Content I

Content II

Content III

Two quotes at the beginning

'I have seen the future and it is very much like the present, only longer.'
– Kehlog Albran, The Profit
This nugget of pseudo-philosophy is actually a concise description of statistical forecasting. We search for statistical properties of a time series that are constant in time–levels, trends, seasonal patterns, correlations and autocorrelations, etc. We then predict that those properties will describe the future as well as the present

'Prediction is very difficult, especially if it's about the future.'
– Nils Bohr, Nobel laureate in Physics
This quote serves as a warning of the importance of validating a forecasting model out-of-sample. It's often easy to find a model that fits the past data well–perhaps too well!–but quite another matter to find a model that correctly identifies those patterns in the past data that will continue to hold in the future.

Why time series analysis and models? I

Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.

Time series models are used for many applications such as Economic Forecasting, Sales Forecasting, Budgetary Analysis, Stock Market Analysis, Yield Projections, Process and Quality Control, Inventory Studies, Workload Projections, Utility Studies, Census Analysis, etc.

The aims are

- to get knowledge about the DGP underlying the time series data in order for example to get reliable forecasts both for the mean as for the variance (e.g. VaR),

Why time series analysis and models? II

- to estimate patterns like trends, seasonality, or cyclical components as economic cycles,
- to test appropriately (i.e. using correct standard errors) economic hypothesis, for example with dynamic models, error correction models, regression analysis with cointegrated time series, ...

Example

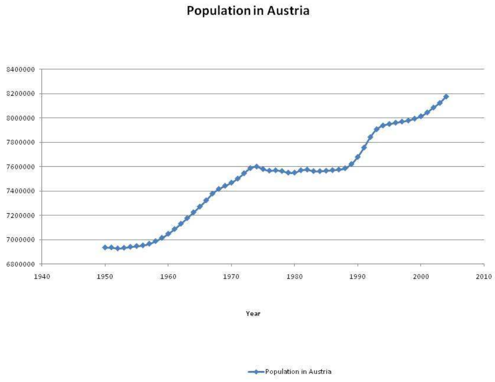


Figure: Population statistics of Austria, source: EUROSTAT, February 28, 2006

Example

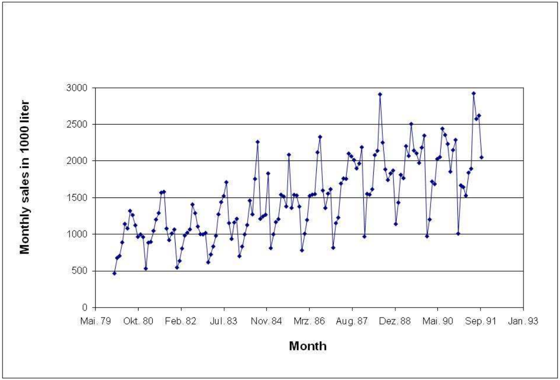


Figure: Red wines sales of vintner in Australia, January 1980 till October 1991

Example

Monthly unemployment rates in Tyrol from Jan. 1995 till Jan. 2007; computed as ratio of unemployed people divided by the sum of unemployed people and wage earners of age 14 up to 65 years.

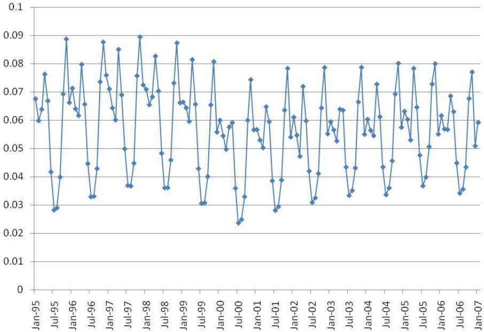


Figure: Manpower capacity in Tyrol, source: statistical bureau of the province Tyrol (Landesstatistik)

General considerations

- The sequence of observations, $\{y_t\}_{t=-\infty}^{t=\infty}$ is a time series process which is characterized by its time ordering and ist systematic correlation between observations in the sequence.
- Let $\{y_t\}_{t\in T}$ be the observed time series. A time series model regards the observations $\{y_t\}$ as realizations of random variables $\{Y_t\}_{t\in T}$ and specifies the joint probability distribution (or at least first and second moments).
- Joint distribution of the random variables $\{Y_t\}_{t\in T}$ is $F(y_{i_1}, y_{i_2}, \ldots, y_{i_n}) = P(Y_{i_1} \leq y_{i_1}, Y_{i_2} \leq y_{i_2}, \ldots, Y_{i_n} \leq y_{i_n})$, for $i_1, i_2, \ldots, i_n \in T$ and $n \in \mathbb{N}$.
  Often one specifies first and second moments of the joint distribution of $\{Y_t\}_{t\in T}$, $E(Y_t)$ and $Cov(Y_t, Y_s)\ \forall s, t \in T$.
- If all involved distributions are multivariate normal distributions then the joint distribution is uniquely determined by the first and second moments. Therefore the second-order properties describe the underlying stochastic process completely.

Strong stationarity I

- Typically a time-series is **one realization** (one path) of the underlying stochastic process.
- Even if we know that the stochastic process is subject to the normal distribution, given n observations theoretically
$$n + \frac{n(n+1)}{2}$$
parameters have to be estimated.
- Assuming the same distribution over time the number of parameters is reduced to $n + 1$.

Strong stationarity II

- Further restrictions reduce the number of parameters till the stochastic process gets estimable.
- If the stochastic process has identical distributions over time (not only identical marginal distributions also all joint distributions are identical) it is called stationary process.

A time series process, $\{y_t\}_{t=-\infty}^{t=\infty}$ is strongly stationary if the joint probability distribution of any set of k observations in the sequence $\{y_t, y_{t+1}, ..., y_{t+k}\}$ is is the same regardless of the origin, t, in the time scale.

Weak stationarity

A stochastic process y_t is weakly stationary or covariance stationary if it satisfies the following requirements:

1. $E[y_t]$ is independent of t
2. $\text{Var}[y_t]$ is a finite, positive constant, independent of t.
3. $\text{Cov}[y_t, y_{t-k}]$ is a finite function of k, but not of the absolute location of either observation on the time scale.

Weak stationarity is obviously implied by strong stationarity (if $E(y_t^2) < \infty \; \forall t$), although it requires less because the distribution can, at least in principle, be changing on the time axis.

Autocorrelations of a stationary stochastic process

For $\{y_t\}_{t \in T}$ let $E(y_t^2) < \infty, \forall t \in T$.

- The mean function of $\{y_t\}$ is defined as

$$\mu_{y_t} = E(y_t), \quad t \in T$$

- The autocovariance function is called λ_k and defined as

$$\lambda_k = Cov[y_t, y_{t-k}] = E[(y_t - \mu_{y_t}) \cdot (y_{t-k} - \mu_{y_{t-k}})]$$

- The autocorrelation function, or ACF, is obtained by dividing λ_k by the variance λ_0, to obtain

$$\rho_k = \frac{\lambda_k}{\lambda_0} \text{ with } -1 \leq \rho_k \leq 1$$

.

Example: Is *i.i.d.* noise weakly stationary?

Let u_t be $i.i.d.$ noise with $E(u_t) = 0,\ E(u_t^2) = \sigma^2 < \infty$, then:

1. $\mu_{u_t} = 0$ and is independent of t
2. Due to the independence of u_t

$$\lambda_k = \begin{cases} \sigma^2 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases}$$

Therefore u_t is weakly stationary.

Example: Is white noise weakly stationary?

Let u_t be a sequence of random variables with the following characteristics:

1. $\mu_{u_t} = 0$ and therefore independent of t,
2. $Var[u_t] = \sigma^2$ for all t,
3. $Cor[u_r, u_s] = 0$ for all $r \neq s$,

Therefore u_t is weakly stationary with ACVF:

$$\lambda_k = \begin{cases} \sigma^2 & \text{for } k = 0 \\ 0 & \text{for } k \neq 0 \end{cases}$$

and is called white noise with expectation value 0 and variance σ^2.

Notation: $u_t \sim WN(0, \sigma^2)$

Note: $u_t \sim i.i.d.(0, \sigma^2) \Rightarrow u_t \sim WN(0, \sigma^2)$, not vice versa! (cf. examples)

Example: An uncorrelated but not independent sequence of random variables

Let u_t be a sequence of $i.i.d.N(0,1)$ random variables and let y_t be defined as:

$$y_t = \begin{cases} u_t & \text{if } t \text{ even} \\ \frac{u_{t-1}^2 - 1}{\sqrt{2}} & \text{if } t \text{ odd} \end{cases}$$

Compute $E(y_t)$, $Var(y_t)$, $Cov(y_t, y_{t+1})$.

Calculate $E(y_{t+1}|y_1, y_2, \ldots, y_t)$ for even t and odd t. Show that y_t is $WN(0,1)$ but not $i.i.d.(0,1)$.

Example, cont.

Expectation:

- t odd: $E(y_t)$

$$E\left(\frac{u_{t-1}^2 - 1}{\sqrt{2}}\right) = \frac{1-1}{\sqrt{2}} = 0$$

- t even: $E(y_t)$

$$E(u_t) = 0$$

Variance:

- t odd: $E(y_t^2)$

$$E\left(\left(\frac{u_{t-1}^2 - 1}{\sqrt{2}}\right)^2\right) = E\left(\frac{u_{t-1}^4 - 2u_{t-1}^2 + 1}{2}\right) = \frac{3-2+1}{2} = 1$$

- t even: $E(y_t^2)$

$$E(u_t^2) = 1$$

Example, cont.

Covariance: y_i's are independent with one exception:

- t even: $Cov(y_t, y_{t+1})$

$$E\left(u_t \cdot \frac{u_t^2 - 1}{\sqrt{2}}\right) = E\left(\frac{u_t^3 - u_n}{\sqrt{2}}\right) = \frac{0-0}{\sqrt{2}} = 0$$

Therefore the conditions for white noise are fulfilled $Cov(y_t, y_{t+1}) = 0 \rightarrow y_t \sim WN$, but not all y_i are independent of each other:

- t even: $E(y_{t+1}|y_1, y_2, \dots, y_t)$

$$E(y_{t+1}|y_1, y_2, \dots, y_t) \Rightarrow E\left(\frac{u_t^2 - 1}{\sqrt{2}}|u_1, u_2, \dots, u_t\right) = \frac{u_t^2 - 1}{\sqrt{2}}$$

Is the Random Walk weakly stationary?

Let $u_t \sim i.i.d.(0, \sigma^2)$.

The Random Walk y_t with mean 0 is defined as:

$$y_t = \begin{cases} 0 & \text{for } t = 0 \\ \sum_{i=1}^{t} u_i & \text{for } t > 0 \end{cases}$$

Note that $\Delta y_t = y_t - y_{t-1} = u_t$.

