Lecture 4
Autoregressive Models

CHEN Ying
FE5209 Financial Econometrics

Outline

- Stationary process and autocorrelation.
- AutoRegressive (AR) models: Properties and estimation

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + \varepsilon_t$$

- Moving-Average (MA) models
- Box-Jenkins methodology

Readings

SDA chapter 9
FTS chapter 2
SFM chapter 11

Quiz: Are they time series?

1.Monthly number of airline tickets sold by Chan Brothers Travel Pte Ltd, a travel agency in SG.

2.Singapore quarterly unemployment rate between 2008:Q2 and 2013:Q4.

3.Quarterly number of home mortgage loan applications to DBS.

4.The annual number of road accidents reported to Land Transport Authority and Traffic Police.

5.Time spent in training by workers in Microsoft Corp.

6.The dates on which a particular employee was absent from work due to illness over the past two years.

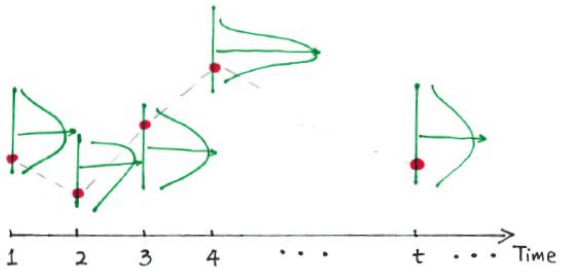
What is a time series …

Financial data are time series and have patterns related to time.

Whenever data is recorded sequentially over time and Time is considered to be an important aspect, we have a Time Series.

- Time sequence of data is important.
- Most time series are equally spaced at roughly regular intervals, such as daily, monthly, quarterly, or annually.

A time series can be considered as a sample from the stochastic process. A stochastic process is a sequence of random variables and can be viewed as the “theoretical" or “population" of a time series. *“Stochastic" is a synonym for random.*

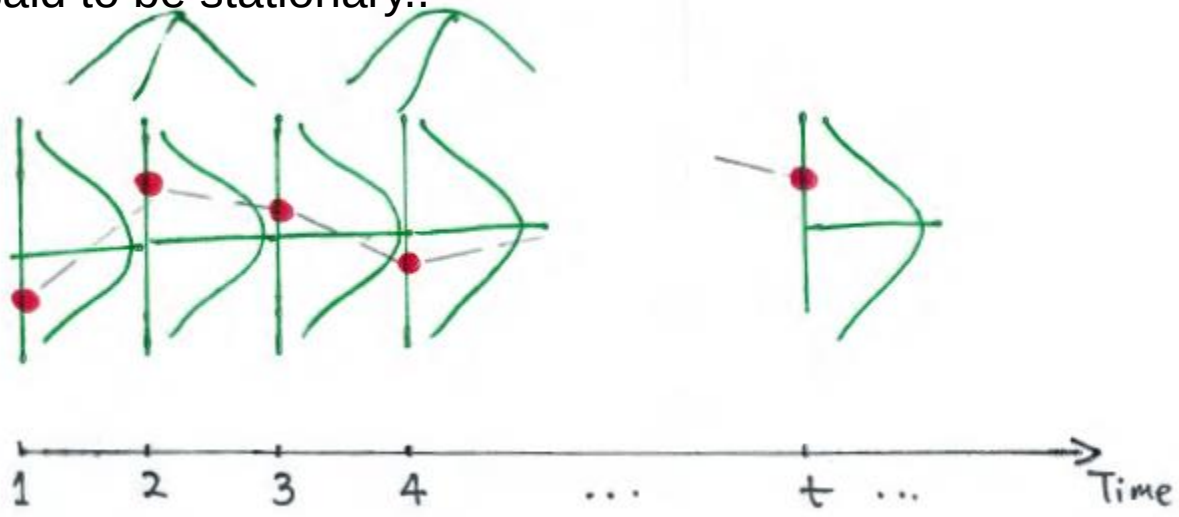


Population: Stochastic Process. Sample: Time series Data

Stationary process

When we observe a time series, the fluctuations appear random, but often with the same type of stochastic behaviour from one time period to the next. A stable estimation requires stable relationship between current values and their lagged values over time!

Stationary stochastic processes are probability models for time series with time-invariant behaviour. Mathematically, If the properties of a stochastic process is unaffected by a change of time origin, that is, for any time s and t, the probability distributions of a sequence $Y_1, \cdots, Y_t$ and $Y_{1+s}, \cdots, Y_{t+s}$ are the same then the process is said to be stationary..



Population: Stationary Stochastic Process. Sample: Stationary Time Series

Why stationarity?

The beauty of a stationary process is that it can be modelled with **relatively few parameters**. For example, we do not need a different expectation for each observation Y_t; rather they all have a common expectation.

A stationary series should show oscillation around some fixed level, a phenomenon called **mean-reversion**.

An AR(1) process is stationary if and only if $-1 < \phi < 1$.

Example 1 (stationary): $Y_t = 80 + 0.2\, Y_{t-1} + e_t$, *with* $Y_1 = 120$. *The mean of time series* $\mu = \frac{80}{1-0.2} = 100$. *According to the AR(1) model, we obtain:*

$$Y_2 = 80 + 0.2 \times 120 = 104$$
$$Y_3 = 80 + 0.2 \times 104 = 100.8$$
$$\vdots$$

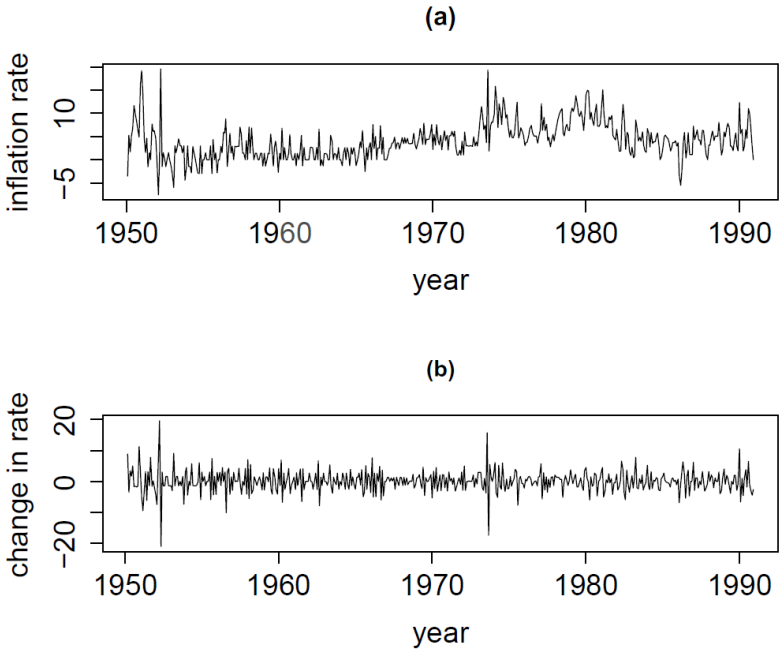
Stationary time series converges to the mean. (Mean reversion!)

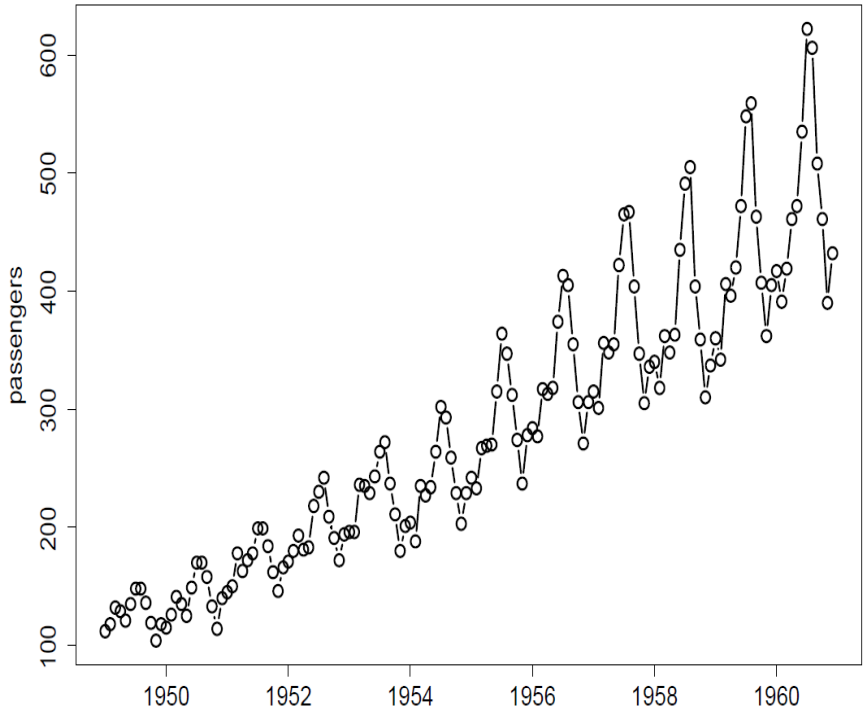
Example 2 (non-stationary): $Y_t = 80 + 1.2\, Y_{t-1} + e_t$, *with* $Y_1 = 120$. *The mean of time series* $\mu = \frac{80}{1-1.2} = -400$. *According to the AR(1) model, we obtain:*

$$Y_2 = 80 + 1.2 \times 120 = 224$$
$$Y_3 = 80 + 1.2 \times 224 = 348.8$$
$$\vdots$$

Non-stationary time series is explosive.

Example: stationary?





Left: One-month inflation rate (in percent, annual rate) and its first difference. The differenced series certainly oscillate around a fixed mean of 0%.The differenced series is clearly stationary, but whether or not the original series is stationary needs further investigation.

Right: Monthly totals of international airline passengers for the years 1949 to 1960. There are three types of nonstationarity seen in the plot: 1) upward trend, 2) seasonal variation, and 3) increase over time in the size of the seasonal oscillations.

4_examples.R

Weak stationarity

Weak stationarity: If all the moments up to some order f are unaffected by a change of origin, the process is said to be weakly stationary of order f

A process is covariance stationary (weakly stationary) if its mean, variance, and covariance are unchanged by time shifts:

- constant mean $E(Y_t) = \mu$
- constant variance $\mathrm{var}(\mathrm{Y_t}) = E(Y_t - \mu)^2 = \sigma^2$
- constant autocovariance structure $\mathrm{corr}(Y_s, Y_t) = \rho_{|s-t|}, \forall s, t$

The mean and variance do not change with time and the correlation between two observations depends only on the lag, the time distance between them.

For example, the auto-covariance between $Y_{t-1}$and Y_{t-2} with time lag 1 being the same as Y_{t-5} and Y_{t-6} with time lag 1. However the autocorrelations $Y_{t-1}$and Y_{t-2} with time lag 1 may differ from the autocorrelations $Y_{t-1}$and Y_{t-3} with time lag 2 .

White-Noise (WN) process

The sequence of random variables X_1, X_2 ... is called IID noise if the observations of the time series are independent and identically distributed (IID) random variables.

Let ϵ_t , $t = \pm 0, \pm 1, \pm 2, \ldots$, be a zero-mean, IID sequence $\{\epsilon_t\}$ with

- $E(\epsilon_t) = 0, E(\epsilon_s \epsilon_t) = \begin{cases} \sigma^2 \,, if\ s = t \\ 0 \,, if\ s \neq t \end{cases}$ for all t and s

Sequence $\{\epsilon_t\}$ is called a purely random process, ***IID noise* or simply *strict white noise*** and we write $\epsilon_t \sim IID(0,\sigma^2)$.

If successive values follow a normal (Gaussian) distribution, then Gaussian white noise is a strict white noise, denoted $\epsilon_t \sim IID\ N(0,\sigma^2)$.

If the sequence $\{\epsilon_t\}$ is only uncorrelated and not necessarily independent, then $\{\epsilon_t\}$ is known as an ***uncorrelated white noise process*** or ***weak white noise***, $\epsilon_t \sim WN(0,\sigma^2)$.

- A weak white noise process is weakly stationary with $\rho_0 = 1$ and $\rho_k = 0, \forall\ k \neq 0.$
- Because of the lack of correlation, past values of a white noise process contain no information that can be used to predict future values. One cannot predict the future values of a white noise process.

How to check stationarity of a stochastic process?

--Lag operator, AR polynomial and stationarity

Let the value of a time series at time t, y_t , is a linear function of the last p values of y and of exogenous terms, denoted by ϵ_t:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \epsilon_t$$

Expressions of this type are called difference equations.

The lag operator, denoted by L (or B), is an operator that shifts the time index backward by one unit.

Applying L the variable at time t will lead to the variable at $t-1$: $Ly_t = y_{t-1}$.

Applying L^2, $L^2 y_t = L(Ly_t) = Ly_{t-1} = y_{t-2}$.

- Formally, the lag operator transforms one time series, say $\{y_t\}_{t=-\infty}^{\infty}$ into another series, say $\{x_t\}_{t=-\infty}^{\infty}$where $x_t = y_{t-1}$.
- A constant c can be viewed as a special series, $\{y_t\}_{t=-\infty}^{\infty}$ with $y_t = c$ for all t, and we can apply the lag operator to a constant obtaining $Lc = c$.
- By raising L to a negative power, we obtain a lead operator:

$$L^{-k} y_t = y_{t+k} .$$

Difference equations

The difference equation for an ARMA(p, q) process:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \epsilon_t + b_1 \epsilon_{t-1} + \cdots + b_q \epsilon_{t-q}$$

can be written as

$$a(L) y_t = b(L) \epsilon_t .$$

where $a(L) = 1 - a_1 L - a_2 L^2 - \cdots - a_p L^p$ is called AR polynomial, and b$(L) = 1 + b_1 L + b_2 L^2 + \ldots + b_q L^q$ is called MA polynomial.

The difference equation for an AR(3) process

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + \epsilon_t$$

can be written as

$$y_t = a_1 L y_t + a_2 L^2 y_t + a_3 L^3 y_t + \epsilon_t$$
$$(1 - a_1 L - a_2 L^2 - a_3 L^3) y_t = \epsilon_t$$

or, in compact form

$$a(L) y_t = \epsilon_t$$

where $a(L) = 1 - a_1 L - a_2 L^2 - a_3 L^3$ is called AR polynomial.

Characteristic equation

The reverse ***characteristic equation*** associated with the difference equation:

$$a(\lambda) = 0$$

Any value of λ which satisfies the reverse characteristic equation is called a root of polynomial $a(\lambda)$. A polynomial of degree p has p roots λ_k, $k = 1, \ldots, p$. In general, roots are complex numbers: $\lambda_k = a_k + b_k i$.

The coefficient form of a reverse characteristic equation:

$$1 - a_1\lambda - \cdots - a_p\lambda^p = 0$$

An alternative is the root form given by

$$(\lambda_1 - \lambda)(\lambda_2 - \lambda) \ldots (\lambda_p - \lambda) = \prod_{i=1}^{p} (\lambda_i - \lambda) = 0$$

The latter form reveals the roots directly.

The ARMA(p,q) process is stationary if the roots of the AR polynomial lie outside the unit circle.

