STAT478_Project_ARIMA_Model_Forecast

Yiqun Hu Z1834885

10/28/2018

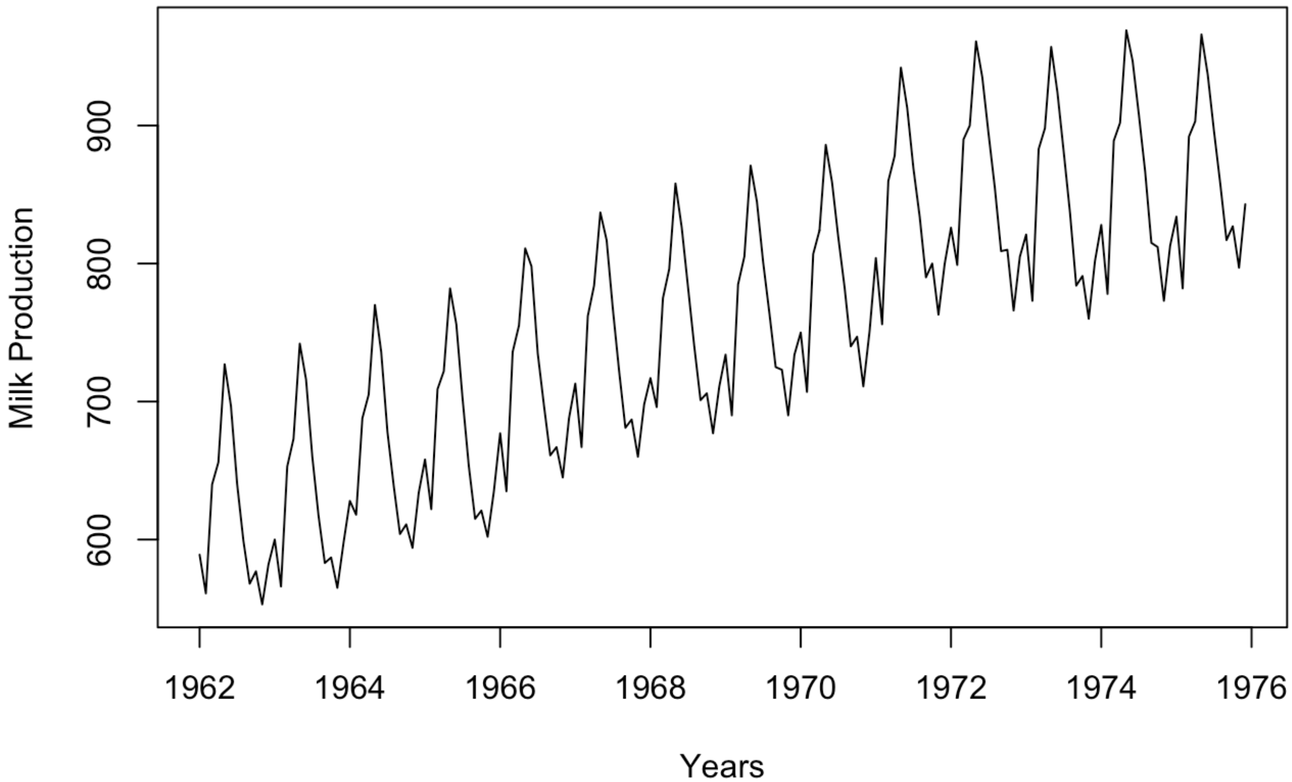
Step 1 Description of data

Dataset Title: Monthly milk production: pounds per cow. Jan 62 – Dec 75 Per cow monthly milk production from Jan 1962 to Dec 1975.(URL:https://datamarket.com/data/set/22ox/monthly-milk-production-pounds-per-cow-jan-62-dec-75#!ds=22ox&display=line (https://datamarket.com/data/set/22ox/monthly-milk-production-pounds-per-cow-jan-62-dec-75#!ds=22ox&display=line))

```
data<-read.csv(file = "~/desktop/monthly-milk-production-pounds-p.csv") ##read data
data = ts(data[,2],start = c(1962,1),frequency = 12)
```

