Model Fitting & Residual Analysis  
for Annual Earthquake Frequency (1916-2015)

Pranav Gopalkrishna, 1940223

2022-04-05

# INTRODUCTION & AIMS

Model identification and fitting for a given time series data is key in helping predict future values for the time series i.e. forecasting. In this assignment, the aim is to fit an ARMA model (ARIMA with order of differencing as 0) to the data using the **auto.arima** function (which compares many possible ARIMA models and selects the best one).

Having identified and fitted the best ARMA model, we will perform residual analysis i.e. analysis of the residuals (i.e. deviations between model estimates and actual values of the dataset), and conclude whether the obtained ARMA model is ideal for the dataset. To conclude this, we will verify the two assumptions of residuals i.e. uncorrelatedness and normality. If both can be said to be satisfied, we may conclude that the obtained model is ideal.

# DATASET

This dataset contains annual information about earthquake frequency in an unknown region or set of regions, from 1916 to 2015.

setwd("~/Documents/Study/computerScience/programming/r/data/")  
data = read.table("earthquakesTimeSeries.txt", header = TRUE, sep = "\t")  
head(data)

**Year** | **Quakes**  
1916 | 2  
1917 | 5  
1918 | 12  
1919 | 8  
1920 | 7  
1921 | 9

summary(data)

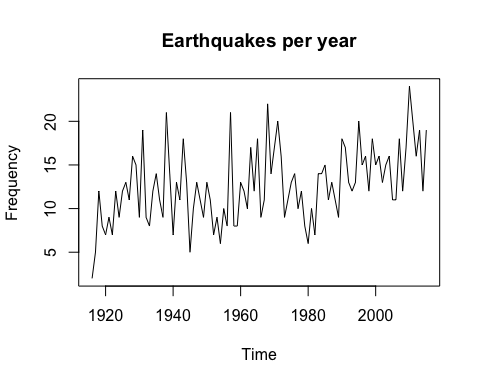
**Year** | **Quakes**   
Min. :1916 | Min. : 2.00   
1st Qu.:1941 | 1st Qu.: 9.00   
Median :1966 | Median :12.00   
Mean :1966 | Mean :12.61   
3rd Qu.:1990 | 3rd Qu.:15.25   
Max. :2015 | Max. :24.00

# CONVERTING EARTHQUAKE FREQUENCY TO TIME SERIES OBJECT

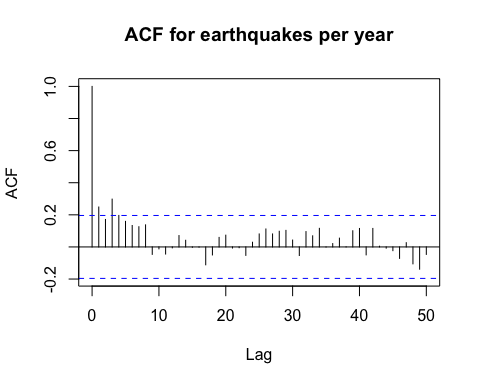
z = ts(data$Quakes, start=c(1916, 1, 1), end=c(2015, 1, 1), frequency = 1)

# BASIC ANALYSIS

ts.plot(z,  
 main="Earthquakes per year",  
 ylab="Frequency")



acf(z, 50, main="ACF for earthquakes per year")



We see strong positive autocorrelation for lags 1 and 3. We see no clear evidence of periodic fluctuations, including seasonality.

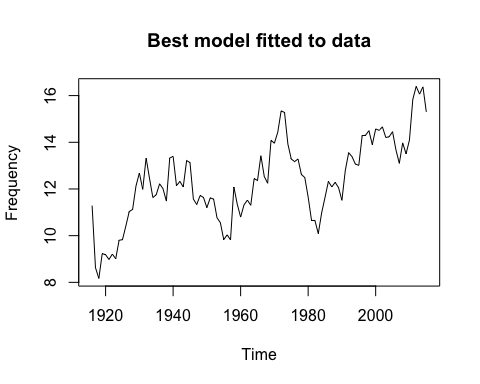
# FINDING THE BEST ARIMA MODEL

Model shows no clear trend or periodicity, hence we give 'seasonal=FALSE'. Since this assignment requires fitting for an ARMA model, we put the order of differencing as 0, using 'max.d=0'. The ARMA model finally selected is the one that gives the least residual sum of squares (residual => difference between model's estimate and actual value).

library(forecast)  
model = auto.arima(z, seasonal = FALSE)  
summary(model)

Series: z   
**ARIMA(1,0,1) with non-zero mean**   
***Coefficients:***  
 **| ar1 | ma1 | mean**  
 |0.9340 | -0.7472 | 12.4831  
**s.e.** |0.0708 | 0.1084 | 1.3524  
  
**sigma^2** = 15.63: **log likelihood** = -278.09  
**AIC**=564.17, **AICc**=564.59, **BIC**=574.59  
  
***Training set error measures:***  
**ME** |**RMSE** |**MAE** |**MPE** |**MAPE** |**MASE** |**ACF1**  
0.2522322 |3.89431 |3.023926|-11.18801|30.09992 |0.7319527|-0.04483419

plot(model$fitted,  
 main="Best model fitted to data",  
 ylab="Frequency")



