Model Fitting & Residual Analysis  
for Annual Stock Prices (US, 1871-1969)

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 first aim is to try fitting the autoregressive moving average model with autoregression and moving average both of the 1st order i.e. ARMA(1, 1) for the given dataset.

The next aim is to obtain the best autoregressive integrated moving average model (ARIMA) for given dataset, using the **auto.arima** function (which compares many possible ARIMA models and selects the best one).

Having identified and fitted the best ARIMA 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 ARIMA 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 US stock prices (for some unnamed enterprise) from 1871 to 1969.

data = read.csv("~/Documents/Study/computerScience/programming/r/data/annualCommonStockPrice (US).csv")  
year = data$Year  
annualStockPrice = data$Annual.common.stock.price..US.  
data = data.frame(year, annualStockPrice)  
head(data)

**year** | **annualStockPrice**  
1871 | 5.03  
1872 | 4.80  
1873 | 4.57  
1874 | 4.45  
1875 | 4.06  
1876 | 3.14

summary(data)

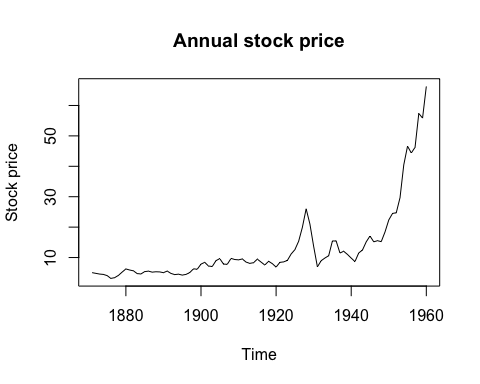
**year** | **annualStockPrice**  
Min. :1871 | Min. : 3.140   
1st Qu.:1896 | 1st Qu.: 5.765   
Median :1920 | Median : 9.050   
Mean :1920 | Mean :19.435   
3rd Qu.:1944 | 3rd Qu.:16.305   
Max. :1969 | Max. :98.700

# CONVERTING ANNUAL STOCK PRICE TO TIME SERIES OBJECT

z = ts(annualStockPrice, start=c(1871, 1, 1), end=c(1960, 1, 1), frequency = 1)

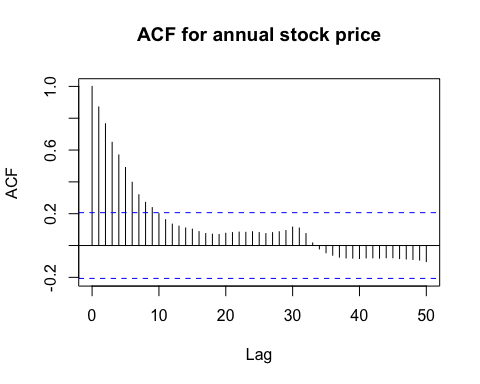
# BASIC ANALYSIS

ts.plot(z,  
 main="Annual stock price",  
 ylab="Stock price")



We observe some inconsistent but overall positive trend, and no clear periodicity.

acf(z, 50, main="ACF for annual stock price")



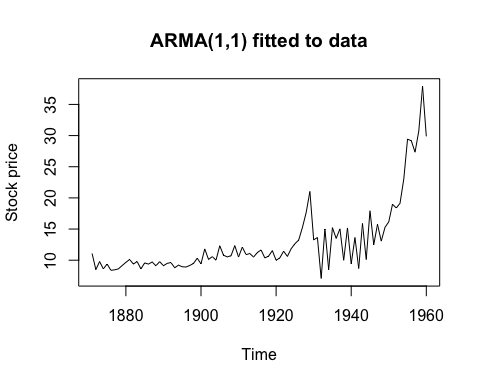
We see strong positive autocorrelation up to 9 lags. Hence, we can infer some degree of autoregression. Furthermore, we see no clear evidence of periodic fluctuations.

# TRYING TO FIT ARMA(1, 1)

library(forecast)  
model = auto.arima(z,  
 max.d = 0,  
 start.p = 1, max.p = 1,  
 start.q = 1, max.q = 1,  
 seasonal = FALSE)  
summary(model)

Series: z   
**ARIMA(0,0,1) with non-zero mean**   
***Coefficients:***  
 |**ma1** |**mean**  
 |0.9245|13.2701  
 **s.e.**|0.0323| 1.4738  
  
**sigma^2** = 54.55: **log likelihood** = -307.62  
**AIC**=621.24, **AICc**=621.52, **BIC**=628.74  
  
***Training set error measures:***  
**ME** |**RMSE** |**MAE** |**MPE** |**MAPE** |**MASE** |**ACF1**  
0.02992014|7.303281|4.824229|-34.00636|48.57019|2.758651|0.7185878

plot(model$fitted, main="ARMA(1,1) fitted to data", ylab="Stock price")



