Macroeconometrics

Lecture 3 Maximum Likelihood Estimation

Tomasz Woźniak

Department of Economics
University of Melbourne

Likelihood function

Estimation: analytical solution

Properties of the maximum likelihood estimator

Maximum likelihood inference

Useful readings:

Harris, Hurn, & Martin (2012) Chapter 1: The Maximum Likelihood Principle, Econometric Modelling with Time Series

Harris, Hurn, & Martin (2012) Chapter 2: Properties of Maximum Likelihood Estimators, Econometric Modelling with Time Series

Objectives.

- To learn the basics of the maximum likelihood method
- To derive analytical solutions for a simple model
- To look at theoretical results that enable hypothesis testing

Learning outcomes.

- Setting up an optimisation problem
- Using calculus to provide closed-form solutions
- Constructing a statistical test of a hypothesis

Likelihood function

A simple model

Univariate linear regression model.

$$y_t = \beta x_t + \epsilon_t$$
$$\epsilon_t | x_t \sim iid\mathcal{N}\left(0, \sigma^2\right)$$

y_t – dependent variable

$\theta = \left(\beta, \sigma^2\right)'$ – a 2×1 vector of unknown parameters

x_t – explanatory variable

ϵ_t – error term

T – sample size and $t \in (1, \ldots, T)$

A simple model

The model in matrix notation.

$$Y = \beta X + E$$
$$E|X \sim \mathcal{N}\left(\mathbf{0}_T, \sigma^2 I_T\right)$$

Data matrices.

$$\underset{(T\times 1)}{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \qquad \underset{(T\times 1)}{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix} \qquad \underset{(T\times 1)}{E} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{bmatrix}$$

Predictive density

Assumptions about the model and the conditional distribution of the error term determine the predictive distribution of data given the parameters and explanatory variables:

$$\begin{aligned} Y &= \beta X + E \\ E|X &\sim \mathcal{N}\left(\mathbf{0}_T, \sigma^2 I_T\right) \end{aligned} \quad \Rightarrow \quad \begin{aligned} Y &= \beta X + E \\ Y|X &\sim \mathcal{N}\left(\beta X, \sigma^2 I_T\right) \end{aligned}$$

Predictive density

Linear transformation of a normal vector.
Let a random vector Y follow an N-variate normal distribution with the mean vector μ and the covariance matrix Σ:

$$Y \sim \mathcal{N}_N(\mu, \Sigma)$$

Let $Z = AY + b$. Then:

$$Z \sim \mathcal{N}_N(A\mu + b, A\Sigma A')$$

Likelihood function

A likelihood function is equivalent to the conditional distribution of the data, given the parameters of the model.

However, for the purpose of the estimation and after plugging in data Y and X we treat it as a function of unknown parameters θ.

$$
\begin{aligned}
L(\theta|Y,X) &= L\left(\beta,\sigma^2|Y,X\right) = p\left(Y|X,\beta,\sigma^2\right) = \mathcal{N}_T\left(\beta X,\sigma^2 I_T\right)\\
&= (2\pi)^{-\frac{T}{2}}\det\left(\sigma^2 I_T\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(Y-\beta X)'\left(\sigma^2 I_T\right)^{-1}(Y-\beta X)\right\}\\
&= (2\pi)^{-\frac{T}{2}}\left(\sigma^2\right)^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\frac{1}{\sigma^2}(Y-\beta X)'(Y-\beta X)\right\}
\end{aligned}
$$

Useful operations.
Let c be a scalar and X an $N \times N$ matrix. Then $\det(cX) = c^N \det(X)$.

The likelihood principle

All of the information about the parameters of the model θ that is embedded in the dataset Y is captured by the likelihood function.

log-likelihood function

To derive the analytical solution and to be able to evaluate the likelihood function for any values of $\theta \in \boldsymbol{\Theta}$, the logarithmic transformation is applied through which the log-likelihood function is obtained. $\boldsymbol{\Theta}$ denotes the parameter space, that is, a set of all admissible values of the parameters.

$$
\begin{aligned}
l\left(\theta|Y,X\right) &= \ln L\left(\theta|Y,X\right) \\
&= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\frac{1}{\sigma^2}(Y-\beta X)'(Y-\beta X) \\
&= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\frac{1}{\sigma^2}(Y'Y-\beta 2X'Y+\beta^2X'X)
\end{aligned}
$$

The maximum likelihood estimator

The maximum likelihood estimator (MLE) of θ, denoted by $\hat{\theta}$, is found where the log-likelihood function is at its maximum:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, l(\theta | Y, X)$$

Finding the maximum of the log-likelihood function is equivalent to finding the maximum of the likelihood function as the logarithm is a monotonic transformation that preserves local optima.

