Estimation of Stationary Time Series

Now that we have learned the (population) model properties of the stationary time series models, our next goal is to learn how to fit such models. That is, how to determine the model type (AR, MA, ARMA), the model order (the p & q values in AR(p), MA(q) and ARMA(p,q)), and to estimate the model parameters.

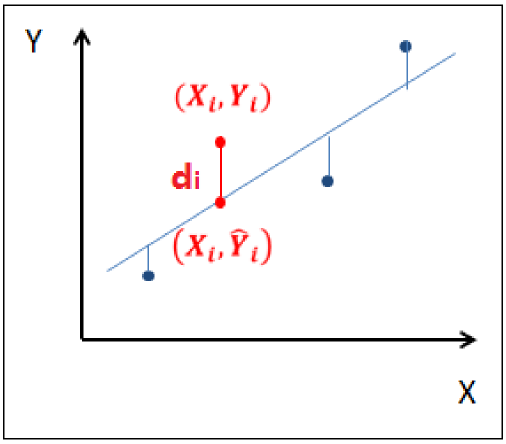
The three typical estimation methods are: (1) the method of moment estimator (MOME), (2) the least squares estimator (LSE), and (3) the maximum likelihood estimator (MLE). In the following, we will first provide a general review of these three estimation methods using non-time series models.

0. Review of Three Estimation Methods.

The Least Squares Estimators

Approach: To minimize the sum of the squared vertical distances (or the squared deviations/errors)
Example: Simple Linear Regression

The aim of simple linear regression is to find the linear relationship between two variables. This is in turn translated into a mathematical problem of finding the equation of the line that is closest to all points observed. Consider the **scatter plot** below. The **vertical distance** each point is above or below the line has been added to the diagram. These distances are called *deviations* or *errors* – they are symbolised as $d_i = |y_i - \hat{y}_i|, i = 1, \cdots n$.



The **least-squares regression line** will minimize the **sum of the squared vertical distance from every point to the line**, i.e. we minimise $\sum d_i^2$.

** The statistical equation of the simple linear regression line, when only the response variable Y is random, is: $Y = \beta_0 + \beta_1 x + \varepsilon$ (or in terms of each point: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$)

Here β_0 is called the intercept, β_1 the regression slope, ε is the random error with mean 0, x is the regressor (independent variable), and Y the response variable (dependent variable).

** The least squares regression line

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \cdots, n$$

is obtained by finding the values of β_0 and β_1 values (denoted in the solutions as $\hat{\beta}_0$ & $\hat{\beta}_1$) that will minimize the sum of the squared vertical distances from all points to the line: $\Delta = \sum \boldsymbol{d}_i^2 = \sum (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2 = \sum \left(\boldsymbol{y}_i - \hat{\beta}_0 - \hat{\beta}_1 \boldsymbol{x}_i\right)^2$

The solutions are found by solving the equations: $\frac{\partial \Delta}{\partial \hat{\beta}_0} = 0$ and $\frac{\partial \Delta}{\partial \hat{\beta}_1} = 0$

** As shown above, the equation of the fitted least squares regression line is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (or in terms of each point: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) ----- For simplicity of notations, many books denote the fitted regression equation as: $\hat{Y} = b_0 + b_1 x$

(* you can see that for some examples, we will use this simpler notation.)

where $\hat{\beta}_1 = \frac{\boldsymbol{S}_{xy}}{\boldsymbol{S}_{xx}}$ and $\hat{\beta}_0 = \overline{\boldsymbol{y}} - \hat{\beta}_1 \overline{\boldsymbol{x}}$.

Notations:

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = \sum (x_i - \bar{x})(y_i - \bar{y}) \quad ;$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = \sum (x_i - \bar{x})^2 \; ;$$

$\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the mean values of *x* and *y* respectively.

<u>Note 1</u>: Please notice that **in finding the least squares regression line, we do not need to assume any distribution for the random errors ε_i. However, for statistical inference on the model parameters (**β_0 and β_1), it is often assumed that the errors have the following three properties:

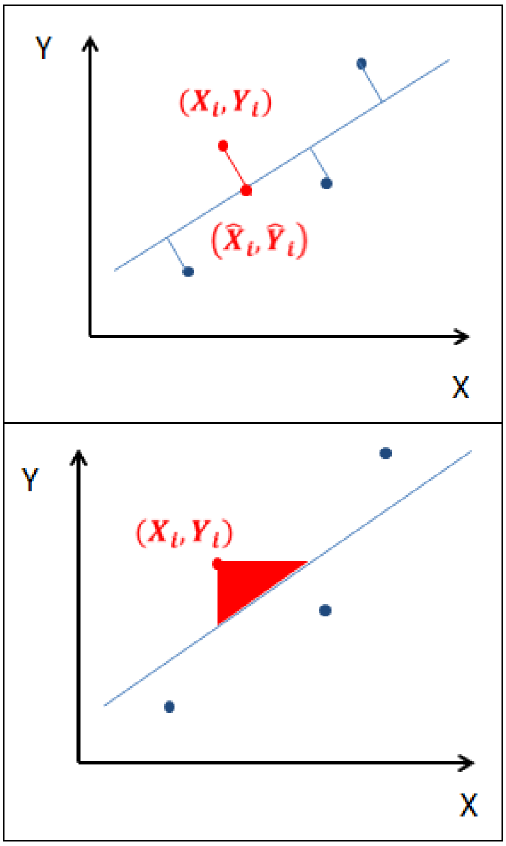
□ (1) Normally distributed errors

□ (2) Homoscedasticity (constant error variance $\text{var}(\varepsilon_i) = \sigma^2$ for Y at all levels of X)

□ (3) Independent errors (usually checked when data collected over time or space)

***The above three properties can be summarized as: $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \; i = 1, \cdots, n$

<u>Note 2</u>: Please notice that the least squares regression is only suitable when the random errors exist in the dependent variable Y only. If the regression X is also random – it is then referred to as the **Errors in Variable (EIV) regression**. One can find a good summary of the EIV regression in section 12.2 of the book: "Statistical Inference" (2nd edition) by George Casella and Roger Berger. In the figures below, we illustrate two commonly used EIV regression lines, the orthogonal regression line (*obtained by minimizing the sum of the squared orthogonal distances) and the geometric mean regression line (*obtained by minimizing the sum of the straight triangular areas).



Note: there are infinite many possible errors in regression lines – they all pass through the point $(\bar{x}, \bar{y})$ and bounded by the two ordinary least squares regression lines of X regress on Y, and Y regression on X.

Finance Application: Market Model

- One of the most important applications of linear regression is the *market model.*
- It is assumed that rate of return on a stock (R) is linearly related to the rate of return on the overall market.

  $$\mathbf{R = \beta_0 + \beta_1 R_m + \varepsilon}$$

  R: Rate of return on a particular stock

  R_m: Rate of return on some major stock index

  β_1: The beta coefficient measures how sensitive the stock's rate of return is to changes in the level of the overall market.

The Maximum Likelihood Estimators (MLE)

Approach: To estimate model parameters by maximizing the likelihood

By maximizing the likelihood, which is the joint probability density function of a random sample, the resulting point estimators found can be regarded as yielded by the most likely underlying distribution from which we have drawn the given sample.

Example 1. $X_i \sim N(\mu, \sigma^2), \text{i.i.d.}, i = 1,2,\dots,n$; Derive the MLE for μ and σ^2.

Solution.

