Slides for the course

Statistics and econometrics

Part 2: Estimation

European University Institute

Andrea Ichino

September 18, 2014

Outline

Section 1

The problem of estimation

The problem of estimation

In the pre-course you have seen how to model data and phenomena with statistical distributions that depend on unknown parameters:

Example: Bernoulli distribution.

$$X = \begin{cases} 1 & \text{with probability } q \\ 0 & \text{with probability } 1-q \end{cases} \tag{1}$$

$$X \sim p_X(X = x|q) = q^x(1-q)^{1-x} = \begin{cases} q & \text{for } x = 1 \\ 1-q & \text{for } x = 0 \end{cases} \tag{2}$$

How can we use a random sample of data to get information on the parameter q in a given population to study the random variable X?

(Note: pay attention to the meanings of X, x, 1, 0, q and $p_X(.)$ ...)

Some notational conventions

- $f_X(X = x|\theta) = f_X(x|\theta)$ is the probability density function (pdf) of the random variable X evaluated at the realization x, given the parameter θ.
- $F_X(X = x|\theta) = F_X(x|\theta)$ is the correspondent cumulative distribution.
- $p_X(.)$ and $P_X(.)$ are used in case it is necessary to make explicit reference to a discrete pdf.

Estimators and estimates

An *estimator* is any function of a random sample that gives information on the parameters of the distribution of a random variable.

- Analogy: the recipe for a cake ...

A random sample is a set of observed realizations of a random variable, satisfying certain properties (to be discussed below)

- Analogy: the ingredients required by the recipe ...

An *estimate* is the specific value that the estimator takes when it is evaluated using the observed realizations of a specific sample

- Analogy: the cake that you actually get when you mix the ingredients according to the recipe (hopefully a good cake ...)

How do we choose between estimators

Estimators (like recipes for cakes) have properties that:

- translate into a quality of the estimate (how good is the cake)
- for given quality of the data (how good are the ingredients).

The estimator is a random variable (with a distribution) because it is a function of the sample observations which are random variables.

The estimate is a number: a function of the sample realizations.

To choose between estimators we will have to study their properties:

- Unbiasedness
- Efficiency
- Consistency
- Other asymptotic properties

Understanding "randomness" (and independence)

If $\{X_1 ... X_i ... X_n\}$ are independent draws from a population with density function $f(X|\theta)$, then

- $\{X_1 = x_1 ... X_i = x_i ... X_n = x_n\}$ is a random sample from the population defined by $f(X|\theta)$.
- each draw X_i is a random variable and x_i is its sample realization.

Exercise questions:

- Is your class a random sample of PhD students in economics?
- Subjects with A-L names, are a random sample of Italians?
- If I toss a fair coin for each of you, those who get a tail are a random sample of the class?

Knowing that $X_i = x_i$ should not have info on the realization of X_j.

Section 2

The method of Maximum Likelihood

The likelihood function

Let $x_1, ...x_i, ...x_n$ be a random sample of size n from the random variable $X \sim f_X(x|\theta)$, where θ is an unknown (set of) parameters. The *likelihood function* is the product of the pdf evaluated at the n realizations x_i:

$$L(X|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta) \tag{3}$$

Note that the likelihood is:

- a function of the parameter θ, given the realizations x_i ;
- the pdf of observing the sample realizations as a function of θ;

The method of Maximum Likelihood estimates θ as the value of the parameter θ that maximizes the likelihood *given the realized sample*.

Regularity conditions

Under some regularity conditions, the ML estimator is obtained by standard maximisation of the likelihood:

- first derivative equal to zero (first order condition);
- negative second derivative (second order condition).

These conditions are:

- $f(X|\theta)$ must be
  - continuous
  - with continuous first order and second order derivatives
- the set of values X for which $f(X|\theta) \neq 0$ must not depend on θ, i.e. the support of the underlying distribution cannot depend on the parameter to be estimated.

We will see cases in which these conditions are violated and still we can derive the ML estimator of the parameter of interest.

