Lecture 4: Estimation of ARIMA models

Florian Pelgrin

University of Lausanne, École des HEC
Department of mathematics (IMEA-Nice)

Sept. 2011 - Dec. 2011

Road map

1. Introduction
1.1. Overview

- Provide an overview of some estimation methods for linear time series models :
  - Sample moments estimation ;
  - (Nonlinear) Linear least squares method ;
  - Maximum likelihood estimation
  - (Generalized) Method of moments
- Other methods are available (e.g., Bayesian estimation or Kalman filtering through state-space models) !

- Broadly speaking, these methods consist in estimating the parameters of interest (autoregressive coefficients, moving average coefficients, and variance of the innovation process) as follows :

$$\begin{aligned} &\underset{\delta}{\text{Optimize}} \qquad Q_T(\delta) \\ &\text{s.t. (nonlinear) linear equality and inequality constraints} \end{aligned}$$

where Q_T is the objective function and δ is the vector of parameters of interest.

Remark : (Nonlinear) Linear equality and inequality constraints may represent the fundamentalness conditions.

- Examples

  - (Nonlinear) Linear least squares methods aims at **minimizing** the sum of squared residuals ;
  - Maximum likelihood estimation aims at **maximizing** the (log-) likelihood function ;
  - Generalized method of moments aims at **minimizing** the distance between the theoretical moments and zero (using a weighting matrix).

- These methods differ in terms of :
  - Assumptions ;
  - "Nature" of the model (e.g., MA *versus* AR) ;
  - Finite and large sample properties ;
  - Etc.
- High quality software programs (Eviews, SAS, Splus, Stata, etc) are available ! However, it is important to know the estimation options (default procedure, optimization algorithm, choice of initial conditions) and to keep in mind that all these estimation techniques do not perform equally and do depend on the nature of the model...

1.2. Framework

- Suppose that (X_t) is **correctly specified** as an ARIMA(p,d,q) model

$$\Phi(L)\Delta^d X_t = \Theta(L)\epsilon_t$$

where ϵ_t is a **weak white noise** $(0, \sigma_\epsilon^2)$ and

$$\Phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p \qquad \text{with } \phi_p \neq 0$$
$$\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q \qquad \text{with } \theta_q \neq 0.$$

- Assume that
  1. The model order (p,d, and q) is known ;
  2. The data has zero mean ;
  3. The series is weakly stationary (e.g., after d-difference transformation, etc).

2. Sample moments

- Let $(X_1, \cdots, X_T)$ represent a sample of size T from a covariance-stationary process with

$$\mathbb{E}[X_t] = m_X \qquad \text{for all } t$$
$$\mathbb{C}ov[X_t, X_{t-h}] = \gamma_X(h) \qquad \text{for all } t$$

- If nothing is known about the distribution of the time series, (non-parametric) estimation can be provided by :
  - Mean

$$\hat{m}_X \equiv \bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t$$

  - Autocovariance

$$\hat{\gamma}_X(h) = \frac{1}{T}\sum_{t=1}^{T-|h|}(X_t - \bar{X}_T)(X_{t+|h|} - \bar{X}_T)$$

    or

$$\hat{\gamma}_X(h) = \frac{1}{T-|h|}\sum_{t=1}^{T-|h|}(X_t - \bar{X}_T)(X_{t+|h|} - \bar{X}_T).$$

  - Autocorrelation

$$\hat{\rho}_X(h) = \frac{\hat{\gamma}_X(h)}{\hat{\gamma}_X(0)}.$$

3. (Nonlinear) Least squares method
3.1. Least squares estimation

- Consider the linear regression model $(t = 1, \cdots, T)$:

$$y_t = x_t' b_0 + \epsilon_t$$

where (y_t, x_t') are i.i.d. (H1), the regressors are stochastic (H2), x_t' is the t$^{\text{th}}$-row of the X matrix, and the following assumptions hold :

1. Exogeneity (H3)

$$\mathbb{E}[\epsilon_t | x_t] = 0 \qquad \text{for all } t;$$

2. Spherical error terms (H4)

$$\mathbb{V}[\epsilon_t | x_t] = \sigma_0^2 \qquad \text{for all } t;$$
$$\mathbb{C}ov[\epsilon_t, \epsilon_{t'} \mid x_t, x_{t'}] = 0 \qquad \text{for } t \neq t'$$