[i]

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right],$$

$$x_i \in R, i = 1, 2, \dots, n$$

[ii] likelihood function= $L = f(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_i)$

$$= \prod_{i=1}^{n} \left\{ (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right]$$

[iii] log likelihood function

$$l = \ln L = \left(-\frac{n}{2}\right)\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

[iv]

$$\begin{cases} \dfrac{dl}{d\mu} = \dfrac{2\sum_{i=1}^{n}(x_i-\mu)}{2\sigma^2} = 0 \\ \dfrac{dl}{d\sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \bar{X} \\ \widehat{\sigma^2} = \dfrac{\sum(X_i-\bar{X})^2}{n} \end{cases}$$

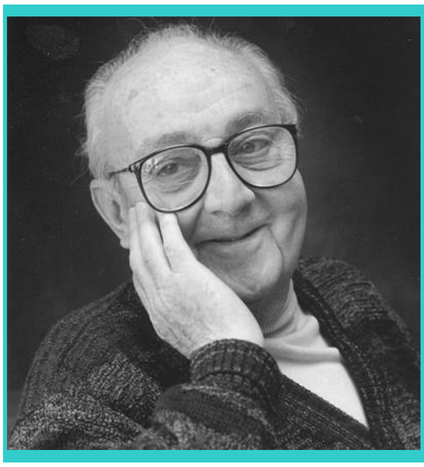
R. A. Fisher (http://en.wikipedia.org/wiki/Ronald_Fisher)

Plate 10. R. A. Fisher, 1938, with sons George (aged 18) and Harry (14).

R A Fisher, with his sons George (18) and Harry (14), 1938

Plate 11. Mrs. Fisher 1938, with daughters, in order of age, Margaret (top right), Joan (bottom right), Phyllis (top left), Elizabeth (bottom left), Rose standing beside her chair, and June in her lap.

Mrs Fisher in 1938, with daughters, left to right, Elizabeth, Phyllis, Rose, June, Margaret, Joan



George Box, FRS
Born 18 October 1919
Gravesend, Kent, England
Died 28 March 2013 (aged 93)
Madison, Wisconsin

Doctoral advisor
Egon Pearson
H. O. Hartley

Box married **Joan Fisher**, the second of Ronald Fisher's five daughters.

In time series analysis, the **Box–Jenkins method**, named after the statisticians **George Box** and **Gwilym Jenkins**, applies autoregressive moving average (ARMA) or autoregressive integrated moving average (ARIMA) models to find the best fit of a time-series model to past values of a time series.

The Method of Moment Estimators (MOME)

Approach: To estimate model parameters by equating the population moments to the sample moments

Order	Population Moment	Sample Moment
1st	$E(X)$	$= \dfrac{X_1 + X_2 + \cdots + X_n}{n}$
2nd	$E(X^2)$	$= \dfrac{{X_1}^2 + {X_2}^2 + \cdots + {X_n}^2}{n}$
$\vdots$	$\vdots$	$\vdots$
kth	$E(X^k)$	$= \dfrac{{X_1}^k + {X_2}^k + \cdots + {X_n}^k}{n}$
$\vdots$	$\vdots$	$\vdots$

Example 1 (continued). $X_i \sim N(\mu, \sigma^2), \text{i.i.d.}, i = 1,2,\ldots,n$; Derive the MOME for μ and σ^2.

Solution.

$$E(X) = \mu = \bar{X}$$

$$E(X^2) = \mu^2 + \sigma^2 = \frac{\sum_{i=1}^n {X_i}^2}{n}$$

$$\Rightarrow \begin{cases} \hat{\mu} = \bar{X} \\ \widehat{\sigma^2} = \dfrac{\sum_{i=1}^n {X_i}^2}{n} - (\bar{X})^2 \end{cases}$$

$$\frac{\sum_{i=1}^n {X_i}^2}{n} - (\bar{X})^2 = \frac{\sum_{i=1}^n (X_i - \bar{X} + \bar{X})^2}{n} - (\bar{X})^2$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n 2\bar{X}(X_i - \bar{X}) + \sum_{i=1}^n (\bar{X})^2}{n} - (\bar{X})^2$$

$$= \frac{\sum (X_i - \bar{X})^2}{n}$$

Therefore, the MLE and MOME for σ^2 are the same for the normal population.

$$E(\widehat{\sigma^2}) = E\left[\frac{\sum (X_i - \bar{X})^2}{n}\right] = E\left[\frac{n-1}{n}\frac{\sum (X_i - \bar{X})^2}{n-1}\right] = \frac{n-1}{n}E(S^2) =$$

$$\frac{n-1}{n}\sigma^2 \overset{n\to\infty}{\Longrightarrow} \sigma^2$$

(asymptotically unbiased)

Example 2. Let $X_i \sim Bernoulli(p)$, i. i. d. $i = 1, \dots, n.$

Please derive 1. The MLE of p 2. The MOME of p.

Solution.

1. MLE

[i] $f(x_i) = p^{x_i}(1-p)^{1-x_i}, i = 1, \dots, n$

[ii] $L = \prod f(x_i) = p^{\sum x_i}(1-p)^{\,n-\sum x_i}$

[iii] $l = \ln L = (\sum x_i) \ln p + (n - \sum x_i) \ln(1-p)$

[iv] $\frac{dl}{dp} = \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = 0 \;\Rightarrow \hat{p} = \frac{\sum x_i}{n}$

2. MOME

$$E(X) = p = \frac{X_1 + X_2 + \cdots + X_n}{n} \Rightarrow \hat{p} = \frac{\sum x_i}{n}$$

Example 3. Let $X_1, X_2, \dots, X_n$ be a random sample from exp(λ)

Please derive 1. The MLE of λ 2. The MOME of λ.

Solution:

1. MLE:

$$f(x) = \lambda \exp(-\lambda x)$$

$$L = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i)$$

$$l = lnL = nln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Thus $\hat{\lambda} = \frac{1}{\bar{X}}$

2. MOME:

$$E(X) = \frac{1}{\lambda}$$

Thus setting:

$E(X) = \frac{1}{\lambda} = \bar{X},$ we have: $\hat{\lambda} = \frac{1}{\bar{X}}$

Example 4. $Y_1,\cdots,Y_n \overset{i.i.d.}{\sim} U[0,\theta]$

(a) Find the MOME for θ

(b) Find the MLE for θ

(c) Are the MOME and MLE unbiased estimators of θ?

Solution:

(a) $f(y)=\frac{1}{\theta},\ 0\le y\le\theta$

$$E(Y)=\int_0^\theta y\cdot\frac{1}{\theta}dy=[\frac{y^2}{2\theta}]_0^\theta=\frac{1}{2\theta}[\theta^2-0]=\frac{\theta}{2}$$

$$E(Y)=\bar{Y}$$

$$\frac{\theta}{2}=\bar{Y} => \hat{\theta}_1=2\bar{Y}$$

(b) $L=\prod_{i=1}^n f(y_i)=(\frac{1}{\theta})^n,\ 0\le y_1,\cdots,y_n\le\theta$

$$\mathrm{l\,n}L=-n\ \mathrm{l\,n}\theta,\ 0\le y_1,\cdots,y_n\le\theta$$

$\frac{d\ln L}{d\theta}=\frac{-n}{\theta}=0 => \hat{\theta}=\pm\infty$? This is not going to lead to any good answers.

