Macroeconometrics

Lecture 4

ML estimation

Convenor: Stefano Grassi

Motivation

- The method of maximum likelihood (ML) is a useful estimation method when the parameters of a model cannot be estimated by OLS. Tests of restrictions are also available (likelihood ratio test principle).
- The method requires to specify fully the probability distribution of the observed sample, not just for example the conditional mean of the sample (like for the linear regression model when we estimate it by OLS).
- Thus, in the linear regression model, for ML estimation, we must also specify the dependence structure of the sample through the error terms (such as "they are i.i.d."), and the distribution of the error term, (such as $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$).
- We could also choose another distribution.

ML for linear regression

- Let $y_t = \beta' x_t + \epsilon_t,\ (t = 1, \ldots, n)$ with $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$.
- By independence, the joint distribution of the ϵ_t's is

$$f(\epsilon_1, \epsilon_2, \ldots, \epsilon_n | \beta, \sigma^2) = \prod_{t=1}^{n} f(\epsilon_t | \beta, \sigma^2), \qquad (1)$$

  where $f(\epsilon_t | \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\epsilon_t^2}{2\sigma^2})$.
- We do not observe ϵ_t but y_t, so we just replace ϵ_t by $y_t - \beta' x_t$. Said differently: $f(y_t | x_t) \sim N(\beta' x_t, \sigma^2)$, and the joint probability distribution of the observed sample is

$$f(y_1, y_2, \ldots, y_n | \beta, \sigma^2) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \beta' x_t)^2}{2\sigma^2} \right). \qquad (2)$$

Likelihood function: idea

- The likelihood function (LF) of the sample $y_1, y_2, \ldots, y_n$ is the joint probability distribution of that sample, viewed as a function of the parameters of the model (β and σ^2).
- Indeed, for a given value of β and σ^2, we can compute (2): it gives us the "probability" of having observed the sample $y_1, y_2, \ldots, y_n$ for the given value of the parameters.
- If we take another possible value of β and σ^2, we obtain another value of that probability.
- The **ML ESTIMATE (MLE)** is the value of the parameters that maximizes the LF; **that is, it is the value of β and σ^2 for which the sample is the most likely (or probable) to have been observed**.

Likelihood function: notations

- To emphasize what is the LF, we use a special notation: $L(\beta, \sigma^2|y_1, y_2, \ldots, y_n)$ designates the LF, where the arguments of the function are β and σ^2, and not the observed data $y_1, y_2, \ldots, y_n$. Formally,

$$L(\beta, \sigma^2|y_1, y_2, \ldots, y_n) = f(y_1, y_2, \ldots, y_n|\beta, \sigma^2). \tag{3}$$

  $L(\beta, \sigma^2|y_1, y_2, \ldots, y_n)$ maps from R^k to R_+, whereas $f(y_1, y_2, \ldots, y_n|\beta, \sigma^2)$ maps from R^n to R_+.
- When it is clear from the context what the data are, we even write $L(\beta, \sigma^2)$ instead of $L(\beta, \sigma^2|y_1, y_2, \ldots, y_n)$.
- We also use the more synthetic and generic notations $L(\theta)$ and $L(\theta|y)$: θ for the parameters (thus $\theta = (\beta', \ \sigma^2)$ in the regression case), and y for the sample of size n.

Steps for ML estimation

Finding the ML estimate of the parameters of a model involves two steps:

- Step 1: from the model formulation and assumptions, obtain the likelihood function expression.
- Step 2: find the value of the parameters that maximize this function.

  Thus, we must solve the maximization problem:

  $$\max_{\theta \in \Theta} L(\theta | y),$$

  where Θ is the set of admissible values of the parameters θ.

Solving the max problem

- It is much easier, and equivalent, to solve

$$\max_{\theta \in \Theta} l(\theta|y),$$

  where $l(\theta|y) = \log L(\theta|y)$ (natural logarithm). Since $L(\theta|y) = \prod_{t=1}^{n} f(y_t|\theta)$, in logarithm $l(\theta|y) = \sum_{t=1}^{n} \log f(y_t|\theta) = \sum_{t=1}^{n} l_t(\theta)$.

- Thus, for the linear regression example, we must solve

$$\max_{\beta \in \mathrm{R}^k, \sigma^2 > 0} \left(-\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (y_t - \beta' x_t)^2 \right),$$

  where we have neglected the term $-n \log(\sqrt{2\pi})$ because it does not depend on β and σ^2.

