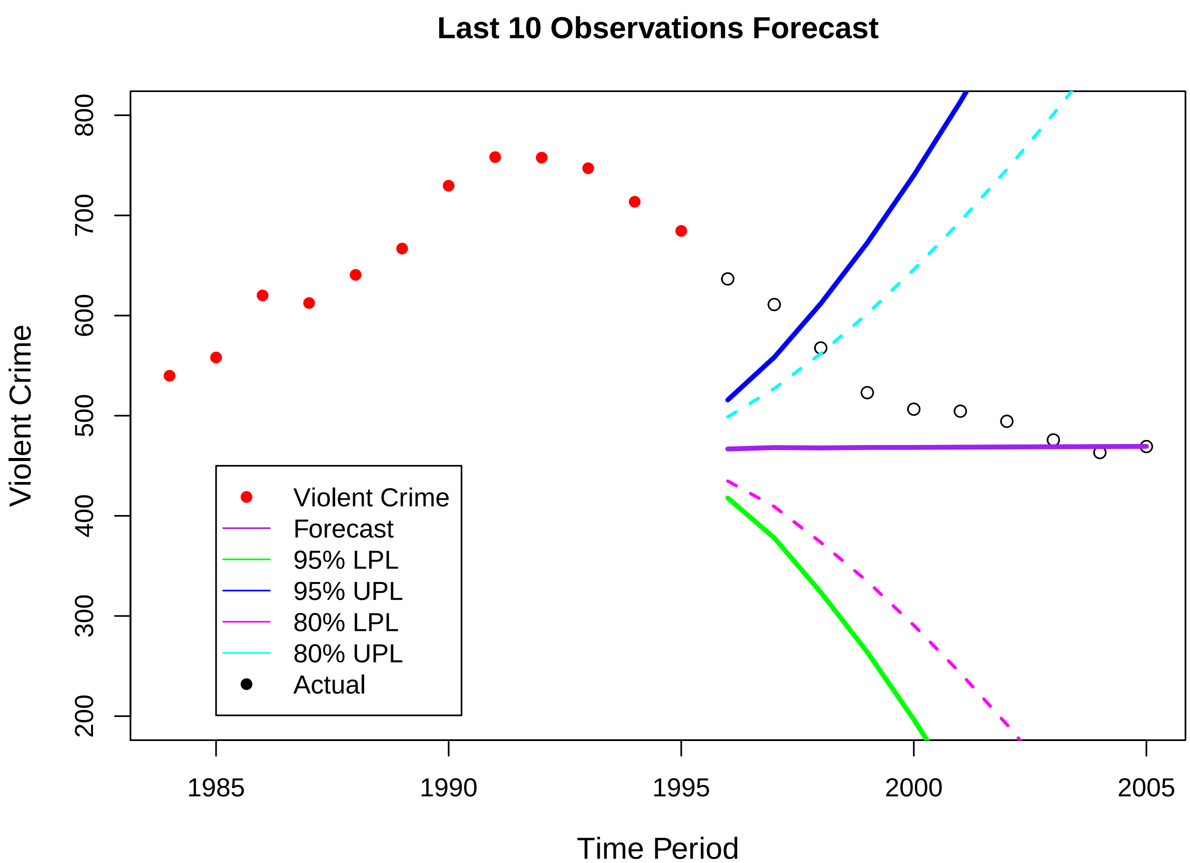
To calculate the forecast ψ weights should be obtained. The ψ weights for the general ARIMA(p,d,q) model may be obtained by equation like powers of B in the expansion of

$$(\psi_0 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 ...)(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)(1 - B)^d = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q) \tag{4}$$

For ARIMA(1,2,0) model, equation 4 becomes:

$$(\psi_0 + \psi_1 B + \psi_2 B^2 + \psi_3 B^3 ...)(1 - B)^2(1 - \phi_1 B) = 1 \tag{5}$$

The $\psi's$ can be estimated equating the power of B, and used in equation 3 for forecasting. The detail calculation of $\psi's$ has not been presented here. However, from the r-output, last 10 observations forecast has been presented in the figure below. The purple line is the forecast, and green and blue line represent 95% lower and upper prediction levels. Magenta and cyan dotted lines are the 80% lower and upper prediction levels. It can be seen that the mean forecast (purple line) cannot forecast the 1996 to 2002 well, though 8 of the data points out of 10 are within 95% prediction interval.



For obtaining prediction intervals, we need to know the variance of the forecast error, which can be obtained by:

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

The $100(1-\alpha)$ prediction interval for $\hat{y}_{T+\tau}(T)$ can be expressed as:

$$\hat{y}_{T+\tau}(T) \pm Z_{\alpha/2}\sqrt{Var[e_T(\tau)]}$$

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

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

Here, σ^2 has been estimated to be 623.5. The $\psi's$ can be estimated from equation 5. Then, prediction intervals can be calculated using above equation. The forecast function gives 95% and 80% prediction interval. The figure displayed above shows the prediction intervals, the dotted line represent 80% prediction interval.The table below presents the 95% and 80% prediction intervals.

```
##    Point Forecast     Lo 80    Hi 80       Lo 95     Hi 95
## 23       466.7685 434.76854 498.7685  417.82878  515.7082
## 24       468.1591 409.22121 527.0969  378.02140  558.2967
## 25       467.8171 373.55509 562.0790  323.65581  611.9783
## 26       468.2604 334.67715 601.8437  263.96243  672.5584
## 27       468.3478 290.70970 645.9859  196.67377  740.0218
## 28       468.5965 243.08759 694.1055  123.71037  813.4827
## 29       468.7721 191.67339 745.8709   44.98618  892.5581
## 30       468.9809 136.90442 801.0573  -38.88624  976.8480
## 31       469.1746  78.89640 859.4528 -127.70440 1066.0536
## 32       469.3751  17.85539 920.8949 -221.16472 1159.9150
```