Other estimation principles: Method of Moments and Maximum Likelihood (2 of 2)

Ragnar Nymoen

University of Oslo

2 May 2013
References

- HGL, Ch 10.3-10.3.3, 10.3.5, C8.C8.2
- BN Kap 6-6.3, 9.2.7
Three estimation principles used in econometrics I

1. Minimization of distance: *Ordinary least squares* (OLS)
3. Estimate parameters of assumed probability distribution: *Maximum likelihood* (ML)

- OLS gives BLUE estimators for the parameters of the conditional expectation function.
- MM gives consistent but not efficient estimators of the parameters of the conditional expectation function.
### OLS and MM and parameters of interest

<table>
<thead>
<tr>
<th>Cond. expectation function</th>
<th>Structural equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>BLUE</td>
</tr>
<tr>
<td>MM</td>
<td>Consistent, but not efficient</td>
</tr>
</tbody>
</table>
The maximum likelihood principle I

- ML estimation requires that we are able to write down the *likelihood-function* for the random variables included in the econometric model.

- Unlike OLS, ML estimation starts with precise and complete distributional assumptions, rather than appending those assumptions to the classical assumptions about the disturbance.

- The ML principle is to choose those values of the parameters of the likelihood function that maximise the probability of observing the sample values of the random variables.

- We show the principle for the case where random variables are normally distributed and the parameters of interest are in the conditional expectation function of $Y$ given $X$. 
Conditional normal distribution I

We have from *Lecture 6* that when $Y$ and $X$ have a joint normal probability density function (pdf):

$$f_{XY}(y, x) = \frac{1}{\sigma_Y \sigma_X 2\pi \sqrt{1 - \rho_{XY}^2}} \times \exp \left[ -\frac{1}{2} \frac{(z_Y^2 - 2\rho_{XY} z_Y z_X + z_X^2)}{(1 - \rho_{XY}^2)} \right]$$

$$z_Y = \frac{(y - \mu_Y)}{\sigma_Y} \text{ and } z_X = \frac{(x - \mu_X)}{\sigma_X}$$

the conditional pdf for $Y$ given $X$ is:

$$f_{Y|X}(y | x) = \frac{1}{\sqrt{2\pi\sigma_Y^2_{|X}}} \times \exp \left\{ -\frac{1}{2} \frac{[y - \mu_{Y|X}]^2}{\sigma_Y^2_{|X}} \right\} \quad (1)$$
Conditional normal distribution II

where

$$
\mu_{Y|X} = E(Y \mid X) = \beta_0 + \beta_1 X
$$

(2)

$$
\beta_0 = \mu_Y - \frac{\sigma_{YX}}{\sigma_X^2} \mu_X
$$

(3)

$$
\beta_1 = \frac{\sigma_{YX}}{\sigma_X^2}
$$

(4)

Remember that we can write this result as

$$
Y \mid X \sim N \left( \beta_0 + \beta_1 X, \sigma_{Y|X}^2 \right)
$$

(5)
Conditional normal distribution III

By defining

\[ \varepsilon \mid X = Y - (\beta_0 + \beta_1 X) \sim N \left( 0, \sigma_{Y \mid X}^2 \right) \]  

(6)

the distributional assumptions of \( Y \) and \( X \) that we have made can be written in **conditional model form** as:

\[ Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \mid X \sim N \left( 0, \sigma_{Y \mid X}^2 \right). \]  

(7)
We next assume that we have \( n \) independent and identically normally distributed variables \((Y_1 \mid X_1), (Y_2 \mid X_2), \ldots, (Y_n \mid X_n)\) from the same distribution.

Then the conditional expectations will be different, but the conditional variance will be the same for each \( i \): 

\[
Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \mid X_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n.
\]

where it is understood that \( \sigma^2 \) is the conditional variance.
Likelihood-function II

Since \( Y_i \mid X_i, \ i = 1, \ldots, n \), are independent, their joint distribution is a product of all \( n \) conditional distributions:

\[
f ( Y_1, \ldots, Y_n \mid X_1, \ldots, X_n ) = \prod_{i=1}^{n} \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} e^\left[ -\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right] = L ( \beta_0, \beta_1, \sigma^2 )
\]

Let us now turn the interpretation of this joint density “around” and ask: Which values of \( \beta_0, \beta_1 \) and \( \sigma^2 \) maximise the likelihood of observing the realizations of \( Y_i \mid X_i \)?

It is this interpretation that motivates the definition to the right and \( L ( \beta_0, \beta_1, \sigma^2 ) \) is called the likelihood-function.