plot milk production as time series

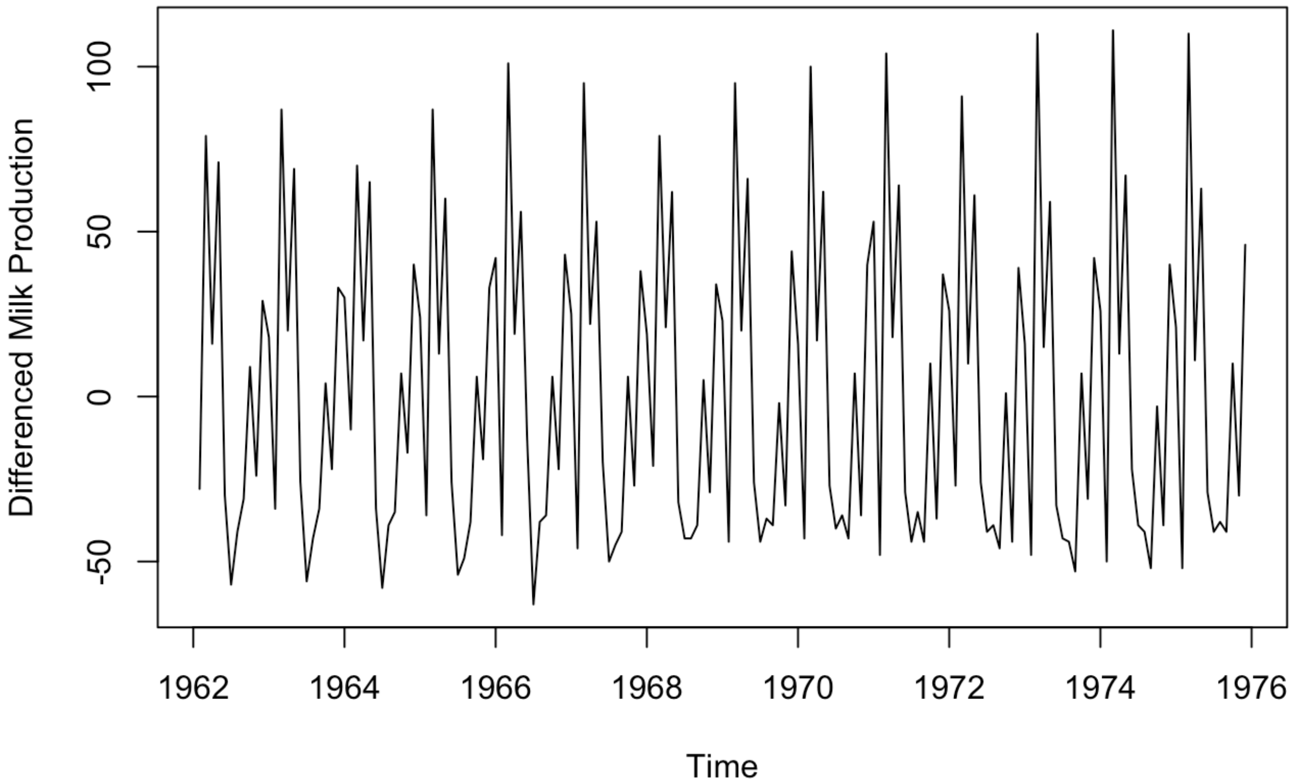
```
plot(data, xlab = "Years", ylab = "Milk Production")
```



Clearly, the above chart has an upward trend for milk production.