Estimated parameters are  
**φ** =0  
**θ** =0.9245

# FINDING THE BEST ARIMA MODEL

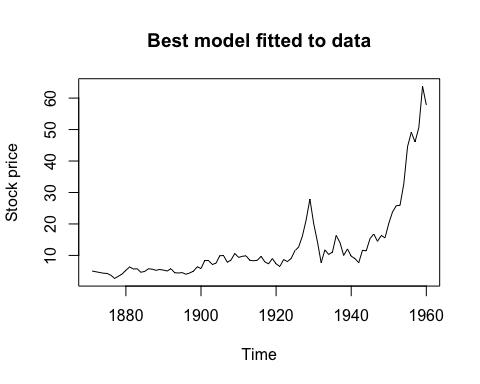
Model shows trend but no periodicity, hence we give ‘seasonal=FALSE’. To see the various models that have been fitted by the function to our time series we put ‘trace=TRUE’. The model finally selected is the one that gives the least residual sum of squares (residual => difference between model’s estimate and actual value).

bestModel = auto.arima(z, seasonal = FALSE)  
summary(bestModel)

Series: z   
**ARIMA(1,2,3)**   
***Coefficients:***  
 |**ar1** |**ma1** |**ma2** |**ma3**  
 |0.6488|-1.3707|0.1305|0.3335  
**s.e.**|0.1891| 0.2042|0.2783|0.1264  
  
**sigma^2** = 7.172: **log likelihood** = -210.42  
**AIC**=430.85, **AICc**=431.58, **BIC**=443.23

***Training set error measures:***  
**ME** |**RMSE** |**MAE** |**MPE** |**MAPE** |**MASE** |**ACF1**  
0.2905446|2.587317|1.610867|0.4426576|12.27738|0.9211462|-0.01616025

plot(bestModel$fitted,  
 main="Best model fitted to data",  
 ylab="Stock price")



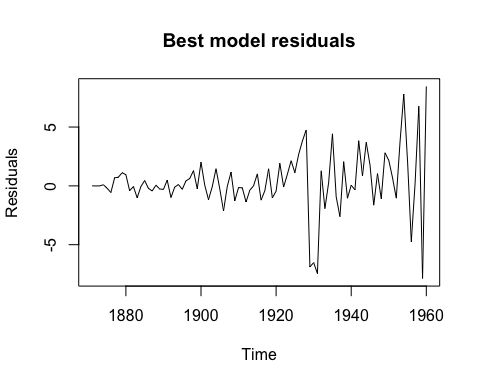
As we can see, while the ARMA(1, 1) model roughly fits the data, the best model derived here, which is an ARIMA(1, 2, 3) model, fits the data much more closely

# RESIDUAL ANALYSIS FOR THE BEST ARIMA MODEL

residuals = bestModel$residuals

## Basic analysis

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

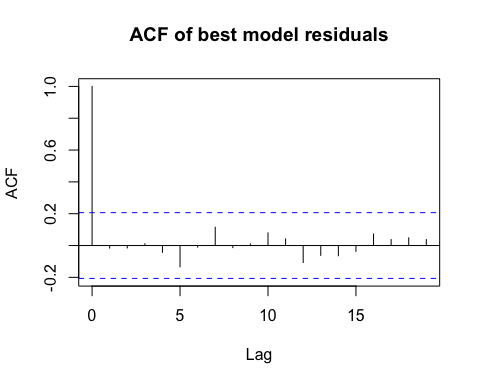


We see that the residuals are centered around the mean and have a limited range.

## 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.

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

***Hypotheses...***

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

Box.test(residuals)

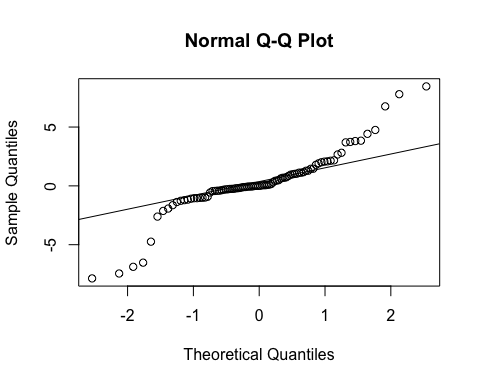
**Box-Pierce test**  
**data**: residuals  
**X-squared** = 0.023504, **df** = 1, **p-value** = 0.8782

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 do not fall on the line. This indicates that the residuals may not have been drawn representatively from a normal distribution. Hence, we may conclude that the residuals are not 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** = 0.87912, **p-value** = 5.28e-07

p-value of the test statistic is below 0.05. Hence, for a 0.05 significance level, we may reject the null hypothesis and conclude that the residuals are not normally distributed.

## CONCLUSIONS & INTERPRETATIONS

The residuals from the best model (as decided by the **auto.arima** function) produces residuals that are uncorrelated but not normally distributed. Hence, this indicates that the obtained model may not be ideal for the given time series, possibly due to overfitting, since we saw that the plot of the estimates from the model closely matched the plot of the actual observations.