$E(y_t) = 0$ for all t, but $\lambda_0 = Var(y_t) = Var(\sum_{i=1}^{t} u_t) = T \cdot \sigma^2$ and therefore not independent of t. y_t is not weakly stationary.

Random Walk

The Random Walk is defined by:

$$y_t = y_{t-1} + u_t$$

Note $\Delta y_t = y_t - y_{t-1} = u_t$.

$$\begin{aligned} E(y_t|y_0) &= y_0 \\ Var(y_t|y_0) &= E((y_t - E(y_t|y_0))^2|y_0) \\ &= E((u_t + u_{t-1} + u_{t-2} + \ldots + u_1)^2|y_0) \\ &= t \cdot \sigma^2 \end{aligned}$$

Therefore, y_t is not weakly stationary.

Random walk with drift

The random walk wit drift ist defined as

$$y_t = \beta_0 + y_{t-1} + u_t$$

$$\begin{aligned} E(y_t|y_0) &= \beta_0 \cdot t + y_0 \\ Var(y_t|y_0) &= E((y_t - E(y_t|y_0))^2|y_0) \\ &= E((u_t + u_{t-1} + u_{t-2} + \ldots + u_1)^2|y_0) \\ &= t \cdot \sigma^2 \end{aligned}$$

y_t is not weakly stationary.

Note: Generally a random walk is defined as a sum of *i.i.d.* random variables, the distribution is chosen accordingly to the problem (e. g. symmetric random walk).
Note: $\Delta y_t = y_t - y_{t-1} = \beta_0 + u_t$. Compare $y_t = \beta_0 + \beta_1 t + u_t$ and $\Delta y_t = \beta_1 + u_t - u_{t-1}$.

Ergodicity, I

Stationarity is a crucial characteristic for estimation purposes however we require another characteristic of a time series: Ergodicity.

In practice we do not observe multiple observations at a specific time ('Ensemble' of time series) we jut have one single path through time and additionally a finite number of observations. Therefore only a part of the underlying stochastic process is observable. That is exactly the complex of the problem of using a single time series in order to estimate the entire stochastic process.

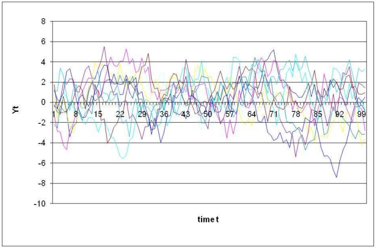


Figure: Ergodicity

Ergodicity, II

The question is which conditions suffice to make sure that by estimating the moments (mean, variance, covariance) just using a single realization (time series) of the process gives us results for the entire stochastic process.

- The property of ergodicity makes this approach possible (together with stationarity of course). Ergodicity ensures that the moments computed from one realization converge to the moments of the ensemble as T gets larger.
- For the mean this implies for example that the mean over time $\bar{y}_T = (1/T)\sum_{t=1}^{T} y_t$ equals the ensemble mean $\mu = E[y_t]$ as T goes to infinity.
- An ergodic process must be stationary but not each stationary process must be ergodic.

Ergodicity, III

Definition 19.3 Ergodicity

A strongly stationary time-series process $\{y_t\}_{t=-\infty}^{t=\infty}$ is ergodic if for any two bounded functions that map vectors in the a and b dimensional real vector spaces to real scalars, $f : \mathbb{R}^a \rightarrow \mathbb{R}^1$ and $g : \mathbb{R}^b \rightarrow \mathbb{R}^1$,

$$lim_{k\rightarrow\infty} \left|E[f(y_t, y_{t+1}, ..., y_{t+a})g(y_{t+k}, y_{t+k+1}, ..., y_{t+k+b})]\right|$$
$$= \left|E[f(y_t, y_{t+1}, ..., y_{t+a})]\right| \left|E[g(y_{t+k}, y_{t+k+1}, ..., y_{t+k+b})]\right| .$$

Theorem 19.1 The Ergodic Theorem

If $\{y_t\}_{t=-\infty}^{t=\infty}$ is a time series process that is strongly stationary and ergodic and $E[|y_t|]$ is a finite constant, and if $\bar{y}_T = (1/T)\sum_{t=1}^{T} y_t$, then $\bar{y}_T \rightarrow \mu$ almost surely, where $\mu = E[y_t]$.

Example

Using a fair coin the random variable X_t is defined as:

$$X_t = \begin{cases} 0 & \text{count} \\ 2 & \text{head} \end{cases}$$

Is X_t weakly stationary? Is this stochastic process ergodic?

- Stationarity:

$$E(X_t) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 2 = 1$$

$$Var(X_t) = E((X_t - E(X_t))^2) = \frac{1}{2}(0-1)^2 + \frac{1}{2}(2-1)^2 = 1$$

- Ergodicity: The mean over time depends on the corresponding path or on the realization of the random experiment: e. g. only counts then the mean is 0 or for just heads the mean is 2. Therefore this process is not ergodic.

Estimation of the first and second moments

Assume a stationary and ergodic process. Let $y_1, \ldots, y_T$ be the observations.

- The sample mean is defined as

$$\overline{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$$

- The sample autocovariance function is

$$\hat{\lambda}_k = \frac{1}{n} \cdot \sum_{t=1}^{T-|k|} (y_{t+|k|} - \overline{y}) \cdot (y_t - \overline{y}), \quad -T < k < T$$

Note: For $k \geq 0$ $\hat{\lambda}$ equals almost the sample covariance of the $(T-k)$ observations $(y_1, y_{1+k}), (y_2, y_{2+k}), \ldots, (y_{T-k}, y_T)$. The difference is in the use of T in the denominator and of the overall mean $\bar{y}$.

- The sample autocorrelation function (ACF, correlogramm):

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

- Sample autocovariance matrix:

$$\hat{\lambda} = \begin{pmatrix} \hat{\lambda}_0 & \hat{\lambda}_1 & \cdots & \hat{\lambda}_{T-1} \\ \hat{\lambda}_1 & \hat{\lambda}_0 & \cdots & \hat{\lambda}_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\lambda}_{T-1} & \hat{\lambda}_{n-2} & \cdots & \hat{\lambda}_0 \end{pmatrix}$$

Definition of an Autoregressive Process

Let $\{y_t\}$ be defined as

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \quad t \in T$$

where

- $\{u_t\} \sim WN(0, \sigma^2)$,
- $|\beta_1| < 1$ and
- $Cor(y_t, u_t) = 0$ for all $s < t$,

then $\{y_t\}$ is called an autoregressive process of order 1 or short $AR(1)$-process.

Characteristics of an AR(1), I

Insering y_{t-1} in y_t and so forth one obtains:

$$y_t = \beta_0(1 + \beta_1 + \beta_1^2 + \ldots) + (u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \ldots)$$

1. Expectation value of $\{y_t\}$

$$E(y_t) = \beta_0(1 + \beta_1 + \beta_1^2 + \ldots) = \frac{\beta_0}{1 - \beta_1}$$

(Hint: Sum of an infinite geometric series)

Characteristics of an AR(1), II

2. Variance of $\{y_t\}$
$y_t = \beta_0(1 + \beta_1 + \beta_1^2 + \ldots) + (u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \ldots)$

$$
\begin{aligned}
y_t - E(y_t) &= u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \ldots \\
E(y_t - E(y_t))^2 &= E(u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \ldots)^2 \\
Var(y_t) &= E(u_t^2 + \beta_1^2 u_{t-1}^2 + \beta_1^4 u_{t-2}^2 + \ldots + 2\beta_1 u_t u_{t-1} + \\
&\quad +2\beta_1^2 u_t u_{t-2} + \ldots) \\
Var(y_t) &= \frac{\sigma^2}{1 - \beta_1^2}
\end{aligned}
$$

Characteristics of an AR(1), III

3. Autocovariance function of $\{y_t\}$

$$\begin{aligned}
\lambda_1 &= Cov(y_t, y_{t-1}) \\
&= E\left[(y_t - E(y_t))(y_{t-1} - E(y_t))\right] \\
&= E\left[(u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \ldots)(u_{t-1} + \beta_1 u_{t-2} + \beta_1^2 u_{t-3} + \ldots \right. \\
&= \beta_1 \sigma^2 + \beta_1^3 \sigma^2 + \beta_1^5 \sigma^2 + \ldots \\
&= \beta_1 \frac{\sigma^2}{1 - \beta_1^2} = \beta_1 \lambda_0
\end{aligned}$$

$$\lambda_k = Cov(y_t, y_{t-k}) = \beta_1^{|k|} \lambda_0$$

Characteristics of an AR(1), IV

4. Autocorrelation function of $\{y_t\}$

$$\rho_k = \frac{\lambda_k}{\lambda_0} = \beta_1^{|k|}, \quad k \in Z$$

For stationary processes the ACF describes the time dependence of the time series $\{y_t\}$. If one assumes that the observations $\{y_1, \ldots, y_T\}$ are the realizations of a stationary process then the sample analogon of the ACF can be used to obtain hints for the choice of the appropriate model.

Definition: lag operator

The lag operator L is defined by

$$LX_t \equiv X_{t-1}$$

for each time series X_t.

Consequently it follows that

- $L^2X_t = L(LX_t) = LX_{t-1} = X_{t-2}$ or
- $L^jX_t = X_{t-j}\ \forall j \in \mathbb{N}$
- Further $L^jc = c\ \forall j \in \mathbb{N}$, $c \ldots$ constant.

Note: The lag operator is often called backward-shift operator and therefore abbreviated as B.

Simple notation of dynamic models

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \\
&\qquad + \gamma_0 X_t + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \cdots + \gamma_s X_{t-s} + U_t \\
Y_t &- \beta_1 Y_{t-1} - \beta_2 Y_{t-2} - \cdots - \beta_p Y_{t-p} = \\
&\qquad \beta_0 + \gamma_0 X_t + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \cdots + \gamma_s X_{t-s} + U_t \\
Y_t &- \beta_1 L Y_t - \beta_2 L^2 Y_t - \cdots - \beta_p L^p Y_t = \\
&\qquad \beta_0 + \gamma_0 X_t + \gamma_1 L X_t + \gamma_2 L^2 X_t + \cdots + \gamma_s L^s X_t + U_t \\
(1 &- \beta_1 L - \beta_2 L^2 - \cdots - \beta_p L^p) Y_t = \beta_0 + (\gamma_0 + \gamma_1 L + \gamma_2 L^2 + \cdots + \gamma_s L^s) X_t + U_t
\end{aligned}
$$

Define the two lag polynomials $\beta(L)$ of order p and $\gamma(L)$ of order s:

$$\beta_p(L) \equiv (1 - \beta_1 L - \beta_2 L^2 - \cdots - \beta_p L^p)$$

$$\gamma_s(L) \equiv (\gamma_0 + \gamma_1 L + \gamma_2 L^2 + \cdots + \gamma_s L^s)$$

Now the general dynamic model can be written as:

$$\beta_p(L) Y_t = \beta_0 + \gamma_s(L) X_t + U_t$$

Showing stationarity using the lag polynomial

Both lag polynomials, $\beta(L)$ and $\gamma(L)$, can be considered as polynomials in z, where $z \in \mathbb{C}$, with p or s roots.

