**Analysis of Time Series Data “Annual Swedish Fertility rates” Using R**

Sattajit Sutradhar

Instructor: Dr. Tharanga Wickramarachchi

**Abstract:** Time series data analysis plays a substantial role in our prediction process. Analysis of the changes of the data values over time has a significant impact on the future predictions. How the society is going be influenced by the happening of the Past and present events is analyzed in times series data analysis. So with the analysis of the data overtime we could predict the future events to a certain level of accuracy and take actions accordingly. Many methods have been developed to analyze different set of time series data i.e. yearly, quarterly, monthly ,weekly daily etc. with various characteristics such as trend, seasonality etc. In our analysis we are going to analyze Annual Swedish Fertility rates over the year period 1750-1849.Using training data set and validation data set we are going predict fertility rates and compare with the actual one’s. First we will select an initial model to fit and then search the neighborhood for the model that fits the best. Then we will perform some diagnostic checking, hypothesis testing of the model and make a prediction.

**Introduction:** There are different kinds of data sets. They can be stationary, nonstationary, seasonal etc. We use different approach for analyzing data set with different characteristics. Initially. By plotting the data set into a graph we can recognize these characteristics. However, for a stationary data we can also look into ACF and PACF for primary model guessing. For a non-stationary data set that contains trend or seasonal pattern or a non-constant variance we have different options to get rid of those characteristics such as (1) Differencing (2)Transform (3) Seasonal model (4) Unit root model. We can use differencing to eliminate any trend or a seasonal component. If the differenced series produces an ARMA process the original series is said to be ARIMA process. Now if the data display characteristics suggesting non-stationary then we might need to make a transformation .So as to produce a new series that is more capable with the assumption of stationary. One of the most popular methods of transforming is box-cox transformation. For seasonal model we know that a seasonal component in a series can be removed by differencing it at a certain lag where it’s the period of seasonal component. Then we look at the ACF and PACF to guess an initial model and search the neighborhood for better fitted model.

Behavior of the ACF and PACF for an ARMA model:

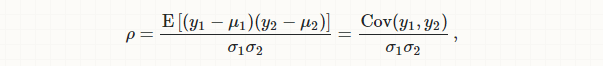
|  |  |  |  |
| --- | --- | --- | --- |
|  | AR(p) | MA(q) | ARMA(p,q) |
| ACF | Tails off(decays gradually) | Cuts off after leg q | Tails off |
| PACF | Cuts off after leg p | Tails off | Tails off |

Thus after the final model selection we are going to diagnostic checking and hypothesis testing to be certain .Hence make prediction.

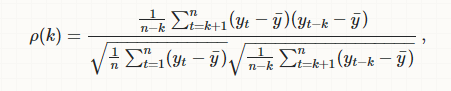
**Methods:**

**Autocorrelations**

The correlation between two variables y1, y2 is defined as:



where E is the expectation operator, μ1 and μ2 are the means respectively for y1 and y2 and σ1 ,σ2 are their standard deviations.



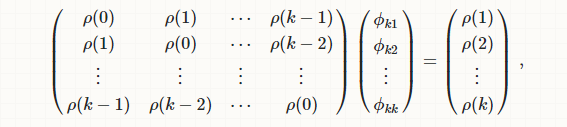
**Partial Autocorrelation:**

 The partial autocorrelation of order measures the effect (linear dependence) of yt−2 on yt after removing the effect of yt−1 on both yt and yt−2.

Partial autocorrelation could be obtained as a series of regressions of the form:

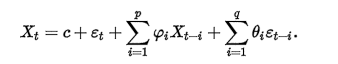


An alternative way to compute the sample partial autocorrelations by solving the following system for each order k.