The log-likelihood and the maximization problem

Under the regularity conditions it is simpler to maximize the log likelihood (given monotonicity of log)

$$\max_{\theta} l(X|\theta) = \ln(L(X|\theta)) = \sum_{i=1}^{n} \ln\left(f_X(x_i|\theta)\right) \tag{4}$$

so that the ML estimator of θ is:

$$\hat{\theta} = arg\max_{\theta} \left\{ \ln L(X|\theta) = \sum_{i=1}^{n} \ln(f_X(x_i|\theta)) \right\} \tag{5}$$

which has to satisfy the first and second order conditions

$$\frac{d\ln(L(X|\theta))}{d\theta} = \sum_{i=1}^{n} \frac{d\ln\left(f_X(x_i|\theta)\right)}{d\theta} = \sum_{i=1}^{n} \frac{1}{f_X(x_i|\theta)} \frac{df_X(x_i,\theta)}{d\theta} = 0 \tag{6}$$

$$\frac{d^2\ln(L(X|\theta))}{(d\theta)^2} < 0 \tag{7}$$

Subsection 1

The ML estimator for the normal distribution

Example: the Normal distribution

$$X \sim f_X(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

The maximization problem is:

$$\max_{\mu,\sigma^2} L(X|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{9}$$

$$\max_{\mu,\sigma^2} l(X|\mu,\sigma^2) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{10}$$

The two first order condition are:

$$\frac{d\ln(L(X|\mu,\sigma^2))}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \tag{11}$$

$$\frac{d\ln(L(X|\mu,\sigma^2))}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\left(\frac{1}{\sigma^2}\right)^2 \sum_{i=1}^{n}(x_i-\mu)^2 = 0 \tag{12}$$

ML estimator for the normal distribution

Using (11), the sample mean is the ML estimator of the mean:

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{13}$$

The corresponding estimate is

$$\mu_e = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{14}$$

Using (12), the sample variance is the ML estimator of the variance:

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2_{ML} \tag{15}$$

The corresponding estimate is

$$\sigma^2_e = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{16}$$

Moments of the ML estimator of the mean of a normal

The ML estimator (as any estimator) is a random variable with moments:

$$E(\hat{\mu}) = \mu \qquad \text{and} \qquad Var(\hat{\mu}) = \frac{\sigma^2}{n} \tag{17}$$

$$E(\hat{\mu}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu \tag{18}$$

$$\begin{aligned} Var(\hat{\mu}) &= E(\hat{\mu} - \mu)^2 = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2 = E\left(\frac{1}{n^2}\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2\right) \\ &= \frac{1}{n^2} E\left(\sum_{i=1}^{n}\sum_{i=j}^{n}(X_i - \mu)(X_j - \mu)\right) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n} \end{aligned} \tag{19}$$

where $E(.) = n\sigma^2$ when $i = j$ and $E(.) = 0$ when $i \neq j$ because of X_i and X_j are independent.

Subsection 2

The Score and the Fisher Information

The Score of the likelihood function

The Score of the (log) likelihood $l(X|\theta)$ is the gradient (i.e. the vector of partial derivatives), with respect to the parameters θ

$$S(\theta, X) = \frac{\partial \ln(L(X|\theta))}{\partial \theta} = \frac{1}{L(X|\theta)} \frac{\partial L(X|\theta)}{\partial \theta} \tag{20}$$

The score

- measures the sensitivity of the likelihood to changes of the parameters for given X;
- plays an important role in many applications of ML estimation.

Note that under the regularity conditions stated above, the first order condition to obtain the ML estimator can be written as:

$$S(\hat{\theta}_{ML}, X) = 0 \tag{21}$$

The mean of the score

A useful property of the score is that its mean (integrating over X at the true θ) is zero

$$\begin{aligned} E_X\left(S(\theta, X)\right) &= \int_{-\infty}^{+\infty} \frac{\partial \ln(L(X|\theta))}{\partial \theta} L(X|\theta) dX \\ &= \int_{-\infty}^{+\infty} \frac{1}{L(X|\theta)} \frac{\partial L(X|\theta)}{\partial \theta} L(X|\theta) dX \\ &= \int_{-\infty}^{+\infty} \frac{\partial L(X|\theta)}{\partial \theta} dX \\ &= \frac{\partial \left(\int_{-\infty}^{+\infty} L(X|\theta) dX \right)}{\partial \theta} = \frac{\partial (1)}{\partial \theta} = 0 \end{aligned}$$

Thus the ML estimator of θ is the value $\hat{\theta}_{ML}$ that makes the realization of the score equal to its expected value at the true θ (under regularity conditions).