Example

Given an AR(2) process

$$y_t = \frac{3}{2} y_{t-1} - \frac{1}{2} y_{t-2} + \epsilon_t$$

The reverse characteristic equation in coefficient form is given by

$1 - \frac{3}{2}\lambda + \frac{1}{2}\lambda^2 = 0$ or $2 - 3\lambda + 1\lambda^2 = 0$

that can be written in ***root form*** as

$$(1 - \lambda)(2 - \lambda) = 0$$

Here, $\lambda_1 = 1$ and $\lambda_2 = 2$ represent the set of possible solutions for λ satisfying the reverse characteristic equation

$$1 - \frac{3}{2}\lambda + \frac{1}{2}\lambda^2 = 0$$

Serial dependence of a stationary process

Given a stationary linear time series, we are interested to know:

- Expectation: $\mu = \mathrm{E}(Y_t)$.
- Variance: $\gamma_0 = var\,(Y_t)$.
- Auto-covariance (serial covariance):

$$\gamma_{t-s} = cov\,(Y_s, Y_t) = E(Y_s - \mu)(Y_t - \mu)$$

that is symmetric with $\gamma_{t-s} = \gamma_{s-t}$.

- Autocorrelation: $\rho_{t-s} = \frac{cov\,(Y_s, Y_t)}{var\,(Y_t)} = \frac{\gamma_{t-s}}{\gamma_0}$

Lagged scatterplot is a simple graphical summary of serial dependence in a time series, which is a scatterplot of the time series against itself offset in time by one to several time steps.

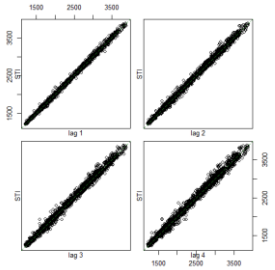
Let the time series of length T be $y_1, \ldots, y_T$:

the lagged scatterplot for lag k is a scatterplot of

the **last** $\mathrm{T} - k$ observations $y_{k+1}, \ldots, y_T$ (concurrent values Y_t)

against

the **first** $\mathrm{T} - k$ observations $y_1, \ldots y_{T-k}$. (lagged values Y_{t-k})



Autocorrelation

The (serial) dependence between values of a time series and their own past values is measured by autocorrelations.

The theoretical autocorrelation function is defined:

$$\rho_k = \frac{E[(Y_t-\mu)(Y_{t-k}-\mu)]}{\sigma^2}.$$

- ACF measures linear relations of Y_t and Y_{t-k}, for different values of k.
- $-1 \leq \rho_k \leq 1$ and $\rho_k = \rho_{-k}$.

The covariance between Y_t and Y_{t-k} is denoted by γ_k and $\gamma(\cdot)$ is called the autocovariance function.

- $\gamma_k = \sigma^2 \rho_k$
- $\gamma_0 = \sigma^2$
- $\rho_k = \frac{\gamma_k}{\sigma^2} = \frac{\gamma_k}{\gamma_0}$

Estimate autocorrelations

The theoretical autocorrelation function is defined:

$$\rho_k = \frac{E[(Y_t-\mu)(Y_{t-k}-\mu)]}{\sigma^2}.$$

The first order (or lag 1) autocorrelation measures the correlation between two successive observations in a time series. Given data, the sample estimator of lag-1 autocorrelation is computed as:

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{T}(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

where $\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$ is the common sample mean of the whole time series.

Higher order autocorrelations: its sample function for lag k is

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

where $\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$ is the common sample mean of the whole time series.

- *impossible to estimate $\hat{\rho}_k$ for $k \geq T$;*
- *$\hat{\rho}_k$ can not be estimated accurately for large k;*
- *rule of thumb: $T \geq 50$ and $k \leq T/4$.*

Sample autocorrelation function (ACF)

A plot of sample autocorrelations $\hat{\rho}_k$ against lags k for $k = 1,2, \ldots$ is called sample autocorrelation function (ACF) or correlogram, typically plotted for the first $T/4$ lags or thereabouts.

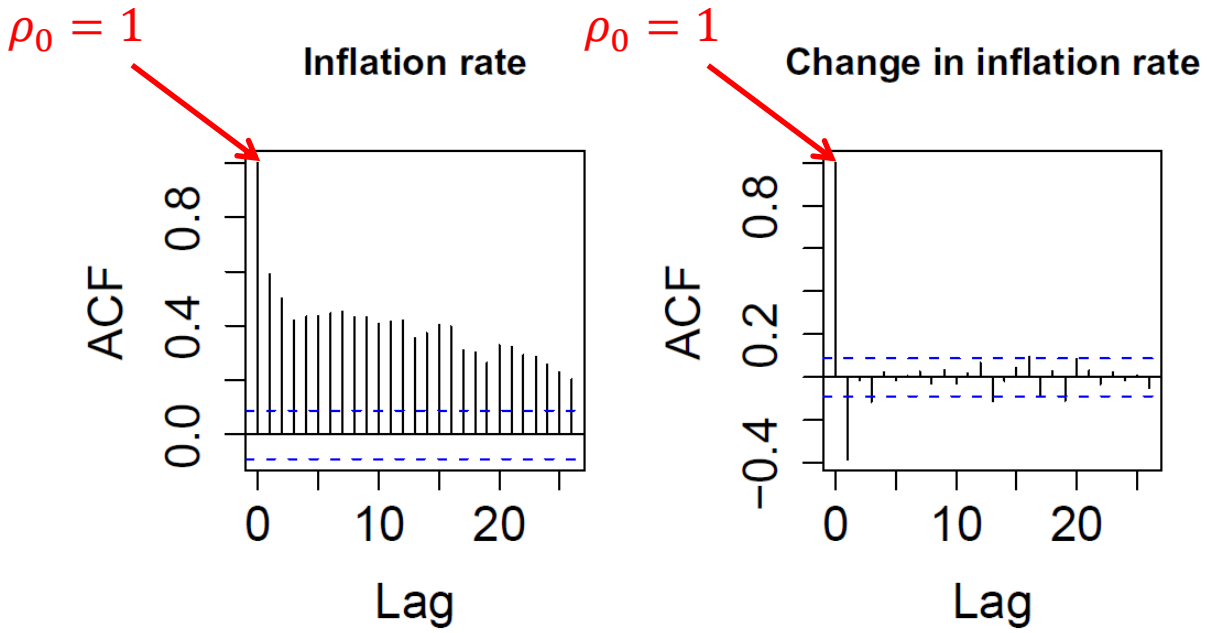


Fig. 9.3. *Sample ACF plots of the one-month inflation rate (a) and changes in this rate (b).*

4_ACF.R

Sampling distribution of AC estimator

Sampling distribution: For a given sample of a stationary time series $(y_t)_{t=1}^{T}$, let $\bar{y}$ be the sample mean. Then the lag-k sample autocorrelation is:

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

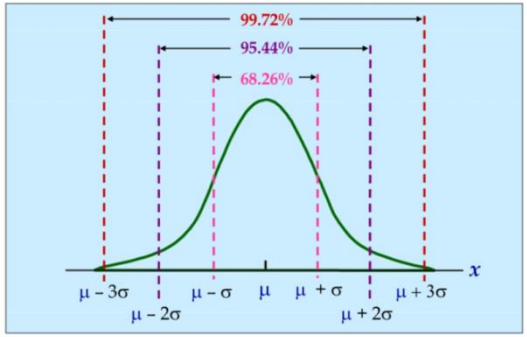
If y_t is an IID sequence with finite variance, $\hat{\rho}_k \sim N(\rho_k, \frac{1}{T})$

If y_t is a weakly stationary series, $\hat{\rho}_k \sim N(\rho_k, (1 + 2\sum_{j=1}^{k-1}\rho_j^2)/T)$

Facts of normal random variable

- 68.26% 2/3 of values of **X** are within 1 std deviation of its mean
- 95.44% 95% of values of **X** are within 2 std deviations of its mean
- 99.72% almost all values of **X** are within 3 std deviations of its mean

$$\boldsymbol{X} \sim N(\mu, \sigma^2)$$



Test for randomness (individual test for autocorrelation)

If autocorrelations are always zero, it indicates that past values are uncorrelated to future values. On the contrary, nonzero autocorrelation(s) implies that past values can be used to forecast future values. *However the sample autocorrelations may not be identical to the theoretical ones due to random errors.*

Test for autocorrelation at any lag k: $H_0: \rho_k = 0$
The lag-k autocorrelation is considered as significant,
if the sample autocorrelation at lag order k is larger than
two standard deviations in magnitude,

$$|\hat{\rho}_k| > 2/\sqrt{T}$$

it is significant at 5% level. We reject the null hypothesis of $H_0: \rho_k = 0$ and conclude that the time series is not random. Otherwise, if all the considered autocorrelations are insignificant, we don't reject the null hypothesis of randomness of the time series.

- *The critical value is 1.96 at 5% significance level. Nevertheless, 2 is used to simplify the computation.*

Correlogram (Sample ACF)

The plot is normally supplemented with 5% significance limits (dashed lines at $\pm 2/\sqrt{T}$) to enable a graphical check of whether serial dependence exists at a particular lag. Any bar beyond the dashed lines indicates significant autocorrelations.

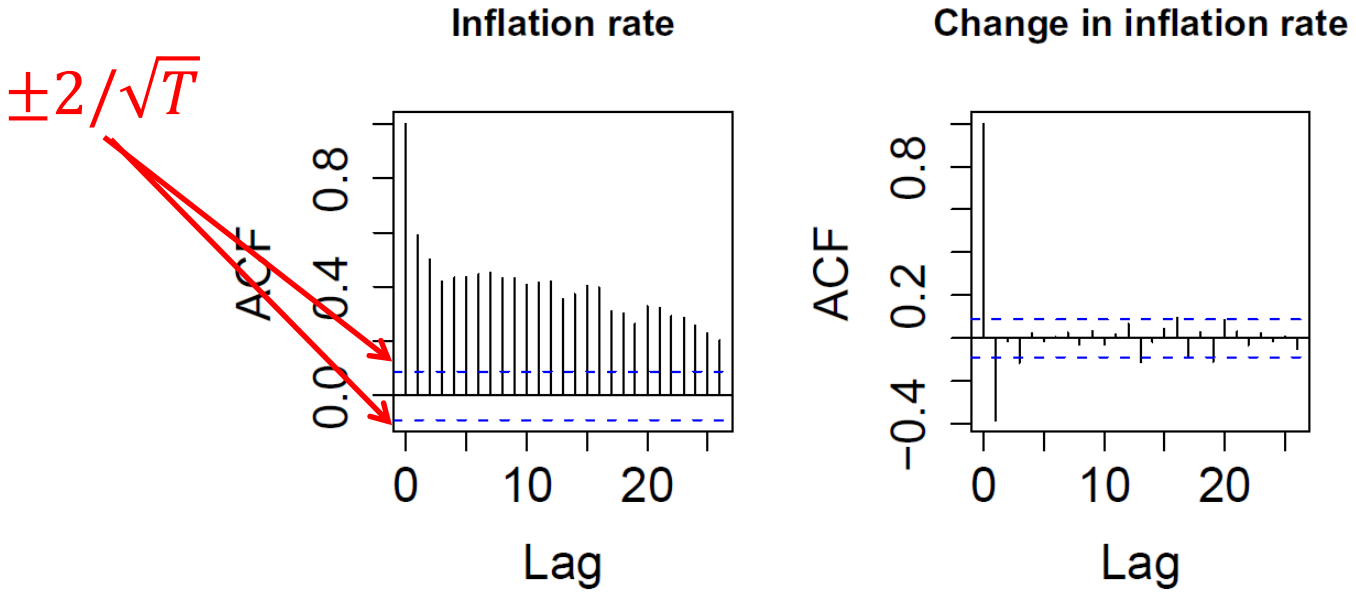


Fig. 9.3. *Sample ACF plots of the one-month inflation rate (a) and changes in this rate (b).*

Joint test for autocorrelations

The Ljung-Box test statistic, also called Q test statistic, can be used to determine if the first m ACFs are jointly equal to zero.

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_m = 0 \;\; vs \;\; H_a: \rho_j \neq 0.$$

$$Q(m) = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k{}^2}{(T-k)} \sim x^2(m)$$

which is asymptotically chi-squared distributed with m degrees of freedom. Decision rule: Reject H_0 if $Q(m) > critical\ value\ of\ \mathcal{X}_m^2\ (\alpha)$ or P-value is less than α.

Remark: 1. The autocorrelation tests, including the individual sampling distribution test and the joint Q-test, can be applied to the original time series and residuals.

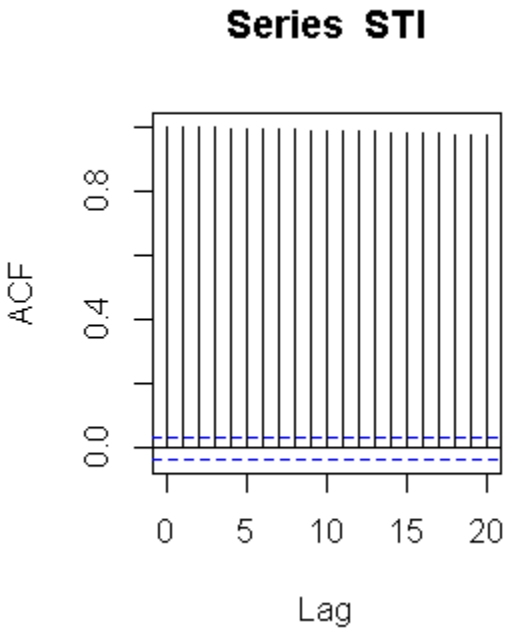
2. For residuals of a fitted model, we use the same test statistics to check the significance of autocorrelations. In this case the $Q(m)$ statistic is asymptotically χ^2_{m-g} distributed, where g is the no. of estimated parameters in the fitted model.

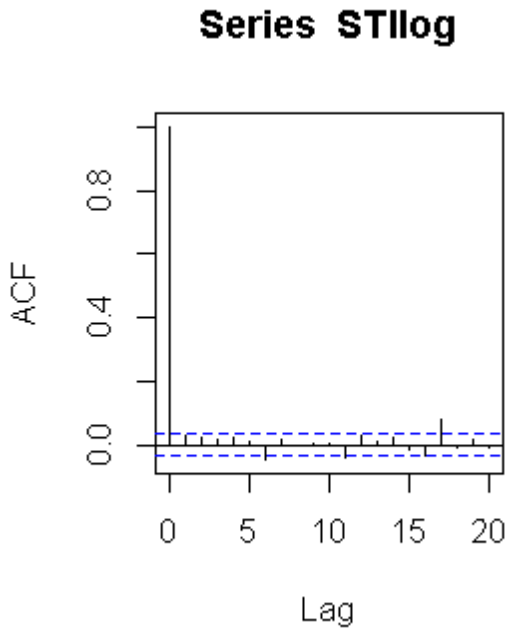
Example

Daily prices of STI from 4th January, 2000 to 26th October, 2012.
Q(5) = 16459.81, df = 5, p-value < 2.2e-16.

Daily log returns of STI: $Q(5) = 7.723(P - value: 0.172)$ and $Q(10) = 15.936(0.102)$

Implication: Daily prices of STI depend on their own past values.
However stock returns do not have significant serial correlations.





Autoregressive (AR) model of order 1

The autoregressive model of lag 1, written as AR(1) is:

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t, \qquad e_t \sim WN(0, \sigma_e^2)$$

or $$Y_t = \delta + \phi Y_{t-1} + e_t, \qquad \text{where } \delta = (1-\phi)\mu$$

Current value of Y_t can be predicted using its past value Y_{t-1}. The deviation between the realized value and the fitted value (from the model) is due to the existence of a random shock e_t.

- The AR coefficient ϕ is zero: Y depends purely on the random component (error), and there is no serial dependence.
- If ϕ is large, past values strongly influence future values.

Assumptions regarding the error term: Zero mean, constant variance (σ_e^2), and mutually uncorrelated (random).