3. Rank condition (H5)

$$\text{rank}[\mathbb{E}(x_t x_t')] = k.$$

Definition

The ordinary least squares estimation of b_0, b_{ols}, defined from the following minimization program

$$\begin{aligned} \hat{b}_{ols} &= \underset{b_0}{\operatorname{argmin}} \sum_{t=1}^{T} \epsilon_t^2 \\ &= \underset{b_0}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_t - x_t' b_0 \right)^2 \end{aligned}$$

is given by

$$\hat{b}_{ols} = \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \left(\sum_{t=1}^{T} x_t y_t \right).$$

Proposition

Under H1-H5, the ordinary least squares estimator of b is weakly consistent

$$\hat{b}_{ols} \xrightarrow[T\to\infty]{p} b_0$$

and is asymptotically normally distributed

$$\sqrt{T}\left(\hat{b}_{ols} - b_0\right) \xrightarrow{\ell} \mathcal{N}\left(0_{k\times 1}, \sigma_0^2 \mathbb{E}\left[x_t x_t'\right]^{-1}\right).$$

Example : AR(1) estimation

- Let (X_t) be a covariance-stationary process defined by the fundamental representation ($|\phi| < 1$) :

$$X_t = \phi X_{t-1} + \epsilon_t$$

where (ϵ_t) is the innovation process of (X_t).

- The ordinary least squares estimation of ϕ is defined to be :

$$\hat{\phi}_{ols} = \left(\sum_{t=2}^{T} x_{t-1}^2 \right)^{-1} \left(\sum_{t=2}^{T} x_{t-1} x_t \right)$$

- Moreover

$$\sqrt{T} \left(\hat{\phi}_{ols} - \phi \right) \xrightarrow{\ell} \mathcal{N} \left(0, 1 - \phi^2 \right).$$

3.2. Nonlinear least squares estimation

Definition

Consider the nonlinear regression model defined by $(t = 1, \cdots, T)$:

$$y_t = h(x_t'; b_0) + \epsilon_t$$

where h is a nonlinear function (w.r.t. b_0). The nonlinear least squares estimator of b_0, $\hat{b}_{nls}$, satisfies the following minimization program :

$$\begin{aligned} \hat{b}_{nls} &= \underset{b_0}{\operatorname{argmin}} \sum_{t=1}^{T} \epsilon_t^2 \\ &= \underset{b_0}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_t - h(x_t'; b_0)\right)^2 . \end{aligned}$$

Example : MA(1) estimation

- Let (X_t) be a covariance-stationary process defined by the fundamental representation ($|\theta| < 1$) :

$$X_t = \epsilon_t + \theta \epsilon_{t-1}$$

where (ϵ_t) is the innovation process of (X_t).

- The ordinary least squares estimation of θ, which is defined by

$$\hat{\theta}_{ols} = \underset{\theta}{\operatorname{argmin}} \sum_{t=2}^{T} (x_t - \theta \epsilon_{t-1})^2 ,$$

is not feasible (since ϵ_{t-1} is not observable !).

Example : MA(1) estimation (cont'd)

- Conditionally on ϵ_0,

$$\begin{aligned} x_1 &= \epsilon_1 + \theta\epsilon_0 && \Leftrightarrow \epsilon_1 = x_1 - \theta\epsilon_0 \\ x_2 &= \epsilon_2 + \theta\epsilon_1 && \Leftrightarrow \epsilon_2 = x_2 - \theta\epsilon_1 \\ & && \Leftrightarrow \epsilon_2 = x_2 - \theta(x_1 - \theta\epsilon_0) \\ & \vdots \\ x_{t-1} &= \epsilon_{t-1} - \theta\epsilon_{t-2} && \Leftrightarrow \epsilon_{t-1} = \sum_{k=0}^{t-2}(-\theta)^k x_{t-1-k} + (-\theta)^{t-1}\epsilon_0. \end{aligned}$$

- The (conditional) nonlinear least squares estimation of θ is defined by (assuming that ϵ_0 is negligible)

$$\hat{\theta}_{nls} = \underset{\theta}{\text{argmin}} \sum_{t=2}^{T} \left(x_t - \theta \sum_{k=0}^{t-2} (-\theta)^k x_{t-1-k} \right)^2$$

and $\sqrt{T}\left(\hat{\theta}_{nls} - \theta\right) \xrightarrow{\ell} \mathcal{N}\left(0, 1-\theta^2\right)$.