A necessary and sufficient condition for stationarity is that all roots of the characteristic equation $\beta(z) = 0$, accordingly to the lag polynomial $\beta(L)$, lie outside the unit circle in $\mathbb{C}$.

Consider an $AR(1)$ process:

$$(1 - \beta_1 L)y_t = \beta_0 + u_t$$

For this example $\beta(L)$ is a lag polynomial of order 1: $\beta(L) = (1 - \beta_1 L)$. Lets compute the root:

$$\beta(z) = 0 \rightarrow (1 - \beta_1 z) = 0 \rightarrow z = \frac{1}{\beta_1}$$

As the root has to lie outside the unit circle to guarantee stationarity ($|z| > 1$), we get $|\beta_1| < 1$ (cf. definition of AR(1)).

$AR(p)$ process

Let y_t be an autoregressive process of order p:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + u_t, \; u_t \sim WN$$

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + u_t$$

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i L^i y_t + u_t$$

$$\beta_p(L) Y_t = \beta_0 + u_t \text{ with } \beta_p(L) \equiv (1 - \beta_1 L - \beta_2 L^2 - \cdots - \beta_p L^p)$$

Without loss of generality and to simplify the following derivations we use $\beta_0 = 0$.
Note: This is always possible as the transformation $\tilde{y}_t \equiv y_t - E(y_t)$ induces a constant of zero.

Hence, we can write:

$$\begin{aligned} y_t &= \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + u_t \\ \beta_p(L) y_t &= u_t \\ &\quad \text{with } \beta_p(L) \equiv (1 - \beta_1 L - \beta_2 L^2 - \cdots - \beta_p L^p) \end{aligned}$$

An $AR(p)$ process is weakly stationary when all roots of the characteristic equation defined by the lag polynomial $\beta_p(L)$ lie outside the unit circle

$$\beta_p(z) = (1 - \beta_1 z - \beta_2 z^2 - \ldots - \beta_p z^p) = 0$$

Example

Is $Y_t = 2.5Y_{t-1} - Y_{t-2} + U_t$ stationary?

$$
\begin{aligned}
Y_t - 2.5Y_{t-1} + Y_{t-2} &= U_t \\
(1 - 2.5L + L^2)Y_t &= U_t \\
1 - 2.5z + z^2 &= 0
\end{aligned}
$$

$\Rightarrow z_1 = 2, z_2 = 0.5$

Autocorrelation function of a stationary $AR(p)$

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t \quad |\text{Trick: } \cdot y_t$$
$$y_t y_t = \beta_1 y_{t-1} y_t + \ldots + \beta_p y_{t-p} y_t + u_t y_t \quad |E(\cdot)$$
$$E(y_t y_t) = \beta_1 E(y_{t-1} y_t) + \ldots + \beta_p E(y_{t-p} y_t) + E(u_t y_t)$$

$$\lambda_0 = \beta_1 \lambda_1 + \beta_2 \lambda_2 + \ldots + \beta_p \lambda_p + \sigma_u^2$$

$$\lambda_1 = \beta_1 \lambda_0 + \beta_2 \lambda_1 + \ldots + \beta_p \lambda_{p-1} \quad | \div \lambda_0$$
$$\rho_1 = \beta_1 + \beta_2 \rho_1 + \ldots + \beta_p \rho_{p-1}$$
$$\rho_k = \beta_1 \rho_{k-1} + \beta_2 \rho_{k-2} + \ldots + \beta_p \rho_{k-p}$$

These equations are called Yule-Walker equations ($\rho_k = \rho_{-k}$).
Note: The autocorrelations follow the same difference equation as the original series.

Yule-Walker equations

Using these equations we get iteratively the autocorrelation function:

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & & \ddots & \vdots \\ \rho_{p-1} & & & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

With these equations we obtain values for $\rho_1, \ldots, \rho_p$, with which we compute iteratively the remaining values
$\rho_k = \beta_1 \rho_{k-1} + \beta_2 \rho_{k-2} + \ldots + \beta_p \rho_{k-p}$.
Equivalently we may use the matrix equation to obtain estimates for $\beta_1, \ldots, \beta_p$ from the estimates of ρ_k.

Example

$AR(2): \; Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + U_t$ Compute the autocorrelation function.

Yule-Walker equations:

$$\begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\begin{aligned} \rho_1 &= \beta_1 + \beta_2 \rho_1 \Rightarrow \rho_1 = \frac{\beta_1}{1-\beta_2} \\ \rho_2 &= \beta_1 \rho_1 + \beta_2 \Rightarrow \rho_2 = \frac{\beta_1^2}{1-\beta_2} + \beta_2 \\ &\Rightarrow \rho_3 = \beta_1 \rho_2 + \beta_2 \rho_1 \\ &\vdots \end{aligned}$$

Example

$AR(2): \;\; Y_t = 0.3Y_{t-1} + 0.04Y_{t-2} + U_t$ Stationary?
Autocorrelation function?

$$
\begin{aligned}
Y_t - 0.3Y_{t-1} - 0.04Y_{t-2} &= U_t \\
(1 - 0.3L - 0.04L^2)Y_t &= U_t \\
1 - 0.3z - 0.04z^2 &= 0 \\
&\Rightarrow z_1 = 2.5, z_2 = 10
\end{aligned}
$$

Autocorrelation function

$$
\begin{aligned}
\rho_1 &= \frac{\beta_1}{1-\beta_2} = 0.3125 \\
\rho_2 &= \frac{\beta_1^2}{1-\beta_2} + \beta_2 = 0.1338 \\
\rho_3 &= \beta_1\rho_2 + \beta_2\rho_1 = 0.0526 \\
\rho_4 &= \beta_1\rho_3 + \beta_2\rho_2 = 0.0211 \;\; \ldots
\end{aligned}
$$

Example

$AR(2): \; Y_t = -0.2\dot{6}Y_{t-1} + 0.2\dot{6}Y_{t-2} + U_t$ Stationarity? Autocorrelation function?

$$
\begin{aligned}
Y_t + 0.2\dot{6}Y_{t-1} - 0.2\dot{6}Y_{t-2} &= U_t \\
(1 + 0.2\dot{6}L - 0.2\dot{6}L^2)Y_t &= U_t \\
1 + 0.2\dot{6}z - 0.2\dot{6}z^2 &= 0 \\
&\Rightarrow z_1 = 2.5, z_2 = 1.5
\end{aligned}
$$

Autocorrelation function

$$
\begin{aligned}
\rho_1 &= \frac{\beta_1}{1-\beta_2} = -0.36 \\
\rho_2 &= \frac{\beta_1^2}{1-\beta_2} + \beta_2 = 0.36 \\
\rho_3 &= \beta_1\rho_2 + \beta_2\rho_1 = -0.19 \\
\rho_4 &= \beta_1\rho_3 + \beta_2\rho_2 = 0.15 \; \ldots
\end{aligned}
$$

Moving-average process of order 1

Let $\{u_t\} \sim i.i.d.(0, \sigma^2)$ and $\theta \in \mathbb{R}$. The process defined by

$$y_t = u_t + \theta u_{t-1} \quad t \in T$$

is called moving-average process of order 1 or short $MA(1)$ process.

Consequently, $E(y_t) = 0$, $Var(y_t) = \sigma^2 + \theta^2\sigma^2 < \infty$ and ACVF:

$$\lambda_k = \begin{cases} \sigma^2(1+\theta^2) & \text{for } k = 0 \\ \sigma^2\theta & \text{for } |k| = 1 \\ 0 & \text{for } |k| > 1 \end{cases}$$

Therefore an $MA(1)$ process is weakly stationary.
The autocorrelation function of $\{y_t\}$ is

$$\rho_k = \begin{cases} 1 & \text{for } k = 0 \\ \frac{\theta}{1+\theta^2} & \text{for } |k| = 1 \\ 0 & \text{for } |k| > 1 \end{cases}$$

Moving-average process of order q

$$y_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q} + u_t, \;\; u_t \sim i.i.d.(0, \sigma^2)$$

$$y_t = \sum_{i=1}^{q} \theta_i u_{t-i} + u_t$$

$$y_t = \sum_{i=1}^{q} \theta_i L^i u_t + u_t$$

$$y_t = \theta_q(L) u_t \text{ with } \theta_q(L) \equiv 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$$

Characteristics of an $MA(q)$ process

1. Expectation of $\{y_t\}$

$$E(y_t) = 0$$

2. Variance of $\{y_t\}$

$$Var(y_t) = \lambda_0 = (1 + \theta_1^2 - \theta_2^2 + \ldots + \theta_q^2)\sigma^2$$

3. Autocovariance function

$$Cov(y_t, y_{t+k}) = \lambda_k = \begin{cases} (\theta_k + \theta_{k+1}\theta_1 + \ldots + \theta_q\theta_{q-k})\sigma^2 & k = 1, 2, \ldots, q \\ 0 & k > q \end{cases}$$

Example

$MA(2)$ process: $y_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2}$, $u_t \sim i.i.d.(0, \sigma^2)$.

1. Compute $E(y_t)$, $Var(y_t)$.
2. Calculate the autocorrelation function.
3. Given $\theta_1 = -0.5$ and $\theta_2 = 0.25$, compute the autocorrelation function? Plot it.

1. $E(y_t) = E(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2}) = 0$

Example, cont.