Alternatively, we look at the domain, $0\le y_1,\cdots,y_n\le\theta\ => \ 0\le y_{(1)},\cdots,y_{(n)}\le\theta$

$\because\theta\ge Y_{(n)}$, $\therefore$ L is maximized when $\theta=Y_{(n)}$

$\therefore$ The MLE for θ is $\hat{\theta}_2=Y_{(n)}$

(c) It is straight-forward to show that

$$E[\hat{\theta}_1]=2E[\bar{Y}]=2E[Y]=\theta$$

Thus the MOME is an unbiased estimator for θ

Now we derive the general formula for the pdf of the last order statistic as follows:

$$P\left(Y_{(n)} \leq y\right) = P(Y_1 \leq y, \ldots, Y_n \leq y) = \prod_{i=1}^{n} P(Y_i \leq y)$$

Therefore we have

$$F_{Y_{(n)}}(y) = \prod_{i=1}^{n} F(y) = [F(y)]^n$$

Differentiating with respect to y at both sides leads to:

$$f_{Y_{(n)}}(y) = nf(y)[F(y)]^{n-1}$$

Now we can derive the pdf of the last order statistic for a random sample from the given exponential family. First, the population cdf is:

$$F(y) = \int_0^y \frac{1}{\theta} du = \left[\frac{u}{\theta}\right]_0^y = \frac{y}{\theta}$$

Now plugging in the population pdf and cdf we have:

$$f_{Y_{(n)}}(y) = n\frac{1}{\theta}\left[\frac{y}{\theta}\right]^{n-1} = \frac{ny^{n-1}}{\theta^n},$$

when $0 \leq y \leq \theta$. Now we are ready to compute the mean of the last order statistic:

$$E\left[Y_{(n)}\right] = \int_0^\theta y\frac{ny^{n-1}}{\theta^n} dy = \left[\frac{ny^{n+1}}{(n+1)\theta^n}\right]_0^\theta = \frac{n}{(n+1)}\theta$$

Now we have shown that the MLE, $\hat{\theta}_2 = Y_{(n)}$ is NOT an unbiased estimator for θ.

Karl Pearson (http://en.wikipedia.org/wiki/Karl_pearson)

Left to Right:

Egon Sharpe Pearson
Karl Pearson
Maria Pearson (née Sharpe)
Sigrid Letitia Sharpe Pearson

Egon Sharpe Pearson, CBE FRS (11 August 1895 – 12 June 1980) was one of three children (Sigrid, Helga, and Egon) and the son of Karl Pearson and, like his father, a leading British statistician.

Now we are ready for the estimation of stationary time series models.

In the following, we will first introduce some relevant sample statistics such as the sample autocorrelation, and population parameters such as the partial autocorrelation. Next we will discuss the estimation methods in detail by the type of stationary models.

1. **Sample autocorrelation (Sample ACF).**

For a stationary time series $\{X_t\}$, what we will have in reality is a realization of this series: $x_1, x_2, \cdots, x_n$. Recall that the usual sample correlation (the Pearson product moment correlation) for a sample of paired data is defined as follows.

Data:

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{array}$$

Sample correlation:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

For a time series data, in order to compute the first order sample autocorrelation, we pair the data as follows:

Data:

$$\begin{array}{cc} x_1 & x_2 \\ x_2 & x_3 \\ \vdots & \vdots \\ x_{n-1} & x_n \end{array}$$

The *first* order sample autocorrelation:

$$\hat{\rho}(1) = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

In general, to compute the sample autocorrelation of order h, we pair the data as follows:

Data:

$$\begin{array}{cc} x_1 & x_{1+h} \\ x_2 & x_{2+h} \\ \vdots & \vdots \\ x_{n-h} & x_n \end{array}$$

The h^{th} order sample autocorrelation:

$$\widehat{\boldsymbol{\rho}}(\mathbf{h}) = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n-h}}(x_i - \bar{x})(x_{i+h} - \bar{x})}{\sum_{\mathbf{i=1}}^{\mathbf{n}}(x_i - \bar{x})^{\mathbf{2}}}$$

2. Population partial autocorrelation (Population PACF).

The partial correlation $\rho_{XY\cdot\mathbf{Z}}$ measures the (linear) association between two random variables $(X \; and \; Y)$, with the (linear) effect of a set of controlling variables $\mathbf{Z} = \{Z_1, Z_2, \cdots, Z_k\}$ removed. Indeed $\rho_{XY\cdot\mathbf{Z}}$ is the correlation between the residuals R_X and R_Y resulting from the linear regression of X with **Z** and Y with **Z**, respectively.

Alternatively one can compute the partial correlation recursively as follows.

The zeroth order partial correlation $\rho_{XY\cdot\emptyset} = \rho_{XY}$, the usual population correlation between X and Y.

The partial correlation controlling for a single variable (univariate) Z is:

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\,\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}}$$

Given that we already have the partial correlations controlling for a set of variables (multivariate) **Z**, and that W (univariate) is a new variable that we also wish to control for in addition to **Z**, the new partial correlation is:

$$\rho_{XY\cdot\mathbf{Z}\&W} = \frac{\rho_{XY\cdot\mathbf{Z}} - \rho_{XW\cdot Z}\,\rho_{WY\cdot Z}}{\sqrt{1-\rho_{XW\cdot\mathbf{Z}}^2}\sqrt{1-\rho_{WY\cdot Z}^2}}$$

In time series analysis, the partial autocorrelation function of lag h, is defined as:

$$\pi(h) = \rho_{X_t X_{t+h}\cdot\{X_{t+1},\cdots,X_{t+h-1}\}}$$

For example:

$$\pi(0) = \rho_{X_tX_t} = 1$$

$$\pi(1) = \rho_{X_tX_{t+1}} = \rho(1), \text{ the first autocorrelation}$$

$$\pi(2) = \rho_{X_tX_{t+2}\cdot\{X_{t+1}\}} = \frac{\rho_{X_tX_{t+2}} - \rho_{X_tX_{t+1}}\rho_{X_{t+1}X_{t+2}}}{\sqrt{1-\rho^2_{X_tX_{t+1}}}\sqrt{1-\rho^2_{X_{t+1}X_{t+2}}}}$$

$$= \frac{\rho(2) - \rho(1)\rho(1)}{\sqrt{1-\rho^2(1)}\sqrt{1-\rho^2(1)}} = \frac{\rho(2) - \rho^2(1)}{1-\rho^2(1)}$$

Estimation of the Moving Average Process (MA)

Approach: LSE, MLE, MOME

3. Moving Average Model

(1) *Order determination.*

The moving average process (MA) can be identified by its autocorrelation function as summarized in Table 1 below:

Type of Model	Typical Pattern of ACF	Typical Pattern of PACF
AR (*p*)	Decays exponentially or with damped sine wave pattern or both	**Cut-off after lags *p***
MA (*q*)	**Cut-off after lags *q***	Declines exponentially
ARMA (*p,q*)	Exponential decay	Exponential decay

Table 1. Theoretical Patterns of ACF and PACF for Stationary Time Series.