The derivations and properties are feasible under regularity conditions.

Regularity conditions

Let θ_0 denote the true values of the parameters θ.

A1 Existence

The following expectation exists:

$$\mathbb{E}[l(\theta|Y,X)] = \int_{-\infty}^{\infty} l(\theta|Y,X) L(\theta_0|Y,X) dY$$

A2 Convergence

$l(\theta|Y,X)$ converges in probability to its expectation uniformly in θ.

$$\text{plim } l(\theta|Y,X) = \mathbb{E}[l(\theta|Y,X)]$$

A3 Continuity

$l(\theta|Y,X)$ is continuous in θ.

A4 Differentiability

$l(\theta|Y,X)$ is at least twice differentiable in an open interval around θ_0.

A5 Interchangeability

The differentiation and integration order of $l(\theta|Y,X)$ is interchangeable.

Estimation: analytical solution

Estimation: analytical solution

To derive the analytical solution of MLE use calculus.

Gradient vector.

$$\underset{(2\times 1)}{G(\theta)} = \frac{\partial l\,(\theta|Y,X)}{\partial \theta} = \begin{bmatrix} \frac{\partial l(\theta|Y,X)}{\partial \beta} \\ \frac{\partial l(\theta|Y,X)}{\partial \sigma^2} \end{bmatrix}$$

The MLE occurs where all of the gradients are equal to zero:

$$G(\hat{\theta}) = \left.\frac{\partial l\,(\theta|Y,X)}{\partial \theta}\right|_{\theta=\hat{\theta}} = \mathbf{0}_2$$

Estimation: analytical solution

Hessian matrix.

$$\underset{(2\times 2)}{H(\theta)} = \frac{\partial^2 l\,(\theta|Y,X)}{\partial\theta\partial\theta'} = \begin{bmatrix} \frac{\partial^2 l(\theta|Y,X)}{\partial^2\beta} & \frac{\partial^2 l(\theta|Y,X)}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 l(\theta|Y,X)}{\partial\sigma^2\partial\beta} & \frac{\partial^2 l(\theta|Y,X)}{\partial^2\sigma^2} \end{bmatrix}$$

The MLE maximizes the log-likelihood function when the Hessian matrix ordinate at the MLE:

$$H(\hat{\theta}) = \left.\frac{\partial^2 l\,(\theta|Y,X)}{\partial\theta\partial\theta'}\right|_{\theta=\hat{\theta}}$$

is negative definite.

Estimation: analytical solution

The gradient.

$$G(\theta) = \begin{bmatrix} \frac{\partial l(\theta|Y,X)}{\partial \beta} \\ \frac{\partial l(\theta|Y,X)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\frac{1}{\sigma^2}\left(-2X'Y + \beta 2X'X\right) \\ -\frac{T}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}(Y-\beta X)'(Y-\beta X) \end{bmatrix}$$

Necessary condition.

$$G(\hat{\theta}) = \begin{bmatrix} \left.\frac{\partial l(\theta|Y,X)}{\partial \beta}\right|_{\theta=\hat{\theta}} \\ \left.\frac{\partial l(\theta|Y,X)}{\partial \sigma^2}\right|_{\theta=\hat{\theta}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -\frac{1}{2}\frac{1}{\hat{\sigma}^2}\left(-2X'Y + \hat{\beta} 2X'X\right) \\ -\frac{T}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2}\frac{1}{(\hat{\sigma}^2)^2}(Y-\hat{\beta} X)'(Y-\hat{\beta} X) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Estimation: analytical solution

The first equation.

$$
\begin{aligned}
0 &= -\frac{1}{2}\frac{1}{\hat{\sigma}^2}\left(-2X'Y + \hat{\beta}2X'X\right) \\
\hat{\beta}X'X &= X'Y \\
\hat{\beta} &= (X'X)^{-1}X'Y
\end{aligned}
$$

The second equation.

$$
\begin{aligned}
0 &= -\frac{T}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2}\frac{1}{\left(\hat{\sigma}^2\right)^2}(Y - \hat{\beta}X)'(Y - \hat{\beta}X)\Bigg/ \cdot \frac{2\left(\hat{\sigma}^2\right)^2}{T} \\
\hat{\sigma}^2 &= \frac{1}{T}(Y - \hat{\beta}X)'(Y - \hat{\beta}X)
\end{aligned}
$$

Estimation: analytical solution

The Hessian matrix.