- If our AR(1) model successfully captures the data's characteristics, then there should be no (significant) autocorrelations in the residuals! To check the adequacy of the fitted AR(1) model, we focus on the residuals from the regression for any "left-over" dependence. Significant?-> Try AR model of higher order (more lagged values)

AR(1) model

The AR(1) model $Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t$ is stationary if and only if $-1 < \phi < 1$.

Stationary models have constant mean, variance and ACFs over time.

Mean: $E(Y_t) = \mu = \delta/(1-\phi)$

Variance: $\gamma_0 = var(Y_t - \mu) = \phi^2 var(Y_{t-1} - \mu) + var(e_t)$

$$\Rightarrow \gamma_0 = \phi^2 \gamma_0 + \sigma_e^2 \Rightarrow var(Y_t) = \gamma_0 = \frac{\sigma_e^2}{1-\phi^2}$$

ACF: $\gamma_k = cov(Y_t, Y_{t-k}) = E[(Y_t - \mu)(Y_{t-k} - \mu)]$

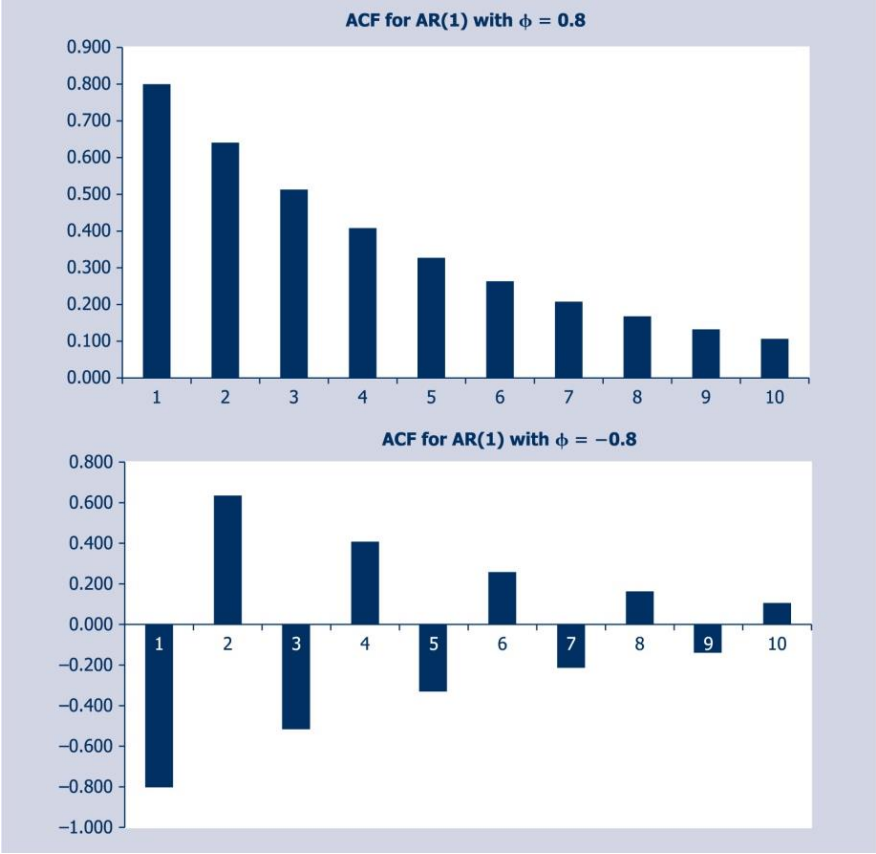
$$= \phi E[(Y_{t-1} - \mu)(Y_{t-k} - \mu)] + E[e_t(Y_{t-k} - \mu)] = \phi\gamma_{k-1}$$

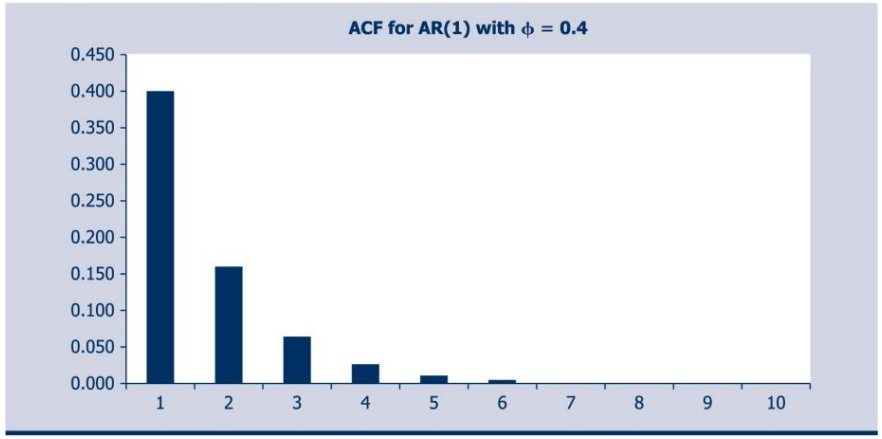
$$\Rightarrow \frac{\gamma_k}{\gamma_0} = \frac{\phi\gamma_{k-1}}{\gamma_0} \Rightarrow \rho_k = \phi\rho_{k-1}$$

$$corr(Y_t, Y_{t-k}) = \rho_k = \phi^k, \qquad k = 1,2, \ldots$$

ACF ρ_k decays exponentially as k increases

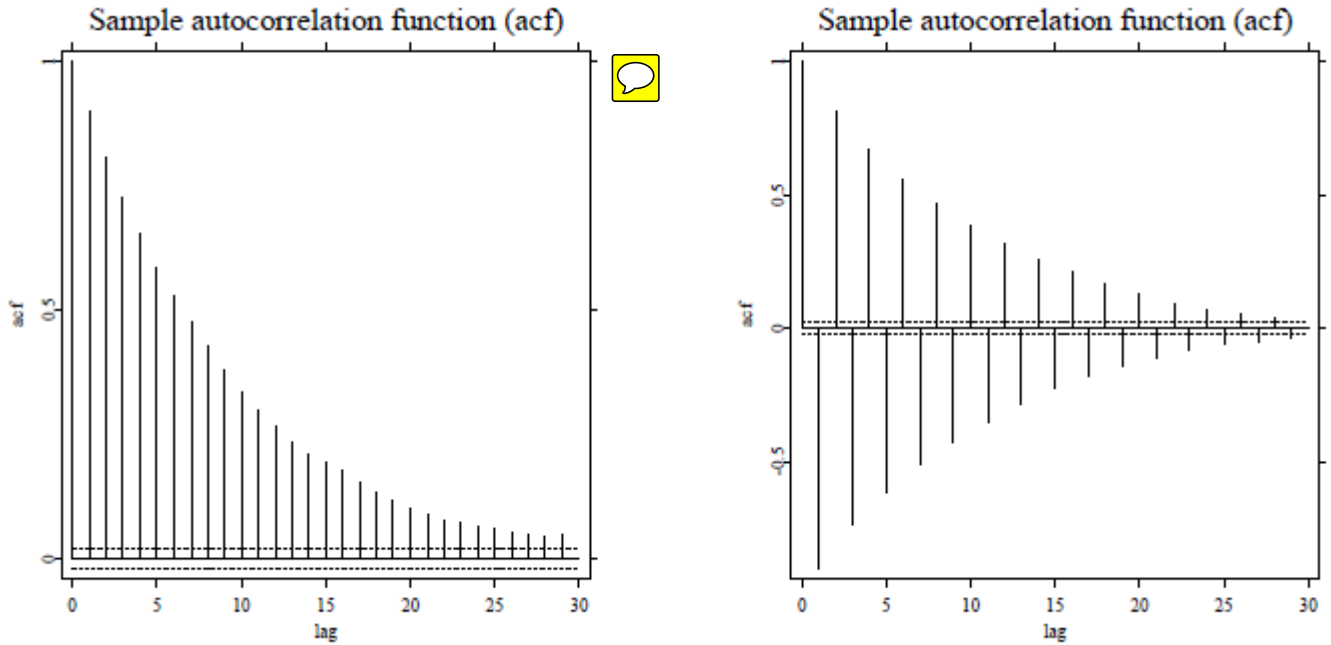
Theoretical ACFs of AR(1) process





Sample ACFs will not look exactly like the theoretical forms due to random noise!

Sample autocorrelations of an AR(1) process



ACF of a AR(1) process with $\phi_1 = 0.9$ (left) and $\phi_1 = -0.9$ (right).

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

Model estimation

There are several approaches to estimating AR models.

1. *The **Yule-Walker estimator*** uses the Yule-Walker equations with $k = 1, \ldots., p$ and estimates the AR parameters of pure *AR* models from the *SACF*.
2. The ***least squares estimator (LSE)*** finds the parameter estimates that minimize the sum of the squared residuals. For pure *AR* models, the *LSE* leads to the linear *OLS* estimator.
3. The ***maximum likelihood estimator (MLE)*** maximizes the (exact or approximate) log-likelihood function associated with the specified model. To do so, explicit distributional assumption for the disturbances, ϵ_t , has to be made.

Maximum Likelihood Estimator

The joint density function of two dependent random variables can be decomposed

$$f(x_2, x_1) = f(x_2|x_1)f(x_1)$$

Accordingly, for three dependent random variables:

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

and

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

For time series data $(y_0, \dots, y_T)$, the joint pdf can be written as

$$f(Y_0, \dots, Y_T) = f(y_T|Y_{T-1})f(y_{T-1}|Y_{T-2}) \dots f(y_0)$$

and the likelihood function becomes

$$L(\theta; Y_0, \dots, Y_T) = f(y_0)\prod_{t=1}^{T} f(y_T|Y_{T-1})$$

Maximizing $L(\theta; Y_T)$ is equivalent to maximizing the log-likelihood function

$$\ln L(\theta; Y_0, \dots, Y_T) = \ln f(y_0) + \sum_{t=0}^{T} \ln f(y_T|Y_{T-1})$$

which is typically maximized in practice.

Maximum Likelihood Estimator

Consider an $AR(1)$ process $Y_t = \phi Y_{t-1} + \epsilon_t$ with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, IID. Thus $f(y_t|Y_{t-1}) \sim \mathrm{N}(\phi y_{t-1}, \sigma_\epsilon^2)$

This gives $L(\phi, \sigma_\epsilon^2; Y_0, \dots, Y_T) = f(y_0) \prod_{t=1}^{T} (2\pi\sigma_\epsilon^2)^{-1/2} exp\left\{-\frac{(y_t - \phi y_{t-1})^2}{2\sigma_\epsilon^2}\right\}$

$$= f(y_0)(2\pi\sigma_\epsilon^2)^{-\frac{T}{2}} \prod_{t=1}^{T} exp\left\{-\frac{(y_t - \phi y_{t-1})^2}{2\sigma_\epsilon^2}\right\}$$

The log-likelihood function is

$$\ln L(\phi, \sigma_\epsilon^2; Y_0, \dots, Y_T) = \ln f(y_0) - \frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}(y_t - \phi y_{t-1})^2$$

The unconditional pdf of initial value y_0 is normal with mean 0 and variance $\sigma_\epsilon^2/(1-\phi^2)$:

$$y_0 \sim N\left(0, \frac{\sigma_\epsilon^2}{1-\phi^2}\right)$$

Therefore

$$\ln f(y_0) = -\frac{1}{2}ln(2\pi) + ln(1-\phi^2) - ln\,\sigma_\epsilon^2 - \frac{y_0^2(1-\phi^2)}{2\sigma_\epsilon^2}$$

Maximization with respect to ϕ and σ^2 yields exact maximum likelihood estimates.

Conditional MLE and conditional LSE

Conditional MLE of AR(1) is obtained by conditioning $y_1, \ldots, y_T$ on pre-sample realizations y_0. It is a simpler estimator that deletes the marginal density of y_0 from the likelihood and maximizes

$$-\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}(y_t - \phi y_{t-1})^2$$

This estimator is called the conditional least-squares estimator. It is a least-squares estimator because it minimize

$$\sum_{t=1}^{T}(y_t - \phi y_{t-1})^2$$

The default method for the function arima in R is to use the conditional least-squares estimates as starting values for maximum likelihood.

Example: BMW log returns

Box-Ljung test
data: bmw
X-squared = 44.987, df = 5, p-value = 1.460e-08

```
Call:
arima(x = bmw, order = c(1, 0, 0))

Coefficients:
          ar1  intercept
      0.081116   0.000340
s.e.  0.012722   0.000205

sigma^2 estimated as 0.000[illegible]6260:
aic = -34418.68
```

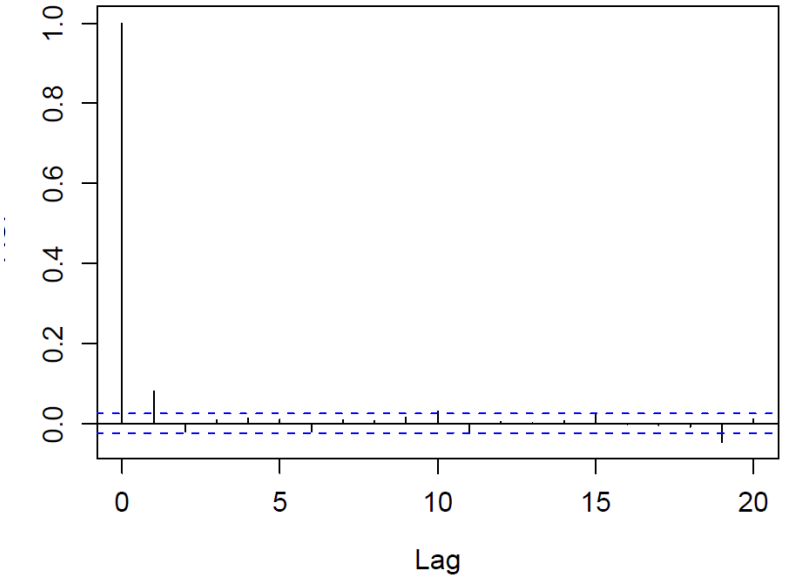


Fig. 9.6. *Sample ACF of BMW log returns.*

A positive value of ϕ means that there is some information in today's return that could be used for prediction of tomorrow's return, but a small value of ϕ means that the prediction will not be very accurate. The potential for profit might be negated by trading costs. 4_BMW.R

Diagnostic checking

We need to check the adequacy of the fitted model.

- The part of the data unexplained by the model (i.e., the residuals) should be small and not exhibit any systematic or predictable patterns.
- One could design diagnostic checking procedures that take the modeling objectives explicitly into account.

Why residual autocorrelation indicate problem?

Suppose that we are fitting an AR(1) model, $Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t$,
but the true model is an AR(2) process given by

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + e_t$$

Since we are fitting the incorrect AR(1) model, there is no hope of estimating ϕ_2 since it is not in the model. Moreover, $\hat{\phi}$ does not necessarily estimate ϕ because of bias caused by model misspecification. Let ϕ^* be the expected value of $\hat{\phi}$.

Residual: $\widehat{e_t} = (Y_t - \mu) - \phi^*(Y_{t-1} - \mu)$

$$= \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + e_t - \phi^*(Y_{t-1} - \mu)$$

$$= (\phi_1 - \phi^*)(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + e_t$$

- The residuals do not estimate the white noise process as they would if the correct AR(2) model were used.
- The presence of $\phi_2(Y_{t-2} - \mu)$ in the residuals causes them to be autocorrelated.

Testing for whiteness of residuals

A standard assumption in econometrics modeling is that the white noise assumption:

$$E(\epsilon_t) = 0, E(\epsilon_s \epsilon_t) = \begin{cases} \sigma_\epsilon^2, if\ s = t \\ 0\ , if\ s \neq t \end{cases}$$

Any departure from whiteness indicates that the residuals still contain serial dependent information that the model has not extracted from the data.

