MSc Development Economics: Quantitative Methods
Maximum Likelihood Estimation

Selma Telalagić

University of Oxford

31st October 2013

My Lecturing for the QM Course

- This lecture
- Friday's class on MLE
- 2 lectures on structural modelling next term
- 1 class on structural modelling next term

Structure of the Lecture

- Intuition behind MLE
- Recap of OLS
- Deriving the ML estimates
- Inference
- *Reading: Hendry and Nielsen ch. 1-3,5*
- Thanks to Simon Quinn for sharing his slides and notes with me

A Statistical Model

- We will use this as a running example throughout the lecture and class
- What is the relationship between wages and schooling?
- A statistical model describing this is

$$W_i = \alpha + \beta S_i + u_i$$

- Assumptions discussed later
- This is the true relationship; how do we estimate α and β? We label the estimates as $\hat{\alpha}$ and $\hat{\beta}$.

What is Maximum Likelihood Estimation?

- A way of estimating the parameters of a statistical model i.e. α and β
- An alternative method to OLS
- Based on a likelihood approach
- What is the most likely value of a parameter that is consistent with the observed data?
- A chosen value for a parameter will change the value of the error terms: $\hat{u}_i = W_i - \hat{\alpha} - \hat{\beta} S_i$
- Maximisation takes into account changing errors and parameters jointly to yield the parameter that gives the maximum likelihood

Recap: Ordinary Least Squares

- Assumptions: Linear relationship; Conditional exogeneity; Homoscedasticity; Normal distribution of errors, Independent & identical distribution of observations
- Minimise the sum of squared errors:
-

$$\begin{aligned} & \min_{\hat{\alpha},\hat{\beta}} \sum \hat{u}_i^2 \\ \Leftrightarrow \quad & \min_{\hat{\alpha},\hat{\beta}} \sum (W_i - \hat{\alpha} - \hat{\beta} S_i)^2 \end{aligned}$$

- Differentiating this with respect to $\hat{\alpha}$ and $\hat{\beta}$ yields the following expressions:
-

$$\begin{aligned} \hat{\alpha} &= \bar{W} - \hat{\beta}\bar{S} \\ \hat{\beta} &= \frac{\sum (S_i - \bar{S})(W_i - \bar{W})}{\sum (S_i - \bar{S})^2} \end{aligned}$$

Maximum Likelihood - assumptions

- Linear relationship
- Independent & identical distribution of observations: $(W_1, S_1), ..., (W_n, S_n)$ are mutually independent
- Conditional exogeneity
- **Conditional normality**: $(W_i|S_i) \sim N(\alpha + \beta S_i, \sigma^2)$
- Parameter space: $\alpha, \beta, \sigma^2 \in \Re^2 x \Re_+$
- These assumptions yield the model
-

$$\begin{aligned} W_i &= \alpha + \beta S_i + u_i, \\ & (u_i|S_i) \sim N(0, \sigma^2) \end{aligned}$$

- In expectations,
-

$$E(W_i|S_i) = \alpha + \beta S_i$$

The Likelihood Function

- How do we find the values of $\hat{\alpha}$ and $\hat{\beta}$ that are most likely to yield the set of observations $(W_1, S_1), ..., (W_n, S_n)$?
- We need to define a function that measures this likelihood, and maximise it with respect to $\hat{\alpha}$ and $\hat{\beta}$
- Due to the independence assumption, we can write

$$f_{\alpha,\beta,\sigma^2}(w_1, ..., w_n | s_1, ..., s_n) = \prod_{i=1}^{n} f_{\alpha,\beta,\sigma^2}(w_i | s_i)$$

- So the probability of jointly observing all these observations is just the product of observing each individual observation
- Assuming conditional normality, we can substitute for the density function f,
-

$$\prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\{-\frac{1}{2\sigma^2}(w_i - \alpha - \beta s_i)^2\}$$

The Likelihood Function cont.

- We can simplify the product,
- $$(2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(w_i-\alpha-\beta s_i)^2\}$$
- The conditional density given the realisations of the random variables:
- $$\begin{aligned} & L_{W_1,\ldots,Wn|S_1,\ldots,S_n}(\alpha,\beta,\sigma^2) \\ = \; & f_{\alpha,\beta,\sigma^2}(W_1,\ldots,W_n|S_1,\ldots,S_n) \\ = \; & (2\pi\sigma^2)^{-\frac{n}{2}}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(W_i-\alpha-\beta S_i)^2\} \end{aligned}$$
- Finally, the log-likelihood function is
- $$\ell_{W_1,\ldots,Wn|S_1,\ldots,S_n}(\alpha,\beta,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(W_i-\alpha-\beta S_i)^2$$

How do we maximise the likelihood function?