2.

$$
\begin{aligned}
\lambda_0 \equiv Var(y_t) &= E\left[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})^2\right] \\
&= E(u_t^2) + \theta_1^2 E(u_{t-1}^2) + \theta_2^2 E(u_{t-2}^2) \\
&= \sigma^2 + \theta_1^2\sigma^2 + \theta_2^2\sigma^2 = (1 + \theta_1^2 + \theta_2^2)\sigma^2 \\
\lambda_1 &= E\left[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})(u_{t-1} + \theta_1 u_{t-2} + \theta_2 u_{t-3})\right] \\
&= \theta_1 E(u_{t-1}^2) + \theta_1\theta_2 E(u_{t-2}^2) \\
&= \theta_1\sigma^2 + \theta_1\theta_2\sigma^2 = (\theta_1 + \theta_1\theta_2)\sigma^2 \\
\lambda_2 &= E\left[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})(u_{t-2} + \theta_1 u_{t-3} + \theta_2 u_{t-4})\right] \\
&= \theta_2 E(u_{t-2}^2) \\
&= \theta_2\sigma^2 \\
\lambda_3 &= E\left[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})(U_{t-3} + \theta_1 u_{t-4} + \theta_2 u_{t-5})\right] \\
&= 0
\end{aligned}
$$

Invertibility of an $MA(q)$ process

$$y_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} + u_t, \; u_t \sim i.i.d.(0, \sigma^2)$$
$$y_t = \theta_q(L) u_t \qquad \text{with } \theta(L) \equiv 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q$$

An $MA(q)$ process is invertible if all roots of the characteristic equation defined by the corresponding lag polynomial $\theta_q(L)$ lie outside the unit circle:

$$\theta_q(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q = 0$$

$z \in \mathbb{C}$ and $|z_i| > 1, \; \forall i.$

Note: $u_t = \theta_q(L)^{-1} y_t$, $u_t = \theta_q(L)^{-1} y_t = \sum_{i=0}^{\infty} a_i L^i y_t$ is an $AR(\infty)$: $u_t - \sum_{i=1}^{\infty} a_i L^i y_t = y_t$

Example

Under which conditions is the following $MA(1)$ process $y_t = u_t + \theta_1 u_{t-1}$ invertible?

$$y_t = u_t + \theta_1 u_{t-1} = (1 + \theta_1 L)u_t$$

$$\theta(L) = (1 + \theta_1 L)$$

$$\theta(z) = (1 + \theta_1 z) = 0 \rightarrow z = -\frac{1}{\theta_1}$$

$$|z| > 1 \rightarrow \left|-\frac{1}{\theta_1}\right| > 1 \rightarrow 1 > |\theta_1|$$

Note: $-\frac{1}{2} < \rho_k < \frac{1}{2}\ \forall \theta_1 \in \mathbb{R}$, as $\rho_1 = \frac{\theta_1}{1+\theta_1^2}$.

Example

Show that the following two processes:

$$X_t = U_t + \theta U_{t-1} \quad U_t \sim WN(0, \sigma^2)$$
$$Y_t = \tilde{U}_t + \frac{1}{\theta}\tilde{U}_{t-1} \quad \tilde{U}_t \sim WN(0, \sigma^2\theta^2)$$

have the same autocovariance function. Are the processes invertible?

Example

Given: $Y_t = U_t + 0.5U_{t-1}$. Compute the corresponding $AR(\infty)$?

$$Y_t = (1 + 0.5L)U_t = (1 + 0.5L)(a_0 + a_1 L + a_2 L^2 + \ldots)Y_t$$

Comparison of the coefficients:

$$L^0 : 1 = a_0$$

$$L^1 : 0 = a_1 + 0.5a_0 \qquad \rightarrow a_1 = -0.5$$

$$L^2 : 0 = a_2 + 0.5a_1 \qquad \rightarrow a_2 = (-0.5)^2$$

$$L^j : \ldots \qquad \rightarrow a_j = (-0.5)^j$$

$$\Rightarrow U_t = Y_t + \sum_{j=1}^{\infty} (-0.5)^2 L^j Y_t$$

$$Y_t = -\sum_{j=1}^{\infty} (-0.5)^j L^j Y_t + U_t$$

$MA(\infty)$ of the finite $AR(p)$ process

$$\beta_p(L)y_t = u_t \qquad AR(p)$$

$$y_t = \beta(L)^{-1}u_t = u_t + a_1 u_{t-1} + a_2 u_{t-2} + \dots \qquad MA(\infty)$$

Example: Given $Y_t = 0.25Y_{t-1} + U_t$, compute the corresponding $MA(\infty)$.

Stationary? $1 - 0.25z = 0 \rightarrow z = 4$

$$(1 - 0.25L)Y_t = U_t$$

$$(1 - 0.25L)(a_0 + a_1 L + a_2 L^2 + \dots)U_t = U_t$$

Example, cont.

Comparison of the coefficients:

$$(1 - 0.25L)(a_0 + a_1 L + a_2 L^2 + \ldots)U_t = U_t$$

$$L^0 : 1 = a_0$$

$$L^1 : 0 = a_1 - 0.25 a_0 \qquad \rightarrow a_1 = 0.25$$

$$L^2 : 0 = a_2 - 0.25 a_1 \qquad \rightarrow a_2 = 0.25^2$$

$$L^j : \ldots \qquad \rightarrow a_j = 0.25^j$$

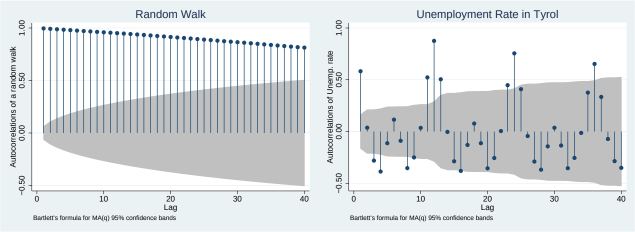
$$\Rightarrow Y_t = 0.25 Y_{t-1} + U_t$$

$$Y_t = \sum_{j=1}^{\infty} 0.25^j L^j Y_t + U_t$$

Sample ACF with non-stationary time series

Sample ACF can be computed for any kind of time series - not only for realizations of a stationary process.

- A time trend becomes obvious if $|r_k|$ decreases very slowly with increasing k.
- If a deterministic periodic fluctuation like seasonality is a part of the stochastic process then r_k should show a similar periodical pattern.



Partial autocorrelation function, I

Consider the stationary AR(1) process $y_t = \beta_1 y_{t-1} + u_t$, where $E[u_t] = 0$ so $E[y_t] = 0$. The second autocorrelation is $\rho_2 = \beta_1^2$.

We might ask what is the autocorrelation between y_t and y_{t-2} net of the intervening effect of y_{t-1}?

In this model if we remove the effect of y_{t-1} from y_t, then only u_t remains and this disturbance is uncorrelated wit y_{t-2}. We would conclude that the partial autocorrelation between y_t and y_{t-2} in this model is zero.

Partial autocorrelation function, II

The partial correlation between y_t and y_{t-k} is the simple correlation between y_{t-k} and y_t minus that part explained by the intervening lags. That is,

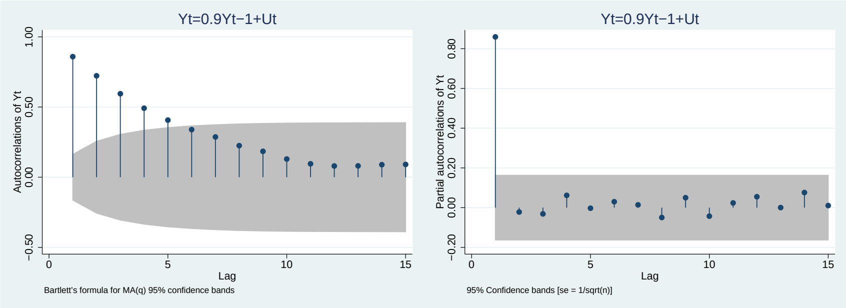
$$\rho_k^* = Corr\left[y_t - E^*(y_t | y_{t-1}, ..., y_{t-k+1}), y_{t-k}\right]$$

where $E^*(y_t | y_{t-1}, ..., y_{t-k+1})$ is the minimum mean-squared error predictor od y_t by $y_{t-1}, ..., y_{t-k+1}$.

Note: The first partial autocorrelation coefficient for any process equals the first autocorrelation coefficient.

Example

$AR(1)$ with $\beta_1 = 0.9$



Confidence intervals for the ACF and PACF

For $\{Y_t\} \sim i.i.d.(0,1)$ is

$$\rho_k = \begin{cases} 1 & \text{for } k = 0 \\ 0 & \text{for } k > 0 \end{cases}$$

One can show that for $k > 0$: $r_k \sim N\left(0, \frac{1}{T}\right)$ approximately for large T. Hence, $\left[-1.96 \cdot \frac{1}{\sqrt{T}}, 1.96 \cdot \frac{1}{\sqrt{T}}\right]$ is an approximate 95% confidence interval for r_k:

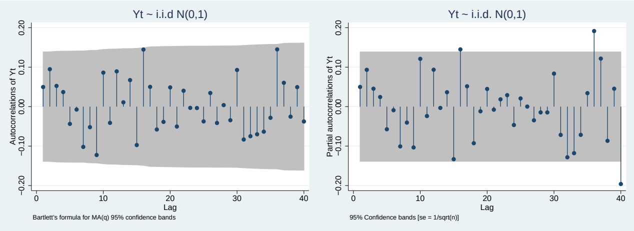


Figure: 200 simulated values for $Y_t \sim i.i.d.N(0,1)$ process

Ljung-Box statistic

The Box-Pierce statistic

$$Q = T \sum_{k=1}^{p} r_k^2$$

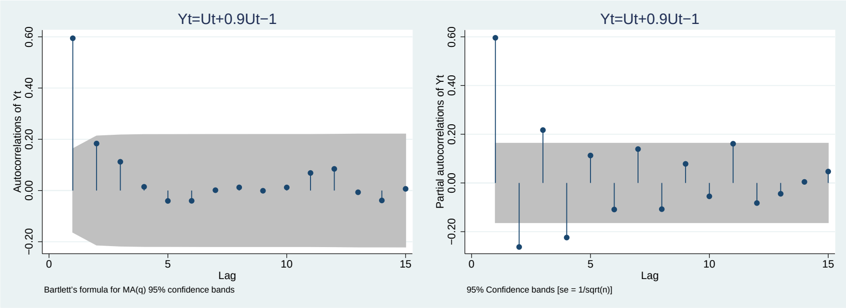
is commonly used to test whether a series is white noise. Under the assumption that $y_1, \ldots, y_T$ are realizations of independent and identically distributed random variables with finite variance Q has a limiting chi-squared distribution with p degrees of freedom.

The Box-Pierce statistic may give very conservative tests for 'short' time series. A refinement that appears to have better finite-sample properties is the Ljung-Box statistic:

$$Q_{LB} = T(T+2) \sum_{k=1}^{p} \frac{r_k^2}{T-k} \sim^a \chi^2(df = p)$$

Example

$MA(1)$ with $\theta_1 = 0.9$



$ARMA(p, q)$ process

$$\beta_p(L)Y_t = \beta_0 + \theta_q(L)U_t$$

$$\beta_p(L) \equiv 1 + \beta_1 L - \beta_2 L^2 - \ldots - \beta_q L^q$$

$$\theta_q(L) \equiv 1 + \theta_1 L - \theta_2 L^2 - \ldots - \theta_q L^q$$

with $E(U_t) = 0$, $E(U_t^2) = \sigma^2$, $E(U_t U_s) = 0 \; t \neq s$.