An example: Binomial likelihood

Consider the likelihood of a binomial sample realisation ($n = 1$):

$$X \sim= q^x(1-q)^{1-x} = L(X|q) \qquad \text{where } x = \{0, 1\}. \tag{22}$$

The log likelihood is

$$l(X|q) = xln(q) + (1-x)ln(1-q) \tag{23}$$

and the score is

$$\frac{\partial l(X|q)}{\partial q} = \frac{x}{q} - \frac{1-x}{1-q} \tag{24}$$

The expected value of the score is

$$\begin{aligned} E\left(\frac{\partial l(X|q)}{\partial q}\right) &= E\left(\frac{x}{q} - \frac{1-x}{1-q}|x=1\right)q + E\left(\frac{x}{q} - \frac{1-x}{1-q}|x=0\right)(1-q) \\ &= \frac{1}{q}q - \frac{1}{1-q}(1-q) = 0 \end{aligned} \tag{25}$$

The variance of the score: Fisher Information

Since the mean of the score is zero, its variance can be written as

$$\begin{aligned}
\mathcal{I}_n(\theta) &= E_X\left((S(\theta, X))^2\right) = E_X\left(\left(\frac{\partial \ln(L(X|\theta))}{\partial \theta}\right)^2\right) \\
&= \int_{-\infty}^{+\infty}\left(\frac{\partial \ln(L(x|\theta))}{\partial \theta}\right)^2 L(X|\theta)dX \\
&= -\int_{-\infty}^{+\infty}\left(\frac{\partial^2 \ln(L(x|\theta))}{\partial \theta^2}\right) L(X|\theta)dX \\
&= -E_X\left(\frac{\partial^2 \ln(L(X|\theta))}{\partial \theta^2}\right) \qquad (26)
\end{aligned}$$

- it increases with the (absolute value of the) second derivative;
- it measures concavity, and thus how precise is ML estimation;
- the subscript n indicates that $\mathcal{I}_n(\theta)$ is the Fisher Information for the likelihood of a sample of n random variables.
- it is a $k \times k$ symmetric matrix for k parameters θ.

Equivalence of the expressions for $\mathcal{I}_n(\theta)$ in (26)

Using the simplified notation:

$$S(\theta, X) = l_\theta(\theta, X) \tag{27}$$

we can demonstrate the second to third line step in previous slide:

$$E_X\left((l_\theta(\theta, X))^2\right) = -E_X\left(l_{\theta\theta}(\theta, X)\right) \tag{28}$$

Given

$$\int_{-\infty}^{+\infty} e^{l(\theta,X)} dX = 1 \tag{29}$$

take the derivative with respect to θ on both sides

$$\int_{-\infty}^{+\infty} l_\theta(\theta, X) e^{l(\theta,X)} dX = 0 \tag{30}$$

...

Expressions for $\mathcal{I}_n(\theta)$ in (26) (cont.)

and do it again on both sides

$$\int_{-\infty}^{+\infty} \left[l_{\theta\theta}(\theta, X) + (l_\theta(\theta, X))^2 \right] e^{l(\theta,X)} dX = 0 \tag{31}$$

rearranging

$$-\int_{-\infty}^{+\infty} l_{\theta\theta}(\theta, X) e^{l(\theta,X)} dX = \int_{-\infty}^{+\infty} (l_\theta(\theta, X))^2 e^{l(\theta,X)} dX \tag{32}$$

which proves the result

$$-E_X \left(l_{\theta\theta}(\theta, X) \right) = E_X \left((l_\theta(\theta, X))^2 \right)$$

Fisher Information for samples of size n and 1

The Fisher Information increases with the size of the sample.

