ECON 61001: *Lecture 9*

Alastair R. Hall

The University of Manchester

Outline of today's lecture

- Maximum Likelihood Estimation

- Binary response models
  - Linear probability model
  - Probit and Logit models
  - Empirical application

Notation

For the purposes of introducing ML, we follow the convention that capital letters are random variables/vectors and small letters denote the observed outcomes.

However, having introduced ML, we revert to the convention - more common in econometrics - that small letters denote either random variables (vectors) or their outcomes with the interpretation defined by context.

Intuition behind ML

Suppose $\{V_i;\ i = 1, 2, \ldots N\}$ is a sequence of discrete random variables with joint probability distribution function:

$$P(V_1 = v_1, V_2 = v_2, \ldots, V_N = v_N;\ \theta_0) = p(v_1, v_2, \ldots, v_N;\ \theta_0), \text{ say,}$$

where θ_0 is a vector of parameters.

If we know θ_0 then the probability of observing a particular sample $\{V_i = v_i;\ i = 1, 2, \ldots, N\}$ is given by

$$p(v_1, v_2, \ldots, v_N;\ \theta_0)$$

Intuition behind ML

The situation we face in estimation is the other way round: *given the observed sample* we wish to work out θ_0.

Basic idea of ML: estimate θ_0 by the parameter value that maximizes the probability of observing the particular sample we have.

To express this mathematically, we now write:

$$LF(\theta;\, v_1, v_2, \ldots v_N) \stackrel{d}{=} p(v_1, v_2, \ldots v_N; \theta)$$

and ML estimator (MLE) of θ is:

$$\hat{\theta} \;=\; arg \max_{\theta \in \Theta} LF(\theta;\, v_1, v_2, \ldots v_N)$$

$LF(\theta;\, v_1, v_2, \ldots v_N)$ is known as the likelihood function.

Intuition behind ML

How can we extend this to continuous random variables?

If $\{V_i;\ i = 1, 2, \ldots N\}$ is a sequence of continuous random variables with joint probability density function (pdf) $f(v_1, v_2, \ldots, v_N; \theta)$ then the likelihood function is defined as:

$$LF(\theta; v_1, v_2, \ldots v_N) \stackrel{d}{=} f(v_1, v_2, \ldots, v_N;\ \theta)$$

We will now focus our attention on the discrete case as this is the scenario relevant to binary response models.

Intuition behind ML

Clearly to implement ML, we require the joint probability distribution function.

Sometimes we specify the model explicitly in terms of the joint distribution and so $p(v_1, v_2, \ldots, v_N;\, \theta)$ is readily available.

In others, we specify model for V_i from which we can deduce joint probability distribution function. For example:

- $\{V_i\}$ is independently and identically distributed, with $P(V_i = v_i;\, \theta_0) = p(v_i; \theta_0)$.

So that

$$p(v_1, v_2, \ldots, v_N;\, \theta) \;=\; \prod_{i=1}^{N} p(v_i; \theta_0)$$

Likelihood function for iid data

This means that:

$$LF(\theta; v_1, v_2, \dots v_N) = \prod_{i=1}^{N} p(v_i; \theta).$$

In this case, easier to equivalently define MLE via:

$$\hat{\theta}_N = arg \max_{\theta \in \Theta} LLF_N(\theta)$$

where

$$LLF_N(\theta) = ln[LF(\theta; v_1, v_2, \dots v_N)] \sim \text{log likelihood function}$$

MLE and score equations

Under relatively mild conditions on the distribution, the MLE can be obtained by solving the first order conditions:

$$\left.\frac{\partial LLF_N(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_N} = 0$$

These equations are known as the score equations.

Sometimes, we can obtain closed form solution for $\hat{\theta}_N$ (in terms of data alone) from score equations. Other times, MLE is found by using computer-based numerical optimization routines.

Example: Bernouilli distribution

- Let $\{V_i\}_{i=1}^N$ be a sequence of i.i.d. Bernoulli random variables with $P(V_i = 1; \theta_0) = \theta_0$.