Summary of behavior of ACF and PACF for the time series models:

Process	**ACF**	**PACF**
$MA(q)$	number of ρ's different from zero = order of MA process	decreasing
$AR(p)$	decreasing	number of ρ^*'s different from zero = order of AR process
$ARMA(p, q)$	decreasing	decreasing

Systems with feedback

$$X_t = \alpha_1 Y_{t-1} + U_t$$
$$Y_t = \alpha_2 X_{t-1} + U'_t$$

with $U_t \sim WN(0, \sigma^2)$ and $U'_t \sim WN(0, \sigma'^2)$.

$$X_t = \alpha_1[\alpha_2 X_{t-2} + U'_{t-1}] + U_t$$
$$X_t = \alpha_1\alpha_2 X_{t-2} + \alpha_1 U'_{t-1} + U_t$$

where $\alpha_3 \equiv \alpha_1 \frac{\sigma'}{\sigma}$, then $\alpha_3 U_{t-1} \equiv \alpha_1 \frac{\sigma'}{\sigma} U_{t-1} = \alpha_1 U'_{t-1}$.

$$(1 - \alpha_1\alpha_2 L^2)X_t = (1 + \alpha_3 L)U_t$$
$$\Rightarrow X_t \sim ARMA(2, 1)$$

Overlapping of *ARMA* processes

X_t and Y_t are two independent *ARMA* processes of order (p_1, q_1) and (p_2, q_2). Then the sum of X_t and Y_t, $Z_t = X_t + Y_t$, is again an *ARMA* process of order (p, q). For (p, q) we know:

$$p \leq p_1 + p_2$$
$$q \leq \max(p_1 + q_2, p_2 + q_1)$$

Except for specific parameter values of p and q one uses the upper limits. Therefore we have:

$$ARMA(p_1, q_1) + ARMA(p_2, q_2) = ARMA(p_1 + p_2, \max(p_1 + q_2, p_2 + q_1))$$

Hence, the sum of two *MA* processes is again an *MA* process, the sum of two *AR* processes is in general an *ARMA* process with a non zero *MA* component (cf. aggregated time series).

Wold's decomposition theorem

Theorem 21.1

Every zero-mean covariance stationary stochastic process can be represented in the form

$$y_t = E^*[y_t|y_{t-1}, y_{t-2}, ..., y_{t-p}] + \sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i},$$

where ε_t is white noise, $\pi_0 = 1$, and the weights are square summable, that is $\sum_{i=1}^{\infty} \pi_i^2 < \infty$.
$E_t^* = E^*[y_t|y_{t-1}, y_{t-2}, ..., y_{t-p}]$ is the optimal linear predictor of y_t based on its lagged values, and the predictor is uncorrelated with ε_{t-i}.

Thus, the theorem decomposes the process generating y_t into

- $E_t^* = E^*[y_t|y_{t-1}, y_{t-2}, ..., y_{t-p}]$ the linearly deterministic component
- $\sum_{i=0}^{\infty} \pi_i \varepsilon_{t-i}$ the linearly indeterministic component.

Characteristics of *ARMA* processes

$\{y_t\}$ is an $ARMA(p,q)$ process if $\{y_t\}$ is stationary and $\forall t : \beta(L)y_t = \beta_0 + \theta(L)u_t$, where $u_t \sim WN(0, \sigma^2)$.

A stationary (and unique) solution to $\beta(L)y_t = \beta_0 + \theta(L)u_t$ exists if all roots of the characteristic equation

$$(1 - \beta_1 z - \beta_2 z^2 - \ldots - \beta_p z^p) = 0$$

lie outside the unit circle.

An $ARMA(p,q)$ process $\{y_t\}$ is invertible if all roots of the characteristic equation $(1 + \theta_1 z - \theta_2 z^2 - \ldots - \theta_q z^q) = 0$ lie outside the unit circle.

Example I

Is the following $ARMA(1,1)$ process stationary?

$$Y_t - 0.5Y_{t-1} = U_t + 0.4U_{t-1}, \; \{U_t\} \sim WN(0, \sigma^2)$$

How does the MA presentation of $\{Y_t\}$ look like? Is it invertible? Give the AR description.

1.

$$Y_t - 0.5Y_{t-1}$$
$$(1 - 0.5L)Y_t$$
$$1 - 0.5z = 0 \rightarrow z = 2$$

Example II

2.

$$(1-0.5L)Y_t = (1+0.4L)U_t$$
$$(1-0.5L)(b_0 + b_1 L + b_2 L^2 + \ldots)U_t = (1+0.4L)U_t$$

Comparison of the coefficients:

$$L^0 : b_0 = 1$$
$$L^1 : b_1 - 0.5 b_0 = 0.4 \qquad \rightarrow b_1 = 0.9$$
$$L^2 : b_2 - 0.5 b_1 = 0 \qquad \rightarrow b_2 = 0.45$$
$$L^j : b_j - 0.5 b_{j-1} = 0 \qquad \rightarrow b_j = 0.5 b_{j-1} = 0.9 \cdot 0.5^{j-1}$$
$$\Rightarrow Y_t = U_t + \sum_{j=1}^{\infty} 0.9 \cdot 0.5^{j-1} L^j U_t$$

Example III

3.

$$
\begin{aligned}
&U_t + 0.4U_{t-1} \\
&1 + 0.4z = 0 \rightarrow z = -2.5
\end{aligned}
$$

4.

$$
\begin{aligned}
(1 - 0.5L)Y_t &= (1 + 0.4L)U_t \\
(1 - 0.5L) &= (1 + 0.4L)(b_0 + b_1 L + b_2 L^2 + \ldots)Y_t
\end{aligned}
$$

Example IV

Comparison of the coefficients:

$$
\begin{aligned}
L^0 &: b_0 = 1 \\
L^1 &: -0.5 = 0.4b_0 + b_1 && \rightarrow b_1 = -0.9 \\
L^2 &: 0 = b_2 + 0.4b_1 && \rightarrow b_2 = -0.4 \cdot (-0.9) \\
L^j &: 0 = b_j + 0.4b_{j-1} && \rightarrow b_j = -0.4b_{j-1} \\
& && = -0.9 \cdot (-0.4)^{j-1}
\end{aligned}
$$

$$\Rightarrow U_t = Y_t - \sum_{j=1}^{\infty} 0.9 \cdot (-0.4)^{j-1} L^j Y_t$$

$$Y_t = -\sum_{j=1}^{\infty} 0.9 \cdot (-0.4)^{j-1} L^j Y_t + U_t$$

Parameter parsimony of *AR*, *MA* and *ARMA* models

An *ARMA* process can be approximated with an *AR* process or an *MA* process with a sufficiently high number of parameters (of high order). However, *ARMA* models are preferred due to the fact that they need less parameters (Box/Jenkins principle of parsimonious models).

If a time series is modeled using an $MA(q)$, an $AR(p)$ and an $ARMA(p', q')$ process, then we get in general $p' + q' \leq q$ and $p' + q' \leq p$ with comparable model quality. Consequently the *ARMA* model needs the lowest number of parameters.

Box-Jenkins Approach for modeling stochastic processes

1. **Stationarity, Model Specification:** Determining the order (p, q) of the model. Graphical methods, autocorrelation function and partial autocorrelation function are used, checking stationarity.
2. **Estimation:** Estimation of the parameters using nonlinear least squares or maximum likelihood.
3. **Diagnosis, Model Quality:** Checking the model quality, overfitting and residual diagnostic.

The aim is to establish a model as parsimonious as possible because the variance of the estimates is inverse proportional to the number of degrees of freedom and to avoid overfitting and its consequences.

Information criteria, I

- Akaike information criterion (1974): $AIC = -2LL + 2k$
- Schwarz Bayesian information criterion (1978):$SBIC = -2LL + k\ln(T)$
- Hannan-Quinn criterion (1979): $HQIC = -2LL + 2k\ln(\ln(T))$

with $T \ldots$ number of observations (sample size), LL value of the optimized loglikelihood function of the model and k the number of estimated parameters (including the constant and $\hat{\sigma}^2$).

Information criteria - Simulation Study

We investigate the following $AR(4)$ process:
$Y_t = 0.5Y_{t-1} - 0.3Y_{t-2} + 0.5Y_{t-4} + U_t, \; U_t \sim i.i.d.N(0,1)$,
$T = 50, 100, 200, \; MC = 1000$. The following table shows the frequency of the obtained order of the model in dependence which criteria was used to estimate the order.

T	p	0	1	2	3	4	5	6	7	8	9	10
50	AIC	1	1	58	39	756	90	27	11	8	6	3
	SBIC	9	3	225	80	663	20	0	0	0	0	0
	HQ	5	3	113	60	775	44	13	2	4	1	0
100	AIC	0	0	1	0	800	106	39	26	16	9	3
	SBIC	0	0	5	1	967	25	1	1	0	0	0
	HQ	0	0	2	1	910	61	16	9	1	0	0
200	AIC	0	0	0	0	752	112	56	37	18	13	12
	SBIC	0	0	0	0	981	16	2	1	0	0	0
	HQ	0	0	0	0	919	50	20	7	1	3	0

$SBIC$ is consistent but inefficient and AIC is not consistent (overestimation of the number of model parameters) but in general more efficient than $SBIC$.

Comments regarding forecasting

- Considerations
  - In-sample versus out-of-sample model quality
  - One-step ahead versus multi-step ahead
  - Recursive versus rolling window
- The conditional expectation value $E(Y_{t+h}|\Omega_t)$ is used as forecast $f_{t,h}$, Ω_t is the set of information available at time t. The conditional expectation value minimizes the MSE:

$$\min_{f_{t,h}}\{MSE(f_{t,h}) \equiv E(Y_{t,h} - f_{t,h})^2\} \Rightarrow f_{t,h} = E[Y_{t,h}|\Omega_t]$$

- The optimal forecast for white noise with expectation zero is by definition: $E(U_{t,h}|\Omega_t) = 0 \, \forall h > 0$.
- The 'Naive' forecast is defined as: $E(Y_{t,h}|\Omega_t) = Y_t$. This is the best forecast approach for a random walk.