The MA(q) model is characterized by its autocorrelation function as follows:

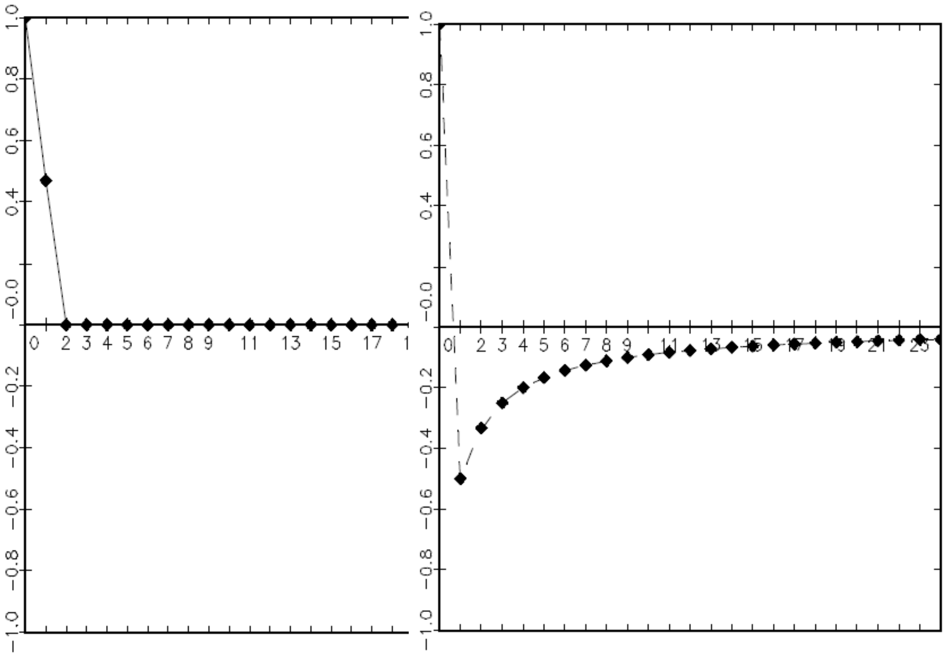
$$\rho(h) \neq 0, \text{for } h \leq q; \text{while } \rho(h) = 0, \text{for } h > q.$$

For a realized stationary time series $\{x_t\}, t = 1,2,\cdots,n$, we can determine whether the (population) autocorrelation of order h is zero or not by looking at its approximate 95% confidence interval based on the corresponding sample autocorrelation of order h as follows:

$$\left(\hat{\boldsymbol{\rho}}(\mathbf{h}) - \frac{\mathbf{2}}{\sqrt{\mathbf{n}}}, \hat{\boldsymbol{\rho}}(\mathbf{h}) + \frac{\mathbf{2}}{\sqrt{\mathbf{n}}}\right)$$

If this interval does not cover 0, then we claim that we are 95% sure that

$$\rho(\mathrm{h}) \neq 0.$$



Population ACF of an MA(1) series

Population PACF of an MA(1) series

Figure 1. Left: Population (theoretical) autocorrelation function (ACF) of an MA(1) series.

Right: Population (theoretical) partial autocorrelation (PACF) of an MA(1) series.

(2) *Parameter Estimation.*

We will discuss three estimation methods starting with the simplest (but not necessarily the best) for the MA models – the method of moment estimator (MOME).

(a) ***MOME*****.**

Example: MA(1).

Suppose we found that the first population autocorrelation is non-zero, however, other auto-correlations of order higher than 1 are zero. Now we fit the underlying MA(1) of the form:

$$X_t = \mu + Z_t + \beta_1 Z_{t-1}$$

where $\{Z_t\}$ is a series of white noise, using the MOME method by equating the sample mean to the population mean, and the first sample autocorrelation to the first population autocorrelation as follows:

$$\mu = \bar{x}$$

$$\rho(1) = \frac{\beta_1}{1+\beta_1^2} = \widehat{\boldsymbol{\rho}}(1)$$

The resulting method of moment estimators are:

$$\hat{\mu} = \bar{x}$$

$\widehat{\beta}_1$ is the root satisfying $|\beta_1| < 1$ (invertible) of the quadratic function:

$$\widehat{\boldsymbol{\rho}}(1)\beta_1^2 - \beta_1 + \widehat{\boldsymbol{\rho}}(1) = 0$$

(b) ***LSE*** **– an iterative procedure:**

One iterative procedure is as follows using the MA(1) as an example:

(i) Select suitable starting values for $\boldsymbol{\mu}$ **and** $\boldsymbol{\beta_1}$ – one easy choice would be the crude MOME presented above;

(ii) Calculate the residual recursively using the relation

$$\boldsymbol{Z_t = X_t - \mu - \beta_1 Z_{t-1}}$$

as follows:

Taking $\boldsymbol{Z_0 = 0}$, we compute $\boldsymbol{\widehat{Z}_1 = X_1 - \widehat{\mu}}$, and then

$$\boldsymbol{\widehat{Z}_2 = X_2 - \widehat{\mu} - \widehat{\beta}_1 \widehat{Z}_1}$$

until

$$\boldsymbol{\widehat{Z}_n = X_n - \widehat{\mu} - \widehat{\beta}_1 \widehat{Z}_{n-1}}$$

Then we compute the residual sum of squares:

$$\boldsymbol{\sum_{i=1}^{n} \widehat{Z}_i^2}$$

(iii) Repeat the above procedures for neighboring values of $\boldsymbol{\mu}$ **and** $\boldsymbol{\beta_1}$ – until we find the values that will minimize the residual sum of squares. These are the LSE's and when the white noise is Gaussian, the MLE's.

Similar recursive procedures can be designed for higher order MA(q) process using the equation:

$$\boldsymbol{Z_t = X_t - \mu - \beta_1 Z_{t-1} - \cdots - \beta_q Z_{t-q}}$$

And suitable starting values.

(c) Maximum Likelihood Estimation (MLE) of MA models

For iid data (sample) with marginal pdf $f(x_t;\boldsymbol{\theta})$, the joint pdf for a sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ is
$f(\boldsymbol{x};\boldsymbol{\theta}) = f(x_1, \cdots, x_n;\boldsymbol{\theta}) = \prod_{t=1}^{n} f(x_t;\boldsymbol{\theta})$
The likelihood function of the sample is: $L(\boldsymbol{\theta};\boldsymbol{x}) = f(x_1, \cdots, x_n;\boldsymbol{\theta}) = \prod_{t=1}^{n} f(x_t;\boldsymbol{\theta})$
The loglikelihood function of the sample is: $l(\boldsymbol{\theta};\boldsymbol{x}) = lnL(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{t=1}^{n} ln\, f(x_t;\boldsymbol{\theta})$

For a sample from a (weak) stationary time series $\{x_t\}$, the random variables in the sample $\{x_1, \cdots, x_n\}$ are not iid – however we can construct the likelihood using the *Multivariate Normal probability density function directly because the likelihood is simply the joint distribution of the entire sample.*

Multivariate normal distribution

Let $\underline{X} = (X_1, \ldots, X_n)'$, the general formula for the

n-dimensional multivariate normal density function is

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\underline{x} - \underline{\mu}\right)\right]$$

where $E\left(\underline{X}\right) = \underline{\mu}$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\underline{X}$.

Given that we have i.i.d. Gaussian white noise, we can write down the exact likelihood of our time series $\{X_1, \ldots, X_n\}$ as a multivariate normal density function directly as follows:

$$L(\boldsymbol{x};\boldsymbol{\theta}) = f(x_1, \cdots, x_n;\boldsymbol{\theta})$$

Example: MLE for an invertible MA(1)

$$\mathrm{X_t = \mu + Z_t + \beta_1 Z_{t-1}}$$

where $\{Z_t\}$ is a series of **Gaussian** white noise (that is, i.i.d. $N(0, \sigma^2), t = 1, \cdots, n$),
$\boldsymbol{\theta} = (\mu, \beta_1, \sigma^2)', |\beta_1| < 1.$

$$E\left(\underline{X}\right) = \underline{\mu} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \mu\underline{1}$$

$$Var\left(\underline{X}\right) = \Sigma = \sigma^2(1+\beta_1^2)\begin{bmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ 0 & \rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

where $\rho = \beta_1/[\sigma^2(1+\beta_1^2)]$.