A systematic way of checking the whiteness: if the SACF and SPACF of the residuals have no significant elements, we conclude that they resemble white noise; otherwise, there is still information in the residuals.

One problem with checking the significance of individual elements of any of the identification functions is that each element might be individually insignificant, but all (or a subset) of the elements taken together may be jointly significant.

Portmanteau test

A popular goodness of fit test is the *Box-Pierce Q-statistic*, also known as the *portmanteau test*, which tests the joint hypothesis

$$H_0 : \rho_{\epsilon,1} = \rho_{\epsilon,2} = \cdots = \rho_{\epsilon,m} = 0$$

The *Q*-statistic is computed by

$$Q = T \sum_{k=1}^{m} \hat{\rho}_{\epsilon,k}^2$$

The sum of squared autocorrelations is intended to capture deviations from zero in either direction and at all lags m.

For data generated by a white noise process, *Q* has an asymptotic chi-square (χ^2) distribution with $(m - p - q)$ degrees of freedom, where p and q refer to the number of model parameters. In the AR(1) case, $p + q = 1$.

Portmanteau test

Ljung and Box adapt the Box-Pierce test for finite sample by modifying the *Q*-statistic to obtain the *Q**-statistic, such that

$$Q^* = T(T+2)\sum_{k=1}^{m}(T-k)^{-1}\hat{\rho}_{\epsilon,k}^2$$

which constitutes the *Ljung-Box test*. However, for moderate sample sizes, also the Ljung-Box test has low power and may fail to detect model misspecifications.

Both versions of the portmanteau test check only for *uncorrelatedness* of the residuals and not for *independence* or "true" whiteness.

The detection of more complex temporal dependencies in the absence of autocorrelations indicates that the class of linear ARMA models is inappropriate for the data at hand.

.

Portmanteau test

Adapting the Ljung-Box test, *McLeod and Li test* the joint hypothesis on the ACF of squared residuals:

$$H_0 : \rho_{\epsilon^2,1} = \rho_{\epsilon^2,2} = \cdots = \rho_{\epsilon^2,m} = 0$$

by performing a *Q* test on the squared residuals:

$$Q_2^* = T(T+2)\sum_{k=1}^{m}(T-k)^{-1}\hat{\rho}_{\epsilon^2,k}^2$$

Under the null hypothesis of no autocorrelation, Q_2^* has a χ^2 distribution with m degrees of freedom.

Alternatively, a goodness of fit test based on residual partial autocorrelation can be used. If is the $\hat{\psi}_{\epsilon,k}$ is the *k*-th order residual partial autocorrelation coefficients, then the statistic

$$Q_M = T(T+2)\sum_{k=1}^{m}(T-k)^{-1}\hat{\psi}_{\epsilon,k}^2$$

is asymptotically χ^2 distributed with $(m-p-q)$ degrees of freedom if the model fitted is appropriate.

Testing for normality

In the common case of a normal assumption, the residuals have to be tested for normality.

The **Jarque-Bera test** accomplishes this. It is based on the third and fourth sample moments of the residuals. Since the normal distribution is symmetric,

the third moment, denoted by μ_3, should be zero;

and the fourth moment μ_4, should satisfy $\mu_4 = 3\sigma_4$.

The measure of third moment or skewness, $\hat{S}$, and the measure of fourth moment or kurtosis, $\widehat{K}$, can be calculated as

$$\hat{S} = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\epsilon}_t^3}{\hat{\sigma}^3}$$

$$\widehat{K} = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\epsilon}_t^4}{3\hat{\sigma}^4} - 1$$

JB test

The Jarque-Bera test tests the null hypothesis

$$H_0 : \frac{\mu_3}{\sigma^3} = 0 \; and \; \frac{\mu_4}{\sigma^4} - 3 = 0$$

The sample statistics

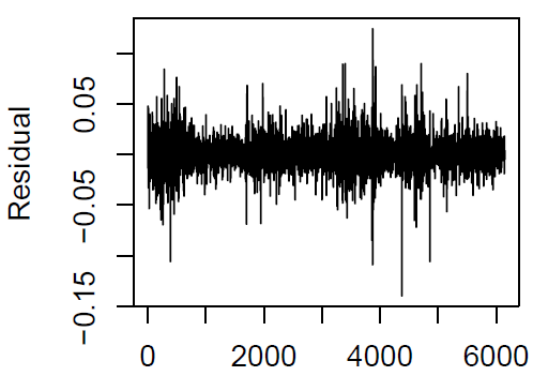
$$\lambda_1 = \frac{1}{6T} \sum_{t=1}^{T} \left(\frac{\hat{\epsilon}_t^3}{\hat{\sigma}^3} \right)^2 \text{ and } \lambda_2 = \frac{1}{24T} \sum_{t=1}^{T} \left(\frac{\hat{\epsilon}_t^4}{\hat{\sigma}^4} - 3 \right)^2$$

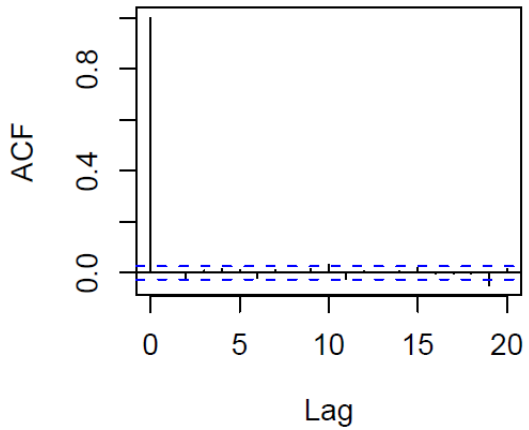
are asymptotically $\chi^2(1)$ distributed, respectively.

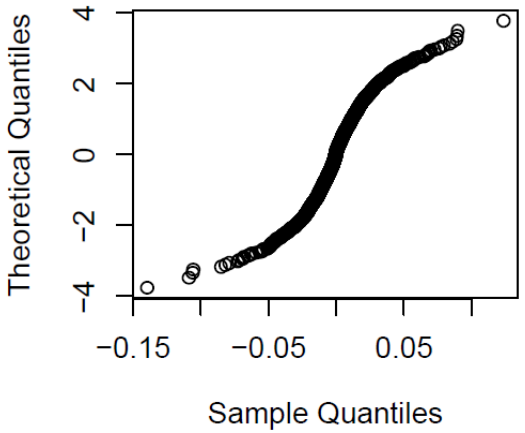
The null hypothesis, H_0, as stated above consists of a joint test for λ_1 and λ_2 being zero and can be tested via $\lambda_3 = \lambda_1 + \lambda_2$ which is asymptotically $\chi^2(2)$ distributed.

Example: BMW log returns

The sample ACF of the residuals is plotted. None of the autocorrelations at low lags is outside the test bounds.
A few at higher lags are outside the bounds, but this type of behaviour is expected to occur by chance or because, with a large sample size, very small but nonzero true correlations can be detected.







Box-Ljung test
data: residuals(fitAR1)
X-squared = 6.8669, df = 5, p-value = 0.1431

The large p-value indicates that we should accept the null hypothesis that the residuals are uncorrelated, at least at small lags. This is a sign that the AR(1) model provides an adequate fit. 4_BMW.R

Example: Inflation rate – AR(1) fit

One might try fitting an AR(1) to the changes in the inflation rate, since this series is stationary. However, the AR(1) model does not adequately fit the changes in the inflation rate.

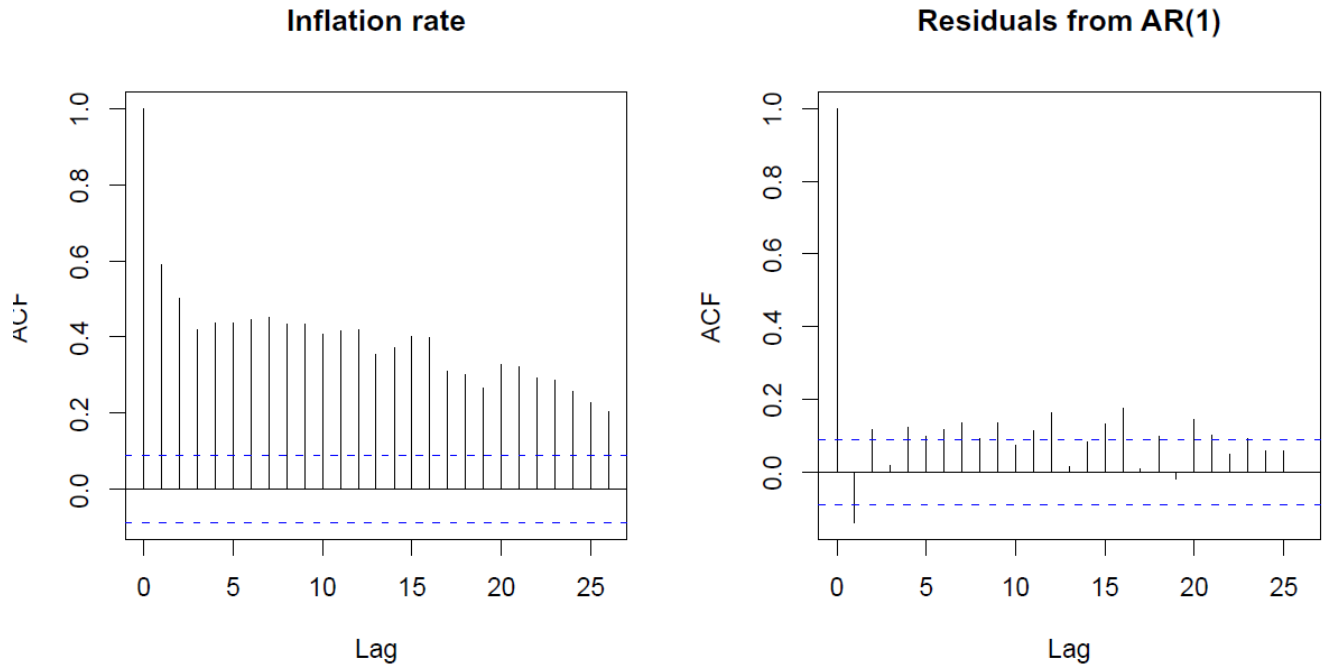


Fig. 9.8. *ACF of the inflation rate time series and residuals from an AR(1) fit.*

Box-Ljung test
data: fit$resid
X-squared = 46.1752, df = 12, p-value = 3.011e-06.
4_inflation.R

Forecasting

Let the data observed up to period t be collected in the information set

$$F_t = \{y_\tau : \tau \leq t\}$$

Having observed a time series up to period t, we would like to forecast a future value y_{t+h} for period $t + h, h = 1, \ldots$.

We distinguish between the one-step-ahead predictor, $\hat{y}_t(1)$ for y_{t+1}, and the multi-step-ahead predictor, $\hat{y}_t(h)$ for y_{t+h}, given the forecasting horizon h and forecasting origin t.

To characterize the forecasts, three quantities are needed:

- Forecast function $\hat{y}_t(h)$
- Forecast error $\hat{\epsilon}_t(h)$
- Variance of the forecast error

Forecasting: Loss Function

Instead of considering the "true cost" of wrong predictions, we consider a purely statistical criterion, the ***mean-squared prediction error (MSE)***. Given the information set, we can also define the *conditional expectation* of y_{t+h}:

$$E_t(y_{t+h}) \coloneqq E(y_{t+h}|F_t) = E(y_{t+h}|y_t, y_{t-1}, \dots)$$

We would like to find the estimate of y_{t+h}, $\hat{y}_t(h)$,which has the smallest possible MSE:

$$MSE\left(\hat{y}_t(h)\right) = \mathrm{E}\left[\left(y_{t+h} - \hat{\mathrm{y}}_\mathrm{t}(\mathrm{h})\right)^2\right] = E\left[\left(y_{t+h} - E_t(y_{t+h}) + E_t(y_{t+h}) - \hat{\mathrm{y}}_\mathrm{t}(\mathrm{h})\right)^2\right]$$

Squaring the expression in brackets and using the fact that

$$E\left[\left(y_{t+h} - E_t(y_{t+h})\right)\left(E_t(y_{t+h}) - \hat{\mathrm{y}}_\mathrm{t}(\mathrm{h})\right)\right] = 0$$

Then, we obtain

$$MSE\left(\hat{y}_t(h)\right) = MSE\left(E_t(y_{t+h})\right) + E[E_t(y_{t+h}) - \hat{\mathrm{y}}_\mathrm{t}(\mathrm{h})]^2$$

We see quantity $MSE\left(\hat{y}_t(h)\right)$ is minimized , if

$$\hat{y}_t(h) = E_t(y_{t+h})$$

Example: AR(1)

1-step ahead forecast at time t, the forecast origin:

$$\hat{y}_{t+1} = a_0 + a_1 y_t$$

1-step ahead forecast error: $\hat{\epsilon}_t(1) \coloneqq y_{t+1} - \hat{y}_{t+1} = \epsilon_{t+1}$ Thus, ϵ_{t+1} is the un-predictable part of y_{t+1}. It is the shock at time $t + 1$!

Variance of 1-step ahead forecast error: $Var[\hat{\epsilon}_t(1)] = Var(\epsilon_{t+1}) = \sigma_\epsilon^2$

2-step ahead forecast: $\hat{y}_{t+2} = a_0 + a_1 \hat{y}_{t+1}$

2-step ahead forecast error: $\hat{\epsilon}_t(2) \coloneqq y_{t+2} - \hat{y}_{t+2} = (a_0 + a_1 y_{t+1} + \epsilon_{t+2}) - (a_0 + a_1 \hat{y}_{t+1}) = \epsilon_{t+2} + a_1 \epsilon_{t+1}$

Variance of 2-step ahead forecast error: $Var[\hat{\epsilon}_t(2)] = Var(\epsilon_{t+2} + a_1 \epsilon_{t+1}) = (1 + a_1^2)\sigma_\epsilon^2$, which is greater than or equal to $Var[\hat{\epsilon}_t(1)]$, implying that uncertainty in forecasts increases as the number of steps increases.

h-step ahead forecast: $\hat{y}_{t+h} = a_0 + a_1 \hat{y}_{t+h-1}$

h-step ahead forecast error: $\hat{\epsilon}_t(h) \coloneqq y_{t+h} - \hat{y}_{t+h} = (a_0 + a_1 y_{t+h-1} + \epsilon_{t+h}) - (a_0 + a_1 \hat{y}_{t+h-1}) = \epsilon_{t+h} + a_1 \epsilon_{t+h-1} + \cdots + a_1^{h-1} \epsilon_{t+1}$

Variance of h-step ahead forecast error: $Var[\hat{\epsilon}_t(h)] = \sigma_\epsilon^2 \sum_{k=0}^{h-1} a^{2k}$

Forecasting: Prediction Intervals

To assess the uncertainty associated with this prediction, we compute the confidence or prediction interval. The distributions of $\hat{y}_t(h)$ and prediction error $\hat{\epsilon}_t(h)$ are determined by the distribution of the ϵ_t .

Let $z_{(\alpha)}$ denote $\alpha \times 100\%$ quantile of standard normal distribution. We have $z_{(\alpha/2)} = -z_{(1-\alpha/2)}$. Then

$$
\begin{aligned}
1-\alpha &= \Pr\left\{-z_{(1-\alpha/2)} \le \frac{\hat{\epsilon}_t(h)}{\sigma_\epsilon(h)} \le z_{(1-\alpha/2)}\right\} \\
&= \Pr\left\{-z_{(1-\alpha/2)} \le \frac{y_{t+h} - \hat{y}_t(h)}{\sigma_\epsilon(h)} \le z_{(1-\alpha/2)}\right\} \\
&= \Pr\left\{\hat{y}_t(h) - z_{(1-\alpha/2)}\sigma_\epsilon(h) \le y_{t+h} \le \hat{y}_t(h) + z_{(1-\alpha/2)}\sigma_\epsilon(h)\right\}
\end{aligned}
$$

Interval

$$\hat{y}_t(h) \pm z_{(1-\alpha/2)}\sigma_\epsilon(h)$$

is called the $(1-\alpha) \times 100\%$ ***h-step prediction interval***.