First order conditions

- If $l(\theta)$ is differentiable in θ, the MLE is the solution of the system of first order conditions (FOC)

  $$q(\theta) = 0 \quad \text{where} \quad q(\theta) = \frac{\partial l(\theta)}{\partial \theta},$$

  provided that the second order condition holds; that is, the matrix of second derivatives, evaluated at the solution of the FOC must be negative definite.
  NB: $q(\theta)$ is called the score function.
- Let us call $\hat{\theta}$ the solution of the FOC: then $\hat{\theta}$ is the MLE if

  $$q(\hat{\theta}) = 0 \quad \text{and} \quad H(\hat{\theta}) < 0.$$

  where $H(\hat{\theta})$ is the Hessian matrix $\frac{\partial l(\theta)}{\partial \theta \partial \theta'}$ evaluated at $\hat{\theta}$.

FOC for β

- For the linear regression example, the FOC are

$$\frac{1}{\sigma^2}\sum_{t=1}^{n} x_t(y_t - \beta' x_t) = 0 \quad (k \text{ equations}) \tag{4}$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{t=1}^{n}(y_t - \beta' x_t)^2 = 0 \quad (1 \text{ equation}) \tag{5}$$

- The solution of (4) does not depend on σ^2 and is the solution of $\sum_{t=1}^{n} x_t(y_t - \beta' x_t) = 0$: this is the system of "normal equations" of OLS estimation.
- The MLE, assuming normality of the error terms, is the same as the OLS estimate.

FOC for σ^2

- Having solved (4), we just have to solve

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{t=1}^{n}(y_t - \hat{\beta}' x_t)^2 = 0,$$

  which gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{\beta}' x_t)^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2. \qquad (6)$$

- The MLE of σ^2 is not the "usual" estimator

$$s^2 = \tfrac{1}{n-k}\textstyle\sum_{t=1}^{n} e_t^2,$$

  but if n is large, the difference is small.

Second order conditions

- Computations give

$$H(\hat{\beta}, \hat{\sigma}^2) = \begin{pmatrix} -\frac{\sum_{t=1}^n x_t x_t'}{\hat{\sigma}^2} & 0 \\ 0' & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix} = \begin{pmatrix} -\frac{X'X}{\hat{\sigma}^2} & 0 \\ 0' & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix},$$

  which is negative definite since $X'X$ is positive definite (if X is of rank k).
- $\hat{\beta}$ and $\hat{\sigma}^2$ are the MLEs of β and σ^2.

Properties of the MLE

- Consistency: $\hat{\theta}_n \xrightarrow{p} \theta$ where θ is the parameter value in the DGP (called "true value").
- Asymptotic normality: we can proceed as if $\hat{\theta}_n \sim \mathrm{N}(\theta, \hat{V})$ if n is large, knowing that the larger n, the better the approximation.
- $\hat{V}$ is an estimate of the exact variance-covariance matrix V because the latter depends in general on some unknown parameters (in particular, but not exclusively, the true θ itself).
- There is no unique choice of $\hat{V}$ but we give the usual choices available in econometric software packages (like OxMetrics, TSP, EVIEWS, RATS, SAS, STATA...)

Formulas for $\hat{V}$

- Based on the Hessian matrix:

$$\hat{V}_H = -[H(\hat{\theta})]^{-1} = -\left[\sum_{t=1}^{n} H_t(\hat{\theta})\right]^{-1},$$

  where $H_t(\hat{\theta}) = \frac{\partial l_t(\theta)}{\partial\theta\partial\theta'}$ evaluated at $\hat{\theta}$.

- Based on the score contributions:

$$\hat{V}_G = \left[\sum_{t=1}^{n} q_t(\hat{\theta})q_t(\hat{\theta})'\right]^{-1},$$

  where $q_t(\hat{\theta}) = \frac{\partial l_t(\theta)}{\partial\theta}$ evaluated at $\hat{\theta}$ is the contribution of observation t to the score function.
  NB: $q(\theta) = \sum_{t=1}^{n} q_t(\theta)$.

The likelihood ratio test

Under the regularity conditions, one way to test restrictions, say $H_0 : R\theta = r$ is as follows:

- Step 1: estimate the model by ML without imposing the restrictions. Save $l(\hat{\theta})$, the value of the LLF at the maximum.
- Step 2: estimate this time by imposing the restrictions. Save $l(\tilde{\theta})$, the value of the LLF at the constrained maximum. So, $\tilde{\theta} = arg \max_\theta l(\theta)$ subject to $R\theta = r$.
- Step 3: Compute the likelihood ratio (LR) test statistic

  $$LR = 2[l(\hat{\theta}) - l(\tilde{\theta})].$$

  Reject H_0 at level α if the p-value of LR is smaller than α, where the p-value is obtained using the $\chi^2(m)$. Notice that by construction $LR \geq 0$.

The likelihood ratio test

- Example: for the MA(1,2,5) model of the log(USGDP), the value of the maximized log-likelihood is -330.95578.
- For the MA(1,2) model, the value is -332.468523 (smaller since there is one parameter less).
- The LR test statistic for the null hypothesis that the MA(5) coefficient is equal to 0 is given by

  $2[-330.95578 - (-332.468523)] = 3.025486 < 3.84,$

  the $\chi^2(1)$ critical value at 5%.