Example : Least squares estimation using backcasting procedure

- Consider the fundamental representation of a MA(q) :

$$\begin{aligned} X_t &= u_t \\ u_t &= \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \end{aligned}$$

- Given some initial values $\delta^{(0)} = \left(\theta_1^{(0)}, \cdots, \theta_q^{(0)}\right)'$,
  - Compute the **unconditional residuals** $\hat{u}_t$

$$\hat{u}_t = X_t !$$

  - Backcast values of ϵ_t for $t = -(q-1), \cdots, T$ using the backward recursion

$$\tilde{\epsilon}_t = \hat{u}_t - \theta_1^{(0)} \tilde{\epsilon}_{t+1} - \cdots - \theta_q^{(0)} \tilde{\epsilon}_{t+q}$$

    where the q values for ϵ_t beyond the estimation sample are set to zero : $\tilde{\epsilon}_{T+1} = \cdots = \tilde{\epsilon}_{T+q} = 0$, and $\hat{u}_t = 0$ for $t < 1$.

Example : Least squares estimation using backcasting procedure (cont'd)

- Estimate the values of ϵ_t using a forward recursion

$$\hat{\epsilon}_t = \hat{u}_t - \theta_1^{(0)}\tilde{\epsilon}_{t-1} - \cdots - \theta_q^{(0)}\tilde{\epsilon}_{t-q}$$

- Minimize the sum of squared residuals using the fitted values $\hat{\epsilon}_t$:

$$\sum_{q+1}^{T}(X_t - \hat{\epsilon}_t - \theta_1\hat{\epsilon}_{t-1} - \cdots - \theta_q\hat{\epsilon}_{t-q})^2$$

  and find $\hat{\delta}^{(1)} = \left(\theta_1^{(1)}, \cdots, \theta_q^{(1)}\right)'$.

- Repeat the backcast step, forward recursion, and minimization procedures until the estimates of the moving average part converge.

Remark : The backcast step can be turned off ($\tilde{\epsilon}_{-(q-1)} = \cdots = \tilde{\epsilon}_0 = 0$) and the forward recursion is only used.

3.3. Discussion

- (Linear and Nonlinear) Least squares estimation are often used in practise. These methods lead to simple algorithms (e.g., backcasting procedure for ARMA models) and can often be used to initialize more "sophisticated" estimation methods.
- These methods are opposed to exact methods (see likelihood estimation).
- Nonlinear estimation requires numerical optimization algorithms :
  - Initial conditions ;
  - Number of iterations and stopping criteria ;
  - Local and global convergence ;
  - Fundamental representation and constraints.

4. (Generalized) Method of moments
4.1. Methods of moments and Yule-Walker estimation

Definition

Suppose there is a set of k conditions

$$S_T - g(\delta) = 0_{k \times 1}$$

where $S_T \in \mathbb{R}^k$ denotes a vector of theoretical moments , $\delta \in \mathbb{R}^k$ is a vector of parameters, and $g : \mathbb{R}^k \to \mathbb{R}^k$ defines a (bijective) mapping between S_T and δ.
The method of moments estimation of δ, $\hat{\delta}_{mm}$, is defined to be the value of δ such that

$$\hat{S}_T - g\left(\hat{\delta}_{mm}\right) = 0_{k \times 1}$$

where $\hat{S}_T$ is the estimation (empirical counterpart) of S_T.

Example 1 : MA(1) estimation

- Let (X_t) be a covariance-stationary process defined by the fundamental representation ($|\theta| < 1$) :

$$X_t = \epsilon_t + \theta \epsilon_{t-1}$$

where (ϵ_t) is the innovation process of (X_t).

- One has

$$S_T - g(\delta) = \begin{pmatrix} \gamma_X(0) - (1+\theta^2)\sigma_\epsilon^2 \\ \gamma_X(1) - \theta\sigma_\epsilon^2 \end{pmatrix} = 0_{2\times 1}$$

where $\delta = (\theta, \sigma_\epsilon^2)'$ and $S_T = (\gamma_X(0), \gamma_X(1))' = (\mathbb{E}\left[X_t^2\right], \mathbb{E}\left[X_t X_{t-1}\right])'$.