- The probability distribution function of V_i is:

$$P(V_i = v_i; \theta_0) \;=\; \theta_0^{v_i}(1-\theta_0)^{1-v_i}.$$

- Likelihood function is:

$$LF_N(\theta) \;=\; \prod_{i=1}^{N} \theta^{v_i}(1-\theta)^{1-v_i}.$$

- Log likelihood function is

$$LLF_N(\theta) \;=\; \sum_{i=1}^{N} \{\, v_i ln[\theta] \;+\; (1-v_i) ln[1-\theta] \,\}\,.$$

Example: Bernouilli distribution

- And so:

$$\frac{\partial LLF_N(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{N} v_i}{\theta} - \frac{\sum_{i=1}^{N}(1 - v_i)}{1-\theta}.$$

- Defining $\sum_{i=1}^{N} v_i = N_1$, score equation is:

$$\frac{N_1}{\hat{\theta}_N} - \frac{N - N_1}{1 - \hat{\theta}_N} = 0.$$

- MLE is: $\hat{\theta}_N = N_1/N$.

Overview of statistical properties of MLE

MLE has appealing intuition but what can be said about its properties?

- Consistency: $\hat{\theta}_N \xrightarrow{p} \theta_0$.

- Asymptotic normality: $N^{1/2}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, V_\theta)$ where

$$V_\theta = \{\lim_{N\to\infty} N^{-1}\mathcal{I}_{\theta,N}\}^{-1}$$

and

$$\mathcal{I}_{\theta,N} = -E\left[\frac{\partial^2 LLF_N(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta_0}\right].$$

$\mathcal{I}_{\theta,N}$ is known as the information matrix.

Overview of properties of MLE - continued

- Invariance: The MLE of $h(\theta_0)$ is $h(\hat{\theta}_N)$.

These properties imply that the MLE is optimal in the sense that it is asymptotically efficient (i.e. minimum variance) in the class of consistent uniformly asymptotically normal (CUAN) estimators of θ_0.

But note:

- justification is via large sample properties
- optimality depends crucially on our correctly specifying the joint distribution.

Hypothesis testing

Suppose we wish to test:

$$H_0 : g(\theta_0) = 0 \quad vs. \quad H_1 : g(\theta_0) \neq 0$$

where

- $g(\cdot)$ is a $n_g \times 1$ vector of continuous differentiable functions
- $G(\bar{\theta}) = \frac{\partial g(\theta)}{\partial \theta'}\Big|_{\theta=\bar{\theta}}$ with $rank\{G(\theta_0)\} = n_g$.

Three fundamental test principles associated with ML estimation: Wald, Likelihood Ratio and Lagrange Multiplier (or Score).

Hypothesis testing

To present the forms of the three test statistics we need the following notation/terminology:

- Unrestricted MLE is:

$$\hat{\theta}_N = \arg\max_{\theta \in \Theta} LLF_N(\theta)$$

- Restricted MLE is:

$$\tilde{\theta}_N = \arg\max_{\theta \in \Theta_R} LLF_N(\theta)$$

where $\Theta_R = \{\theta : g(\theta) = 0, \theta \in \Theta\}$.

Hypothesis testing

Wald test:

$$W_N = N\, g(\hat{\theta}_N)'[G(\hat{\theta}_N)'\hat{V}_\theta G(\hat{\theta}_N)]^{-1} g(\hat{\theta}_N)$$

where $\hat{V}_\theta \xrightarrow{p} V_\theta$.

Likelihood ratio (LR) test:

$$LR_N = 2\{LLF_N(\hat{\theta}_N) - LLF_N(\tilde{\theta}_N)\}$$

Hypothesis testing

Lagrange Multiplier (LM) test:

$$LM_N \;=\; N\bar{s}_N(\tilde{\theta}_N)'\tilde{V}_\theta\bar{s}_N(\tilde{\theta}_N)$$

- $\bar{s}_N(\bar{\theta}) = N^{-1}\frac{\partial LLF_N(\theta)}{\partial\theta}\Big|_{\theta=\bar{\theta}}$.
- $\tilde{V}_\theta \xrightarrow{p} V_\theta$ (under H_0).

Under H_0 then: W_N, LR_N, $LM_N \xrightarrow{d} \chi^2_{n_g}$ and further all three are asymptotically equivalent ($W_N - LR_N \xrightarrow{p} 0$, $W_N - LM_N \xrightarrow{p} 0$).