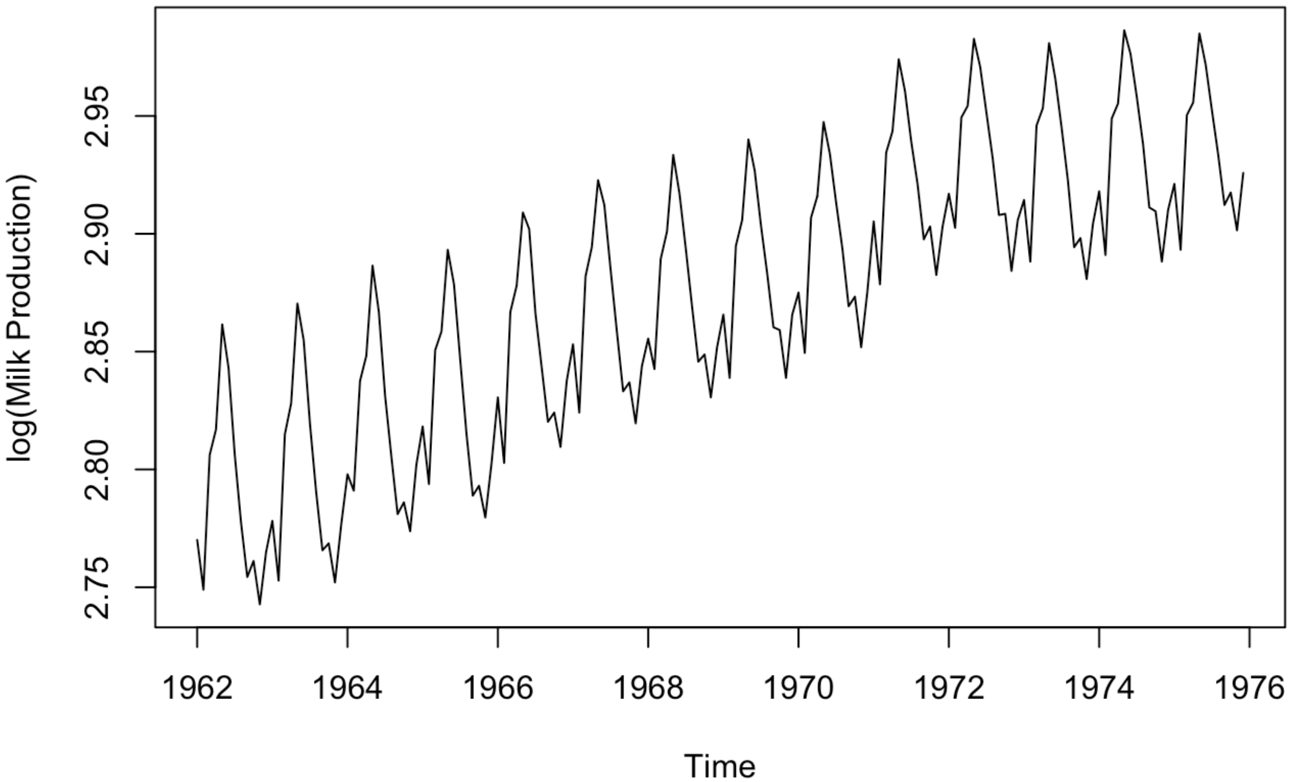
Step 2 Difference data to make data stationary on mean

```
plot(diff(data),ylab = "Differenced Milk Production")
```



Step 3 log transform data to make data stationary on variance

```
plot(log10(data), ylab = "log(Milk Production)")
```