Forecasting with an underlying $MA(3)$ process I

$$
\begin{aligned}
Y_t &= \theta_1 U_{t-1} + \theta_2 U_{t-2} + \theta_3 U_{t-3} + U_t \\
Y_{t+1} &= \theta_1 U_t + \theta_2 U_{t-1} + \theta_3 U_{t-2} + U_{t+1} \\
Y_{t+2} &= \theta_1 U_{t+1} + \theta_2 U_t + \theta_3 U_{t-1} + U_{t+2} \\
Y_{t+3} &= \theta_1 U_{t+2} + \theta_2 U_{t+1} + \theta_3 U_t + U_{t+3}
\end{aligned}
$$

$$
\begin{aligned}
f_{t,1} &\equiv E(Y_{t,1}|\Omega_t) = E(\theta_1 U_t + \theta_2 U_{t-1} + \theta_3 U_{t-2} + U_{t+1}|\Omega_t) \\
&= \theta_1 U_t + \theta_2 U_{t-1} + \theta_3 U_{t-2} \\
f_{t,2} &\equiv E(Y_{t,2}|\Omega_t) = E(\theta_1 U_{t+1} + \theta_2 U_t + \theta_3 U_{t-1} + U_{t+2}|\Omega_t) \\
&= \theta_2 U_t + \theta_3 U_{t-1} \\
f_{t,3} &\equiv E(Y_{t,3}|\Omega_t) = E(\theta_1 U_{t+2} + \theta_2 U_{t+1} + \theta_3 U_t + U_{t+3}|\Omega_t) = \theta_3 U_t \\
f_{t,4} &\equiv E(Y_{t,4}|\Omega_t) = E(\theta_1 U_{t+3} + \theta_2 U_{t+2} + \theta_3 U_{t+1} + U_{t+4}|\Omega_t) = 0 \\
f_{t,h} &= 0 \ \forall h \geq 4
\end{aligned}
$$

Forecasting with an underlying $MA(3)$ process II

The corresponding MSE is:

$$
\begin{aligned}
E(Y_{t+1} - E(Y_{t+1}|\Omega_t))^2 &= E(\theta_1 U_t + \theta_2 U_{t-1} + \theta_3 U_{t-2} + U_{t+1} - \theta_1 U_t \\
&- \theta_2 U_{t-1} - \theta_3 U_{t-2})^2 = E(U_{t+1})^2 = \sigma^2 \\
E(Y_{t+2} - E(Y_{t+2}|\Omega_t))^2 &= E(\theta_1 U_{t+1} + \theta_2 U_t + \theta_3 U_{t-1} + U_{t+2} - \theta_2 U_t \\
&- \theta_3 U_{t-1})^2 \\
&= E(\theta_1 U_{t+1} + U_{t+2})^2 = \theta_1^2\sigma^2 + \sigma^2 = \sigma^2(1+\theta_1^2) \\
E(Y_{t+3} - E(Y_{t+3}|\Omega_t))^2 &= E(\theta_1 U_{t+2} + \theta_2 U_{t+1} + \theta_3 U_t + U_{t+3} - \theta_3 U_t)^2 \\
&= E(\theta_1 U_{t+2} + \theta_2 U_{t+1} + U_{t+3})^2 = \sigma^2(1+\theta_1^2+\theta_2^2) \\
E(Y_{t+4} - E(Y_{t+4}|\Omega_t))^2 &= E(\theta_1 U_{t+2} + \theta_2 U_{t+1} + \theta_3 U_{t+3} + U_{t+4})^2 \\
&= \sigma^2(1+\theta_1^2+\theta_2^2+\theta_3^2) = E(Y_t)^2 \; \forall h \geq 4
\end{aligned}
$$

Forecasting with an underlying $MA(3)$ process III

Estimates:

$$\hat{f}_{t,1} = \hat{\theta}_1 \hat{u}_t + \hat{\theta}_2 \hat{u}_{t-1} + \hat{\theta}_3 \hat{u}_{t-2}$$
$$\hat{f}_{t,2} = \hat{\theta}_2 \hat{u}_t + \hat{\theta}_3 \hat{u}_{t-1}$$
$$\hat{f}_{t,3} = \hat{\theta}_3 \hat{u}_t$$
$$\hat{f}_{t,4} = 0$$
$$\hat{f}_{t,0} = 0 \;\forall h \geq 4$$

Memory up to the order q!

Forecasting with an underlying $AR(2)$ process, I

$$
\begin{aligned}
Y_t &= \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + U_t \\
Y_{t+1} &= \beta_1 Y_t + \beta_2 Y_{t-1} + U_{t+1} \\
Y_{t+2} &= \beta_1 Y_{t+1} + \beta_2 Y_t + U_{t+2} \\
Y_{t+3} &= \beta_1 Y_{t+2} + \beta_2 Y_{t+1} + U_{t+3}
\end{aligned}
$$

$$
\begin{aligned}
f_{t,1} &\equiv E(Y_{t,1}|\Omega_t) = E(\beta_1 Y_t + \beta_2 Y_{t-1} + U_{t+1}|\Omega_t) = \beta_1 Y_t + \beta_2 Y_{t-1} \\
f_{t,2} &\equiv E(Y_{t,1}|\Omega_t) = E(\beta_1 Y_{t+1} + \beta_2 Y_t + U_{t+2}|\Omega_t) = \beta_1 f_{t,1} + \beta_2 Y_t \\
f_{t,3} &\equiv E(Y_{t,3}|\Omega_t) = E(\beta_1 Y_{t+2} + \beta_2 Y_{t+1} + U_{t+3}|\Omega_t) = \beta_1 f_{t,2} + \beta_2 f_{t,1} \\
f_{t,4} &\equiv E(Y_{4,3}|\Omega_t) = E(\beta_1 Y_{t+3} + \beta_2 Y_{t+2} + U_{t+4}|\Omega_t) = \beta_1 f_{t,3} + \beta_2 f_{t,2} \\
&\vdots \\
f_{t,h} &= \beta_1 f_{t,h-1} + \beta_2 f_{t,h-2}
\end{aligned}
$$

Memory goes to infinity!

Forecasting with an underlying $AR(p)$ process, II

MSE for AR(1):

$$
\begin{aligned}
E(Y_{t+1} - E(Y_{t+1}|\Omega_t))^2 &= E(\beta_1 Y_t + U_{t+1} - \beta_1 Y_t)^2 = \sigma^2 \\
E(Y_{t+2} - E(Y_{t+2}|\Omega_t))^2 &= E(\beta_1 Y_{t+1} + U_{t+2} - \beta_1^2 Y_t)^2 \\
&= E(\beta_1^2 Y_t + \beta_1 U_{t+1} + U_{t+2} - \beta_1^2 Y_t)^2 \\
&= \sigma^2(1 + \beta_1^2) \\
E(Y_{t+3} - E(Y_{t+3}|\Omega_t))^2 &= E(\beta_1^3 Y_t + \beta_1^2 U_{t+1} + \beta_1 U_{t+2} + U_{t+3} \\
&- \beta_1^3 Y_t)^2 = \sigma^2(1 + \beta_1^2 + \beta_1^4) \\
&\vdots \\
E(Y_{t+h} - E(Y_{t+h}|\Omega_t))^2 &= \sigma^2(1 + \beta_1^2 + \beta_1^4 + \ldots + \beta_1^{2(h-1)}) \\
\lim_{h\to\infty} E(Y_{t+h} - E(Y_{t+h}|\Omega_t))^2 &= \frac{\sigma^2}{1 - \beta_1^2} = E(Y_t)^2
\end{aligned}
$$

Quality criteria for prediction I

Let T be the number of available observations, all data till T_1 are employed to estimate the process and the remaining observations are used to investigate the quality of the estimation (of the model).

- Mean squared error:

$$MSE = \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} (y_{t,h} - \hat{f}_{t,h})^2$$

- Mean absolute error:

$$MAE = \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} |y_{t,h} - \hat{f}_{t,h}|$$

Quality criteria for prediction II

- Mean percentage error:

$$MAPE = \frac{1}{T - T_1} \sum_{t=n_1+1}^{T} \left| \frac{y_{t,h} - \hat{f}_{t,h}}{y_{t,h}} \right| \cdot 100$$

- Mean symmetric absolute percentage error:

$$AMAPE = \frac{1}{T - T_1} \sum_{t=n_1+1}^{T} \left| \frac{y_{t,h} - \hat{f}_{t,h}}{y_{t,h} + \hat{f}_{t,h}} \right| \cdot 100$$

Quality criteria for prediction III

Note: *AMAPE* is the same (symmetric) regardless the forecast value is 0.5 and the true value is 0.3 or vice versa. That is not true for *MAPE*.

Note: *AMAPE* gives invalid values if y_{t+h} and $\hat{f}_{t,h}$ are of same size but have different signs. Similar reacts *MAPE* in the following case: $\hat{f}_{t,h} = 1$, $y_{t,h} = 0.0001$.

- Theil's U statistic:

$$U = \sqrt{\frac{\sum_{t=T_1+1}^{T}\left(\frac{y_{t,h}-\hat{f}_{t,h}}{y_{t,h}}\right)^2}{\sum_{t=T_1+1}^{T}\left(\frac{y_{t,h}-\hat{f}^*_{t,h}}{y_{t,h}}\right)^2}} \text{ or } U = \sqrt{\frac{\sum_{t=T_1+1}^{T}\left(y_{t,h}-\hat{f}_{t,h}\right)^2}{\sum_{t=T_1+1}^{T}\left(y_{t,h}-\hat{f}^*_{t,h}\right)^2}}$$

Quality criteria for prediction IV

where $f^*_{t,h}$ is the forecast via the benchmark model (e. g. the naive forecast, random walk). $U < 1$ denotes that the investigated approach is better than the benchmark model.

- Compare Theil's inequality coefficient:

$$TIC = \frac{\sqrt{\sum_{t=T_1}^{T}\left(y_{t,h} - \hat{f}_{t,h}\right)^2}}{\sqrt{\sum_{t=T_1}^{T}\left(\hat{f}_{t,h}\right)^2} + \sqrt{\sum_{t=T_1}^{T}\left(y_{t,h}\right)^2}}, \quad 0 \leq TIC \leq 1$$

Characteristics of OLS estimator for $AR(1)$ I

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t,\ U_t \sim WN(0, \sigma^2)$$

- Bias:

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + U) = \beta + (X'X)^{-1}X'U$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} n & \sum_{t=1}^T y_{t-1} \\ \sum_{t=1}^T y_{t-1} & \sum_{t=1}^T y_{t-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^T u_t \\ \sum_{t=1}^T y_{t-1} u_t \end{pmatrix}$$

$$E\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + E\left[\begin{pmatrix} T & \sum_{t=1}^T y_{t-1} \\ \sum_{t=1}^n y_{t-1} & \sum_{t=1}^T y_{t-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^T u_t \\ \sum_{t=1}^T y_{t-1} u_t \end{pmatrix}\right]$$

generally $\neq 0$

Therefore the **OLS estimator is biased** for the $AR(1)$ process!

Characteristics of OLS estimator for $AR(1)$ II

- Consistency:

$$p\lim \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} +$$

$$+p\lim \left(\frac{1}{T} \begin{pmatrix} T & \sum_{t=1}^{T} y_{t-1} \\ \sum_{t=1}^{T} y_{t-1} & \sum_{t=1}^{T} y_{t-1}^2 \end{pmatrix} \right)^{-1} \cdot p\lim \left(\frac{1}{T} \begin{pmatrix} \sum_{t=1}^{T} u_t \\ \sum_{t=1}^{T} y_{t-1} u_t \end{pmatrix} \right),$$

with $p\lim \left(\frac{1}{T} \begin{pmatrix} \sum_{t=1}^{T} u_t \\ \sum_{t=1}^{T} y_{t-1} u_t \end{pmatrix} \right) = 0$, as y_{t-1} and u_t are uncorrelated!