Usually, for α the values of 0.05 or 0.01 are chosen.

Autoregressive models of order p

An ***autoregressive model*** is a regression model that is based on ***an weighted average of prior values*** in the series, ***weighted*** according to a regression on lagged version of the series.

The autoregressive model of lag p, written as AR(p) is:

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots + \phi_p\left(Y_{t-p} - \mu\right) + e_t \text{ or}$$

$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t, \text{ where } \delta = \left(1 - \phi_1 - \phi_2 - \cdots - \phi_p\right)\mu$$

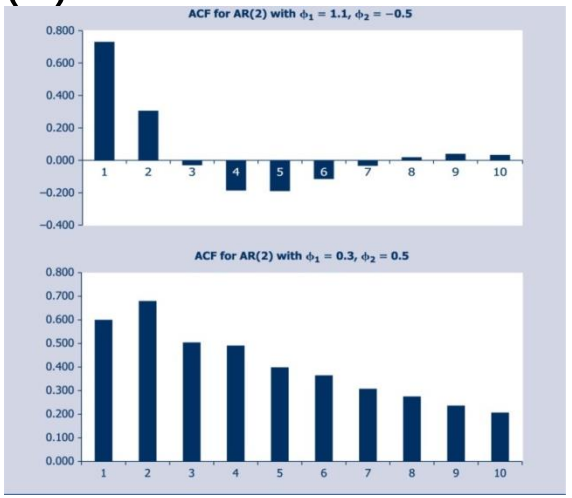
Assumptions regarding the error term are the same as before: Zero mean, constant variance, and mutually uncorrelated.

Most of the concepts we have discussed for AR(1) models generalize easily to AR(p) models.

AR(2) model

An autoregressive time series of order $p = 2$, or AR(2) model:

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$$



Properties of AR(2):

- Mean: $E(Y_t) = \frac{\delta}{1-\phi_1-\phi_2}$
- ACF: $\rho_0 = 1, \rho_1 = \frac{\phi_1}{1-\phi_2}, \rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2},$ $k \geq 2$.
- Forecasts: Similar to AR(1) models
- Stationarity condition: The AR(2) model is stationary, if and only if the characteristic equation: $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$

$$|\phi(z)| \neq 0 \qquad when \ |z| \leq 1$$

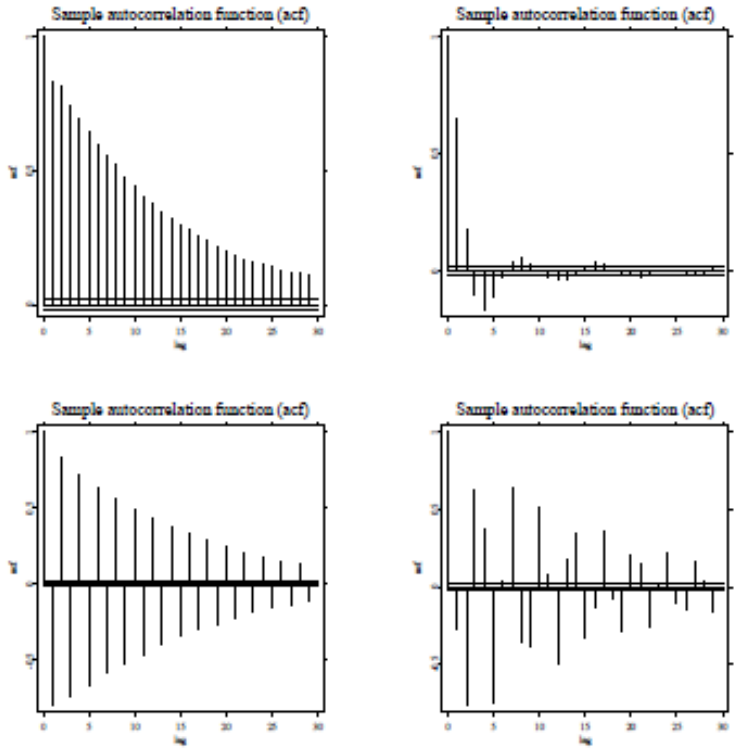
i.e. all roots including complex roots of $\phi(z)$ lie outside the unit circle.

The condition is equivalent to:

$$\phi_2 + \phi_1 < 1, \qquad \phi_2 - \phi_1 < 1, \qquad -1 < \phi_2 < 1.$$

Proof tip: $\rho_1 = \phi_1 + \phi_2\rho_1, |\rho_1| < 1, |z_1 z_2| > 1$

Sample autocorrelations of an AR(2) process



ACF of a AR(2) process with ($\phi_1 = 0.5,\ \phi_2 = 0.4$) (upper left), ($\phi_1 = 0.9, \phi_2 = -0.4$) (upper right), ($\phi_1 = -0.4, \phi_2 = 0.5$) (lower left) and ($\phi_1 = -0.5, \phi_2 = -0.9$) (lower right).

How to identify AR model and its order?

Partial autocorrelations

AR(p) model has a distinct rubric: Its partial autocorrelations for lag order higher than p are zero: $\pi_{kk} = 0, for\ all\ k > p.$

A **partial *auto*correlation** is the amount of correlation between a variable and a lag of itself that is not explained by correlations at all *lower-order*-lags. In other words, it measures the dependence between Y_t and Y_{t+k} after correcting Y_t and Y_{t+k} for the linear influence of the variables $Y_{t+1}, \cdots, Y_{t+k-1}$

$$\pi_{kk} = Corr(Y_t - \boldsymbol{\mathcal{P}}(Y_t|Y_{t+1}, \cdots, Y_{t+k-1}), Y_{t+k} - \boldsymbol{\mathcal{P}}(Y_{t+k}|Y_{t+1}, \cdots, Y_{t+k-1}))$$

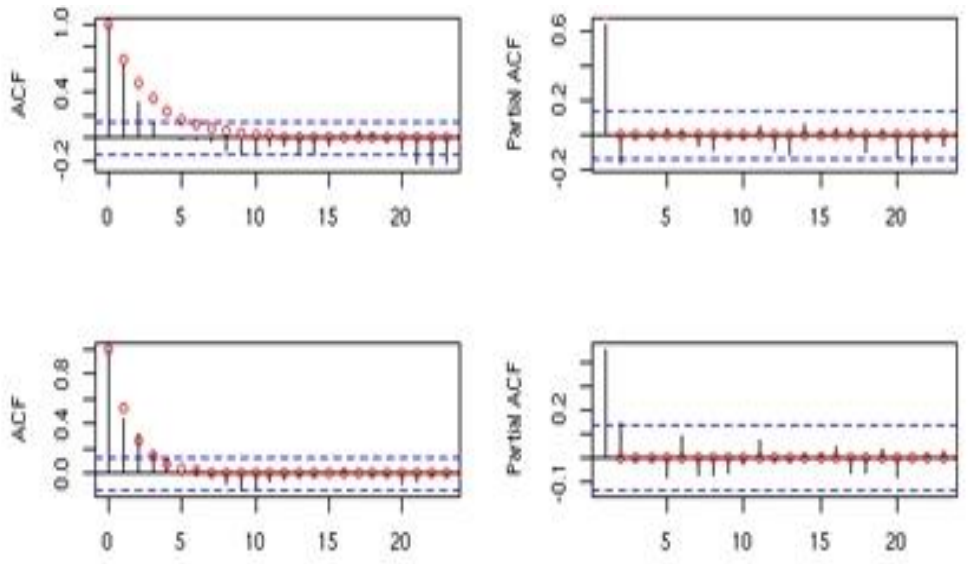
where $\boldsymbol{\mathcal{P}}(W|Z)$ is the 'best linear projection' of W on Z.

The partial autocorrelations at all lags can be computed by fitting a succession of autoregressive models with increasing numbers of lags. In particular, the partial autocorrelation at lag k is equal to the estimated AR(k) coefficient in an autoregressive model with k terms--i.e., a multiple regression model in which Y is regressed on lag-1, lag-2, etc., up to lag-k.

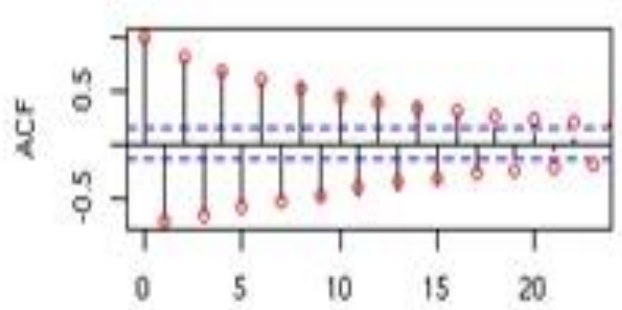
ACF and PACF of AR(1) models

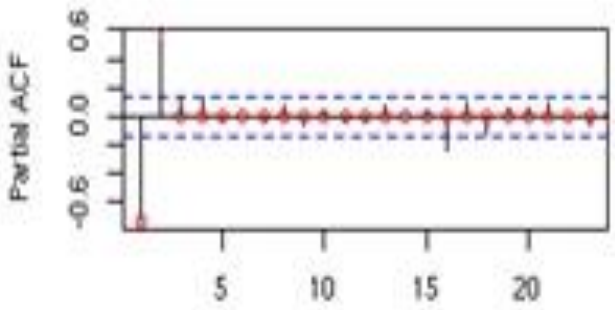
PACF and ACF of order $k = 1$ are identical: $\pi_{11} = \rho_1$

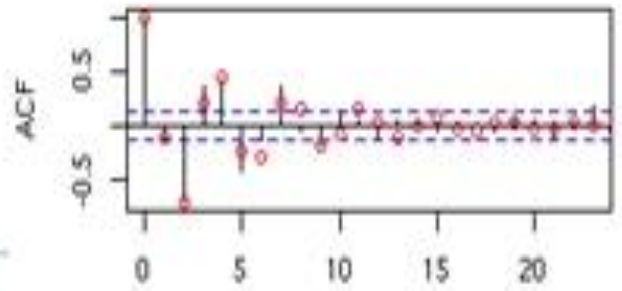
By observing correlogram and sample PAC plot can help to select model type. Moreover, by mere inspection of the PACF we can determine how many AR terms are needed to explain the autocorrelation pattern in a time series: if the partial autocorrelation is significant at lag *k* and not significant at any higher order lags, i.e., if the PACF "cuts off" at lag *p*, then this suggests that you should try fitting an autoregressive model of order *p. Note that sample PACF is also asymptotically normal distributed with SD of* $1/\sqrt{T}$.

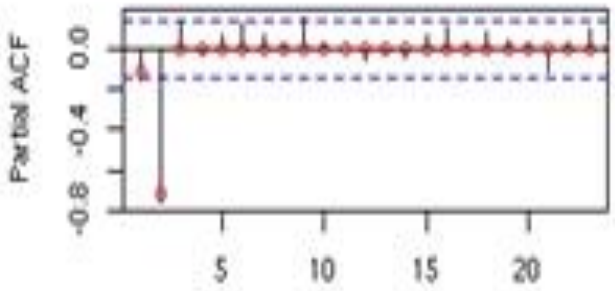


ACF and PACF of AR(2) models









PACF cuts off at lag 2, while the ACF decays slowly and may have significant values at ' igher lags. We say nat the series robably displays n "AR signature“ ith order 2.

Yule-Walker equation

For an AR(p) process $Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + \varepsilon_t$ we have

$$\gamma_k = a_1 \gamma_{k-1} + a_2 \gamma_{k-2} + \cdots + a_p \gamma_{k-p}, \qquad k = 0,1,2, \ldots$$

which carries over to the ACF, namely,

$$\rho_k = a_1 \rho_{k-1} + a_2 \rho_{k-2} + \cdots + a_p \rho_{k-p}, \qquad k = 0,1,2, \ldots$$

These relations are called ***Yule-Walker equations***.

Using sample autocorrelations and collecting the first p equations in matrix form we obtain

$$\begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \hat{\rho}_3 \\ \vdots \\ \hat{\rho}_{p-1} \\ \hat{\rho}_p \end{bmatrix} = \begin{bmatrix} 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{p-1} \\ \hat{\rho}_1 & 1 & & \hat{\rho}_{p-2} \\ \hat{\rho}_2 & \hat{\rho}_1 & & \hat{\rho}_{p-3} \\ \vdots & & \ddots & \vdots \\ \hat{\rho}_{p-2} & & & \hat{\rho}_1 \\ \hat{\rho}_{p-1} & \hat{\rho}_{p-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{p-1} \\ a_p \end{bmatrix}$$

or, in short $\hat{\boldsymbol{\rho}}_{\boldsymbol{p}} = \hat{\boldsymbol{T}}\boldsymbol{a}.$

Yule-Walker estimation

YW estimator is then given by

$$\hat{\boldsymbol{a}} = \hat{\boldsymbol{T}}^{-1}\hat{\boldsymbol{\rho}}_{\boldsymbol{p}}$$

If SACF is estimated by

$$\hat{\rho}_k = \frac{\sum_{i=k+1}^{T}(y_i - \hat{\mu})(y_{i-k} - \hat{\mu})}{\sum_{i=1}^{T}(y_i - \hat{\mu})^2}$$

all roots of the YW-estimated AR polynomial will be $1 - \hat{a}_1 L - \cdots - \hat{a}_p L^p$ *greater than unity*.

Let's estimate an *AR*(2) model, where we have over-identified YW:

$$\begin{bmatrix} 1 & \hat{\rho}_1 \\ \hat{\rho}_1 & 1 \\ \hat{\rho}_2 & \hat{\rho}_1 \\ \hat{\rho}_3 & \hat{\rho}_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \hat{\rho}_3 \\ \hat{\rho}_4 \end{bmatrix}$$

The over-identified YW estimator is then given by the least squares solution:

$$\hat{\boldsymbol{a}} = \left(\hat{\boldsymbol{T}}_{\boldsymbol{4}}'\hat{\boldsymbol{T}}_{\boldsymbol{4}}\right)^{-1}\hat{\boldsymbol{T}}_{\boldsymbol{4}}'\hat{\boldsymbol{\rho}}_{\boldsymbol{4}}$$

Extended Yule-Walker equations

In the case of a pure MA process, simplifies to

$$\gamma_k = \begin{cases} \sigma^2 \sum_{j=k}^{2} b_j b_{j-k} & if\ k = 0, 1, \ldots, q \\ 0 & if\ k > q \end{cases}$$

More generally, for an *ARMA*(*p*,*q*) process,

$$\gamma_k = a_1\gamma_{k-1} + a_2\gamma_{k-2} + \cdots + a_p\gamma_{k-p}, \qquad k = q+1, q+2, \ldots$$

or

$$\rho_k = a_1\rho_{k-1} + a_2\rho_{k-2} + \cdots + a_p\rho_{k-p}, \qquad k = q+1, q+2, \ldots$$

The latter recursions are sometimes called ***extended Yule-Walker equations***.