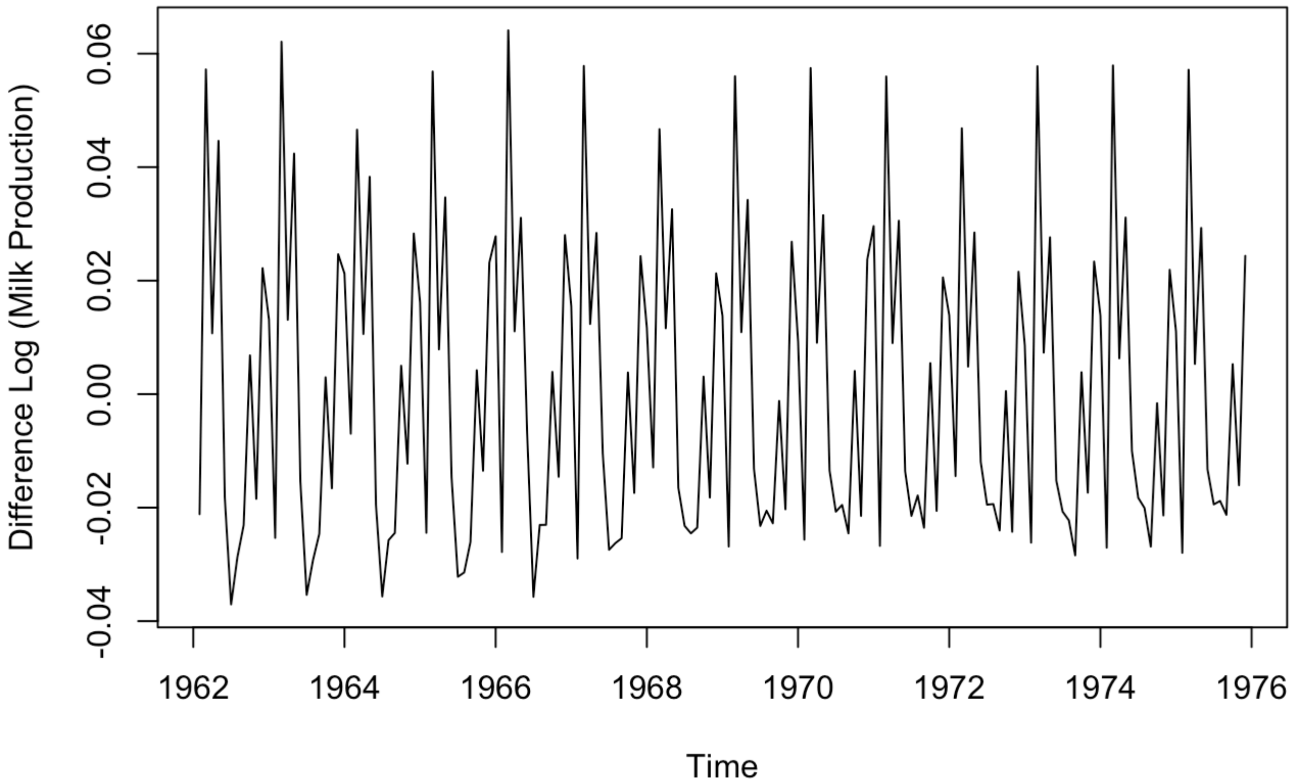


Step 4 Difference log transform data to make data stationary on both mean and variance

Now the series looks stationary on both mean and variance. 1st Differencing (d=1) of log of production