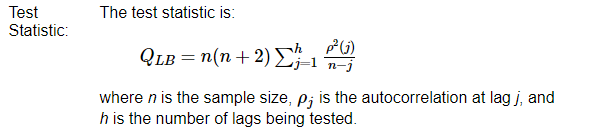


**ARMA(p,q) model:**

The notation ARMA(*p*, *q*) refers to the model with *p* autoregressive terms and *q* moving-average terms. This model contains the AR(*p*) and MA(*q*) models.



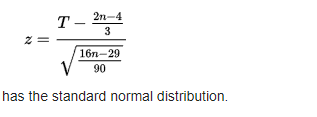
**LJUNG-BOX TEST**



**Turning Point test:**

We say *i* is a turning point if the vector *X*1, *X*2, ..., *Xi*, ..., *Xn* is not monotonic at index *i*. The number of turning points is the number of maxima and minima in the series.

Letting *T* be the number of turning points then for large *n*, *T* is approximately [normally distributed](https://en.wikipedia.org/wiki/Normal_distribution) with mean (2*n* − 4)/3 and variance (16*n* − 29)/90. The test statistic



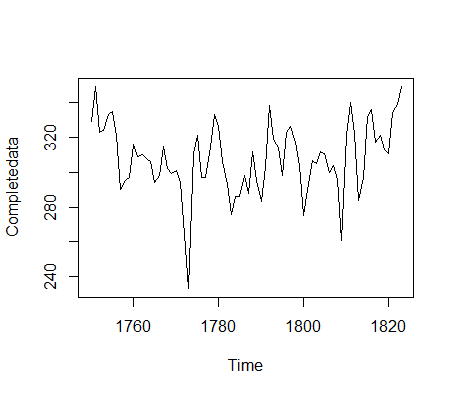
**Results:**

We have fertility rate of Swedish women form year 1750-1849. The name of the text file of the data set anualfertilityrate.txt so

# Read data as a time series

Completedata = ts(scan("anualfertilityrate.txt"), start=c(1750, 1), frequency=1)

plot(Completedata)

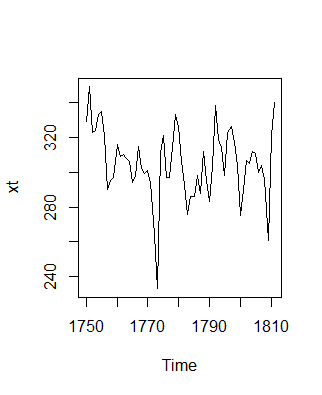


training = ts(scan("anualfertilityrate.txt")[-(63:74)], start=c(1750, 1), frequency=1)

validation=ts(scan("anualfertilityrate.txt")[(63:74)], start=c(1750, 1), frequency=1)

xt=training

plot(xt)



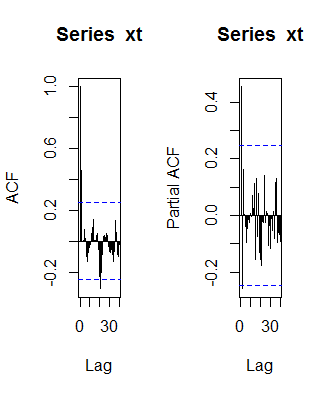
Looking at it we can say that there is no significant trend or seasonal pattern. So now we look at the ACF and PACF.

## Viewing acf and pacf

par(mfrow=c(1,2))

acf(xt, lag.max=40)

pacf(xt, lag.max=40)



Notice that (a) the correlation at lag 1 is significant and positive indicating MA(1), and (b)the PACF has only 2 legs indicating AR(2) model. So let’s guess the initial model is ARMA (2,0, 1). Let's fit a model assuming there is no trend. We should be able to see it at the end if it this assumption is not appropriate. Now we look in the neighborhood for best fitted model that gives us the smallest AICC.