Therefore we obtain **consistent estimates using OLS**!

Characteristics of OLS estimator for $AR(1)$ III

For T sufficiently large one can show that the estimator is normally distributed given that the AR process is stationary. The test statistic holds asymptotically, no more finite sample distributions are available.

In practice one may take y_1 as given and compute the OLS estimate using the remaining $2, \ldots, T$ observations. Sample size reduces to $T - 1$ for an $AR(1)$ process.

Conditional ML-Estimator

The OLS estimator can be regarded as an ML estimator:

$$L_C = P(y_2, y_3, \dots, y_T | y_1) = P(y_2|y_1) \cdot P(y_3|y_2) \cdot \dots \cdot P(y_T|y_{T-1})$$

$$P(y_t|y_{t-1}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_t - \beta_0 - \beta_1 y_{t-1})^2\right)$$

$$LL_C = \ln L_C = (T-1)\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - \beta_0 - \beta_1 y_{t-1})^2$$

$$= (T-1)\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{T-1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - \beta_0 - \beta_1 y_{t-1})^2$$

$$\max LL_C(\beta_0, \beta_1, \sigma^2; y_t) \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$$

Unconditional/full ML-Estimator, I

$$L = p(y_1)L_c$$

$$Y_1 \sim N\left(\frac{\beta_0}{1-\beta_1}, \frac{\sigma^2}{1-\beta_1^2}\right)$$

$$p(x_t) = \frac{1}{\tilde{\sigma}\sqrt{2\pi}} \exp\left(-\frac{1}{2\tilde{\sigma}^2}(x_t - \tilde{\mu})^2\right)$$

$$\ln p(y_1) = \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \ln \frac{\sigma^2}{1-\beta_1^2} - \frac{1-\beta_1^2}{2\sigma^2}\left(y_1 - \frac{\beta_0}{1-\beta_1}\right)^2$$

$$= \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\ln \sigma^2 + \frac{1}{2}\ln(1-\beta_1^2) - \frac{1-\beta_1^2}{2\sigma^2}\left(y_1 - \frac{\beta_0}{1-\beta_1}\right)^2$$

Unconditional/full ML-Estimator, II

$$LL = \ln p(y_1) + LL_c =$$

$$= T \ln \frac{1}{\sqrt{2\pi}} - \frac{T}{2} \ln \sigma^2 + \frac{1}{2} \ln(1 - \beta_1^2) - \frac{1 - \beta_1^2}{2\sigma^2} \left(y_1 - \frac{\beta_0}{1 - \beta_1} \right)^2$$

$$- \frac{1}{2\sigma^2} \sum_{t=2}^{T} (y_t - \beta_0 - \beta_1 y_{t-1})^2$$

$$\max LL(\beta_0, \beta_1, \sigma^2) \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$$

$MA(q)$ models

$$MA(1):\ y_t = \beta_0 + u_t + \theta_1 u_{t-1}$$

Conditional Estimator:

$$\begin{aligned} u_0 &= 0 \\ u_1 &= y_1 - \beta_0 \\ u_2 &= y_2 - \beta_0 - \theta_1 u_1 \\ &= \vdots \\ u_i &= fkt(\beta_0, \theta_1) \end{aligned}$$

$\min \sum_{i=1}^{T} u_i^2 \Rightarrow \hat{\beta}_0, \hat{\theta}_1$ For the optimization of this nonlinear function with respect to the parameters one needs already for the conditional estimator iterative optimization algorithms. Also in this case there exists an unconditional/full estimator (cf. Hamilton 1994).

AR(1) disturbances

$$
\begin{aligned}
y_t &= X_t\beta + \varepsilon_t \\
&\quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t \\
&\quad u_t \sim WN(0, \sigma_u^2) \\
&\quad |\rho| < 1
\end{aligned}
$$

$$
Var[\varepsilon_t] = \sigma^2\Omega = \frac{\sigma_u^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \cdots & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{pmatrix}
$$

Least Squares Estimation

The least squares estimator is

$$b = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon$$

and it is

- unbiased,
- consistent, and
- asymptotically normally distributed

However, $s^2(X'X)^{-1}$ is an inappropriate estimator of $\sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$.

We should use the Newey-West autocorrelation consistent covariance estimator as a robust covariance estimator.

Investigating autocorrelation in the residuals

- Lagrange multiplier test:

  H_0 : no autocorrelation versus $H_1 : \varepsilon_t = AR(p)$ or $\varepsilon_t = MA(q)$.

  Regress the OLS residuals e_t on the X and lagged OLS residuals, $e_{t-1}, ..., e_{t-p}$. Compute $TR_0^2 \sim \chi(p)$.
- Autocorrelation function and partial autocorrelation function
- Ljung-Box statistic

Efficient estimation when Ω is known I

GLS estiamtor:

$$\begin{aligned}
\hat{\beta} &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y) \\
Est.Var[\hat{\beta}] &= \sigma_\varepsilon^2(X'\Omega^{-1}X)^{-1} \\
\sigma_\varepsilon^2 &= \frac{1}{T}(y - X\hat{\beta})'\Omega^{-1}(y - X\hat{\beta})
\end{aligned}$$

For the $AR(1)$ case, data for the transformed model are

$$y^* = \begin{pmatrix} \sqrt{1-\rho^2}y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{pmatrix}, \qquad X^* = \begin{pmatrix} \sqrt{1-\rho^2}x_1 \\ x_2 - \rho x_1 \\ x_3 - \rho x_2 \\ \vdots \\ x_T - \rho x_{T-1} \end{pmatrix}$$

Efficient estimation when Ω is known II

Partial differences/quasi differences, or pseudo differences

$$Var[\varepsilon_t^*] = Var[\varepsilon_t - \rho\varepsilon_{t-1}] = Var[u_t] = \sigma_u^2$$

$$Var[\varepsilon_1^*] = Var[\sqrt{1-\rho^2}\varepsilon_1] = (1-\rho^2)\frac{\sigma_u^2}{1-\rho^2} = \sigma_u^2$$

Efficient estimation when Ω is unknown I

All is needed for efficient estimation of β is a consistent estimator of $\Omega(\rho)$.

- Prais and Winsten (1954) estimator
- Cochrane and Orcutt (1949) estimator (omits the first observation)
- Maximum likelihood estimators

Integrated process I

Recall

$$y_t = \mu + y_{t-1} + \varepsilon_t = \sum_{i=0}^{\infty}(\mu + \varepsilon_{t-i})$$

The random walk is clearly a nonstationary process. However, the first differences of y_t

$$\Delta y_t = y_t - y_{t-1} = \mu + \varepsilon_t$$

are stationary.

- The series y_t is said to be integrated of order one, denoted $I(1)$, because taking the first difference produces a stationary process.

Integrated process II

- A nonstationary series is integrated of order d, denoted $I(d)$, if it becomes stationary after being first differenced d times.
- A further generalization of the *ARMA* model would be the autoregressive integrated moving-average model, or $ARIMA(p, d, q)$

$$\beta(L)[(1-L)^d y_t] = \mu + y(L)\varepsilon_t$$

Unit Root

Each of the following series is characterized by a unit root.

- Random walk with drift: $y_t = \mu + y_{t-1} + \varepsilon_t$
- Trend stationary process: $y_t = \mu + \beta \cdot t + \varepsilon_t$
- Random walk: $y_t = y_{t-1} + \varepsilon_t$

In each case, the DGP can be written as

$$(1 - L)y_t = \alpha + v_t,$$

where $\alpha = \mu, \beta$, and 0, respectively, and v_t is a stationary process. Thus the characteristic equation has a single root equal to one, hence the name.

Basic equations for Unit Root tests

Nesting all three models in a single equation gives the basic equation for a variety of unit root tests in economic applications:

$$
\begin{aligned}
z_t &= \mu + \beta \cdot t + \gamma z_{t-1} + \varepsilon_t \\
z_t - z_{t-1} &= \alpha_0 + \alpha_1 t + (\gamma - 1) z_{t-1} + \epsilon_t \\
H_0 &: \gamma - 1 = 0
\end{aligned}
$$

Dickey-Fuller Unit Root Test I

Assume the process $y_t = \gamma y_{t-1} + \varepsilon_t$. Therefore we want to test the hypothesis:

$$H_0 : \gamma = 1 \quad H_1 : |\gamma| < 1$$

Although (downward) biased the OLS estimator is consistent and we would like to use the t-statistic in order to test for a unit root:

$$\hat{y}_t = \hat{\gamma} y_{t-1} \quad t_{\hat{\gamma}} = \frac{\hat{\gamma} - \gamma}{se(\hat{\gamma})}$$

However, this statistic is distributed as:

$$\sqrt{T}(\hat{\gamma} - \gamma) \sim N(0, (1 - \gamma^2)) \quad |\gamma| < 1$$

This distribution is under H_0 useless, as for $\gamma = 1$ die variance equals zero (degenerated distribution).

Dickey-Fuller Unit Root Test II

'Derivation' of the degenerated distribution:

$$y_t = \gamma y_{t-1} + \varepsilon_t \quad \varepsilon_t \sim i.i.N(0, \sigma^2)$$

We know from OLS theory that:

$$(\hat{\beta} - \beta) \sim N(0, \sigma^2 (X'X)^{-1})$$

for an $AR(1)$ process:

$$E(X'X) = E\left(\sum y_{t-1}^2\right) = T \frac{\sigma^2}{1 - \gamma^2}$$

$$\Rightarrow \sqrt{T}(\hat{\gamma} - \gamma) \sim N(0, (1 - \gamma^2)) \; |\gamma| < 1.$$

- This distribution is only valid under H_1 for $|\gamma| < 1$. We have no distribution for $\hat{\gamma}$ under the null hypothesis $H_0 : \gamma = 1$.

Dickey-Fuller Unit Root Test III

- Dickey and Fuller (1979) showed that the distribution for $T(\hat{\gamma} - 1)$ exists but has no analytic expression (density function). Critical values are computable via simulation. Tables with critical values are available, e.g. in Hamilton 1994 or Davidson and McKinnon 1996.

- Because the distribution $T(\hat{\gamma} - 1)$ exists but not the distribution for $\sqrt{T}(\hat{\gamma} - 1)$, the OLS estimator of γ is called super consistent.