Estimation of the Autoregressive Process (AR)

Approach: LSE, MLE, MOME

4. Autoregressive Model

(1) Order determination.

While the moving average process (MA) can be identified by its autocorrelation function, the AR process can be identified through its partial autocorrelation function (PACF) as summarized in Table 1 below:

Type of Model	Typical Pattern of ACF	Typical Pattern of PACF
AR (*p*)	Decays exponentially or with damped sine wave pattern or both	**Cut-off after lags *p***
MA (*q*)	**Cut-off after lags *q***	Declines exponentially
ARMA (*p,q*)	Exponential decay	Exponential decay

Table 1. Theoretical Patterns of ACF and PACF for Stationary Time Series.

That is, the AR(p) model is characterized by its partial autocorrelation function as follows:

$$\pi(\mathrm{h}) \neq 0, \text{for h} \leq \mathrm{p}; \text{while } \pi(\mathrm{h}) = 0, \text{for h} > p.$$

The lag h partial autocorrelation is the last regression coefficient φ_{hh} in the h^th^ order autoregression AR(h):

$$X_t = \varphi_{h1}X_{t-1} + \varphi_{h2}X_{t-2} + \cdots + \varphi_{hh}X_{t-h} + Z_t$$

That is,

$$\pi(h) = \varphi_{hh}$$

We now prove this for the lag 1 and lag 2 partial autocorrelations.

Given the following AR(1) model:

$$X_t = \varphi_{11}X_{t-1} + Z_t$$

$$\mathrm{COV}(X_t, X_{t+1}) = \mathrm{COV}(X_t, \varphi_{11}X_t + Z_{t+1}) = \varphi_{11}\mathrm{Var}(X_t)$$

$$\gamma(1) = \varphi_{11}\sigma_X^2$$

$$\frac{\gamma(1)}{\sigma_X^2} = \frac{\varphi_{11}\sigma_X^2}{\sigma_X^2}$$

$$\varphi_{11} = \rho(1) = \pi(1)$$

Given the following AR(2) model:

$$X_t = \varphi_{21}X_{t-1} + \varphi_{22}X_{t-2} + Z_t$$

$$\begin{aligned} \mathrm{COV}(X_t, X_{t+1}) &= \mathrm{COV}(X_t, \varphi_{21}X_t + \varphi_{22}X_{t-1} + Z_{t+1}) \\ &= \varphi_{21}\mathrm{Var}(X_t) + \varphi_{22}\mathrm{COV}(X_t, X_{t-1}) \end{aligned}$$

$$\gamma(1) = \varphi_{21}\sigma_X^2 + \varphi_{22}\gamma(1)$$

$$\frac{\gamma(1)}{\sigma_X^2} = \frac{\varphi_{21}\sigma_X^2}{\sigma_X^2} + \frac{\varphi_{22}\gamma(1)}{\sigma_X^2}$$

$$\rho(1) = \varphi_{21} + \varphi_{22}\rho(1)$$

$$\begin{aligned} \mathrm{COV}(X_t, X_{t+2}) &= \mathrm{COV}(X_t, \varphi_{21}X_{t+1} + \varphi_{22}X_t + Z_{t+2}) \\ &= \varphi_{21}\mathrm{COV}(X_t, X_{t+1}) + \varphi_{22}\mathrm{Var}(X_t) \end{aligned}$$

$$\gamma(2) = \varphi_{21}\gamma(1) + \varphi_{22}\sigma_X^2$$

$$\frac{\gamma(2)}{\sigma_X^2} = \frac{\varphi_{21}\gamma(1)}{\sigma_X^2} + \frac{\varphi_{22}\sigma_X^2}{\sigma_X^2}$$

$$\rho(2) = \varphi_{21}\rho(1) + \varphi_{22}$$

Solving the two high-lighted equations above, we have:

$$\varphi_{22} = \frac{\rho(2) - \rho^2(1)}{1 - \rho^2(1)} = \pi(2)$$

For a realized stationary time series $\{x_t\}, t = 1,2,\cdots,n$, we can determine whether the (population) partial autocorrelation of order h is zero or not by looking at its approximate 95% confidence interval based on the corresponding sample partial autocorrelation of order h as follows:

$$\left(\hat{\boldsymbol{\pi}}(\mathbf{h}) - \frac{\mathbf{2}}{\sqrt{\mathbf{n}}}, \hat{\boldsymbol{\pi}}(\mathbf{h}) + \frac{\mathbf{2}}{\sqrt{\mathbf{n}}}\right)$$

Or

$$\left(\hat{\varphi}_{hh} - \frac{\mathbf{2}}{\sqrt{\mathbf{n}}}, \hat{\varphi}_{hh} + \frac{\mathbf{2}}{\sqrt{\mathbf{n}}}\right)$$

If this interval does not cover 0, then we claim that we are 95% sure that

$$\pi(h) = \varphi_{hh} \neq 0.$$

(2) *Parameter Estimation.*

We will discuss three estimation methods starting with the simplest (but not necessarily the best) for the AR models – the least squares estimator (LSE).

(b) ***<u>LSE</u>*.**

For a suspected AR process, we often fit the AR models of progressively higher order, and to check whether the lag h partial autocorrelation, which is last regression coefficient φ_{hh} in the hth order autoregression, that is $\pi(h) = \varphi_{hh}$, is zero.

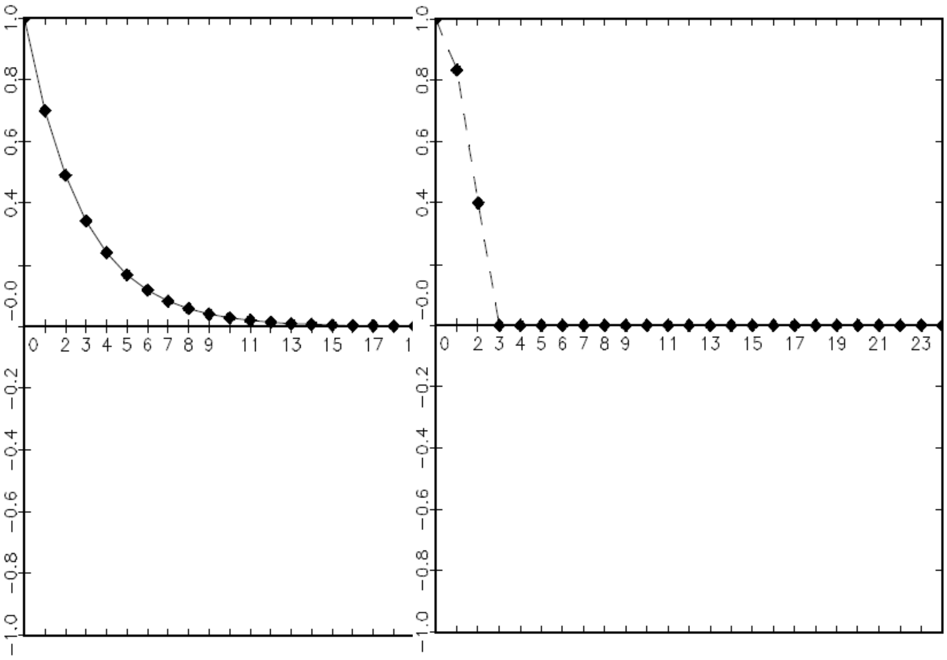
Example: AR(h).