A useful relationship links $\mathcal{I}_n(\theta)$ to $\mathcal{I}_1(\theta)$.

$$\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta) \tag{33}$$

where $\mathcal{I}_1(\theta)$ is the Information computed for one generic observation in the sample.

This distinction is important because, as we will show:

- $\frac{1}{\mathcal{I}_n(\theta)}$ is the variance of an unbiased ML estimator which is also the Cramer-Rao lower bound, i.e. the lowest possible variance of an unbiased estimator when the sample has size n;
- $\frac{1}{\mathcal{I}_1(\theta)}$ is the variance of the (normal) asymptotic distribution of the ML estimator.

Example: $\mathcal{I}_n(\theta)$ for the exponential distribution

$$X \sim f_X(X|\theta) = \theta e^{-\theta X} \tag{34}$$

The maximization problem is:

$$\max_{\theta} L(X|\theta) = \prod_{i=1}^{n} \theta e^{-\theta X} \tag{35}$$

$$\max_{\theta} l(X|\theta) = +n\ln\theta - \theta \sum_{i=1}^{n} X_i \tag{36}$$

which leads to the first order condition and ML estimator:

$$\frac{dl(X|\theta)}{d\theta} = S(\theta, X) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i = 0 \tag{37}$$

$$\hat{\theta}_{ML} = \frac{n}{\sum_{i=1}^{n} X_i} \tag{38}$$

Note: another case of biased ML estimator. What if we had specified the exponential as $X \sim f_X(X|\lambda) = \frac{1}{\lambda} e^{-\frac{1}{\lambda} X}$? (See problem set 1).

Example: $\mathcal{I}_n(\theta)$ for the exponential distribution (cont.)

The Score is

$$S(\theta, X) = l_\theta(\theta, X) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i \tag{39}$$

The Information for the sample of size n can be computed as

$$\mathcal{I}_n(\theta) = -E_X\left(\frac{\partial^2 \ln(L(X|\theta))}{\partial\theta^2}\right) = \frac{n}{\theta^2} \tag{40}$$

while the Information for a generic sample observation X_i

$$\mathcal{I}_1(\theta) = -E_X\left(\frac{\partial^2 \ln(L(X_i|\theta))}{\partial\theta^2}\right) = \frac{1}{\theta^2} \tag{41}$$

In part 3 of the slides, we will see that because of the CLT

$$AsyVar(\hat{\theta}_{ML}) = \frac{1}{\mathcal{I}_1(\theta)} = \theta^2 \tag{42}$$

Section 3

The method of Moments

A convenient method for multiple parameters

The “Methods of Moments" constructs estimators using restrictions of the population’s moments that should be satisfied also in the sample (under random sampling).

Let $x_1, ...x_n$ be a random sample of n draws from the random variable $X \sim f_X(x|\theta_1, ...\theta_k)$, where the parameters θ are unknown.

Suppose that k moments of X exist and let the *jth* moment be

$$E(X^j) = g_j(\theta_1, ..., \theta_k) \tag{43}$$

The MM estimates the k parameters as solutions of the system:

$$E(X^j) \quad = \quad g(\theta_1, ..., \theta_k) = \int_{-\infty}^{\infty} X^j f_X(X|\theta_1, ...\theta_k) dX = \frac{1}{n} \sum_{i=1}^{n} x_i^j$$

with one equation for each $j = 1, ..., k$.

Subsection 1

The MM estimator for the normal distribution

The MM estimator for the normal distribution

$$f_X(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{44}$$

The system that defines the MM estimators of μ and σ^2 is:

$$\begin{aligned} E(X) &= g_1(\hat{\mu},\hat{\sigma}^2) = \int_{-\infty}^{\infty} X f_X(x|\hat{\mu},\hat{\sigma}^2) dX = \frac{1}{n}\sum_{i=1}^{n} x_i \\ E(X^2) &= g_2(\hat{\mu},\hat{\sigma}^2) = \int_{-\infty}^{\infty} X^2 f_X(x|\hat{\mu},\hat{\sigma}^2) dX = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{aligned}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{45}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \tag{46}$$

The MM and ML estimators coincide in this case.