- The method of moment estimation of δ, $\hat{\delta}_{mm}$, solves

$$\hat{S}_T - g(\hat{\delta}_{mm}) = \begin{pmatrix} \hat{\gamma}_X(0) - (1+\hat{\theta}_{mm}^2)\hat{\sigma}_{\epsilon,mm}^2 \\ \hat{\gamma}_X(1) - \hat{\theta}_{mm}\hat{\sigma}_{\epsilon,mm}^2 \end{pmatrix} = 0_{2\times 1}$$

- Therefore (using the constraint $|\theta| < 1$)

$$\hat{\phi}_{mm} = \frac{1 - \sqrt{1 - 4\hat{\rho}_X^2(1)}}{2\hat{\rho}_X(1)}$$

and

$$\hat{\sigma}_{\epsilon,mm}^2 = \frac{2\gamma_X(0)}{1 - \sqrt{1 - 4\hat{\rho}_X^2(1)}}.$$

Example 2 : Yule-Walker estimation

- Let (X_t) be a causal AR(p). The Yule-Walker equations write

$$\mathbb{E}\left[X_{t-k}\left(X_t - \sum_{j=1}^{p}\phi_j X_{t-j}\right)\right] = 0 \qquad \text{for all } k = 0, \cdots, p$$

or

$$S_T - g(\delta) = 0_{(p+1)\times 1} \Leftrightarrow \begin{cases} \gamma_X(0) - \phi'\gamma_p = \sigma_\epsilon^2 \\ \gamma_p - \Gamma_p\phi = 0_{p\times 1} \end{cases}$$

where $\gamma_p = (\gamma_X(1), \cdots, \gamma_X(p))' = (\mathbb{E}[X_t X_{t-1}], \cdots, \mathbb{E}[X_t X_{t-p}])'$,
$S_T = (\gamma_X(0), \gamma_p')'$, $\phi = (\phi_1, \cdots, \phi_p)'$, $\delta = (\phi', \sigma_\epsilon^2)'$.

- The method of moment estimation of $(\phi', \sigma_\epsilon^2)'$ solves

$$\hat{\phi}_{mm} = \hat{\Gamma}_p^{-1}\hat{\gamma}_p$$

and

$$\hat{\sigma}^2_{\epsilon,mm} = \hat{\gamma}_X(0) - \hat{\phi}'_{mm}\hat{\gamma}_p.$$

- Question : Difference(s) with ordinary least squares estimation of an AR(p) ?

Discussion

- It is necessary to use theoretical moments that can be well-estimated from the data (or given the sample size).
- Method of moment estimation depends on the mapping g^{-1} and especially the "smoothness" of the inverse g^{-1} (need to avoid that the inverse map has a "large derivative").
- Summarizing the data through the sample moments incurs a loss of information—the main exception is when the sample moments are sufficient (in a statistical sense).
- Instead of reducing a sample of size T to a "sample" of empirical moments of size k, one may reduce the sample to more "empirical moments" than there are parameters...the so-called generalized method of moments.

4.2. Generalized method of moments

Intuition :

- Consider the linear regression model in an i.i.d. context

$$y_t = x_t' b_0 + \epsilon_t$$

- The exogeneity assumption $\mathbb{E}[\epsilon_t \mid X]$ implies the k moments conditions (for $t = 1, \cdots, T$)

$$\mathbb{E}\left[x_t(y_t - x_t' b_0)\right] = 0_{k\times 1}.$$

- Using the empirical counterpart (analogy principle)

$$\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t' b_0) = 0_{k\times 1}$$

This is a system of k equations with k unknowns (**just-identified**)

$$\hat{b}_{ols} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\sum_{t=1}^{T} x_t y_t\right).$$

- More generally, suppose that there exists a set of **instruments** $z_t \in \mathbb{R}^p$ (with $p \geq k$)—the instruments being (i) uncorrelated with the error terms and (ii) correlated with the explanatory variables—such that

$$\mathbb{E}\left[z_t(y_t - x_t' b_0)\right] = 0_{p\times 1} \qquad \Leftrightarrow \qquad \mathbb{E}\left[h(z_t; \delta_0)\right] = 0_{p\times 1}.$$

- Using the empirical counterpart (analogy principle)

$$\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t' b_0) = 0_{p\times 1} \qquad \Leftrightarrow \qquad \frac{1}{T}\sum_{t=1}^{T} h(z_t; \delta_0) = 0_{p\times 1}.$$

- This system is **overidentified** for $p > k$. How can we find an estimation of b_0 ?