$$Y_t^{new'} = log_{10}(Y_t) - log_{10}(Y_{t-1})$$

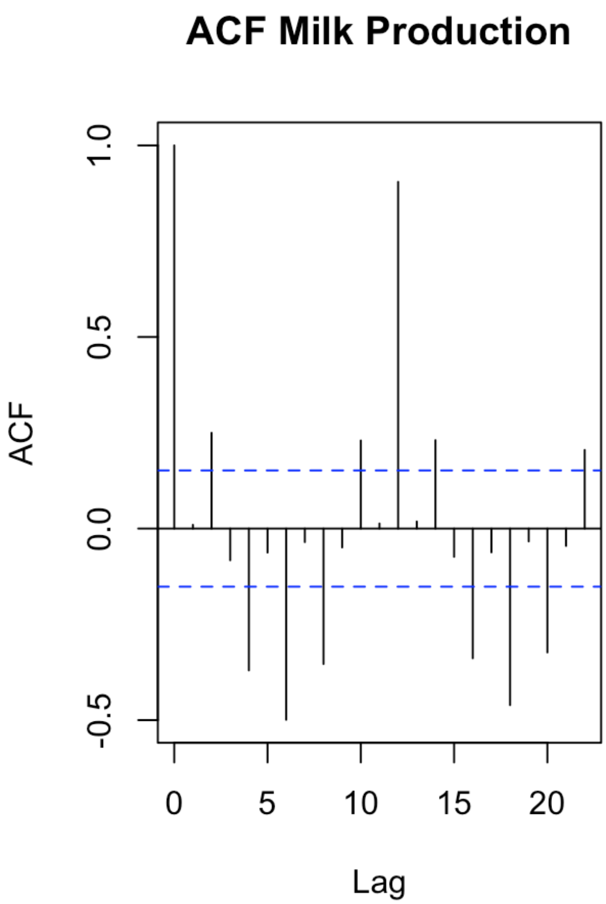
```
plot(diff(log10(data)),ylab = "Difference Log (Milk Production)")
```

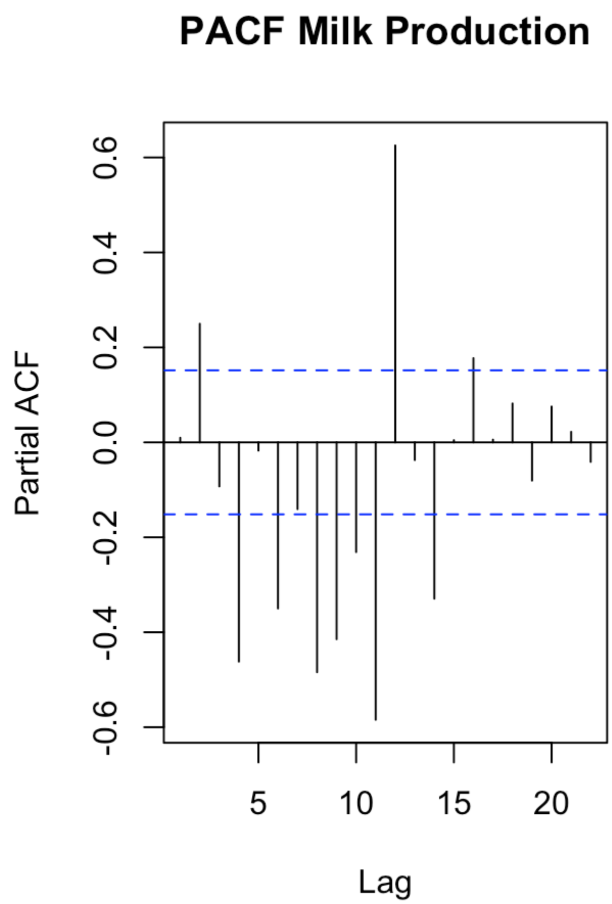


Step 5 plot ACF and PACF to identify potential AR and MA model

Create autocorrelation(ACF) and partial autocorrelation factor(PACF) plots to identify patterns in the above data which is stationary on both mean and variance.

```
par(mfrow = c(1,2))
acf(ts(diff(log10(data))),main = 'ACF Milk Production')
pacf(ts(diff(log10(data))), main = 'PACF Milk Production')
```





Since, there are enough spikes in the plot outside the insignificant zone, we can conclude that the residuals are not random. Also, there is a seasonal component available in the residuals at lag12 (represented by spikes at lag 12).