## ACF and PACF suggests an AR(2) or an ARMA(2,0,1) process so we start with it Arima(xt,order=c(2,0,1), include.mean=TRUE)

Arima(xt,order=c(2,0,0), include.mean=TRUE)

Arima(xt,order=c(1,0,0), include.mean=TRUE)

Arima(xt,order=c(0,0,1), include.mean=TRUE)

Arima(xt,order=c(0,0,2), include.mean=TRUE)

Arima(xt,order=c(1,0,1), include.mean=TRUE)

Arima(xt,order=c(1,0,2), include.mean=TRUE)

Arima(xt,order=c(2,0,2), include.mean=TRUE)

**Winner**

> fit = Arima(xt,order=c(0,0,1), include.mean=TRUE)

> fit

Series: xt

ARIMA(0,0,1) with non-zero mean

Coefficients:

ma1 mean

0.5647 305.4894

s.e. 0.0939 3.3436

sigma^2 estimated as 295.9: log likelihood=-263.54

AIC=533.08 AICc=533.5 BIC=539.46

So our best fitted model is MA(1)

**Xt=Zt+0.0564Z(t-1)+305.4894**

**T test:**

> Tstat = fit$coef/sqrt(diag(fit$var.coef))

> Tstat

ma1 intercept

6.011045 91.364067

> pval = 2\*pt(abs(Tstat), df=60, lower.tail = FALSE)

> pval

ma1 intercept

1.177574e-07 4.142450e-66

Here the p value is very small so we cannot dropped the coefficients from the model. They

Should be included in the model.

**Model Diagnostic:**

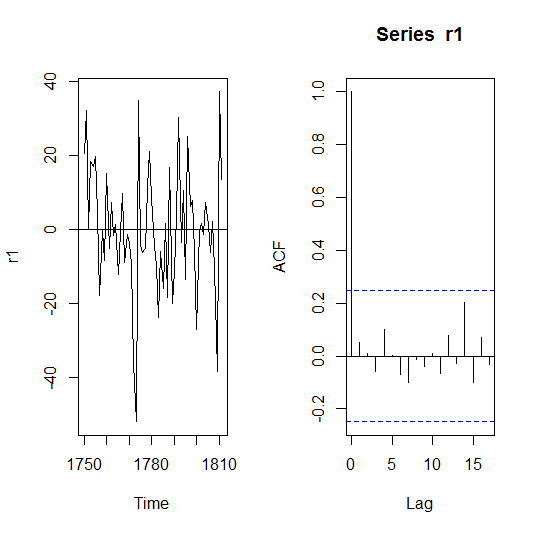
r1 = fit$residuals

plot(r1)

abline(h=0)

## ACF of residuals

acf(r1)



**From the residual plot and ACF we can say that it is indeed MA(1).**

**Box-Ljung test:**

> Box.test(r1, lag=20, type="Ljung-Box")

Box-Ljung test

data: r1

X-squared = 9.709, df = 20, p-value = 0.9731

Here p value is greater than .05 so the residual is WN.

**Few more WN tests:**

> turning.point.test(r1, "two.sided")

**Turning Point Test**

data: r1

statistic = 0.91713, n = 62, p-value = 0.3591

alternative hypothesis: non randomness

> rank.test(r1, "two.sided")

**Mann-Kendall Rank Test**

data: r1

statistic = -1.1116, n = 62, p-value = 0.2663

alternative hypothesis: trend

> qqnorm(r1)

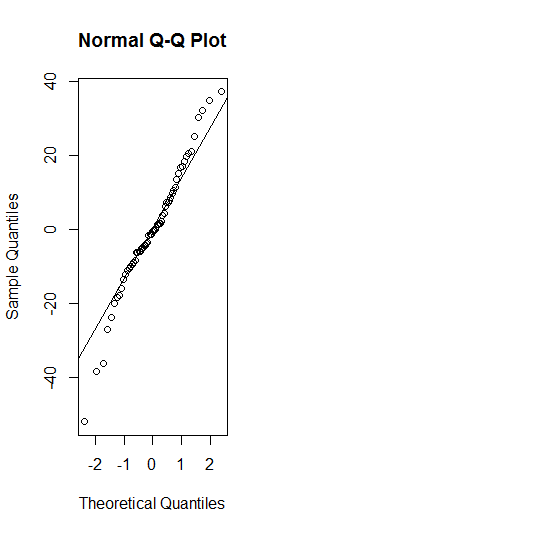
> qqline(r1)

> shapiro.test(r1)

**Shapiro-Wilk normality test**

data: r1

W = 0.97649, p-value = 0.2788



All these WN tests telling us that the residual is a WN.