The Dickey-Fuller tests I

The following three DGPs are underlying the DF test:

- DGP 1:

$$Y_t = \beta_1 Y_{t-1} + U_t$$
$$H_0 : \beta_1 = 1 \quad H_1 : |\beta_1| < 1$$

$$\Delta Y_t = (\beta_1 - 1) Y_{t-1} + U_t$$
$$\Delta Y_t = \theta Y_{t-1} + U_t$$
$$H_0 : \theta = 0 \quad H_1 : \theta \neq 0$$
$$t_\theta = \frac{\hat{\theta}}{se(\hat{\theta})}$$

$$n = 100, \alpha = 5\% \rightarrow t_{DF;crit} = -1.95$$

The Dickey-Fuller tests II

- DGP 2:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + U_t \\ \Delta Y_t &= \beta_0 + (\beta_1 - 1) Y_{t-1} + U_t \\ \Delta Y_t &= \beta_0 + \theta Y_{t-1} + U_t \\ t_\theta &= \frac{\hat{\theta}}{se(\hat{\theta})} \end{aligned}$$

$$n = 100, \alpha = 5\% \rightarrow t_{DF;crit} = -2.89$$

The Dickey-Fuller tests III

- DGP 3:

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \gamma t + U_t \\
\Delta Y_t &= \beta_0 + (\beta_1 - 1) Y_{t-1} + \gamma t + U_t \\
\Delta Y_t &= \beta_0 + \theta Y_{t-1} + \gamma t + U_t \\
t_\theta &= \frac{\hat{\theta}}{se(\hat{\theta})}
\end{aligned}
$$

$$n = 100, \alpha = 5\% \rightarrow t_{DF;crit} = -3.45$$

Power of Dickey-Fuller tests, I

The weakness (as of other unit roots tests too) of Dickey-Fuller tests is their **low power**!

- *Case 1*: Monte Carlo study
  $y_t = 0.95 y_{t-1} + u_t,\ u_t \sim N(0,1),\ T = 50,\ MC = 10,000.$
  The alternative hypothesis of stationarity would be correct.

  The Dickey-Fuller test rejects the null hypothesis of a unit root process only in 14.9% of the cases! If as DGP just the constant is taken H_0 gets rejected in 6.8% of the cases; and with the DGP of a constant and a deterministic time trend H_0 gets rejected just in 5.2% of the cases.

  Therefore if DF tests are used in the majority of the applications 'integrated' time series are employed in econometric models.

Power of Dickey-Fuller tests, II

- *Case 2*: $y_t = \beta_1 y_{t-1} + \gamma t + u_t$, $u_t \sim WN(o, \sigma^2)$
  DGP is a stationary *AR* process ($|\beta_1| < 1$) but with a deterministic time trend.

  Assumption: The time trend is not considered in the estimation $y_t = \beta_0 + \tilde{\beta}_1 y_{t-1} + \tilde{u}_t$.
  If the main part of the variance in the mean of the time series is attributed to the deterministic trend the OLS estimate $\tilde{\beta}_1$ will be near one. The constant gets erroneously interpreted as drift that catches the trend. Therefore the test does not reject H_0 of a unit root, the power of the test is almost zero. One can show that the test in this case is even inconsistent.

Power of Dickey-Fuller tests, III

- *Case 3*: It is problematic to apply a unit root test if a structural break (ω) is in the data:

$$y_t = \omega I(t > t^*) + \beta_1 y_{t-1} + u_t,$$

where $I(t > t^*)$ is a dummy variable taking the value one if $(t > t^*)$ and zero elsewhere.

If the structural break is ignored the OLS estimate for β_1 will be biased towards one as the difference in level is modeled as a permanent chock not dying out.

Power of Dickey-Fuller tests, IV

- *Case 4*: What happens if the noise is not white noise? Let the DGP be $y_t = \beta_1 y_{t-1} + u_t + \theta_1 u_{t-1}$:

$$\hat{\beta}_1 = \frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2} = \beta_1 + \frac{\sum (u_t + \theta_1 u_{t-1}) y_{t-1}}{\sum y_{t-1}^2}$$

$p \lim_{n \to \infty} \hat{\beta}_1 \neq \beta_1$. Consequently the OLS estimator is no longer consistent.

Due to the super consistency of the OLS estimator the test statistic can still be employed but the asymptotic distribution differs from the distribution originally developed by Dickey and Fuller. The new underlying distribution is dependent on the exact specification of the autocorrelation of the residuals. To avoid this problem the following statistics were developed.

Augmented Dickey-Fuller Test

Generally

$$\beta(L)y_t = \beta_0 + u_t$$
$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-1} + \ldots + \beta_p y_{t-p} + u_t$$
$$\Delta y_t = \theta_0 + \theta_1 y_{t-1} + \sum_{i=1}^{p-1} \alpha_i \Delta y_{t-i} + u_t$$

- Choose p so that the residuals are uncorrelated (use sufficiently many lags).
- Under the null hypothesis of a unit root the process y_{t-1} is not stationary but the differenced lagged variables are stationary Δy_{t-i} (avoiding spurious regression).
- Therefore: $t_{\hat{\alpha}} = \frac{\hat{\alpha}}{se(\hat{\alpha}_i)}$ is Student's t distributed and $t_{\hat{\theta}} = \frac{\hat{\theta}_1}{se(\hat{\theta}_1)}$ is Dickey-Fuller t distributed.

Phillips-Perron Test

Phillips and Perron (1988) suggested an alternative (non parametric) strategy in order to consider autocorrelation in the error term testing for a unit root. They computed a factor which corrects the original test statistic of Dickey-Fuller in the presence of autocorrelation (spectral density estimation).

The advantage of this approach is that no order of the $AR(p)$ process must ex-ante be specified as with the augmented Dickey-Fuller test. Unfortunately the empirical significance levels may differ significantly from the nominal level in small samples.

KPSS test

Kwiatkowski, Phillips, Schmidt and Shin (1992) have devised an alternative to the Dickey-Fuller test for stationarity of a time series.

$$\begin{aligned} y_t &= \alpha + \beta t + \gamma \sum_{i=1}^{t} z_i + \varepsilon_t, t = 1, ..., T \\ &= \alpha + \beta t + \gamma Z_t + \varepsilon_t \end{aligned}$$

H_0	:	$\gamma = 0$ and $\beta = 0$	stationarity
H_1	:	$\gamma = 0$ and $\beta \neq 0$	trend stationarity
		$\gamma \neq 0$	y_t is nonstationary

Note: The null and alternative hypothesis switched their meaning.

Spurious Regression

- Studies in empirical macroeconomics almost always involve nonstationary and trending variables such as income, consumption, money demand, the price level, trade flows, and exchange rates.
- In a cross-sectional environment, the phrase 'spurious regression' is used to describe a situation where two variables are related through their correlation with a third variable.
- In time series analysis often a time trend of the dependent variable and the independent variables is responsible for spurious regression results.

Example for spurious regression

Two independent unit root processes are regressed onto each other:

$$
\begin{aligned}
x_t &= x_{t-1} + \varepsilon_t \qquad \text{with } \varepsilon_t \;\; i.i.d.(0, \sigma_\varepsilon^2) \\
y_t &= y_{t-1} + \epsilon_t \qquad \text{with } \epsilon_t \;\; i.i.d.(0, \sigma_\epsilon^2)
\end{aligned}
$$

Let $x_0 = y_0 = 0$ and assume further that ε_t and ϵ_t are independent processes.
This implies that $\{x_t\}$ and $\{y_t\}$ are independent of each other.

What if we regress $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$?
Simulation study by Davidson and MacKinnon (1993): $n = 50$, $H_0 : \beta_1 = 0$ gets rejected about 66.2% instead of 5% (10,000 Monte Carlo data sets)! As sample size increases it gets worse.

Where does this come from? We run the regression $y_t = \beta_0 + \beta_1 x_t + u_t$. Under $H_0 : \beta_1 = 0, y_t = \beta_0 + u_t$; given the assumptions of OLS u_t must be serially uncorrelated, stationary with mean zero but as y_t is a random walk u_t is one.

Including a time trend I

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + \varepsilon_t$$

- Omitting t from the regression will generally yield biased estimators for β_1 and β_2 (Omitted variable bias).
- Detrending interpretation of regressions with a time trend:

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 t + \varepsilon_{yt} \rightarrow e_y \\ x_{1t} &= \gamma_0 + \gamma_1 t + \varepsilon_{x_1 t} \rightarrow e_{x_1} \qquad \Rightarrow e_{yt} = \hat{\beta}_1 e_{x_1 t} + \hat{\beta}_2 e_{x_2 t} + u_t \\ x_{2t} &= \lambda_0 + \lambda_1 t + \varepsilon_{x_2 t} \rightarrow e_{x_2} \end{aligned}$$

Including a time trend II

The estimates of primary interest can be interpreted as coming from a regression without trend, but where we first detrend the dependent variable and all other independent variables.

- The interpretation of $\hat{\beta}_1$ and $\hat{\beta}_2$ shows that it is a good idea to include a trend in the regression if any independent variable is trending, even if y_t is not.

Cointegration

- The notion of cointegration makes regression involving $I(1)$ variables potentially meaningful.
- If $\{y_t\}$ and $\{x_t\}$ are two $I(1)$ processes, then in general $y_t - \beta x_t$ is an $I(1)$ process for any number β.
- It is possible that for some $\beta \neq 0, y_t - \beta x_t$ is an $I(0)$ process. If such a β exists y_t and x_t are said to be cointegrated.
- In such a case, we can distinguish between a long-run relationship between y_t and x_t, that is the manner in which the two variables drift upward together, and the short-run dynamics, that is the relationship between deviations of y_t from its long-run trend and deviations of x_t from its long-run trend (differencing of the data would be counterproductive).

Testing for Cointegration

- Idea: If we know the value of β (sometimes we know from economic theory) we simply define a new variable, $z_t \equiv y_t - \beta x_t$ and apply a unit root test in order to test whether $z_t \sim I(0)$ and therefore y_t and x_t are cointegrated.

- If β is unknown we have to estimate it first: $y_t = \hat{\beta}_0 + \hat{\beta} x_t$. It turns out that if the two series are cointegrated then $\hat{\beta}$ from OLS is a consistent estimate. Employing a Dickey-Fuller test, the null hypothesis is that the series are not cointegrated. Consequently, under H_0 a spurious regression is computed.

- Fortunately, it is possible to tabulate critical values even when β is estimated. $\Rightarrow$ Engle-Granger test.

Error Correction Modell

- In addition to learning about a potential long-run relationship between two series, the concept of cointegration enriches the kinds of dynamic models at our disposal.

- If y_t and x_t are $I(1)$ processes and are not cointegrated, we might estimate a dynamic model in first differences:

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + u_t$$

- If y_t and x_t are cointegrated with parameter β then we have additional $I(0)$ variables that we can include in the model:

$$\begin{aligned} \Delta y_t &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta z_{t-1} u_t \\ \Delta y_t &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta(y_{t-1} - \beta x_{t-1}) + u_t \end{aligned}$$

Error correction term