For an AR(h) with possibly non-zero mean as follows:

$$X_t = \delta + \varphi_{h1}X_{t-1} + \varphi_{h2}X_{t-2} + \cdots + \varphi_{hh}X_{t-h} + Z_t$$

It is often fitted using the (ordinary) least squares regression method after we center the series using the sample mean:

$$X_t - \overline{X} = \varphi_{h1}(X_{t-1} - \overline{X}) + \varphi_{h2}(X_{t-2} - \overline{X}) + \cdots \\ + \varphi_{hh}(X_{t-h} - \overline{X}) + Z_t$$



Population ACF of an AR(1) series

Population PACF of an AR(2) series

Figure 2. Left: Population (theoretical) autocorrelation function (ACF) of an AR(1) series.

Right: Population (theoretical) partial autocorrelation (PACF) of an AR(2) series.

(a) MOME: The AR(p) model

$$X_t = \delta + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + Z_t$$

can also be estimated using the generalized method of moment approach. The (population) mean μ is often estimated by the sample mean as the following:

$$\hat{\mu} = \frac{\hat{\delta}}{1 - \hat{\alpha}_1 - \hat{\alpha}_2 - \cdots - \hat{\alpha}_p} = \bar{X}$$

The other AR(p) regression coefficients α_i 's can be estimated, for example, by plugging the sample ACF as estimates of the population ACF in the Yule-Walker equation below:

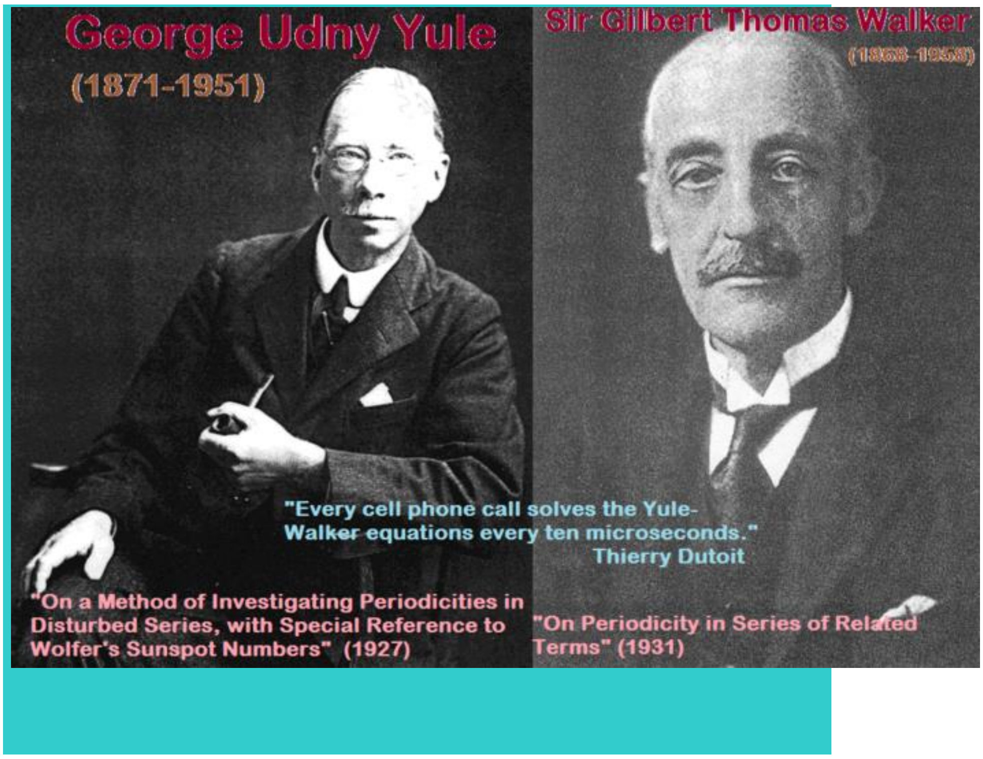
$$\rho(1) = \alpha_1 1 + \alpha_2 \rho(1) + \cdots + \alpha_p \rho(p-1)$$

$$\rho(2) = \alpha_1 \rho(1) + \alpha_2 1 + \cdots + \alpha_p \rho(p-2)$$

$$\cdots$$

$$\rho(p) = \alpha_1 \rho(p-1) + \alpha_2 \rho(p-2) + \cdots + \alpha_p 1$$

In the MOME approach, we substitute the population autocorrelations $\rho(h)$ with the sample autocorrelations $\hat{\rho}(h), h = 1,2, \ldots p$, in the above Yule-Walker equations. These p-equations, along with the one equation for mean estimation above, can be used to obtain the MOME estimators of the (p+1) model parameters.


George Udny Yule
(1871-1951)
Sir Gilbert Thomas Walker
(1868-1958)
"Every cell phone call solves the Yule-Walker equations every ten microseconds."
Thierry Dutoit
"On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers" (1927)
"On Periodicity in Series of Related Terms" (1931)

(c) Maximum Likelihood Estimation (MLE) of AR models

For iid data (sample) with marginal pdf $f(x_t; \boldsymbol{\theta})$, the joint pdf for a sample $\boldsymbol{x} = (x_1, \cdots, x_n)$ is
$f(\boldsymbol{x}; \boldsymbol{\theta}) = f(x_1, \cdots, x_n; \boldsymbol{\theta}) = \prod_{t=1}^{n} f(x_t; \boldsymbol{\theta})$
The likelihood function of the sample is: $L(\boldsymbol{\theta}; \boldsymbol{x}) = f(x_1, \cdots, x_n; \boldsymbol{\theta}) = \prod_{t=1}^{n} f(x_t; \boldsymbol{\theta})$
The loglikelihood function of the sample is: $l(\boldsymbol{\theta}; \boldsymbol{x}) = lnL(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{t=1}^{n} ln\, f(x_t; \boldsymbol{\theta})$

Approach 1. Using the Conditional Distributions to Obtain the Likelihood.

For a sample from a (weak) stationary time series $\{x_t\}$, the random variables in the sample $\{x_1, \cdots, x_n\}$ are not iid – <u>however we can construct the likelihood using the conditional probability density functions as follows.</u>

$$f(x_1, x_2) = f(x_2|x_1) f(x_1)$$
$$f(x_1, x_2, x_3) = f(x_3|x_2, x_1) f(x_2|x_1) f(x_1)$$
etc.

Therefore we have:

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = f(x_1, \cdots, x_n; \boldsymbol{\theta})$$
$$= \prod_{t=p+1}^{n} f(x_t|I_{t-1}; \boldsymbol{\theta}) * f\left(x_1, \cdots, x_p; \boldsymbol{\theta}\right)$$

Here $I_t = \{x_1, \cdots, x_t\}$ represents information available at time *t*, and

$\{x_1, \cdots, x_p\}$ are the initial values. This yields the following:

Exact loglikelihood:

$$lnL(\boldsymbol{\theta}; \boldsymbol{x}) = \sum\nolimits_{t=p+1}^{n} lnf(x_t|I_{t-1}; \boldsymbol{\theta}) + lnf(x_1, \cdots, x_p; \boldsymbol{\theta})$$

Conditional loglikelihood:

$$lnL(\boldsymbol{\theta}; \boldsymbol{x}) = \sum\nolimits_{t=p+1}^{n} lnf(x_t|I_{t-1}; \boldsymbol{\theta})$$

We can subsequently obtain the exact (the usual) MLEs and the conditional MLEs. For stationary time series data, they are both consistent and have the same limiting distribution, but may differ in finite samples.