Conditional models

Often interested in conditional models that is, explaining y_i in terms of x_i.

But our original definition of LF is in terms of joint distribution of $v_i = (y_i, x_i')'$.

However can also work with conditional distribution under certain circumstances.

For example, if v_i is i.i.d and $p(v_i;\theta) = p(y_i|x_i;\phi)p(x_i;\psi)$ so parameters of distribution of $y_i|x_i$ do not appear in marginal distribution of x_i then can obtain ML by maximizing

$$CLLF(\phi) = \sum_{i=1}^{N} ln[p(y_i \,|\, x_i; \phi)]$$

What are binary response models?

In binary response models, we are interested in modeling the probability that an event occurs.

Examples of events that may be of interest:

- whether an individual is employed;
- whether an individual is divorced;
- whether an individual receives a loan;
- whether a firm is taken over.

As economists, we are interested in modeling the conditional probability of the event given characteristics of the individual or firm concerned.

Basic structure of binary response models

Use a dummy variable to model outcome that is

$$\begin{aligned} y_i &= 1, \text{ if event occurs} \\ &= 0, \text{ if event does not occur} \end{aligned}$$

Let x_i be a vector of explanatory variables then we are interested in:

- the conditional probability that an event occurs, $P(y_i = 1|x_i)$.
- how probability changes with changes in the elements of x_i.

Basic structure of binary response models

Example: event of interest is employment

- $y_i = 1$ if individual is employed, $y_i = 0$ if individual is unemployed;

- x_i might contain information about education, experience, demographic information (using dummy variables), economic conditions if sample over disparate area.

Basic structure of binary response models

A common approach to binary response modeling is to use so-called index models for which

$$P(y_i = 1 \,|\, x_i) \;=\; p(x_i'\beta_0)$$

where β_0 is a vector of unknown parameters (as in the multiple linear regression model), $p(.)$ is some function and $x_i'\beta_0$ is known as the "index".

Examples of index models:

- linear probability model (LPM);
- logit model;
- probit model.

These three differ in their choice of $p(.)$ above (and hence in implicit definition of β_0).

Basic structure of LPM

For the LPM, we assume that $p(.)$ is linear so that

$$p(x_i'\beta_0) = x_i'\beta_0$$

How can we estimate β_0? Answer: it turns out we can estimate this model using regression techniques. To uncover why we need to think about the $E[y_i \mid x_i]$ for this model.

Since y_i is a dummy variable:

$$\begin{aligned} E[y_i \mid x_i] &= 0 \times P(y_i = 0|x_i) + 1 \times P(y_i = 1|x_i) \\ &= P(y_i = 1|x_i) \\ &= x_i'\beta_0 \end{aligned}$$

Basic structure of LPM

Therefore, we can write:

$$y_i = x_i'\beta_0 + u_i$$

and u_i satisfies $E[u_i|x_i] = 0$ which is a key condition for OLS to be a consistent estimator of β_0.

This suggests that we can estimate the LPM by OLS.

However, we must use heteroscedasticity robust standard errors; see Tutorial 9 Question 1.

Problem: the predicted probabilities can be outside $[0, 1]$.

Probit model

Define

- $\Phi(z)$ is the cumulative distribution function of a standard normal rv that is, $\Phi(z) = \int_{-\infty}^{z} \phi(v)dv$ where $\phi(v)$ is the pdf of a standard normal distribution.

Then for Probit model set

$$P(y_i = 1 \mid x_i) = \Phi(x_i'\beta_0)$$

- $\Phi(x_i'\beta_0)$ is a nonlinear function of $x_i'\beta_0$;
- If $P(y_i = 1|x_i) = \Phi(x_i'\beta_0)$ then $P(y_i = 0|x_i) = 1 - \Phi(x_i'\beta_0)$.

Latent variable model interpretation

These models look very different from our linear regression model, but we can uncover a connection by taking a latent variable approach to specifying these models.

What is a latent variable? Answer: it is variable that is observed by individual/firm but not observed by the econometrician. In this case, it is some variable that is the sole determinant of whether the event occurs.

For example: if event is whether individual is given loan by bank then latent variable is credit score given by bank.

Now consider math.