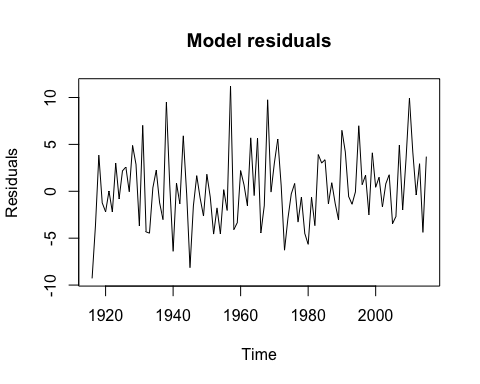
The best model turns out to be an ARMA(1, 0, 1) model, where there is no differencing, and autoregression and moving average are of order 1. We see that while the estimated observations plot does not closely resemble the actual observations plot, this is because of smoothening, which follows the average values and disregards much of the irregular fluctuations. Hence, this model may be more accurate overall, since it has clearly not been overfitted to irregular fluctuations.

# RESIDUAL ANALYSIS FOR THE BEST ARIMA MODEL

residuals = model$residuals

## Basic analysis

plot(residuals,  
 main="Best model residuals",  
 ylab="Residuals")

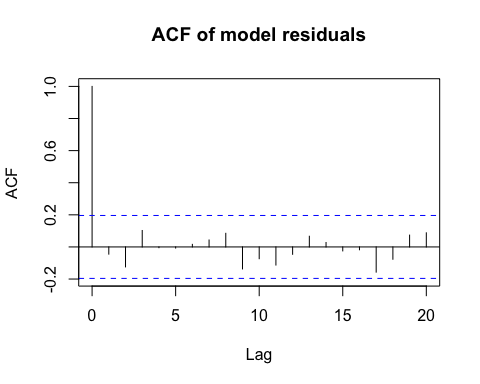


We see that the residuals are centered around the mean and have a limited range.

## TESTING ASSUMPTION 1: Residuals are uncorrelated

### Using ACF plot…

acf(residuals, main="ACF of best model residuals")



From this ACF plot, we see that the residuals are a stationary time series, with insignificant correlation betwee residuals at any lag, since all autocorrelation coefficients lie between the threshold values -0.2 and 0.2. Hence, we can conclude that the residuals are uncorrelated. Furthermore, given that they are centered at zero, we may conclude that the residuals form a white noise sequence.

acf(residuals, main="ACF of best model residuals")

### Using Portmanteau test (Box-Pierce test)…

***Hypotheses...***

H0: No significant autocorrelation  
H1: Significant autocorrelation for at least some lags

Box.test(residuals)

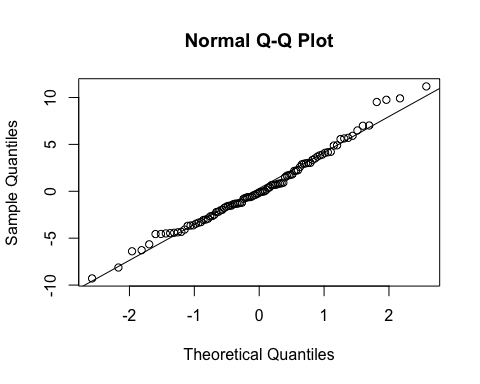
**Box-Pierce test**  
**data**: residuals  
**X-squared** = 0.20101, **df** = 1, **p-value** = 0.6539

p-value of the test statistic exceeds 0.05. Hence, for a 0.05 significance level, we may accept the null hypothesis and conclude that there is no significant autocorrelation in the residuals i.e. they are uncorrelated.

## TESTING ASSUMPTION 2: Residuals are normally distributed

### Using Q-Q plot (quantile-quantile plot) with respect to a normal distribution…

# Plotting sample quantiles against theoretical normal distribution quantiles...  
qqnorm(residuals)  
# Plotting the line formed when sample and theoretical normal distribution quantiles would be equal...  
qqline(residuals)



From the above, we see that many of the points mostly fall on the line. This indicates that the residuals can be considered to be drawn representatively from a normal distribution. Hence, we may conclude that the residuals are normally distributed.

### Using Shapiro-Wilk test…

***Hypotheses...***

H0: No difference between normal distribution and sample distribution  
H1: There is difference between normal distribution and sample distribution

shapiro.test(residuals)

**Shapiro-Wilk normality test**  
**data**: residuals  
**W** = 0.98321, **p-value** = 0.2343

p-value of the test statistic is above 0.05. Hence, for a 0.05 significance level, we may accept the null hypothesis and conclude that the residuals are normally distributed.

## CONCLUSIONS & INTERPRETATIONS

The residuals from the above model (as optimized by the **auto.arima** function) produces residuals that are uncorrelated and normally distributed. Hence, this indicates that the obtained model is reliable overall, since it has neither been overfitted not underfitted.