Another application of YW equation: estimate partial autocorrelation function

To compute the PACF, letting a_{kl} denote the *l*-th autoregressive coefficient of an *AR*(*k*) process, that is,

$y_t = a_{k1}y_{t-1} + a_{k2}y_{t-2} + \cdots + a_{kk-1}y_{t-(k-1)} + a_{kk}y_{t-k} + \epsilon_{k,t}$ then,

$\alpha_k = a_{kk}\,, k = 1, 2, \ldots$

The *k* Yule-Walker equations for the ACF, give rise to the system of linear equations

$$\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & & \rho_{k-2} \\ \rho_2 & \rho_1 & & \rho_{k-3} \\ \vdots & & \ddots & \vdots \\ \rho_{k-2} & & & \rho_1 \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_{k1} \\ a_{k2} \\ a_{k3} \\ \vdots \\ a_{k,k-1} \\ a_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_{k-1} \\ \rho_k \end{bmatrix}$$

or, in short $\boldsymbol{P}_k\boldsymbol{a}_k = \rho_k\,, k = 1, 2, \ldots.$

Estimate PACF

Sample Partial Autocorrelation Function

To estimate the sample PACF (SPACF), we follow the procedure for computing the theoretical PACF described earlier, but replace theoretical autocorrelations, ρ_i, by their estimates, $\hat{\rho}_i$:

$$\hat{a}_{kk} = \frac{|\hat{P}_k^*|}{|\hat{P}_k|}, k = 1, 2, \ldots$$

where P_k^* replaces the k-th column of P_k by $(\rho_1, \ldots, \rho_k)'$. **Cramer's rule**.

$$\boldsymbol{P}_k = \begin{bmatrix} 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{k-2} & \hat{\rho}_{k-1} \\ \hat{\rho}_1 & 1 & & \hat{\rho}_{k-3} & \hat{\rho}_{k-2} \\ \hat{\rho}_2 & \hat{\rho}_1 & & \hat{\rho}_{k-4} & \hat{\rho}_{k-3} \\ \vdots & & \ddots & \vdots & \vdots \\ \hat{\rho}_{k-2} & & & 1 & \hat{\rho}_1 \\ \hat{\rho}_{k-1} & \hat{\rho}_{k-2} & \cdots & \hat{\rho}_1 & 1 \end{bmatrix} \text{ and } \boldsymbol{P}_k^* = \begin{bmatrix} 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{k-2} & \hat{\rho}_1 \\ \hat{\rho}_1 & 1 & & \hat{\rho}_{k-3} & \hat{\rho}_2 \\ \hat{\rho}_2 & \hat{\rho}_1 & & \hat{\rho}_{k-4} & \hat{\rho}_3 \\ \vdots & & \ddots & \vdots & \vdots \\ \hat{\rho}_{k-2} & & & 1 & \hat{\rho}_{k-1} \\ \hat{\rho}_{k-1} & \hat{\rho}_{k-2} & \cdots & \hat{\rho}_1 & \hat{\rho}_k \end{bmatrix}$$

From the Yule-Walker equations it is evident that $|\boldsymbol{P}_k^*| = 0$ for an AR process whose order is less than *k*.

Estimate PACF

A computationally more efficient procedure for estimating the SPACF is the following recursion for $k = 1, 2, ...$

$$\hat{a}_{kk} = \frac{\hat{\rho}_k - \sum_{l=1}^{k-1} \hat{a}_{k-1,l}\hat{\rho}_{k-l}}{1 - \sum_{l=1}^{k-1} \hat{a}_{k-1,l}\hat{\rho}_k}$$

$$\hat{a}_{kl} = \hat{a}_{k-1,l} - \hat{a}_{kk}\hat{a}_{k-1,k-l}, l = 1, 2, \dots, k$$

with $\hat{a}_{ij} = 0$, for $i, j < 1$. For large samples and values of k sufficiently large, the PACF is approximately normally distributed with variance $Var(\hat{a}_k) \approx \frac{1}{T}$. The 95% confidence interval can be approximated by $\pm 2/\sqrt{T}$.

Impact of forecast error

***Considering forecast error* helps for *forecasting* exchange rates.**

With the increasing globalization and liberalization of economies various corporate and firms are expanding their business operations overseas. These organizations usually encounter the risk of currency exposure and need to forecast the exchange rates to hedge against the risk of exchange rate fluctuation. Not only this, corporate needs to forecast exchange rates to take decisions regarding short term investments, long term investments, short term and long term financing and other capital budgeting decisions.

However the exchange rates forecasted are seldom accurate and it is this aspect which gives rise to **forecast error.** The forecast errors are more frequent in such periods where there was a lot of fluctuation in the currency rates. Potential forecasted errors have a great impact on the financial position of the firm, it is due to this that the corporate need to examine and calculate the degree of impact of such potential errors on the financial positions before taking up any financial decision.

Moving average model of order 1

The moving average model of lag 1, written as MA(1) is:

$$Y_t = \mu + e_t + \theta_1 e_{t-1}$$

Current value of Y_t can be found from past shocks/error, plus a new error (e_t).The time series is regarded as a moving average (unevenly weighted, because of different coefficients) of a random error series e_t. Assumptions regarding the error term are the same as before: Zero mean, constant variance σ_e^2, and mutually uncorrelated.

- If θ_1 is zero, Y depends purely on the error or shock (e) at the current time, and there is no serial dependence
- If θ_1 is large, previous errors influence the value of Y_t much.
- If our model successfully captures the dependence structure in the data then the residuals should look random.

Property of MA(1)

The MA(1) process: $Y_t = \mu + e_t + \theta_1 e_{t-1}$ is stationary!

$$E(Y_t) = \mu$$
$$var(Y_t) = \gamma_0 = \left(1 + {\theta_1}^2\right)\sigma_e^2$$
$$cov(Y_t, Y_{t-1}) = \gamma_1 = \theta_1 \sigma_e^2$$
$$\gamma_k = 0, \qquad k > 1$$
$$Corr(Y_t, Y_{t-1}) = \rho_1 = \frac{\theta_1}{1 + {\theta_1}^2}$$
$$Corr(Y_t, Y_{t-k}) = \rho_k = 0 \quad k > 1$$

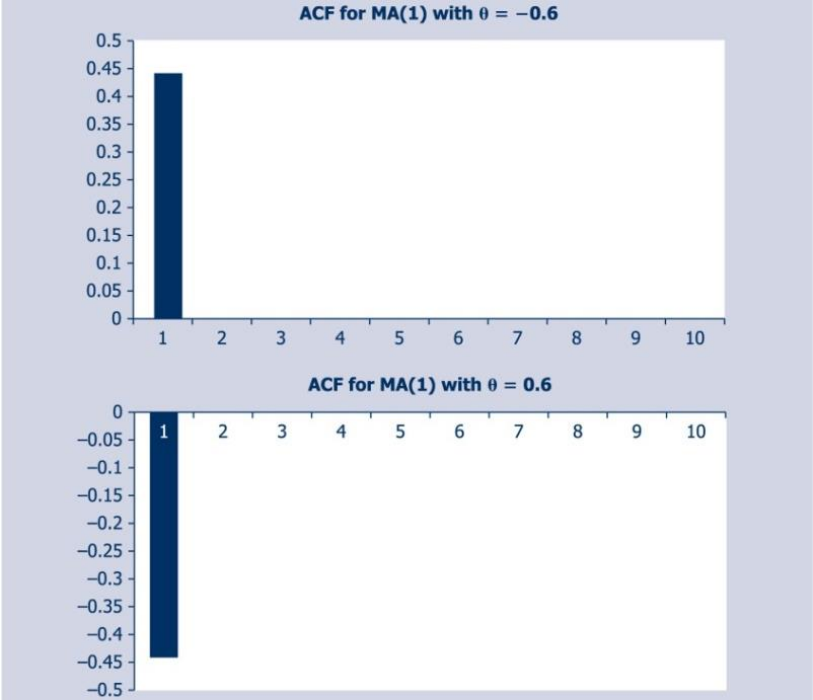
- Finite memory! MA(1) models do not remember what happens two time periods ago. The autocorrelation function has a cut-off after lag 1.
- **Invertibility Condition**: θ_1 can be solved from the equation: $\theta_1^2 - \theta_1/\rho_1 + 1 = 0$. If θ_1 is a solution, so is $\frac{1}{\theta_1}$. Thus it requires $|\theta_1| < 1$.
- What's the link between the AR and MA models? The MA model can be reformulated as an AR(∞). Given MA(1): $Y_t = e_t + \theta_1 e_{t-1}$ and $Y_{t-1} = e_{t-1} + \theta_1 e_{t-2} \Rightarrow e_{t-1} = Y_{t-1} - \theta_1 e_{t-2}$, we have

  $$Y_t = e_t + \theta_1 Y_{t-1} - \theta_1^2 e_{t-2} = e_t + \theta_1 Y_{t-1} - \theta_1^2 Y_{t-2} + \theta_1^3 e_{t-3} = \cdots$$

  Thus the PACF of MA(1) is infinite in contents but damps out geometrically. Its signs alternate if $\theta_1 < 0$.

MA(1) model

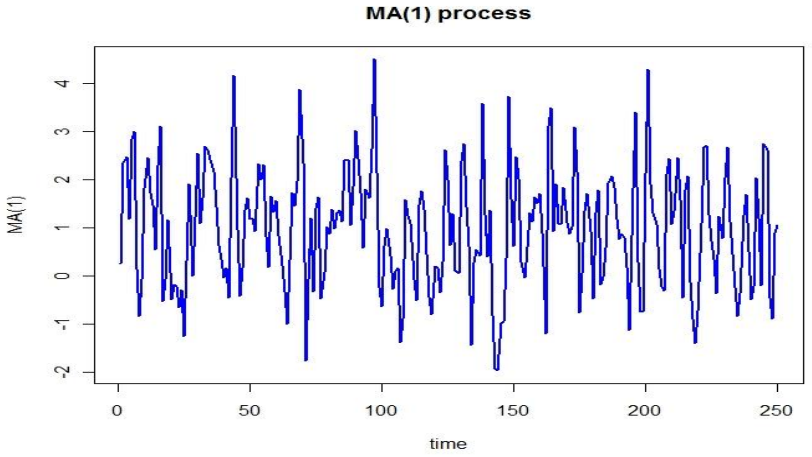
Moving Average models have a simple ACF structure.

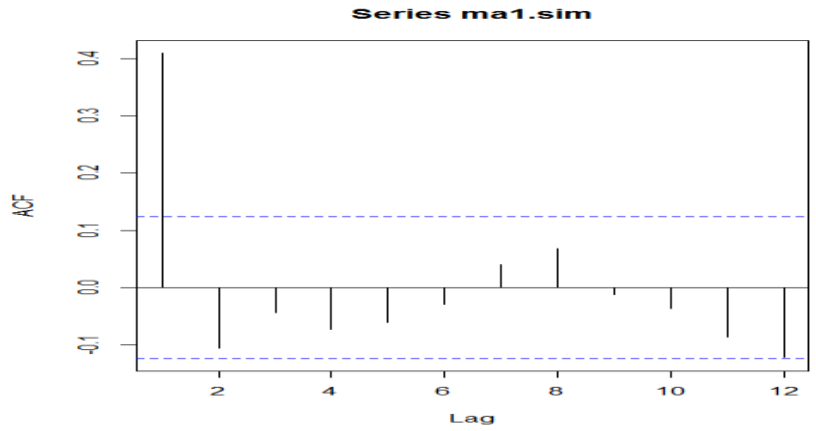


The MA(1) models have nonzero autocorrelations only for k=1.

A simulated MA model of lag 1 and its sample ACF:

$$Y_t = \mu + e_t + \theta e_{t-1}$$





Moving average models

The moving average model of lag q, written as MA(q) is:

$$Y_t = \mu + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}$$

Assumptions regarding the error term are the same as before: Zero mean, constant variance σ^2, and mutually uncorrelated.

- The MA models are stationary!

$$E(Y_t) = \mu$$

$$var(Y_t) = (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma_e^2$$

$$\rho_k = Corr(Y_t, Y_{t-k}) = 0 \quad k > q, \ldots$$

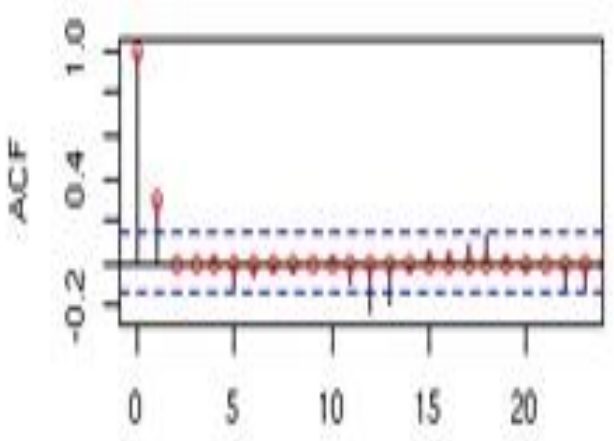
- The MA(q) models have nonzero autocorrelations for $k = 1, \ldots, q$. If the corrlogram "cuts off" at lag *k*, then this suggests that we should try fitting an moving average model of order *q.*
- Invertibility conditions: The MA(q) model is invertible if and only if

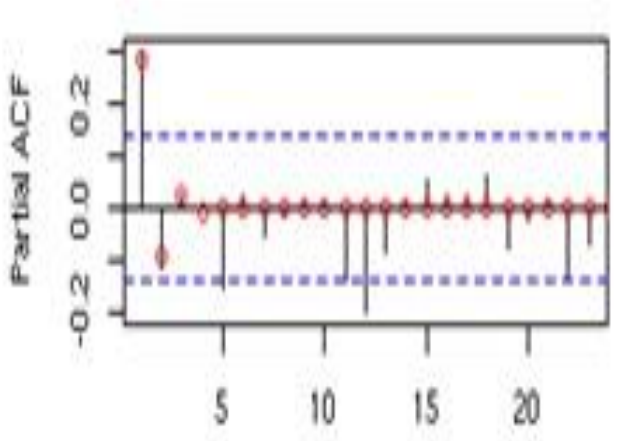
$$|\theta(z)| \neq 0 \qquad when\ |z| \leq 1$$

i.e. all roots including complex roots of $\theta(z)$ lie outside the unit circle.

- PACF: Infinite in contents.

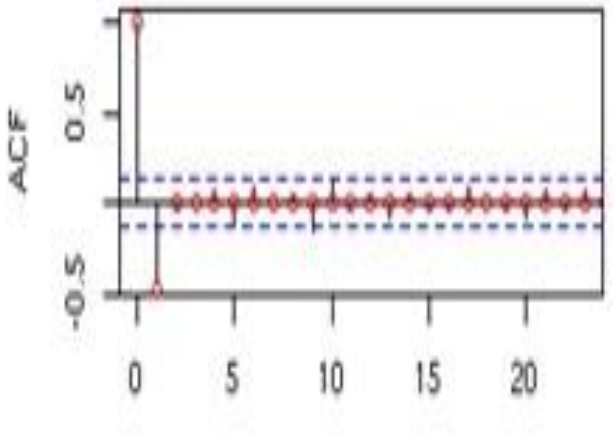
ACF and PACF of MA(1) model

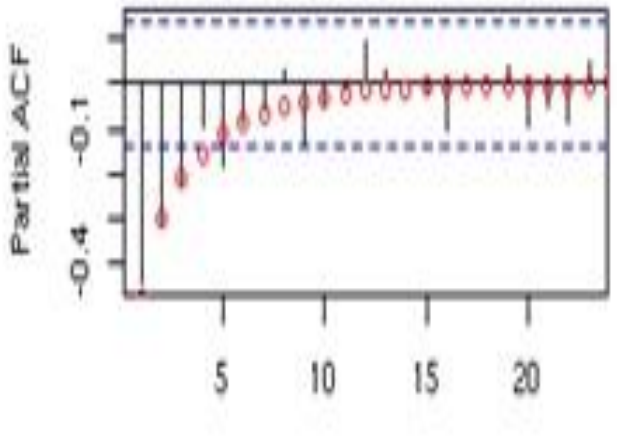




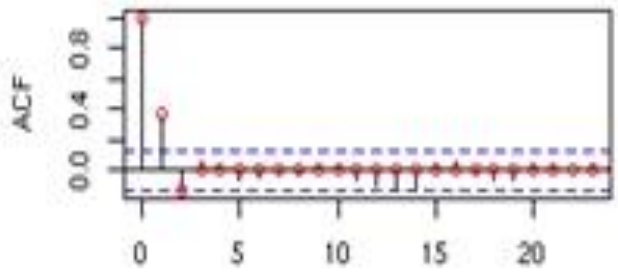
ACF: cuts off at lag 1.