Latent variable model interpretation

Suppose that the latent variable y^* satisfies the linear regression model

$$y_i^* = x_i'\beta_0 + u_i$$

and

$$\begin{aligned} y_i^* &> 0 \Rightarrow y_i = 1 \\ y_i^* &\leq 0 \Rightarrow y_i = 0 \end{aligned}$$

Then

- if u_i has standard normal distribution then y_i follows probit model.

How does change in $x_{i,\ell}$ affect probability?

This is more complicated here than in the LPM because probit model implies the probability is a nonlinear function of the x's.

If $x_{i,\ell}$ is continuous then

$$\frac{\partial P(y_i = 1|x_i)}{\partial x_{i,\ell}} = \phi(x_i'\beta_0)\beta_{0,\ell}.$$

If $x_{i,\ell}$ is a dummy variable then holding values of other elements of x_i constant then

$$\Delta P(y_i = 1|x_i) \;=\; \Phi(x_i'\beta_0, x_{i,\ell} = 1) \;-\; \Phi(x_i'\beta_0, x_{i,\ell} = 0)$$

Note that in both cases response $\Delta P(y_i = 1|x_i)$ depends on x_i.

ML Estimation of probit model

Assume

- $\{(y_i, x_i')\}_{i=1}^N$ are i.i.d
- $P(y_i = 1|x_i) = \Phi(x_i'\beta_0)$
- $p(x_i)$ does not depend on β_0

Then:

$$CLLF_N(\beta) \ = \sum_{i=1}^{N} \{ \, y_i ln[\Phi(x_i'\beta)] + (1 - y_i) ln[1 - \Phi(x_i'\beta)] \, \}$$

Empirical example: Wooldridge Example 7.12 on p.256

y is a dummy variable that takes the value one if a man is arrested during 1986 and is zero, otherwise.

x contains:

- *pcnv* $=$ the proportion of prior arrests that led to conviction;
- *avgsen* $=$ the average sentence served from prior convictions;
- *tottime* $=$ the total time spent in prison since age 18 prior to 1986;
- *ptime*86 $=$ months spent in prison in 1986;
- *qemp*86 $=$ the number of quarters that the man was legally employed in 1986;

Linear probability model:


```
------------------------------------------------------
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.440615   0.018546  23.758  < 2.2e-16 ***
pcnv        -0.162445   0.019220  -8.452  < 2.9e-16 ***
avgsen       0.006113   0.006210   0.984   0.326
tottime     -0.002262   0.004578  -0.494   0.621
ptime86     -0.021966   0.002919  -7.526  7.06e-14 ***
qemp86      -0.042829   0.005466  -7.835  6.66e-15 ***
------------------------------------------------------
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.4373 on 2719 degrees of freedom
Multiple R-squared:  0.04735, Adjusted R-squared:  0.0456
F-statistic: 27.03 on 5 and 2719 DF,  p-value: < 2.2e-16
```

Probit model:

```
---------------------------------------------------------
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -0.101999   0.051462   -1.982    0.0475 *
pcnv        -0.540475   0.069303   -7.799  6.25e-15 ***
avgsen       0.018923   0.020459    0.925    0.3550
tottime     -0.006569   0.016175   -0.406    0.6847
ptime86     -0.078288   0.017055   -4.587  4.49e-06 ***
qemp86      -0.131653   0.016600   -7.931  2.17e-15 ***
---------------------------------------------------------
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 3216.4  on 2724  degrees of freedom
Residual deviance: 3079.2  on 2719  degrees of freedom
AIC: 3091.2

Number of Fisher Scoring iterations: 5
```

Probit model:

```
------------------------------------------------------
Marginal Effects:
             dF/dx     Std. Err.         z      P>|z|
pcnv    -0.1775607   0.0226318   -7.8456  4.308e-15 ***
avgsen   0.0062166   0.0067212    0.9249     0.3550
tottime -0.0021580   0.0053141   -0.4061     0.6847
ptime86 -0.0257034   0.0055866   -4.6009  4.207e-06 ***
qemp86  -0.0432534   0.0054381   -7.9537  1.810e-15 ***
------------------------------------------------------
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Further reading- to come

- Notes: Chapter 6

- Greene:
  - ML: Ch 14.1-14.6 (but goes into more detail on properties of MLE than we do)
  - Binary response model: Ch 17.1-17.3 (but again more detail than this course)