Step 6 Identification of best fit ARIMA model

```
require(forecast)
```

```
## Loading required package: forecast
```

```
ARIMAfit = auto.arima(log10(data), approximation = FALSE, trace = FALSE)
summary(ARIMAfit)
```

```
## Series: log10(data)
## ARIMA(0,1,1)(0,1,1)[12]
##
## Coefficients:
##           ma1     sma1
##       -0.1527  -0.5990
## s.e.   0.0830   0.0644
##
## sigma^2 estimated as 2.134e-05:  log likelihood=614.32
## AIC=-1222.64   AICc=-1222.48   BIC=-1213.51
##
## Training set error measures:
##                         ME        RMSE        MAE          MPE    MAPE
## Training set -5.930383e-05 0.004408474 0.00324883 -0.002049193 0.11311
##                   MASE      ACF1
## Training set 0.2596191 0.0175153
```

The best fit model is selected based on minimum AIC, and BIC values. As expected, our model has I component equal to 1. This represents differencing of order 1. There is additional differencing of lag 12 in the above best fit model. Moreover, the best fit model has MA value of order 1. Alsa, there is seasonal MA with lag 12 of order 1.

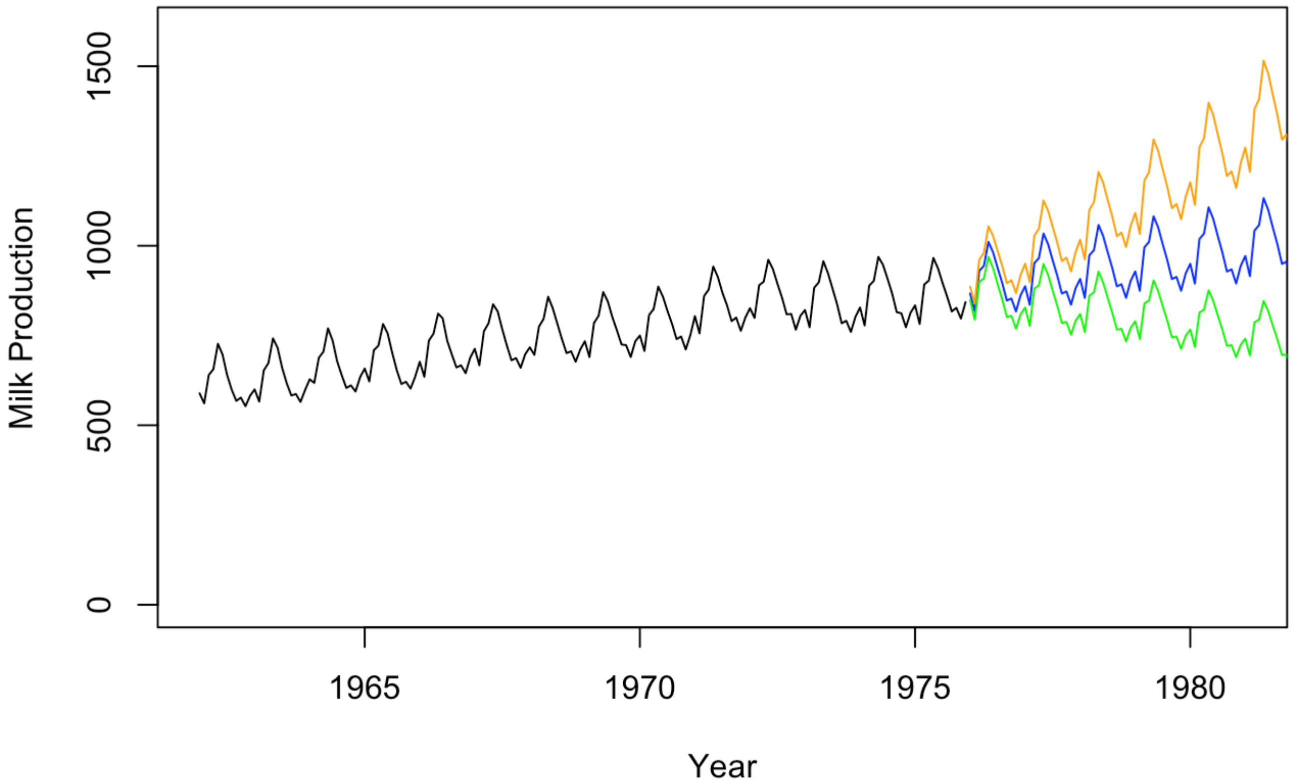
Step 7 Forecast production by using the best fit ARIMA model

```
par(mfrow = c(1,1))
pred = predict (ARIMAfit, n.ahead = 72) ##predict for 6 yrs
pred
```

```
## $pred
##           Jan      Feb      Mar      Apr      May      Jun      Jul
## 1976 2.938125 2.912174 2.968548 2.974928 3.004665 2.992002 2.972204
## 1977 2.948004 2.922052 2.978426 2.984807 3.014544 3.001881 2.982082
## 1978 2.957882 2.931931 2.988305 2.994685 3.024422 3.011759 2.991961
## 1979 2.967760 2.941809 2.998183 3.004563 3.034301 3.021638 3.001839
## 1980 2.977639 2.951688 3.008061 3.014442 3.044179 3.031516 3.011717
## 1981 2.987517 2.961566 3.017940 3.024320 3.054057 3.041395 3.021596