Example: MLE for Stationary AR(1)

$$X_t = \delta + \alpha_1 X_{t-1} + Z_t$$

where $\{Z_t\}$ is a series of **Gaussian** white noise (that is, i.i.d. $N(0, \sigma^2), t = 1, \cdots, n$),

$\boldsymbol{\theta} = (\delta, \alpha_1, \sigma^2)', |\alpha_1| < 1$

Conditional on I_{t-1}, which reduces to only x_{t-1}

$$X_t|I_{t-1} = X_t|x_{t-1} \sim N(\delta + \alpha_1 x_{t-1}, \sigma^2), t = 2, \cdots, n$$

To determine the marginal density for the initial value X_1, recall that for a stationary AR(1) process:

$$\mathrm{E}[X_1] = \mu = \frac{\delta}{1 - \alpha_1}$$

$$\mathrm{var}[X_1] = \frac{\sigma^2}{1 - \alpha_1^2}$$

and therefore:

$$X_1 \sim N\left(\frac{\delta}{1 - \alpha_1}, \frac{\sigma^2}{1 - \alpha_1^2}\right)$$

This yields the following:

Exact loglikelihood:

$$lnL(\boldsymbol{\theta}; \boldsymbol{x}) = \sum\nolimits_{t=2}^{n} lnf(x_t|x_{t-1}; \boldsymbol{\theta}) + lnf(x_1; \boldsymbol{\theta})$$

Conditional loglikelihood:

$$lnL(\boldsymbol{\theta}; \boldsymbol{x}) = \sum\nolimits_{t=2}^{n} lnf(x_t|x_{t-1}; \boldsymbol{\theta})$$

We can subsequently obtain the exact (the usual) MLEs and the conditional MLEs.

Approach 2. Using the Multivariate Normal Distribution to Obtain the Likelihood.

Multivariate normal distribution

Let $\underline{X} = (X_1, \dots, X_n)'$, the general formula for the

n-dimensional multivariate normal density function is

$$f(x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\underline{x} - \underline{\mu}\right)\right]$$

where $E\left(\underline{X}\right) = \underline{\mu}$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\underline{X}$.

Given that we have i.i.d. Gaussian white noise, we can write down the exact likelihood of our time series $\{X_1, \dots, X_n\}$ as a multivariate normal density function directly as follows:

$$L(\boldsymbol{x}; \boldsymbol{\theta}) = f(x_1, \cdots, x_n; \boldsymbol{\theta})$$

Example: MLE for Stationary AR(1)

$$X_t = \delta + \alpha_1 X_{t-1} + Z_t$$

where $\{Z_t\}$ is a series of **Gaussian** white noise (that is, i.i.d. $N(0, \sigma^2), t = 1, \cdots, n$),
$\boldsymbol{\theta} = (\delta, \alpha_1, \sigma^2)', |\alpha_1| < 1$.

$$E\left(\underline{X}\right) = \underline{\mu} = \frac{\delta}{1 - \alpha_1}\underline{1}$$

$$Var\left(\underline{X}\right) = \Sigma = \frac{\sigma^2}{1 - \alpha_1^2}\begin{bmatrix} 1 & \alpha_1 & \cdots & \alpha_1^{n-1} \\ \alpha_1 & 1 & \cdots & \alpha_1^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{n-1} & \alpha_1^{n-2} & \cdots & 1 \end{bmatrix}$$

Estimation of the Autoregressive Moving Average Process (ARMA)

Approach: LSE, MLE, MOME

ARMA(p,q) Model:

$$X_t = \delta + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$$

where $\{Z_t\}$ ***is a series of white noise.***

(1) Order determination.

The order of ARMA(p,q) is not as easily identified as the pure MA or AR process based on the ACF's and the PACF's as we can see below:

Type of Model	Typical Pattern of ACF	Typical Pattern of PACF
AR (*p*)	Decays exponentially or with damped sine wave pattern or both	**Cut-off after lags *p***
MA (*q*)	**Cut-off after lags *q***	Declines exponentially
ARMA (*p,q*)	Exponential decay	Exponential decay

Table 1. Theoretical Patterns of ACF and PACF for Stationary Time Series.

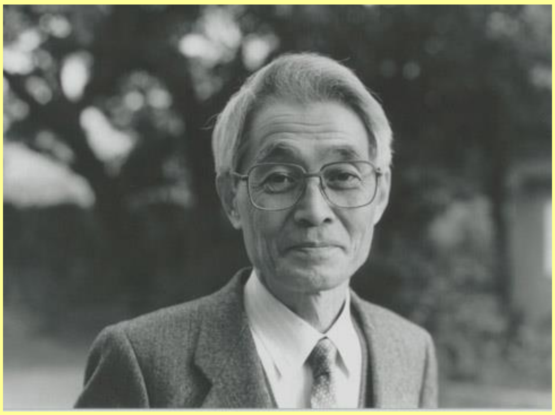
Akaike's Information Criterion:

Instead, we use other model selection tools such as the Akaike's Information Criterion (AIC). The AIC is a function of the maximum likelihood plus twice the number of parameters. The number of parameters in the formula penalizes models with too many parameters.

One often adopted a better, corrected version of the AIC, commonly referred to as (AICC):

$$\mathbf{AICC} = -2\ln\left(L(\hat{\alpha},\hat{\beta})\right) + \frac{2(p+q+1)n}{n-p-q-2}$$

Model that will minimize the AICC is selected.



Hirotugu Akaike
(http://en.wikipedia.org/wiki/Hirotugu_Akaike)

Parsimony:

- **Once principal generally accepted is that models should be parsimonious—having as few parameters as possible**
- **Note that any ARMA model can be represented as a pure AR or pure MA model, but the number of parameters may be infinite**
- **AR models are easier to fit so there is a temptation to fit a less parsimonious AR model when a mixed ARMA model is appropriate**
- **It has been shown, for example, fitting unnecessary extra parameters, or an AR model when a MA model is appropriate, will result in loss of forecast accuracy**

Like the MA(q) model, the ARMA(p,q) is harder to estimate than the AP(p) model because explicit estimators can not be found in general, instead, various numerical methods have been developed to derive the estimators. To fix ideas, here we will consider the simple ARMA(1,1) process:

$$\boldsymbol{X_t - \mu = \alpha_1(X_{t-1} - \mu) + Z_t + \beta_1 Z_{t-1}}$$

(2) **LSE – an iterative procedure:** One iterative procedure is as follows using the ARMA(1,1) as an example:

(i) Select suitable starting values for μ, α_1 **and** β_1

(ii) Calculate the residual recursively using the relation

$$Z_t = X_t - \mu - \alpha_1(X_{t-1} - \mu) - \beta_1 Z_{t-1}$$

as follows:

