Toulouse School of Economics

M1 INTERMEDIATE ECONOMETRICS

Maximum-likelihood estimation

Koen Jochmans François Poinas

2025 — 2026

Binary-outcome example

$Y \in \{0, 1\}$ is Bernoulli

$$\theta = \mathbb{P}(Y = 1) \in (0, 1).$$

So its PMF is

$$f(y; \theta) = \theta^y \times (1 - \theta)^{1-y}, \qquad y \in \{0, 1\}.$$

Consider independent and identically distributed draws $Y_1, \ldots, Y_n$.

These can be any sequence from $\{0, 1\}^n$. Any given sequence $y_1, \ldots, y_n$ occurs with probability

$$\prod_{i=1}^{n} f(y_i; \theta) = \theta^{\sum_{i=1}^n y_i} \times (1 - \theta)^{n - \sum_{i=1}^n y_i} = \theta^{n\bar{y}_n} \times (1 - \theta)^{n(1-\bar{y}_n)}.$$

Given the data, $Y_1, \ldots, Y_n$, we can think about this probability as a function of θ:

$$L_n(\theta) = \prod_{i=1}^{n} f(Y_i; \theta) = \theta^{n\bar{Y}_n} \times (1-\theta)^{n(1-\bar{Y}_n)}.$$

This is the likelihood function.

It traces out how 'likely' it is to draw our sample if sampling is done with success probability θ.

Let θ_0 be the success probability from which we sampled our data (this is unknown to us).

The maximum-likelihood estimator of θ_0 is

$$\hat{\theta} = \arg\max L_n(\theta).$$

It is often easier to maximize the log-likelihood function

$$\ell_n(\theta) = \log L_n(\theta) = \log\left(\prod_{i=1}^{n} f(Y_i;\theta)\right) = \sum_{i=1}^{n} \log f(Y_i;\theta),$$

which is just a monotonic transformation.

In our example

$$\ell_n(\theta) = \log\left(\theta^{n\bar{Y}_n} \times (1-\theta)^{n(1-\bar{Y}_n)}\right) = n\bar{Y}_n \log(\theta) + n(1-\bar{Y}_n)\log(1-\theta),$$

with first-order condition

$$n\frac{\bar{Y}_n - \theta}{\theta\,(1-\theta)} = 0$$

and global maximiser $\hat{\theta} = \bar{Y}_n$, the maximum-likelihood estimator.

Normal distribution

Suppose now that $Y \sim N(\mu, \sigma^2)$. The density function has parameter $\theta = (\mu, \sigma^2)$ and equals

$$f(y;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right).$$

The likelihood function is

$$L_n(\theta) = \prod_{i=1}^{n} f(Y_i;\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{\sigma^2}\right).$$

Note that here we use the density function, as the variable in question is continuous.

The log-likelihood is

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{\sigma^2}.$$

Maximising this with respect to μ amounts to minimising the sum of squares

$$\sum_{i=1}^{n}(Y_i-\mu)^2,$$

which gives $\hat{\mu} = \bar{Y}_n$. Given $s^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\mu})^2$, we have

$$\ell_n(\hat{\mu}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\frac{s^2}{\sigma^2},$$

with first-order condition

$$-\frac{n}{2}\frac{1}{\sigma^2}\left(1-\frac{s^2}{\sigma^2}\right) = 0,$$

and solution $\hat{\sigma}^2 = s^2$.

Justification for maximizing the likelihood

Under regularity conditions,

$$n^{-1}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i;\theta) \underset{p}{\to} \mathbb{E}(\log f(Y;\theta)) = \ell(\theta)$$

in a uniform sense. This function achieves its global maximum at θ_0.

$$\begin{aligned}
\ell(\theta) - \ell(\theta_0) &= \mathbb{E}\left(\log\left(\frac{f(Y;\theta\ )}{f(Y;\theta_0)}\right)\right) \\
&\leq \log \mathbb{E}\left(\frac{f(Y;\theta\ )}{f(Y;\theta_0)}\right) \\
&= \log \int \frac{f(y;\theta\ )}{f(y;\theta_0)}\, f(y;\theta_0)\, dy \\
&= \log \int f(y;\theta)\, dy \\
&= 0,
\end{aligned}$$

because $\int f(y;\theta)\, dy = 1$ for any θ.

Estimating equations

In regular problems the maximum-likelihood estimator is a stationary point of the score equation

$$n^{-1}\frac{\partial \ell_n(\hat{\theta})}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(Y_i;\hat{\theta})}{\partial \theta} = 0.$$

Again, this can be seen as a sample counterpart to the population problem

$$\mathbb{E}\left(\frac{\partial \log f(Y;\theta_0)}{\partial \theta}\right) = 0.$$

The score

Let

$$s(y;\theta) = \frac{\partial \log f(y;\theta)}{\partial \theta}$$

be the score.

Note that

$$\mathbb{E}_\theta(s(Y;\theta)) = \int \frac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta)\, dy = \int \frac{\partial f(y;\theta)}{\partial \theta}\, dy = 0$$

as long as integration and differentiation can be interchanged.