- The answer is provided by the generalized method of moments, i.e. an estimation method based on **estimating equations** that imposes the nullity of the expectation of a vector function of the observations and the parameters of interest b_0.

- The basic idea is to choose a value for δ such that the empirical counterpart is **as close as possible to zero**.

- "**As close as possible to zero**" means that one minimizes the distance between $\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t' b_0)$ and $0_{p\times 1}$ using a **weighting matrix** (say, W_T).

- Two issues
  1. Choice of instruments ;
  2. Choice of the (optimal) weighting matrix.

Definition

Consider the set of moment conditions

$$\mathbb{E}\left[h(z_t, \delta_0)\right] = 0_{p\times 1}$$

where $\delta_0 \in \mathbb{R}^k$, $p \geq k$, and h is a p-dimensional (nonlinear) linear function of the parameters of interest and depends on (i.i.d.) observations. The generalized method of moments estimator of δ_0 associated with W_T (symmetric positive definite matrix) is a solution to the problem

$$\min_{\delta} \left[\frac{1}{T}\sum_{t=1}^{T} h(z_t;\delta)\right]' W_T \left[\frac{1}{T}\sum_{t=1}^{T} h(z_t;\delta)\right].$$

Definition

The generalized method of moments estimator of δ_0 solves the first-order conditions

$$\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\partial h(z_t;\hat{\delta}_{gmm})}{\partial \delta}\right]' W_T \left[\frac{1}{T}\sum_{t=1}^{T} h(z_t;\hat{\delta}_{gmm})\right] = 0_{k\times 1}.$$

This estimator is consistent and asymptotically normally distributed.

Remark : The first-order conditions correspond to k linear combinations of the moments conditions.

- This estimator depends on W_t. There exists a best GMM estimator, i.e. an **optimal weighting matrix**.

- The optimal weighting matrix is given by (i.i.d. context)

$$W_T^* = \mathbb{V}^{-1}\left[h(z_t, \delta_0)\right] \equiv \mathbb{E}^{-1}\left[h(z_t, \delta_0)h(z_t, \delta_0)'\right]$$

  and the **sample analog** is

$$\left[\frac{1}{T}\sum_{t=1}^{T} h(z_t, \delta_0)h(z_t, \delta_0)'\right]^{-1}.$$

  It does depend on δ_0 !

- How can we proceed in practise ?
  - Two-step GMM or iterated GMM method ;
  - One-step method (Continuously updating estimator, exponential tilting estimator, empirical likelihood estimator, etc).

Example : AR(1) estimation

- Let (X_t) be a covariance-stationary process defined by the fundamental representation ($|\phi| < 1$) :

$$X_t = \phi X_{t-1} + \epsilon_t.$$

- Since (ϵ_t) is the innovation process of (X_t), one has the following moment conditions

$$\mathbb{E}\left[x_{t-k}\epsilon_t\right] = 0 \qquad \text{for } k > 0$$

$\Rightarrow$ Past values $x_{t-1}, \cdots, x_{t-h}$ can be used as instruments

- Let z_t denote the set of instruments

$$z_t = \left(x_{t-1}, \cdots, x_{t-h}\right)'.$$

- The moment conditions write

$$\mathbb{E}\left[h(z_t, \delta_0)\right] = 0_{h\times 1}$$

where

$$h(z_t, \delta_0) = z_t(x_t - \phi_1 x_{t-1}).$$

- The empirical counterpart is defined to be

$$\frac{1}{T}\sum_{t=h+1}^{T}\begin{pmatrix} x_{t-1} \\ \vdots \\ x_{t-h} \end{pmatrix}(x_t - \phi_1 x_{t-1}) = 0_{h\times 1}.$$

- The 2S-GMM estimator of ϕ is a solution to the problem

$$\left[\sum_{t=h+1}^{T} \frac{\partial h(z_t;\hat{\delta}_{gmm})}{\partial \delta}\right]' \hat{\Omega}_T^{-1} \left[\sum_{t=h+1}^{T} h(z_t;\hat{\delta}_{gmm})\right] = 0_{h\times 1}.$$

where $\hat{\Omega}_T^{-1}$ is a first-step consistent estimate of the inverse of the variance-covariance matrix of the moments conditions and it accounts for the dependence of the moment conditions.