  The hypothesis is not rejected at that level.

Possible problems with ML

- Solving the FOC is not usually possible analytically, like was done in the regression example. Numerical methods are implemented in econometric softwares.
- The MLE is not always unique: the FOC system may have several solutions. If this is suspected, one must search numerically for several maxima and choose the global one.
- Extreme case of the previous problem: the MLE does not exist (or there is an infinity of solutions: this is called an identification problem). An example of this: linear regression when X is not of full rank.
- The MLE may be on a boundary of the parameter space. Then its asymptotic normality does not hold and the LR test is not asymptotically distributed as χ^2.

Strength and weakness of ML

- Strength: if the assumptions are correct, the MLE is the most efficient estimator asymptotically, in the sense that it is unbiased asymptotically and has the smallest possible variance. Said more intuitively, ML uses in the best way the information in the data to estimate the parameters.
- Weakness: ML requires a choice of probability distribution. So, what are the consequences of a wrong assumption about the probability distribution of the y_t's?
- The answer to this question is:
  - in some cases, the MLE is not even consistent;
  - in other cases, it is still consistent and asymptotically normal but the variance matrix $\hat{V}$ is different from that given previously.

Quasi-ML

- The quasi-ML (QML) estimate is the MLE computed from a LF based on a (possibly wrong) probability distribution for which $\mathrm{E}[q_t(\theta)] = 0$. We denote it by $\hat{\theta}_Q$.
- In many cases, when $y_t \in \mathrm{R}$, that wrong distribution can be taken to be a normal one, with the mean and variance specified in the model. The LF is then called the "quasi-likelihood function".
- The QMLE is consistent and asymptotically normal, so in large samples we can proceed as if $\hat{\theta}_Q \sim \mathrm{N}(\theta, \hat{V}_Q)$, where

$$\hat{V}_Q = \hat{V}_H^{-1} \hat{V}_G \hat{V}_H^{-1}.$$

NB: In the regression model, $\hat{V}_Q$ is $\mathrm{Var}_{HC}(b)$ (White's formula).

Tests based on QML

- Using the result that $\hat{\theta}_Q \sim \mathrm{N}(\theta, \hat{V}_Q)$, we can use Wald tests that is distributed as χ^2 under H_0, or for a single restriction, we can use the $\mathrm{N}(0,1)$ version of the test.
- Note however that we cannot apply the the LR test principle in the QML case, that is $2[l(\hat{\theta}_Q) - l(\tilde{\theta}_Q)]$ is not a valid χ^2 test statistic.

OLS for AR models

- These models are particular ARDL models. They can be cast in the form of a regression equation $y_t = \beta' x_t + \epsilon_t$ after suitably defining x_t and β.
- When we use OLS for an AR(p) model and we have n observations, we "loose" p observations: we use $y_1, y_2, \ldots, y_p$ to construct the X matrix. Its first row being $1, y_1, y_2, \ldots, y_p$, etc.
- We say that we condition on the first p observations. But we cannot use y_1, y_2, . . . , y_p on the left-hand side of the regressions equation (as dependent variable).

Conditional ML for AR models(1)

- The likelihood function conditional on the initial p observations is:

$$L_c(\theta) = \prod_{t=p+1}^{n} f(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \theta)$$

  where $f(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \theta)$ is the conditional density of y_t given $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, with mean $\beta' x_t = \alpha_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p}$ and variance σ^2.
- Here $\theta = (\beta', \sigma^2) = (\alpha_0, \phi_1, \phi_2, \ldots, \phi_p, \sigma^2)$.

Conditional ML for AR models(2)

- This LF is the "probability" of observing $y_{p+1}, y_{p+2}, \ldots y_n$ given that we have observed $y_1, y_2, \ldots, y_p$ and given the value θ of the parameters.
- Assuming $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$ for $t \geq p+1$, we know that $f(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ is the $\mathrm{N}(\beta' x_t, \sigma^2)$ density, thus:

$$l_c(\theta) = -\frac{1}{2}(n-p)\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=p+1}^{n}(y_t - \beta' x_t)^2$$

Maximizing this gives the OLS estimator of β and (6) for σ^2.