So the score gives a valid estimating equation.

It is also the 'optimal' estimating equation, in a well-defined sense.

Cramer-Rao bound

Consider a generic setting where an estimator $\hat{\theta}$ is unbiased for θ.

Unbiasedness means that

$$\mathbb{E}_\theta(\hat{\theta} - \theta) = \int \cdots \int (\hat{\theta} - \theta) \left(\prod_{i=1}^{n} f(y_i; \theta) \right) dy_1 \cdots dy_n = 0.$$

This holds for any θ so we can take derivatives of both sides of this expression to get

$$\int .. \int (\hat{\theta} - \theta) \frac{\partial \left(\prod_{i=1}^{n} f(y_i; \theta) \right)}{\partial \theta} dy_1 \cdots dy_n = \int .. \int \left(\prod_{i=1}^{n} f(y_i; \theta) \right) dy_1 \cdots dy_n.$$

Clearly, the right-hand side is equal to 1.

The left-hand side is equal to

$$\mathrm{cov}_\theta \left(\hat{\theta}, \sum_{i=1}^{n} s(Y_i; \theta) \right).$$

To see this, note that

$$\begin{aligned}\sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial\theta} &= \sum_{i=1}^{n} \frac{1}{f(y_i;\theta)} \frac{\partial f(y_i;\theta)}{\partial\theta} \\ &= \sum_{i=1}^{n} \left(\frac{\prod_{j\neq i} f(y_j;\theta)}{\prod_j f(y_j;\theta)} \right) \frac{\partial f(y_i;\theta)}{\partial\theta} \\ &= \frac{\sum_{i=1}^{n} \prod_{j\neq i} f(y_j;\theta) \frac{\partial f(y_i;\theta)}{\partial\theta}}{\prod_j f(y_j;\theta)} = \frac{\frac{\partial \prod_i f(y_i;\theta)}{\partial\theta}}{\prod_j f(y_j;\theta)}\end{aligned}$$

using the chain rule on the differentiation of a product. Hence,

$$\frac{\partial \prod_{i=1}^{n} f(y_i;\theta)}{\partial\theta} = \left(\sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial\theta} \right) \left(\prod_{j=1}^{n} f(y_j;\theta) \right),$$

and so

$$\int .. \int (\hat{\theta} - \theta) \frac{\partial \left(\prod_{i=1}^{n} f(y_i;\theta) \right)}{\partial\theta} dy_1 \cdots dy_n = \mathbb{E}_\theta \left((\hat{\theta} - \theta) \sum_{i=1}^{n} s(Y_i;\theta) \right).$$

So any unbiased estimator satisfies

$$\mathrm{cov}_\theta\left(\hat{\theta}, \textstyle\sum_{i=1}^n s(Y_i;\theta)\right) = 1.$$

By the Cauchy-Schwarz inequality

$$\mathrm{cov}_\theta\left(\hat{\theta}, \sum_{i=1}^n s(Y_i;\theta)\right)^2 \leq \mathrm{var}_\theta(\hat{\theta}) \times \mathrm{var}_\theta\left(\sum_{i=1}^n s(Y_i;\theta)\right).$$

Now,

$$\mathrm{var}_\theta\left(\sum_{i=1}^n s(Y_i;\theta)\right) = \sum_{i=1}^n \mathrm{var}_\theta\left(s(Y_i;\theta)\right) = n\,\mathrm{var}_\theta\left(s(Y;\theta)\right) =: n\,I_\theta$$

so then

$$\frac{I_\theta^{-1}}{n} \leq \mathrm{var}_\theta(\hat{\theta}).$$

The lower bound is the Cramer-Rao bound and I_θ is the information (matrix).

Cauchy-Schwarz holds with equality iff

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^{n} I_\theta^{-1} s(Y_i; \theta).$$

In the Bernoulli problem

$$s(y; \theta) = \frac{y - \theta}{\theta(1 - \theta)}$$

and

$$I_\theta = \mathbb{E}_\theta \left(\frac{(Y - \theta)^2}{\theta^2 (1 - \theta)^2} \right) = \frac{\mathrm{var}_\theta(Y)}{\theta^2 (1 - \theta)^2} = \frac{1}{\theta(1 - \theta)}$$

Hence, the efficient estimator is

$$\hat{\theta} = \theta + \frac{1}{n} \sum_{i=1}^{n} (\theta(1 - \theta)) \, \frac{Y_i - \theta}{\theta(1 - \theta)} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

which was the maximum-likelihood estimator.

Optimality

In general, the best unbiased estimator does not exist.

So the Cramer-Rao bound cannot be attained except in some simple problems.

The maximum-likelihood estimator quite generally satisfies

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^{n} I_\theta^{-1} s(Y_i; \theta) + o_p(n^{-1/2})$$

and so achieves the bound in large samples.