##           Aug      Sep      Oct      Nov      Dec
## 1976 2.952349 2.928147 2.930971 2.912251 2.935456
## 1977 2.962227 2.938026 2.940850 2.922129 2.945334
## 1978 2.972106 2.947904 2.950728 2.932007 2.955212
## 1979 2.981984 2.957782 2.960607 2.941886 2.965091
## 1980 2.991862 2.967661 2.970485 2.951764 2.974969
## 1981 3.001741 2.977539 2.980363 2.961643 2.984848
##
## $se
##              Jan         Feb         Mar         Apr         May
## 1976 0.004619526 0.006054921 0.007210001 0.008204033 0.009090006
## 1977 0.014937872 0.015912750 0.016831258 0.017702171 0.018532201
## 1978 0.024653583 0.025642870 0.026595384 0.027514942 0.028404748
## 1979 0.035134867 0.036177714 0.037191330 0.038178045 0.039139893
## 1980 0.046500113 0.047604280 0.048683411 0.049739134 0.050772911
## 1981 0.058739640 0.059905924 0.061049931 0.062172891 0.063275926
##              Jun         Jul         Aug         Sep         Oct
## 1976 0.009896982 0.010642946 0.011339944 0.011996515 0.012618971
## 1977 0.019326616 0.020089641 0.020824728 0.021534737 0.022222073
## 1978 0.029267513 0.030105563 0.030920908 0.031715299 0.032490273
## 1979 0.040078664 0.040995944 0.041893144 0.042771528 0.043632232
## 1980 0.051786055 0.052779754 0.053755088 0.054713038 0.055654501
## 1981 0.064360059 0.065426229 0.066475303 0.067508075 0.068525284
##              Nov         Dec
## 1976 0.013212133 0.013779786
## 1977 0.022888778 0.023536604
## 1978 0.033247187 0.033987249
## 1979 0.044476283 0.045304612
## 1980 0.056580301 0.057491195
## 1981 0.069527613 0.070515695
```

The following is the output with forecasted values of milk production is blue. Also, the range of the expected error(2 times standard deviation) is displayed with orange and green line on the either side of predicted blue line.