- Notice that maximising ℓ is equivalent to minimising $\sum_{i=1}^{n}(W_i - \alpha - \beta S_i)^2$.
- Look familiar?
- This is identical to the procedure for OLS. So the ML estimators will be the same as OLS in this special case:
-

$$\begin{aligned} \hat{\alpha}_{MLE} &= \hat{\alpha}_{OLS} \\ \hat{\beta}_{MLE} &= \hat{\beta}_{OLS} \end{aligned}$$

Why are the ML estimators identical to the OLS estimators in our example?

- What have we assumed?
- We assumed a normal distribution for the errors
- However, this is not necessary
- One could assume any distribution for the u_i
- In practise, other distributions will give you forms for ℓ that cannot be maximised analytically, so we use numerical methods
- The most common other example of a distribution for u_i is the t-distribution
- A likelihood function with t-distributed errors can only be maximised numerically

What about the variance?

- The estimator for the variance can be found as a second step, once we have $\hat{\alpha}$ and $\hat{\beta}$
- Step 1: Plug $\hat{\alpha}$ and $\hat{\beta}$ into ℓ
- Step 2: Maximise ℓ with respect to $\hat{\sigma}^2$

Properties of the Maximum Likelihood Estimator

- Consistency: $\hat{\beta}$ converges in probability to β (same for $\hat{\alpha}$):
- $$\text{plim}_{N \to \infty} \hat{\beta} = \beta$$
- Efficiency: $\hat{\beta}$ is the most precise estimator with the smallest possible standard error
- These are both *asymptotic* properties, i.e. they hold as $N \longrightarrow \infty$.
- In finite samples, these don't hold. This is why OLS is preferable to MLE in small samples.
- MLE is not always unbiased
- With normal errors, the parameter estimates are unbiased but the variance of the errors is biased downwards

Inference: The Likelihood Ratio Test

- There are three ways of doing inference with MLE: the Likelihood Ratio Test, the Wald Test and the Score Test
- The Wald test uses only the unrestricted estimates, the Score test the restricted estimates and the LR test both
- Assume we are testing $H_0 : \beta = 0$. Let this be the restricted model with β_0. Then the LRT is
- $$2(\ell(\hat{\beta}_{MLE}) - \ell(\beta_0)) \sim \chi^2(1)$$
- Intuitively: Is the likelihood of observing the data with $\hat{\beta}_{MLE}$ significantly higher than the likelihood with β_0?
- Removing prediction variables almost always lowers the likelihood in practise, but is the fall significant?
- In practise: estimate the model with the constraint $\beta = \beta_0$. This will give you a value for the likelihood: $\ell(\beta_0)$. Use in above formula.

Inference: The Wald Test

- The Wald test looks at whether estimated parameters are significantly different from zero
- The formula for the Wald test is
- $$\frac{(\hat{\beta}_{MLE} - \beta_0)^2}{var(\hat{\beta}_{MLE})} \sim \chi^2(1)$$
- Wald is t^2 which is χ^2 distributed since we know MLE is asymptotically normal
- We should prefer the LRT because the WT uses an estimated value for the variance, which is not entirely accurate.
- Asymptotically they are equivalent.

Inference: The Score Test

- Calculates the slope of the log-likelihood at the constrained value of the parameter
- How quickly is the slope changing at the null hypothesis?
- Is it worth adding more variables?
- A bit like a test for omitted variables
- `http://www.ats.ucla.edu/stat/mult_pkg/faq/general/nested_tests.gif`

Problems with MLE: Endogeneity

- We assumed conditional exogeneity of S_i.
- However, there may be omitted variables such as ability.
- Do more able people spend more time in school and also get higher wages?
- If yes, this invalidates our estimates.

Summary

- MLE is a method of estimating the parameters of a statistical model
- It uses a likelihood approach
- When we assume errors are normally distributed, the resulting parameter estimates are the same as in OLS
- In small samples, better to use OLS
- MLE is asymptotically efficient and consistent
- Inference best done with a likelihood ratio test