- Remark : In a first step, one solves

$$\left[\sum_{t=h+1}^{T} \frac{\partial h(z_t;\hat{\delta}_{1,gmm})}{\partial \delta}\right]' I_T \left[\sum_{t=h+1}^{T} h(z_t;\hat{\delta}_{1,gmm})\right] = 0_{h\times 1}.$$

and then compute $\hat{\Omega}_T^{-1}(\hat{\delta}_{1,gmm}) \equiv \hat{\Omega}_T^{-1}$.

Discussion

- Choice of instruments (optimal instruments ?) and of the weighting matrix ;
- Bias-efficiency trade-off ;
- Numerical procedures can be cumbersome !
- Other methods can be written as a GMM estimation.

5. Maximum likelihood estimation
5.1. Overview

- Consider the linear regression model $(t = 1, \cdots, T)$:

$$y_t = x_t' b_0 + \epsilon_t$$

where (y_t, x_t') are i.i.d., the regressors are **stochastic**, x_t' is the t^{th}-row of the X matrix, and

$$\epsilon_t \mid x_t \sim i.i.d. \mathcal{N}(0, \sigma_\epsilon^2).$$

- It follows that

$$y_t \mid x_t \sim i.i.d. \mathcal{N}(x_t' b_0, \sigma_\epsilon^2).$$

- One observes a sample of (conditionally) i.i.d. observations (y_t, x_t'). These observations are realizations of random variables with a known distribution but unknown parameters.
- The basic idea is to maximize the probability of observing these realizations (given a **known parametric distribution** with **unknown parameters**).
- The probability of observing these realizations is provided by the joint density (or pdf) of the observations (or evaluated at the observations).
- This joint density is the likelihood function, with the exception that it is a **function of the unknown parameters** (and not the observations !).

5.2. Estimation

Definition

The exact likelihood function of the linear regression model is defined to be

$$L(y, x; \delta_0) = \prod_{t=1}^{T} L_t(y_t, x_t; \delta_0)$$

where L_t is the (exact) likelihood function for observation t

$$L_t(y_t, x_t; \delta_0) = f_{Y_t|X_t}(y_t \mid x_t; \delta_0) \times f_{X_t}(x_t).$$

where $f_{Y_t|X_t}$ (respectively, f_{X_t}) is the (conditional) probability density function of $Y_t \mid X_t$ (respectively, of X_t). Accordingly, the exact log-likelihood function is defined to be

$$\ell(y, x; \delta_0) = \sum_{t=1}^{T} \ell_t(y_t, x_t; \delta_0) \equiv \sum_{t=1}^{T} \log L_t(y_t, x_t; \delta_0).$$

- The exact likelihood function is given by

$$L(y, x; \delta_0) = \prod_{t=1}^{T} f_{Y_t|X_t}(y_t \mid x_t; \delta_0) \prod_{t=1}^{T} f_{X_t}(x_t)$$

- The second right-hand side term is often neglected (especially if f_X does not depend on δ_0) and one considers the conditional likelihood function

$$\begin{aligned} L(y \mid x; \delta_0) &= \prod_{t=1}^{T} f_{Y_t|X_t}(y_t \mid x_t; \delta_0) \\ &= \prod_{t=1}^{T} \left(\sigma_\epsilon^2 2\pi\right)^{-1/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\left(y_t - x_t' b_0\right)^2\right) \\ &= \left(\sigma_\epsilon^2 2\pi\right)^{-T/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}\left(y_t - x_t' b_0\right)^2\right) \end{aligned}$$

and the conditional log-likelihood function is given by

$$\ell(y \mid x; \delta_0) = -\frac{T}{2}\log(\sigma_\epsilon^2) - \frac{T}{2}\log(2\pi) - -\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}\left(y_t - x_t' b_0\right)^2.$$

Definition

The maximum likelihood estimator of δ_0 in the linear regression model, which is the solution of the problem

$$\max_{\delta} L(y \mid x; \delta) \Leftrightarrow \max_{\delta} \ell(y \mid x; \delta),$$

is asymptotically unbiased and efficient, and normally distributed

$$\sqrt{T}\left(\hat{\delta}_{ml} - \delta_0\right) \xrightarrow{\ell} \mathcal{N}\left(0_{k\times 1}, I_F^{-1}\right)$$

where I_F is the Fisher information matrix.