Subsection 2

A case in which the MM and ML estimators do not coincide

The Pareto distribution

Let X denote individual income. The Pareto's Law claims that

$$P(X \geq x) = \left(\frac{\nu}{X}\right)^{\theta} \quad \Rightarrow \quad F_X(X|\theta,\nu) = 1 - \left(\frac{\nu}{X}\right)^{\theta} \tag{47}$$

where ν is the (known) minimum income in the population and $\theta > 1$.

Thus by differentiation the pdf is:

$$X \sim f_X(X|\theta,\nu) = \theta\nu^{\theta}\left(\frac{1}{X}\right)^{\theta+1} \quad , \quad X > \nu; \theta > 1 \tag{48}$$

The MM estimator for θ in the Pareto distribution

$$E(X) = \int_{\nu}^{\infty} X\theta\nu^{\theta}\left(\frac{1}{X}\right)^{\theta+1} dX = \theta\nu^{\theta}\int_{\nu}^{\infty} X^{-\theta} dX = \frac{\theta\nu}{\theta - 1} \tag{49}$$

The Method of Moments estimates θ solving for $\hat{\theta}$:

$$E(X) = \frac{\hat{\theta}\nu}{\hat{\theta} - 1} = \frac{1}{n}\sum_{i=1}^{n}(X_i) = \bar{X} \tag{50}$$

which gives the estimator

$$\hat{\theta}^{MM} = \frac{\bar{X}}{\bar{X} - \nu} \tag{51}$$

And given a random sample $x_1, ..., x_n$, an estimate

$$\theta_e^{MM} = \frac{\bar{x}}{\bar{x} - \nu} \tag{52}$$

The ML estimator for θ in the Pareto distribution

The likelihood of the random sample $x_i, ..., x_n$ is

$$L(X|\theta,\nu) = \prod_{i=1}^{n} \theta\nu^{\theta} \left(\frac{1}{X_i}\right)^{\theta+1} \tag{53}$$

$$l(X|\theta,\nu) = \ln L(X|\theta,\nu) = n\ln\theta + n\theta\ln\nu - (\theta+1)\sum_{i=1}^{n} lnX_i \tag{54}$$

The first order condition is

$$\frac{n}{\theta} + n\ln\nu - \sum_{i=1}^{n} lnX_i = 0 \tag{55}$$

Hence the estimator and estimate are :

$$\hat{\theta}^{ML} = \frac{n}{-n\ln\nu + \sum_{i=1}^{n} lnX_i}; \qquad \theta_e^{ML} = \frac{n}{-n\ln\nu + \sum_{i=1}^{n} lnx_i} \tag{56}$$

Section 4

Small sample distribution of an estimator

An estimator is a random variable with a distribution

Estimators are random variables because they are transformations of the sample draws which are random variables.

Distributions and related moments of an estimator can thus be derived using:

- the rules for random variable transformation;
- the Moment Generating Functions;
- the Characteristic Function.

There are however cases in which the small sample distribution of an estimator cannot be derived, while using asymptotic results one can obtain the asymptotic distribution of the estimator.

This possibility is crucially useful in empirical analysis, as we will see.

Example: distribution of the sample mean of a normal

Using the *Moment Generating Function* we can show that:

$$\hat{\mu} \sim \phi\left(\hat{\mu}|\mu, \frac{\sigma^2}{n}\right) \tag{57}$$

where $\phi(.)$ is a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Consider (for $t \in \mathcal{R}$) the MGFs of X_i, $S = \sum_{i=1}^{n} X_i$ and $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$M_{X_i}(t) = E_{X_i}\left(e^{tX_i}\right) = e^{\mu t + \frac{\sigma^2 t^2}{2}} \tag{58}$$

$$M_S(t) = \prod_{i=1}^{n}\left(e^{\mu t + \frac{\sigma^2 t^2}{2}}\right) = e^{n\mu t + n\frac{\sigma^2 t^2}{2}} \tag{59}$$

$$M_{\hat{\mu}}(t) = M_S(\frac{1}{n}t) = e^{\mu t + \frac{\sigma^2 t^2}{n2}} \tag{60}$$

which is the MGF of a normal with mean μ and variance $\frac{\sigma^2}{n}$.