Taking $Z_0 = 0$, and $X_0 = \hat{\mu}$ we compute $\hat{Z}_1 = X_1 - \hat{\mu}$, and then

$$\hat{Z}_2 = X_2 - \hat{\mu} - \hat{\alpha}_1(X_1 - \hat{\mu}) - \hat{\beta}_1 \hat{Z}_1$$

until

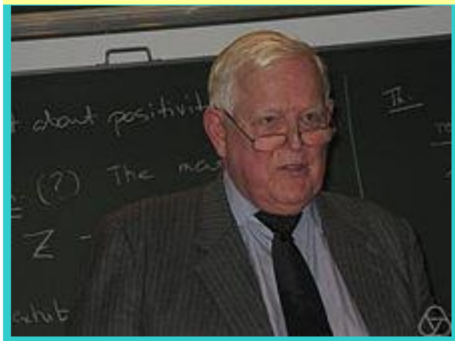
$$\hat{Z}_n = X_n - \hat{\mu} - \hat{\alpha}_1(X_{n-1} - \hat{\mu}) - \hat{\beta}_1 \hat{Z}_{n-1}$$

Then we compute the residual sum of squares:

$$\sum_{i=1}^{n} \hat{Z}_i^2$$

(iii) Repeat the above procedures for neighboring values of μ, α_1 **and** β_1 – until we find the values that will minimize the residual sum of squares.

(3) **MLE –** the MLE can be derived theoretically by writing down the likelihood of the data as the pdf of the multivariate normal distribution the observed time series – with its mean vector and variance-covariance matrix written in terms of the ARMA model parameters as we have done for the MA and AR models previously. Many numerical methods are available to compute the MLE numerically, for example, the **Kalman filter**, to any desired degree of approximation.



Rudolf E. Kálmán (http://en.wikipedia.org/wiki/Rudolf_E._K%C3%A1lm%C3%A1n)

In summary, the Box-Jenkins approach for fitting the ARMA(p, q) models is as follows:

1. Transform the data, if necessary, to a stationarity time series
2. Guess for the (range of) values of p and q by examining the ACF, PACF etc.
3. Estimate the parameters of the possible ARMA(p, q) models
4. Compare model goodness-of-fit using AICC (or AIC, BIC, etc.)
5. Perform residual analysis to confirm that the chosen model adequately describes the data

Residual Analysis

Test for white noise: ***Portmanteau tests***

(1) Residuals:

The residual is defined as

Residual = observation – fitted value

Take a zero-mean AR(1) Model as an example:

$$X_t = \alpha_1 X_{t-1} + Z_t$$

where $\{Z_t\}$ ***is a series of white noise, the residual is:***

$$\widehat{Z}_t = X_t - \widehat{\alpha}_1 X_{t-1}$$

(2) Sample ACF and PACF:

Since for a series of white noise, we would expect zero ACF and zero PACF, so the sample ACF and the sample PACF can be plotted and inspected.

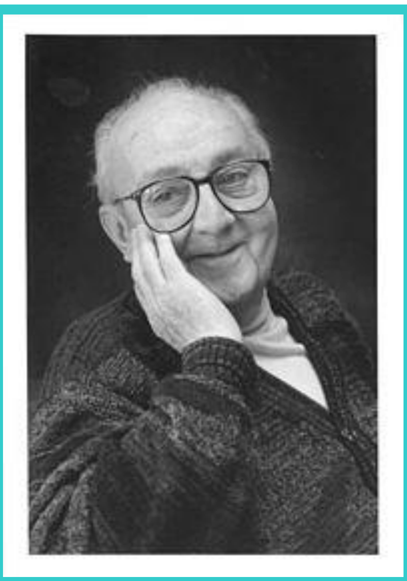
(3) Portmanteau tests:

- **Box and Peirce proposed a statistic which tests the magnitudes of the residual autocorrelations as a group**
- **Their test was to compare Q below with the Chi-Square with K – p – q d.f. when fitting an ARMA(p, q) model**

$$Q = n \sum_{k=1}^{K} \widehat{\rho}_k^2$$

- **Box & Ljung discovered that the test was not good unless n was very large**
- **Instead use modified Box-Pierce or Ljung-Box-Pierce statistic—reject model if Q* is too large**

$$Q^* = n(n+2) \sum_{k=1}^{K} \frac{\widehat{\rho}_k^2}{n-k}$$



Greta M. Ljung and George E. P. Box (http://en.wikipedia.org/wiki/Box%E2%80%93Pierce_test#Box-Pierce_test)
http://genealogy.math.ndsu.nodak.edu/id.php?id=42226

The normality test for Gaussian white noise: ***Shapiro-Wilk Test***, ***Q-Q plot etc.***
http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

Martin B. Wilk (http://en.wikipedia.org/wiki/Martin_Wilk)

Model Diagnostics – Revisit

In summary, there are three major steps in checking how good our model fits.

1. Residuals

The first step of our diagnosis is to look at the residuals obtained from our fitted model. Suppose we are fitting an AR(3) model to a time series X_t, so that

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \emptyset_3 X_{t-3} + Z_t$$

We use the procedures of the last chapter to estimate the parameters $\emptyset_1, \emptyset_2, \emptyset_3$ and hence the residuals from this model are:

$$\widehat{Z_t} = X_t - \widehat{X}_t = X_t - \widehat{\emptyset}_1 X_{t-1} - \widehat{\emptyset}_2 X_{t-2} - \widehat{\emptyset}_3 X_{t-3}$$

If our model was a perfect fit then we would expect that these residuals would be distributed as white noise. To test this we could consider a time series plot of the residuals. We can also look at a normal-probability plot of the residuals to see whether we have Gaussian white noise or not. A third graphical tool is to examine a plot of the residuals versus the fitted values from the model. In all instances we are looking for patterns which would indicate that the residuals are not distributed as white noise.

2. ACF and Portmanteau Tests (for example, the Ljung-Box-Pierce test)

The second stage of the diagnosis is to examine the sample autocorrelation function of the residuals. We plot the sample ACF of the residuals and we look for any significantly non-zero autocorrelations which would indicate deviations from white noise. While examining sample ACF plots of the residuals is informative a better approach is to consider the Ljung-Box-Pierce Q test statistic that tests whether the first K autocorrelations are all zero:

$$\mathrm{Q} = \mathrm{n}(\mathrm{n}+2)\sum_{k=1}^{K}\frac{\hat{r}_k^2}{n-k}$$

Where n is the number of observations in the time series and the maximum lag is K. Ljung and Box showed that the statistic Q is distributed as a chi-squared distribution with K-p-q degrees of freedom if the ARMA(p, q) model is a correct fit.

3. Overfitting and the Model Goodness of Fit Indices

The final diagnostic tool is to overfit the model. Suppose we want to examine whether an AR(2) model is the best model to fit our time series data. We fit a larger model which includes the AR(2) model as a special case. So we could fit either an AR(3) or an ARMA(2,1) model. We then examine the estimates of the parameters in the larger model and we conclude that our original model is to be preferred if

1. The estimates of the additional parameters ($\emptyset_3 or\ \theta_1$) are not significantly different from zero

2. The estimates of the original parameters $\emptyset_1$ and $\emptyset_2$ are not significantly different in the larger models from their values in the AR(2) model.

We should also compare the model goodness-of-fit indices such as the AIC, AICC, and BIC to see which model is the best.

PLANTE TULSA WORLD 8·24·2015
NO, NO STAY ON! EVERYBODY READY TO GO AGAIN?
STOCK MARKET EXPERT
STOCK MARKET