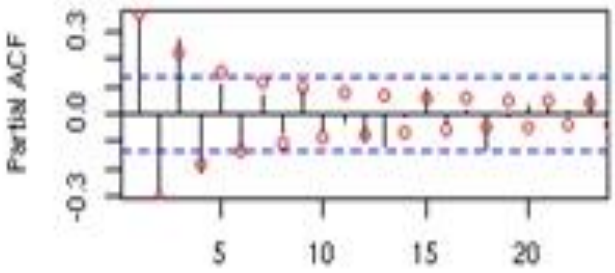
PACF: declines over time.



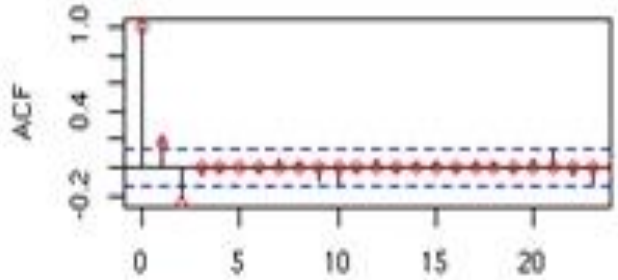


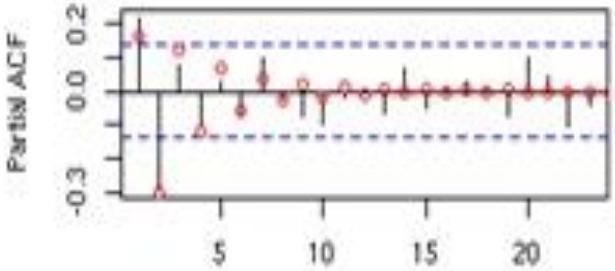
ACF and PACF of MA(2) model

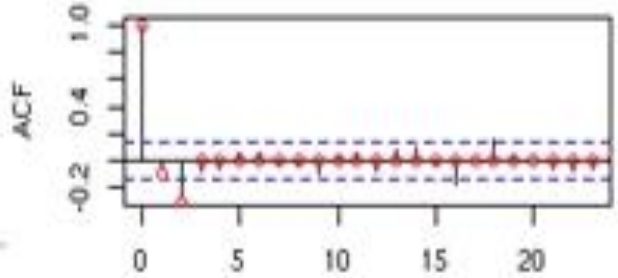


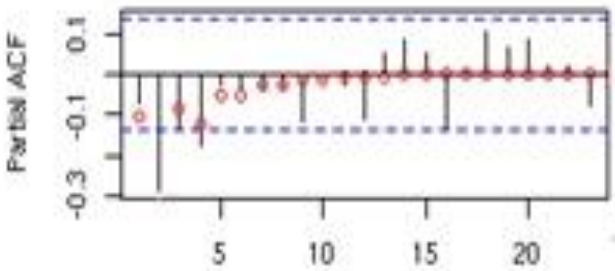


The lag at which the ACF cuts off is the indicated number of possible MA terms.



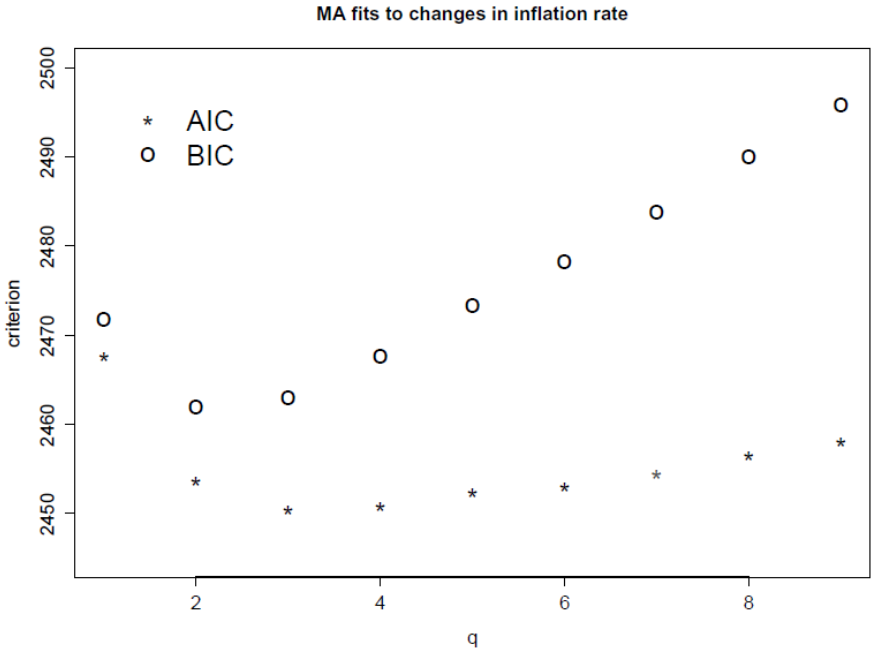






Example: Inflation rate – MA(q) fit

The auto.arima function in R's forecast package found that q = 3 is the first local minimum of AIC, while the first local minimum of BIC is at q = 2.



Call:
arima(x = diff(x), order = c(0, 0, 3))

Coefficients:

	ma1	ma2	ma3	intercept
	-0.632950	-0.102734	-0.108172	-0.000156
s.e.	0.046017	0.051399	0.046985	0.020892

Thus, if an MA model is used, then only two or three MA parameters are needed. This is a strong contrast with AR models, which require far more parameters.

ARMA(1,1) models

A model may have both autoregressive and moving average components. **Autoregressive Moving Average (ARMA)** model is an extension of AR model, where the future values depend on both the historical values (AR part) and the past forecast errors (MA part).

The ARMA(1,1) model has the form: $Y_t = \delta + \phi Y_{t-1} + e_t + \theta e_{t-1}$

Just a combination of MA and AR terms.

Reformulation:

$$Y_t = \delta + \phi Y_{t-1} + e_t + \theta e_{t-1} = \delta + \phi(\delta + \phi Y_{t-2} + e_{t-1} + \theta e_{t-2}) + e_t + \theta e_{t-1} = \cdots$$
$$= A_1 + B_1 Y_{t-k} + e_t + C_1 e_{t-1} + \cdots + C_k e_{t-k}$$
$$Y_t = \delta + \phi Y_{t-1} + e_t + \theta e_{t-1} = \delta + \phi Y_{t-1} + e_t + \theta(Y_{t-1} - \delta - \phi Y_{t-2} - \theta e_{t-2}) = \cdots$$
$$= A_2 + D_1 Y_{t-1} + \cdots + D_m Y_{t-m} + e_t + E_1$$

Where A.,B.,C.,D. and E. are coefficients of the AR and MA terms.

Property of ARMA(1,1) process

From the ARMA(1,1) model: $Y_t = \delta + \phi Y_{t-1} + e_t + \theta e_{t-1}$, we obtain

$$cov(Y_t, e_t) = E(Y_t e_t) = \sigma_e^2,$$

since e_t is independent of e_{t-1} and Y_{t-1}.

Multiplying Y_t on both sides and taking expectation, we have

$$var(Y_t) = \gamma_0 = \phi^2\gamma_0 + (1+\theta^2)\sigma_e^2 + 2\phi\theta\sigma_e^2 \Rightarrow \gamma_0 = (1+\theta^2+2\phi\theta)\sigma_e^2/(1-\phi^2)$$

Similarly, we obtain

$$\rho_1 = \frac{(1+\phi\theta)(\phi+\theta)}{1+\theta^2+2\phi\theta}$$

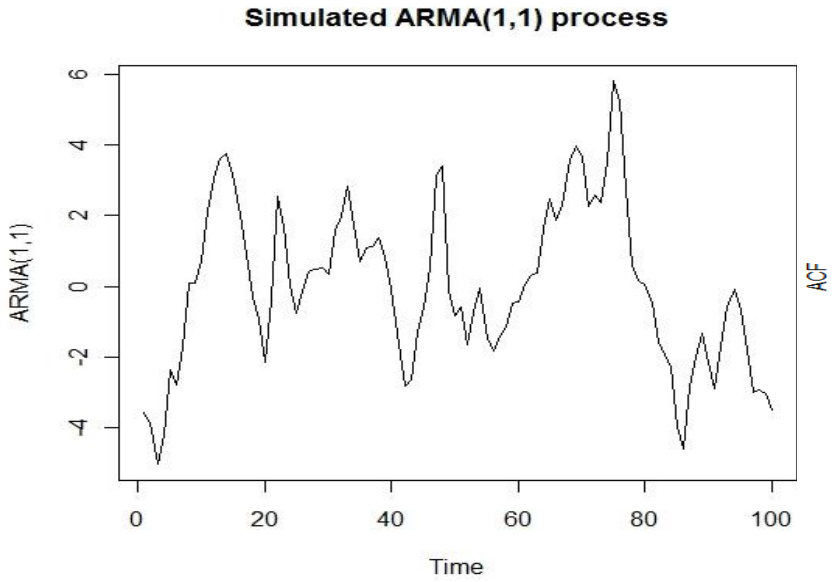
For $k \geq 2$, multiply Y_{t-k} on both sides and take expectation

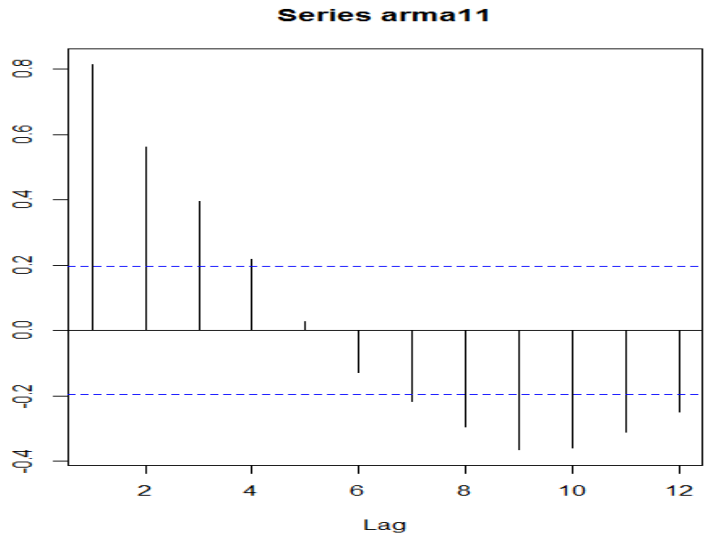
$$\rho_k = \phi\rho_{k-1}, \qquad k \geq 2.$$

After one lag the ACF of an ARMA(1,1) process decays in the same way as the ACF of an AR(1) process with the same ϕ.

ARMA(1,1) model

A simulated ARMA(1,1) model $Y_t = 0.75Y_{t-1} + e_t + 0.75e_{t-1}$ and its sample ACF.





ARMA(p,q) models

The ARMA(p,q) model has the form:

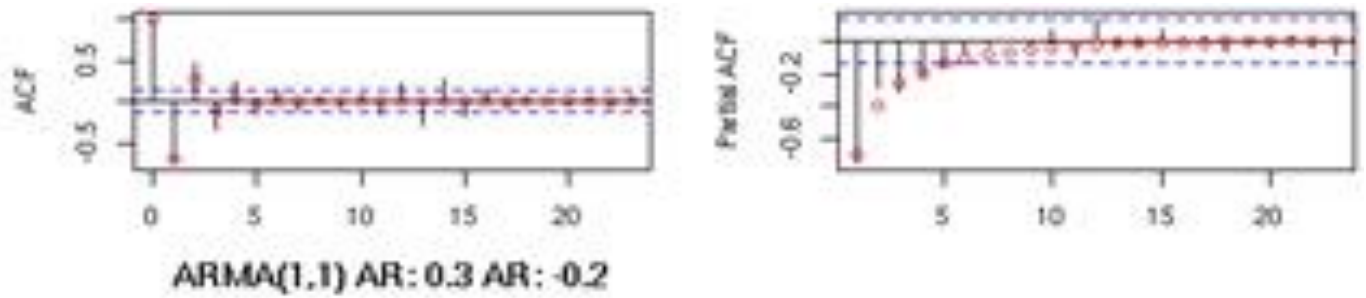
$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots \phi_p Y_{t-p} + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}.$$

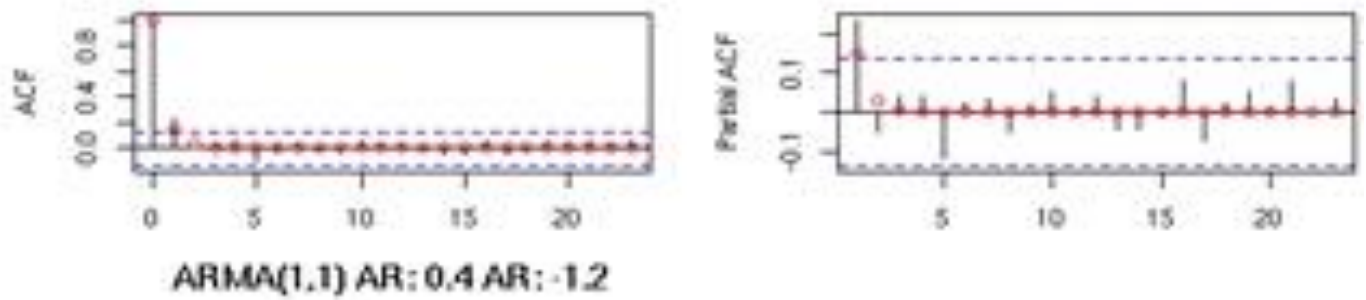
where p indicates the order of the lagged values (AR part) and q refers to the order of the past errors (MA part). By including MA part, the process learns from the error made over time and tries to improve forecast accuracy in the future.

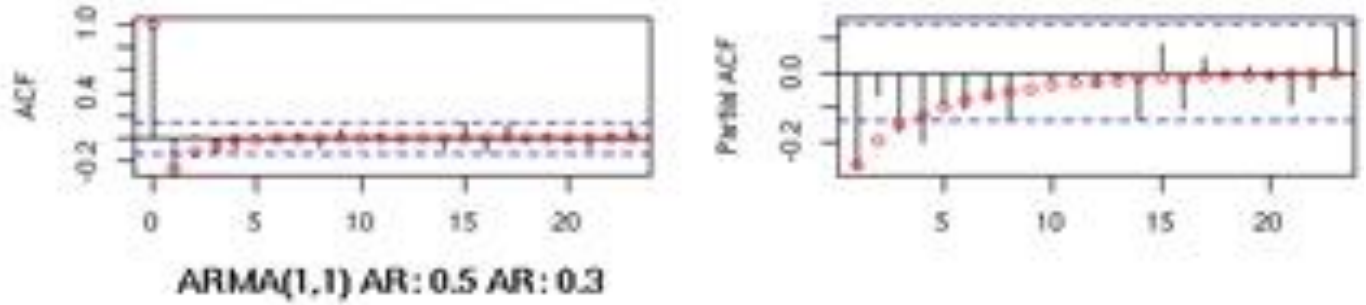
AR(p) can be written as ARMA(p,0).