Example : AR(1) estimation

- Consider the following AR(1) representation ($|\phi| < 1$)

$$X_t = \phi X_{t-1} + \epsilon_t.$$

where ϵ_t or $\epsilon_t | X_{t-1} \sim i.i.d. \mathcal{N}\left(0, \sigma_\epsilon^2\right)$

- Therefore

$$X_t \mid X_{t-1} = x_{t-1} \sim \mathcal{N}\left(\phi x_{t-1}, \sigma_\epsilon^2\right)$$

$$X_1 \sim \mathcal{N}\left(0, \frac{\sigma_\epsilon^2}{1-\phi^2}\right)$$

- Accordingly,

$$f_{X_t|X_{t-1}}(x_t \mid x_{t-1}; \delta_0) = \left(\sigma_\epsilon^2 2\pi\right)^{-1/2} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\left(x_t - \phi x_{t-1}\right)^2\right)$$

$$f_{X_1}(x_1; \delta_0) = \left(\frac{\sigma_\epsilon^2}{1-\phi^2} 2\pi\right)^{-1/2} \exp\left(-\frac{1-\phi^2}{2\sigma_\epsilon^2} x_1^2\right).$$

Exact (log-) likelihood function

- The exact likelihood function is given by

$$\begin{aligned} L(x;\delta_0) &= L_1(x_1;\delta_0) \times \prod_{\tau=2}^{T} L_\tau\left(x_\tau \mid x_{\tau-1};\delta_0\right) \\ &= f_{X_1}(x_1;\delta_0) \times \prod_{\tau=2}^{T} f_{X_\tau|X_{\tau-1}}\left(x_\tau \mid x_{\tau-1};\delta_0\right) \\ &= \left(\frac{\sigma_\epsilon^2}{1-\phi^2}2\pi\right)^{-1/2} \exp\left(-\frac{1-\phi^2}{2\sigma_\epsilon^2}x_1^2\right) \\ &\quad \times \left(\sigma_\epsilon^2 2\pi\right)^{-\frac{T-1}{2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=2}^{T}\left(x_t-\phi x_{t-1}\right)^2\right). \end{aligned}$$

- The exact log-likelihood is

$$\begin{aligned} \ell(x;\delta_0) &= -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{\sigma_\epsilon^2}{1-\phi^2}\right) - \frac{1-\phi^2}{2\sigma_\epsilon^2}x_1^2 \\ &\quad -\frac{T-1}{2}\log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\sum_{t=2}^{T}(x_t-\phi x_{t-1})^2. \end{aligned}$$

Conditional (log-) likelihood function

- The conditional likelihood function is given by

$$\begin{aligned} L(x \mid x_{-1}; \delta_0) &= \prod_{\tau=2}^{T} L_\tau \left(x_\tau \mid x_{\tau-1}; \delta_0\right) \\ &= \prod_{\tau=2}^{T} f_{X_\tau | X_{\tau-1}} \left(x_\tau \mid x_{\tau-1}; \delta_0\right) \\ &= \left(\sigma_\epsilon^2 2\pi\right)^{-\frac{T-1}{2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{T} \left(x_t - \phi x_{t-1}\right)^2\right). \end{aligned}$$

- The conditional log-likelihood is

$$\begin{aligned} \ell(x \mid x_{-1}; \delta_0) &= -\frac{T-1}{2}\log(2\pi) - \frac{T-1}{2}\log(\sigma_\epsilon^2) \\ &\quad -\frac{1}{2\sigma_\epsilon^2} \sum_{t=2}^{T} \left(x_t - \phi x_{t-1}\right)^2. \end{aligned}$$

- The two methods lead to different maximum likelihood estimates

$$\hat{\delta}_{cml} \neq \hat{\delta}_{eml}$$

- The conditional maximum likelihood estimator of ϕ is also the ordinary least squares estimator of ϕ

$$\hat{\phi}_{cml} = \hat{\phi}_{ols}$$

- This also holds for any AR(p).

- The Yule-Walker, conditional maximum likelihood and (ordinary) least squares estimator are asymptotically equivalent.