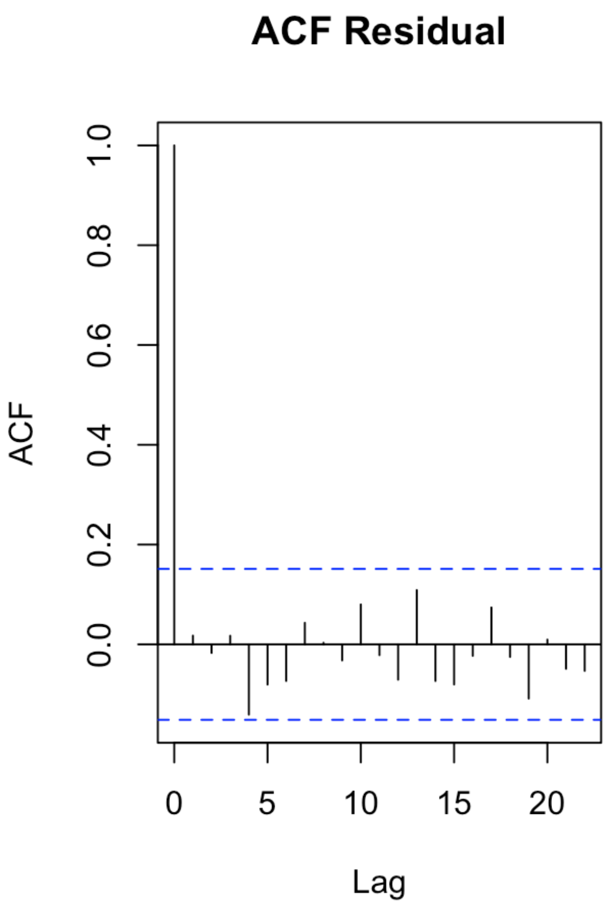
```
plot(data, type = 'l', xlim = c(1962,1981), ylim = c(1,1600), xlab = 'Year', ylab = 'Milk
Production' )
lines(10^(pred$pred),col = 'blue')
lines(10^(pred$pred+2*pred$se), col = 'orange')
lines(10^(pred$pred-2*pred$se), col = 'green')
```

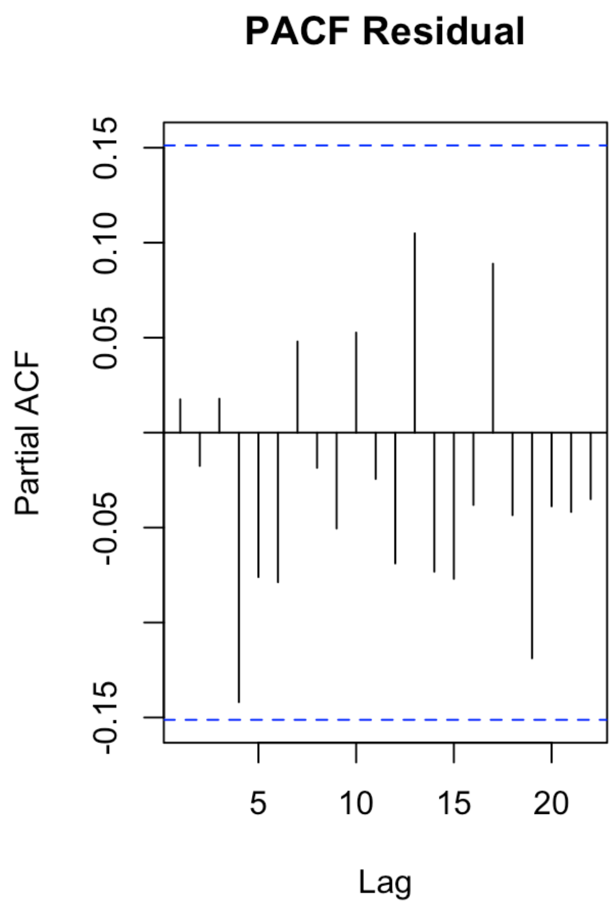


Step 8 Plot ACF and PACF for residuals of ARIMA model

Plot ACF anf PACF of the residual of best fit ARIMA model.

```
par(mfrow = c(1,2))
acf(ts(ARIMAfit$residuals), main = 'ACF Residual')
pacf(ts(ARIMAfit$residuals), main = 'PACF Residual')
```





Since there are no spikes outside the insignificant zone for ACF and PACF plots, we can conclude that residuals are random. Hence the ARIMA model is working fine.