**Forecasting:**

plot(xt, col="red")

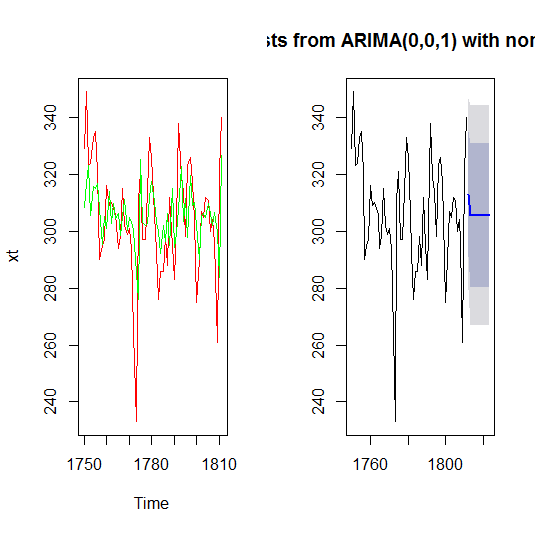
lines(fitted(fit), col="green")

## Forecasting next 12 observations

fcast=forecast(fit,12)

fcast

plot(fcast)



We can see that our prediction is pretty much accurate with 80% and 90% confidence interval.

**Comparing:**

Forecasting Data set

fcast

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

1812 313.1026 291.0575 335.1476 279.3876 346.8175

1813 305.4894 280.1721 330.8067 266.7699 344.2089

1814 305.4894 280.1721 330.8067 266.7699 344.2089

1815 305.4894 280.1721 330.8067 266.7699 344.2089

1816 305.4894 280.1721 330.8067 266.7699 344.2089

1817 305.4894 280.1721 330.8067 266.7699 344.2089

1818 305.4894 280.1721 330.8067 266.7699 344.2089

1819 305.4894 280.1721 330.8067 266.7699 344.2089

1820 305.4894 280.1721 330.8067 266.7699 344.2089

1821 305.4894 280.1721 330.8067 266.7699 344.2089

1822 305.4894 280.1721 330.8067 266.7699 344.2089

1823 305.4894 280.1721 330.8067 266.7699 344.2089

Our Validation Data set:

322 284 297 332 336 317 321 313 311 334 339 349

So we can see that values from these data set almost always falls in 80% confidence interval and even more precisely with 95% confidence interval. Hence our prediction is pretty much accurate.

**Conclusion:** From our analysis of the data set ‘Annual Swedish Fertility rates’ from 1750-1849

We can say that it fits the best with MA (1) which is also assured by Other Diagnostic tests of the residuals .Coefficients of the model is Significant and residual follows WN. The residual plot looks stationary. Forecasting on the validation data set based on the training Data is very accurate.

Therefore our model gives us a reliable prediction for the Annual Fertility rates of Swedish women. So we can use this model to furthermore predict future number of annual fertility rates.This model can be used by statistical farms and the government for greater research.

**References**

## [1]Time Series Data Library

The Time Series Data Library (TSDL) was created by Rob Hyndman, Professor of Statistics at Monash University, Australia.

# [2] " Visual Detection of Additive and Multiplicative Seasonality, [Shishir Shakya's Blog](http://shishirshakya.blogspot.com/)

# [3] ARIMA estimation in R, [Ralf Becker](https://www.youtube.com/channel/UCuBP5noByZUS1505JRrflaQ), Published on Feb 17, 2016.