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 l(\theta|Y,X)}{\partial^2 \beta} & \frac{\partial^2 l(\theta|Y,X)}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 l(\theta|Y,X)}{\partial \sigma^2 \partial \beta} & \frac{\partial^2 l(\theta|Y,X)}{\partial^2 \sigma^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma^2} X'X & -\frac{1}{\left(\sigma^2\right)^2} X'(Y-\beta X) \\ & \frac{T}{2}\frac{1}{\left(\sigma^2\right)^2} - \frac{1}{\left(\sigma^2\right)^3}(Y-\beta X)'(Y-\beta X) \end{bmatrix}$$

Sufficient condition.
$H(\hat{\theta})$ must be negative definite.

$$H(\hat{\theta}) = \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} X'X & -\frac{1}{\left(\hat{\sigma}^2\right)^2} X'\hat{E} \\ & \frac{1}{2}\frac{T}{\left(\hat{\sigma}^2\right)^2} - \frac{1}{\left(\hat{\sigma}^2\right)^3}\hat{E}'\hat{E} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} X'X & 0 \\ 0 & -\frac{1}{2}\frac{T}{\left(\hat{\sigma}^2\right)^2} \end{bmatrix}$$

where $\frac{1}{T} X'\hat{E} = 0$ (exogeneity condition), and $\hat{\sigma}^2 = \frac{1}{T}\hat{E}'\hat{E}$.

Estimation: analytical solution

Negative definite matrix.
A symmetric $N \times N$ matrix Z is negative definite if $z'Zz < 0$ for all $N \times 1$ vectors $z \neq \mathbf{0}_N$.

A symmetric 2×2 matrix Z is negative definite if $Z_{11} < 0$ and $\det(Z) > 0$.

Since $-\frac{1}{\hat{\sigma}^2} X'X < 0$ and $\det(H(\hat{\theta})) = \frac{1}{2} \frac{1}{(\hat{\sigma}^2)^3} X'X > 0$ the Hessian matrix is negative definite.

The MLE:

$$\hat{\theta} = \begin{bmatrix} \hat{\beta} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} (X'X)^{-1} X'Y \\ \frac{1}{T}(Y - \hat{\beta}X)'(Y - \hat{\beta}X) \end{bmatrix}$$

is a point at which the log-likelihood achieves the global maximum.

Properties of the MLE

MLE properties: consistency

Consistency.
The probability limit of the MLE when the sample size increases is the vector of the true parameter values.

$$\text{plim } \hat{\theta} = \theta_0$$

Definition of plim.

$$\lim_{T \to \infty} \Pr\left[|\hat{\theta} - \theta_0| < c \right] = 1, \text{ for any } c > 0$$

MLE properties: asymptotic normality

Normality.

The MLE converges in distribution to the following normal distribution when the sample size goes to infinity.

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Omega(\theta_0)\right)$$

where $\Omega(\theta_0)$ is the inverse of the Fisher information matrix:

$$\Omega(\theta_0) = TI^{-1}(\theta_0) = T\left[-\mathbb{E}\left[H(\theta_0)\right]\right]^{-1}$$

Asymptotic distribution.

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{T}\Omega(\theta_0)\right)$$

MLE properties: asymptotic normality

Estimator of covariance of $\hat{\theta}$.

$$\widehat{Var}(\hat{\theta}) = \frac{1}{T}\Omega(\theta)|_{\theta=\hat{\theta}} = \left[-H(\theta)\right]^{-1}\Big|_{\theta=\hat{\theta}}$$
$$= \begin{bmatrix} \hat{\sigma}^2(X'X)^{-1} & 0 \\ 0 & \frac{2\left(\hat{\sigma}^2\right)^2}{T} \end{bmatrix}$$

Estimation standard erros.

$$\hat{se}\left(\hat{\beta}\right) = \hat{\sigma}(X'X)^{-\frac{1}{2}}$$
$$\hat{se}\left(\hat{\sigma}^2\right) = \sqrt{\frac{2}{T}}\hat{\sigma}^2$$

MLE properties: efficiency and invariance

Efficiency.

The covariance of the MLE hits the Rao-Cramer lower bound:

$$\frac{1}{T}\Omega(\theta_0) = I^{-1}(\theta_0)$$

No other estimator has lower standard errors than the MLE.

Invariance.