Discussion

- Generally the maximum likelihood estimates are built as if the observations are Gaussian, though it is not necessary that (X_t) is Gaussian when doing the estimation.
- If the model is misspecified, the only difference (to some extent...) is that estimates may be less efficient. In this case, one can use the pseudo- or quasi-maximum likelihood estimation.
- In the case of linear models, there might be not so much gain of using exact maximum likelihood against conditional likelihood method (or least squares method).
- Maximum likelihood estimation often requires nonlinear numerical optimization procedures.
- Estimation of the Fisher information matrix can be cumbersome (especially in finite samples).

Appendix : ML estimation of a MA(q)

- Step 1 : Writing $\epsilon = (\epsilon_1, \cdots, \epsilon_T)'$

One has

$$\begin{aligned}
\epsilon_{1-q} &= \epsilon_{1-q} \\
\vdots &= \vdots \\
\epsilon_0 &= \epsilon_0 \\
\epsilon_1 &= x_1 - \theta_1\epsilon_1 - \cdots - \theta_q\epsilon_{1-q} \\
\epsilon_1 &= x_2 - \theta_1\epsilon_2 - \cdots - \theta_q\epsilon_{2-q} \\
\vdots &= \vdots \\
\epsilon_T &= x_T - \theta_1\epsilon_T - \cdots - \theta_q\epsilon_{T-q}
\end{aligned}$$

Therefore

$$\begin{aligned} \epsilon &= NX + Z\epsilon_* \\ &= \begin{pmatrix} Z & N \end{pmatrix} \begin{pmatrix} \epsilon_* \\ X \end{pmatrix} \end{aligned}$$

with $X = (x_1, \cdots, x_T)'$, $\epsilon_* = (\epsilon_{1-q}, \epsilon_{2-q}, \cdots, \epsilon_{-1}, \epsilon_0)'$, and

$$N = \begin{pmatrix} 0_{q \times T} \\ A_1(\theta) \end{pmatrix}, Z = \begin{pmatrix} I_q \\ A_2(\theta) \end{pmatrix}$$

where $A_1(\theta)$ is a lower triangular matrix $(T \times T)$, $A_2(\theta)$ is a $T \times q$ matrix, and I_q is the identity matrix of order q.

- Step 2 : Determination of the joint density of ϵ

The joint probability density function of ϵ is defined to be

$$\left(2\pi\sigma_\epsilon^2\right)^{-\frac{T+q}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(NX+Z\epsilon_*\right)'\left(NX+Z\epsilon_*\right)\right].$$

- Step 3 : Determination of the marginal density of X

The marginal density of X is defined to be

$$\left(2\pi\sigma_\epsilon^2\right)^{-\frac{T}{2}} \left(\det X'X\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(NX+Z\hat{\epsilon}_*\right)'\left(NX+Z\hat{\epsilon}_*\right)\right]$$

with $\hat{\epsilon}_* = -(Z'Z)^{-1}Z'NX$.

- Step 4 : Determination of the log-likelihood funcion

The log-likelihood function is defined to be

$$\ell(x;\delta) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_\epsilon^2) - \frac{T}{2}\log\left(\det X'X\right) - \frac{S(\theta)}{2\sigma_\epsilon^2}$$

where $\delta = (\theta', \sigma_\epsilon^2)'$ and

$$S(\theta) = (NX + Z\hat{\epsilon}_*)'(NX + Z\hat{\epsilon}_*).$$

- Step 5 : Determination of $\hat{\sigma}^2_{\epsilon,ml}$

Using the first-order condition w.r.t. σ_ϵ^2, one has

$$\hat{\sigma}^2_{\epsilon,ml} = \frac{S\left(\hat{\theta}_{ml}\right)}{T}.$$

- Step 6 : Determination of the concentrated log-likelihood funcion

The concentrated log-likelihood function is defined to be

$$\ell^*(x;\theta) = -T\log\left[S(\theta)\right] - \log\left[\det X'X\right].$$

- Step 7 : Determination of $\hat{\theta}_{ml}$

$\hat{\theta}_{ml}$ solves

$$\min_{\theta}\left\{T\log\left[S(\theta)\right] + \log\left[\det X'X\right]\right\}.$$

Remarks :

1. Least squares methods only focus on the first right-hand side term of the log-likelihood function. Exact methods account for both right-hand side terms.
2. This method generalizes to ARMA(p,q) models.