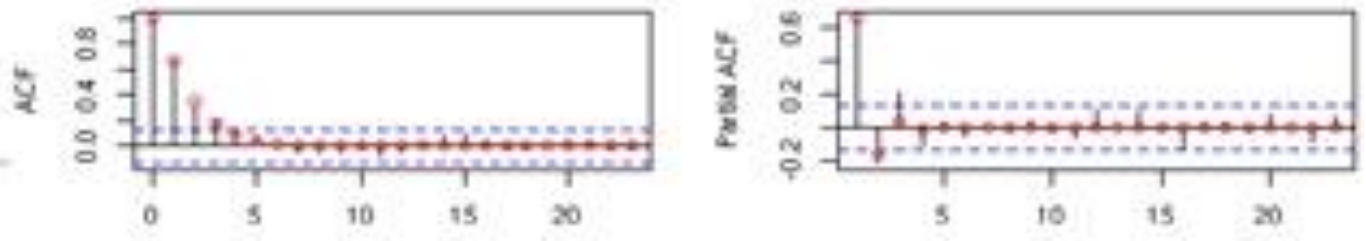
MA(q) can be represented as ARMA(0,q).

ACF and PACF of ARMA(1,1) models









For the ARMA(1,1), both the ACF and the PACF exponentially decrease.

Much of fitting ARMA models is guess work and trial-and-error!

Identification of lag orders: visual inspection

Identification of lag orders: visual inspection of SACF and SPACF

Process	ACF	PACF
White Noise	All 0	All 0
AR(1)	$\rho_s = \phi^s$	0 beyond lag 2
AR(P)	Decays toward zero exponentially	Non-zero through lag P, 0 thereafter
MA(1)	$\rho_1 \neq 0, \rho_s = 0, s > 1$	Decays toward zero exponentially
MA(Q)	Non-zero through lag Q, 0 thereafter	Decays toward zero exponentially
ARMA(P,Q)	Exponential Decay	Exponential Decay

Estimation. By relating the sample autocorrelations and partial autocorrelations to the ACF and PACF of ARMA models, candidates may be identified. The selected model is estimated and its residuals are tested for randomness using the Q test statistics on the residual ACFs. If significance of autocorrelations is detected in the residuals, a new model, normally with lagged values at higher order is considered and the procedure is repeated.

Two principles: forecast accuracy, and parsimony.

Identification of lag orders: model selection criteria

Let the residuals of an estimated *ARMA*(*p,q*) model be denoted by $\hat{\epsilon}_t\,(p,q)$. The estimate of the corresponding residual variance, denoted by $\hat{\sigma}^2_{p,q}$, is

$$\hat{\sigma}^2_{p,q} = \frac{1}{T}\sum_{i=1}^{T} \hat{\epsilon}^2_t(p,q)$$

Larger models tend to fit in-sample better. However, if we use too many parameters we fit noise and obtain poor forecasting capabilities. This phenomenon is called ***overfitting***.

In the extreme, we could achieve a perfect fit by fitting a "model" that has as many parameters as observations. Such models *overfit* the data by also capturing non-systematic features contained in the data. In general, overparameterized models tend to be unreliable.

Log-likelihood:

$$-\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\hat{\sigma}^2_{p,q} - \frac{1}{2\hat{\sigma}^2_{p,q}}\sum_{t=1}^{T}\hat{e}_t{}^2$$

Model selection criteria

Several model-selection criteria attempting to overcome the overparameterization problem have been proposed in the literature:

1. The ***Akaike Information Criterion (AIC)*** is given by

$$AIC_{p,q} = \ln \hat{\sigma}^2_{p,q} + \frac{2}{T}(p+q)$$

   The (*p,q*)-combination that minimizes the AIC should be selected. However, this criterion may give more than one minimum, depends on assumption that the data are normally distributed and tends to overparameterize.

2. The ***Bayesian Information Criterion (BIC)*** is given by

$$BIC_{p,q} = \ln \hat{\sigma}^2_{p,q} + \frac{\ln T}{T}(p+q)$$

   This criterion imposes a more severe penalty for each additional parameter and thereby tends to select lower-order models than the AIC.

Model selection criteria

3. The ***Corrected Akaike Information Criterion (AICC)*** given by

$$AICC_{p,q} = \ln \hat{\sigma}^2_{p,q} + \frac{2}{T-p-q-2}(p+q+1)$$

attempts to correct the bias of the AIC that is causing the overparameterization problem and is especially designed for small samples. For small sample sizes, it tends to select different models.

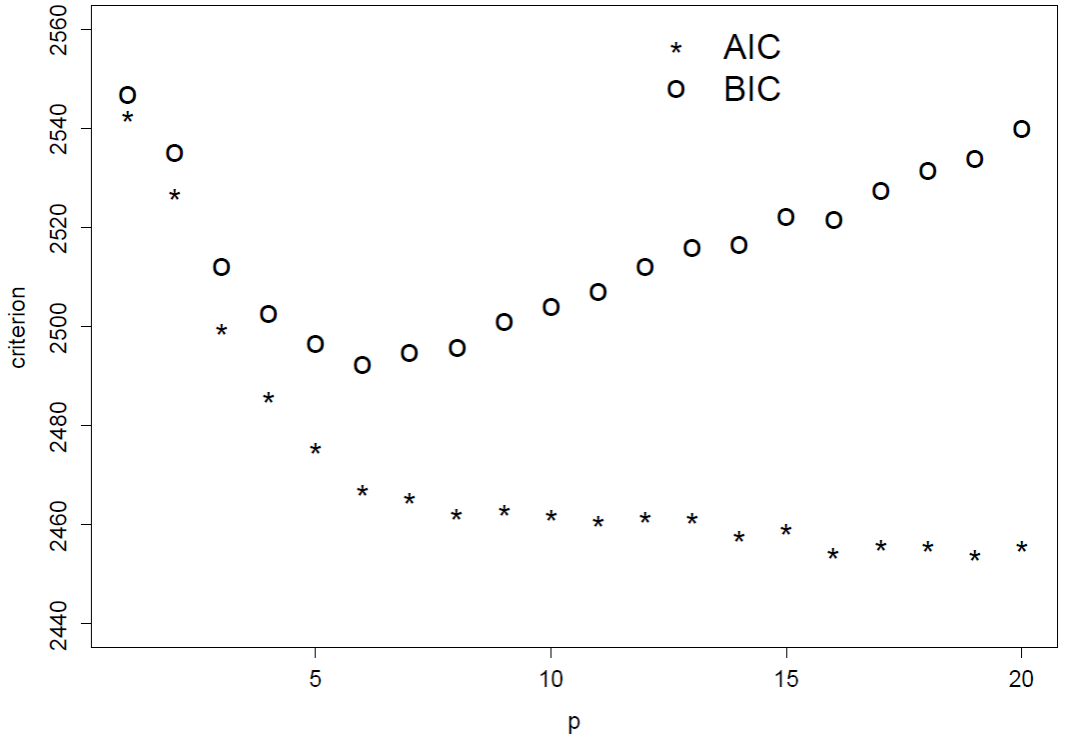
- Let *k* be the number of estimated parameters of a model as recommended by an information criterion. Due to the different penalty terms we have $k_{AIC} \geq k_{AICC} \geq k_{BIC}$.

Model selection criteria

- The BIC is ***strongly consistent*** in selecting the orders of a process; namely, it determines the true model asymptotically.
- In contrast, AIC will always determine an overeparameterized model, independent of the length of the sample.
- In practice, use of information criteria should be viewed as supplementary guidelines to assist in the model selection process rather than as a main model selection criteria.
- There may be several models that produce criterion values that are very close to the minimum value. All reasonable models should remain candidates for the final selection and be subjected to further diagnostic checks (for example, a test for whiteness of the residuals).

Example: Inflation rate – AR(p) fit

The auto.arima function in R's forecast package found that p = 8 is the first local minimum of AIC, while the first local minimum of BIC is at p = 6.



> auto.arima(diff(x),max.p=10,max.q=0,ic="aic")
> auto.arima(diff(x),max.p=10,max.q=0,ic="bic")
4_inflation.R

Example: Inflation rate – AR(7) fit

Here are the results for p =7.

```
Series: x
ARIMA(7,0,0) with non-zero mean

Coefficients:
        ar1    ar2     ar3    ar4    ar5    ar6    ar7  intercept
      0.366  0.129  -0.020  0.099  0.065  0.080  0.119       3.99
s.e.  0.045  0.048   0.048  0.048  0.049  0.048  0.046       0.78

sigma^2 estimated as 8.47:  log-likelih
AIC = 2462   AICc = 2522   BIC = 2467
```

4_inflation.R

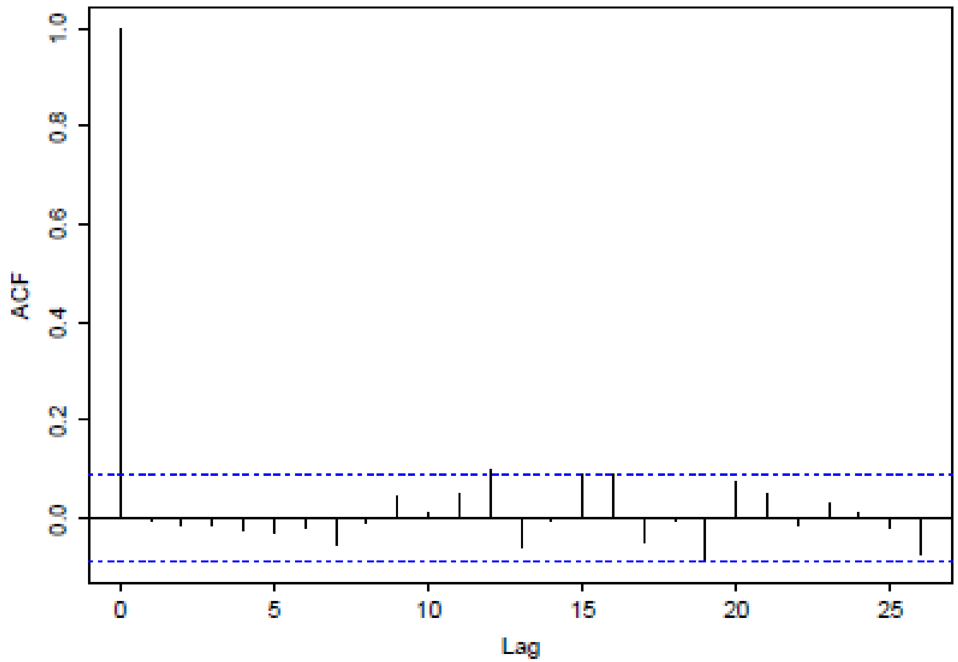
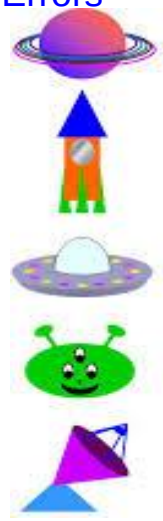


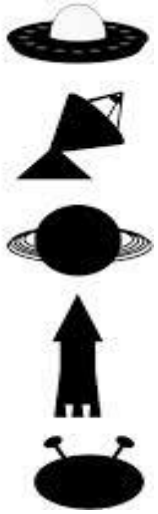
Fig. 9.11. *ACF of residuals from an AR(7) fit to the inflation rates.*

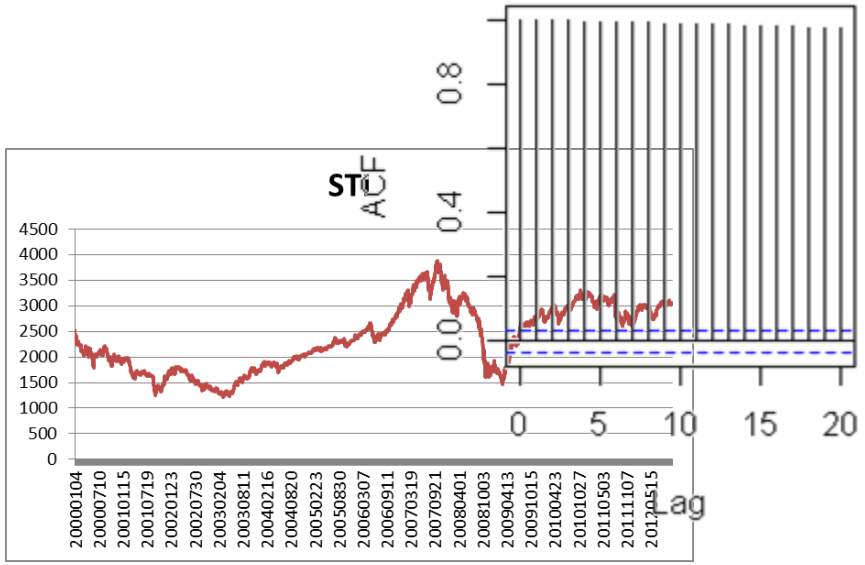
Time series models

Population
Properties
True dynamics
Errors



Sample
Sample ACFs
Model
Residuals





Given data, we conduct statistical analysis to discover:

Static distributional properties such as sample mean, sample variance

Serial dependence analysis such as correlogram (sample ACFs)

(Relationship on exogenous variables)

We then select the model that matches data's characteristics

Q

Model type: time series models AR(p), MA(q) or ARMA(p,q)?

Order: select lag order p and q?

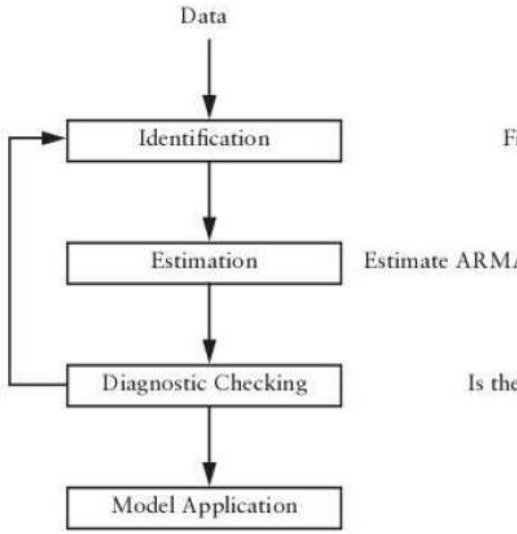
Box-Jenkins procedure

1. The purpose of the *identification step* in the Box-Jenkins approach is to first determine the autoregressive order, *p*, and the moving average order, *q*. These initial guesses are typically not final values for these orders, but they are used to specify one or more tentative (competing) models.

2. Given values for *p* and *q* from the identification step, the parameters of the *ARMA*(*p,q*) model are derived in the *estimation step*. It delivers estimates of the ARMA coefficients for ARMA model formulations and the residual variance σ^2.

3. In the *diagnostic-checking step* we examine the adequacy of the estimated model. Here, the main objective is to determine whether or not all relevant information has been "extracted“ from the data. If so, the residuals should resemble white noise.

4. Forecast: A comparison of the forecasting performance of alternative models over several post sample periods may help to find the most suitable model.



Which model should we choose?

That depends on the assumptions we are comfortable making with respect to the data. The model with lower order is essentially easy to understand and interpret. It is also fairly optimistic about the accuracy with which it can forecast with less information necessary. As a general rule in this kind of situation, I would recommend choosing the model with the lower order, other things being roughly equal.