The MLE of a continuous and differentiable function of parameters $g(\theta)$ is given by:

$$\widehat{g(\theta)} = g(\theta)|_{\theta=\hat{\theta}} = g(\hat{\theta})$$

Example: $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

Maximum likelihood inference

Wald test: notation

Let $\mathbf{R}(\theta) : \mathbb{R}_k \to \mathbb{R}_l$ be a l-variate function of $k \times 1$ vector of parameters, such that:

$$\mathbf{R}(\theta) = \begin{bmatrix} \mathbf{R}_1(\theta) \\ \vdots \\ \mathbf{R}_l(\theta) \end{bmatrix}$$

Let $\mathbf{J}(\theta)$ be a $l \times k$ Jacobian matrix such that:

$$\mathbf{J}(\theta) = \begin{bmatrix} \frac{\partial \mathbf{R}_1(\theta)}{\partial \theta_1} & \dots & \frac{\partial \mathbf{R}_1(\theta)}{\partial \theta_k} \\ \vdots & & \vdots \\ \frac{\partial \mathbf{R}_l(\theta)}{\partial \theta_1} & \dots & \frac{\partial \mathbf{R}_l(\theta)}{\partial \theta_k} \end{bmatrix}$$

delta method

Let $\hat{\theta}$ be a vector of normally distributed parameters with an asymptotic covariance matrix $Var[\hat{\theta}]$.

Problem: What is an asymptotic covariance matrix of an estimator of the function of parameters $\mathbf{R}(\hat{\theta})$?

Soluton: Use delta method to obtain:

$$Var\left[\mathbf{R}(\hat{\theta})\right] = \mathbf{J}(\hat{\theta})Var[\hat{\theta}]\mathbf{J}(\hat{\theta})'$$

Wald test

Estimation of the unrestricted model is required. The estimator of parameters for this model is denoted by $\hat{\theta}$.

Hypotheses:

$$\mathcal{H}_0 : \mathbf{R}(\theta) = \mathbf{r}$$
$$\mathcal{H}_1 : \mathbf{R}(\theta) \neq \mathbf{r}$$

Test statistic:

$$\mathbb{W} = \left(\mathbf{R}(\hat{\theta}) - \mathbf{r}\right)' \left(\mathbf{J}(\hat{\theta}) Var[\hat{\theta}] \mathbf{J}(\hat{\theta})'\right)^{-1} \left(\mathbf{R}(\hat{\theta}) - \mathbf{r}\right)$$

Asymptotic distribution: $\mathbb{W} \sim \chi^2(l)$

Reject the null hypothesis if $\mathbb{W} > \chi^2_\alpha(l)$, where $\chi^2_\alpha(l)$ is a $100(1-\alpha)$th percentile of the χ^2 distribution with l degrees of freedom.

Likelihood ratio test

Estimation of the restricted and unrestricted model is required.

Hypotheses:

$$\mathcal{H}_0 : \mathbf{R}(\theta) = \mathbf{r}$$
$$\mathcal{H}_1 : \mathbf{R}(\theta) \neq \mathbf{r}$$

Test statistic:

$$\mathbb{LR} = 2\Big(l(\hat{\theta}|Y, X) - l_R(\hat{\theta}_R|Y, X)\Big)$$

where $l(\hat{\theta}|Y, X)$ - is the value of the log-likelihood function for the unrestricted model, and $l_R(\hat{\theta}_R|Y, X)$ - is the value of the log-likelihood function for the restricted model

Asymptotic distribution: $\mathbb{LR} \sim \chi^2(l)$

Reject the null hypothesis if $\mathbb{LR} > \chi^2_\alpha(l)$, where $\chi^2_\alpha(l)$ is a $100(1-\alpha)$th percentile of the χ^2 distribution with l degrees of freedom.

Lagrange multiplier test

Estimation of the restricted model is required. The estimated parameter vector for this model is denoted by $\hat{\theta}_R$. Note that it contains the values of restricted parameters.

Hypotheses:

$$\mathcal{H}_0 : \mathbf{R}(\theta) = \mathbf{r}$$
$$\mathcal{H}_1 : \mathbf{R}(\theta) \neq \mathbf{r}$$

Test statistic:

$$\mathbb{LM} = TG(\hat{\theta}_R)'\Omega(\hat{\theta}_R)G(\hat{\theta}_R)$$

where $G(\theta)$ and $\Omega(\theta)$ are the gradient vector and the information matrix derived for the log-likelihood function of the unrestricted model.

Asymptotic distribution: $\mathbb{LM} \sim \chi^2(l)$

Reject the null hypothesis if $\mathbb{LM} > \chi^2_\alpha(l)$, where $\chi^2_\alpha(l)$ is a $100(1-\alpha)$th percentile of the χ^2 distribution with l degrees of freedom.

Maximum Likelihood Estimation

Maximum likelihood estimation and inference is a powerful tool for data analysis.

It is still one of the most frequently used methods in macroeconometrics as long as its application is **numerically feasible**.

Some specialised techniques, such as appropriately set **numerical optimization** and **concentration** of the likelihood function that are presented later during this subject, make its application simpler for some models.