These (most likely) parameters are unknown, but we estimate them by maximizing \( L ( \beta_0, \beta_1, \sigma^2 ) \) with respect \( \beta_0, \beta_1 \) and \( \sigma^2 \).
Log-likelihood 1

- Easier to maximise the log-likelihood function

\[
\ln L (\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^{n} \left[ -\frac{1}{2} \ln (2\pi\sigma^2) - \frac{(Y_i - \beta_0 - \beta_1X_i)^2}{2\sigma^2} \right]
\]

\[
= -\frac{1}{2} \sum_{i=1}^{n} \ln (2\pi\sigma^2) - \frac{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1X_i)^2}{2\sigma^2}
\]

\[
= -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1X_i)^2}{2\sigma^2}
\]

(8)
Log-likelihood II

We can simplify further by writing the log-likelihood in terms of \( \alpha \equiv \beta_0 + \bar{X}\beta_1 \), \( \beta_1 \) and \( \sigma^2 \):

\[
\ln L (\alpha, \beta_1, \sigma^2) = -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left( Y_i - \alpha - \beta_1 (X_i - \bar{X}) \right)^2
\]

\[
(9)
\]
Derivatives of likelihood function I

\[
\frac{\partial \ln L}{\partial \alpha} = \frac{\sum_{i=1}^{n} \{ Y_i - \alpha - \beta_1 (X_i - \bar{X}) \}}{\sigma^2}
\]

\[
\frac{\partial \ln L}{\partial \beta_1} = \frac{\sum_{i=1}^{n} \{ Y_i - \alpha - \beta_1 (X_i - \bar{X}) \} (X_i - \bar{X})}{\sigma^2}
\]

\[
\frac{\partial \ln L}{\partial \sigma^2} = -n \frac{1}{2 \sigma^2} + \frac{\sum_{i=1}^{n} \{ Y_i - \alpha - \beta_1 (X_i - \bar{X}) \}^2}{2\sigma^4}
\]
First order conditions I

- The first two 1oc’s become

\[
\sum_{i=1}^{n} \left\{ Y_i - \hat{\alpha} - \hat{\beta}_1 (X_i - \bar{X}) \right\} = 0
\]

\[
\sum_{i=1}^{n} Y_i (X_i - \bar{X}) - \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})^2 = 0
\]

- They are the same as the OLS normal equations for \( \hat{\alpha} \) and \( \hat{\beta}_1 \).
The Maximum Likelihood (ML) estimators for $\alpha$, and $\beta_1$ (and $\beta_0$) are therefore identical to the OLS estimators.

What about the ML estimator for $\sigma^2$?

From the third partial derivative:

\[-\frac{n}{2}\hat{\sigma}^2 + \frac{\sum_{i=1}^{n} \{ Y_i - \hat{\alpha} - \hat{\beta}_1 (X_i - \bar{X}) \}^2}{2\hat{\sigma}^4} = 0\]

\[-\frac{n}{2} + \frac{\sum_{i=1}^{n} \{ Y_i - \hat{\alpha} - \hat{\beta}_1 (X_i - \bar{X}) \}^2}{2\hat{\sigma}^2} = 0\]
ML estimators II

Note that \( Y_i - \hat{\alpha} - \hat{\beta}_1(X_i - \bar{X}) \) are the OLS residuals \( \hat{\epsilon}_i \).

Therefore

\[
\hat{\sigma}^2 n = \sum_{i=1}^{n} \left( Y_i - \hat{\alpha} - \hat{\beta}_1(X_i - \bar{X}) \right)^2
\]

\[
\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i^2
\]

where we use the subscript \( ML \) to distinguish the Maximum-Likelihood estimator from the unbiased estimator

\[
\hat{\sigma}^2 = \frac{1}{n - 2} \sum_{i=1}^{n} \hat{\epsilon}_i^2
\]  

(10)
The principle is the message! I

- Why bother with ML or MM if they give the same estimators for the parameters of the regression model?
- The point is the different *principles of estimation*, which have relevance for consistent estimation of parameters of interest that are not “in” the linear conditional expectation function
- MM: Models with measurement errors and simultaneous equations models
- ML: Wide range of applications to regression models that are
  - non-linear in parameters
  - have heteroskedastic errors. As shall see in E 4160, weighted least-squares estimators are easiest achieved by ML estimation
- system of equations
The principle is the message! II

- Beyond the “linear probability model for discrete choices: Probit and Logit models (E 4136)
- The common ground of these applications is the maximum likelihood principle, which is illustrated by the our example.