Information matrix equality

Important component that leads to optimality of maximum likelihood is the information equality,

$$I_\theta = \mathrm{var}_\theta\left(\frac{\partial \log f(Y;\theta)}{\partial \theta}\right) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(Y;\theta)}{\partial \theta \partial \theta'}\right).$$

Can show this by differentiation the unbiasedness condition of the score:

$$\mathbb{E}_\theta\left(\frac{\partial \log f(Y;\theta)}{\partial \theta}\right) = \int \frac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta)\, dy = 0.$$

This gives

$$\int \frac{\partial^2 \log f(y;\theta)}{\partial \theta \partial \theta'} f(y;\theta)\, dy + \int \frac{\partial \log f(y;\theta)}{\partial \theta} \frac{\partial f(y;\theta)}{\partial \theta'} dy = 0$$

But note that

$$\int \frac{\partial^2 \log f(y;\theta)}{\partial\theta\partial\theta'} f(y;\theta)\, dy = \mathbb{E}_\theta \left(\frac{\partial^2 \log f(y;\theta)}{\partial\theta\partial\theta'} \right)$$

and that

$$\int \frac{\partial \log f(y;\theta)}{\partial\theta} \frac{\partial f(y;\theta)}{\partial\theta'} dy = \int \frac{\partial \log f(y;\theta)}{\partial\theta} \frac{\partial \log f(y;\theta)}{\partial\theta'} f(y;\theta)\, dy$$

which is

$$\mathbb{E}_\theta \left(\frac{\partial \log f(y;\theta)}{\partial\theta} \frac{\partial \log f(y;\theta)}{\partial\theta'} \right)$$

Limit distribution

In regular problems,

$$n^{-1}\frac{\partial \ell_n(\hat{\theta})}{\partial \theta} = 0.$$

We can expand around θ_0 using the mean-value theorem to write

$$n^{-1}\frac{\partial \ell_n(\hat{\theta})}{\partial \theta} = n^{-1}\frac{\partial \ell_n(\theta_0)}{\partial \theta} + n^{-1}\frac{\partial^2 \ell_n(\theta_*)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0).$$

Combining yields

$$(\hat{\theta} - \theta_0) = \left(-n^{-1}\frac{\partial^2 \ell_n(\theta_*)}{\partial \theta \partial \theta'}\right)^{-1} n^{-1}\frac{\partial \ell_n(\theta_0)}{\partial \theta}$$

Here,

$$-n^{-1}\frac{\partial^2 \ell_n(\theta_*)}{\partial\theta\partial\theta'} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(Y_i;\theta_*)}{\partial\theta\partial\theta'} \underset{p}{\to} -\mathbb{E}\left(\frac{\partial^2 \log f(Y;\theta_0)}{\partial\theta\partial\theta'}\right) = I_\theta$$

while

$$n^{-1/2}\frac{\partial \ell_n(\theta_0)}{\partial\theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(Y_i;\theta_0)}{\partial\theta} \underset{d}{\to} N(0, I_\theta).$$

By Slutsky's theorem,

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, I_\theta^{-1}).$$

Conditional models

Often the model includes conditioning variables.

Can work with the conditional distribution

$$f(y|x;\theta).$$

In this case everything continues to go through as before, with

$$I_\theta = -\mathbb{E}\left(\frac{\partial^2 \log f(Y|X;\theta_0)}{\partial\theta\partial\theta'}\right).$$

Binary outcomes

An example for Bernoulli outcomes is

$$\mathbb{E}(Y|X = x) = F(x'\theta).$$

Here,

$$f(y|x;\theta) = F(x'\theta)^y \times (1 - F(x'\theta))^{1-y}.$$

The log-likelihood function is

$$\sum_{i=1}^{n} Y_i \log(F(X_i'\theta)) + (1 - Y_i) \log(1 - F(X_i'\theta)).$$

This function is nonlinear in θ, with first-order condition

$$\sum_{i=1}^{n} X_i \frac{f(X_i'\theta)}{F(X_i'\theta)(1 - F(X_i'\theta))}(Y_i - F(X_i'\theta)) = 0.$$

Tobit model

Classical linear regression model

$$Y^* = X'\beta + e, \qquad e|X \sim N(0, \sigma^2)$$

but we observe

$$Y = \max(Y^*, 0).$$

The cumulative distribution is

$$\mathbb{P}(Y \leq y|X) = \mathbb{P}(Y = 0|X) + \mathbb{P}(0 < Y \leq y|X)$$

which has

$$\mathbb{P}(Y = 0|X) = \mathbb{P}(Y^* < 0|X) = \Phi(-X'\beta/\sigma) = 1 - \Phi(X'\beta/\sigma)$$

and for $y > 0$

$$\mathbb{P}(Y \leq y|X) = \Phi\left(\frac{y - X'\beta}{\sigma}\right),$$

with density $\frac{1}{\sigma}\phi\left(\frac{y-X'\beta}{\sigma}\right)$.

The Tobit likelihood is

$$\prod_{i=1}^{n}\left(1-\Phi\left(\frac{X_i'\beta}{\sigma}\right)\right)^{\{Y_i=0\}}\left(\frac{1}{\sigma}\phi\left(\frac{Y_i-X_i'\beta}{\sigma}\right)\right)^{\{Y_i>0\}}.$$

It has both a probit component and a normal-regression component.

This is again a nonlinear problem.