Exact ML for AR(1) model

- The exact likelihood function uses all observations. It is equal to the conditional one times the marginal density of the first observation:

$$L(\theta) = f(y_1|\theta) \prod_{t=2}^{n} f(y_t|y_{t-1};\theta).$$

- For the AR(1) model, assuming normality, it can be shown that $f(y_1|\theta) \sim \mathrm{N}\left(\frac{\alpha_0}{1-\phi_1}, \frac{\sigma^2}{1-\phi_1^2}\right)$. The LLF is then

$$\begin{aligned} l(\theta) &= -\frac{1}{2}\log\frac{\sigma^2}{1-\phi_1^2} - \frac{1}{2}\left(y_1 - \frac{\alpha_0}{1-\phi_1}\right)^2 \\ &\quad -\frac{1}{2}(n-1)\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t - \alpha_0 - \phi_1 y_{t-1})^2. \end{aligned}$$

Exact ML for AR(p) models

- The exact likelihood function is equal to the conditional one times the joint density of the first p observations:

$$L(\theta) = f(y_1, y_2, \ldots y_p|\theta) \prod_{t=p+1}^{n} f(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \theta).$$

- The density $f(y_1, y_2, \ldots y_p|\theta)$ is multivariate normal if $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$ but its formula is heavy. Anyway, the solution of the maximization of the LLF is obtained by numerical methods.

Conditional ML for MA models ($y_t = \mu + \theta_1 \epsilon_t$)

- If $\epsilon_t \sim \mathrm{N}(0, \sigma^2)$, then the LLF is

$$l_c(\theta) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} \epsilon_t^2, \qquad (7)$$

  where $\epsilon_t = y_t - \mu - \theta_1 \epsilon_{t-1}$ for $t = 1, 2, \ldots, n$ in the MA(1). Problem: we need ϵ_0 to compute ϵ_1.
- Solution: we condition on a value of ϵ_0: $\epsilon_0 = 0$ (the expectation) is convenient since then we compute successively $\epsilon_1 = y_1 - \mu$, $\epsilon_2 = y_2 - \mu - \theta_1 \epsilon_1$, $\epsilon_3 = y_3 - \mu - \theta_1 \epsilon_2$, ...
- For a MA(q), the LLF is also given by (7). We can fix $\epsilon_0 = \epsilon_{-1} = \ldots = \epsilon_{1-q} = 0$ and compute $\epsilon_t = y_t - \mu - \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$ for $t \geq 1$.

Maximum Likelihood Estimation (State Space Model)

- Let $\theta \in \Theta$ denote a vector containing the so-called hyperparameters, i.e. the vector of structural parameters other than the scale factor σ^2.
- The state space model depends on θ via the system matrices $Z_t = Z_t(\theta), G_t = G_t(\theta), T_t = T_t(\theta), H_t = H_t(\theta)$.

Local Level model

Consider the local level model:

$$\begin{aligned} y_t &= \alpha_t + \epsilon_t \qquad & \epsilon_t \sim \mathrm{N}(0, \sigma_\epsilon^2) \\ \alpha_t &= \alpha_{t-1} + \eta_t \qquad & \eta_t \sim \mathrm{N}(0, \sigma_\eta^2) \end{aligned} \tag{8}$$

The parameter to be estimated are σ_ϵ^2 and σ_η^2. Those parameter are restricted in the region $[0, +\infty)$. It is much better to maximize the function in the domain $(-\infty, +\infty)$.

Reparametrization

The vector of parameters, θ, has two unrestricted elements, which are related to the model's hyperparameters by:

$$\sigma_\eta^2 = \exp(2\theta_1), \qquad \sigma_\epsilon^2 = \exp(2\theta_2),$$

or in the inverse way:

$$\theta_1 = \frac{1}{2}\log(\sigma_\eta^2) \qquad \theta_2 = \frac{1}{2}\log(\sigma_\epsilon^2)$$

Likelihood using the Kalman filter

- Let $\mathrm{L}(Y_n;\theta)$ denote the log-likelihood function, that is the log of the joint density of the sample time series $\{y_1, \dots, y_n\}$ as a function of the parameters θ.
- The log-likelihood can be evaluated by the prediction error decomposition:

$$\mathrm{L}(Y_n;\theta) = \log f(y_1, \dots, y_n;\theta) = \sum_{t=1}^{n} \log f(y_t|Y_{t-1};\theta).$$

- The predictive density $f(y_t|Y_{t-1};\theta)$ is evaluated with the support of the **Kalman Filter**.

Proof

- In this case we assume **NORMALITY**.
  Recall that $F_t = \mathrm{Var}(y_t|Y_{t-1})$ and $\nu_t = y_t - Z_t\tilde{\alpha}_t$.
- Then we can substitute $\mathrm{N}(Z_t\tilde{\alpha}_t, F_t)$ for $f(y_t|Y_{t-1})$ and we get, apart from constant :

$$\log \mathrm{L}(Y_n) \quad = \quad -\tfrac{1}{2}\left(\sum_{t=1}^{n} \log|F_t| + \sum_{t=1}^{n} \nu_t' F_t^{-1} \nu_t\right).$$

- The likelihood function can be maximized numerically by a quasi-Newton optimization routine.