Econometrics (Review)

Lecture 1 EE426 – 2/2014

Chayanee Chawanote

Finite sample properties of estimators

- a random sample: $\{Y_1, Y_2, ..., Y_n\}$ drawn from a population distribution that depends on an unknown parameter θ.
- An estimator of θ is a rule that assigns each possible outcome of the sample a value of θ.
- E.g., a natural estimator of mean μ is an average of the random sample.
- More generally, an estimator *W* of a parameter θ can be expressed in a function *h* (rule): $W = h(Y_1, Y_2, ..., Y_n)$
- When a particular set of numbers, $\{y_1, y_2, ..., y_n\}$, is plugged into the function *h*, we obtain as estimate of θ: $w = h(y_1, y_2, ..., y_n)$
- *W* is a random variable because it depends on the random sample: we obtain different random samples from the population, the value of *W* can change.

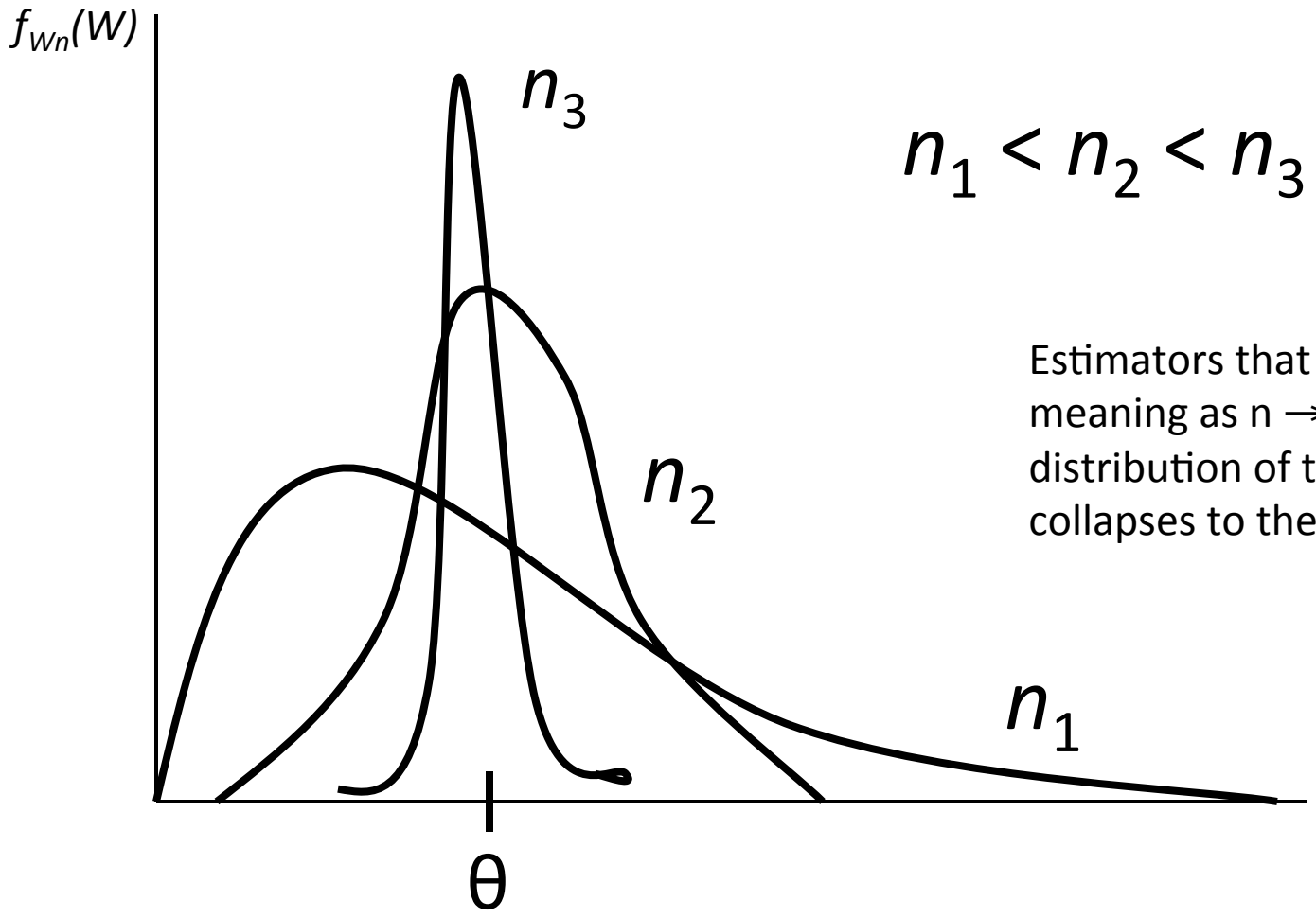
Finite sample properties of estimators

- Unbiased estimator: An estimator, *W* of θ, is an unbiased estimator if $E(W) = 0$, for all possible values of θ
- Bias of an estimator: $\text{Bias}(W) \equiv E(W) - \theta$
- Sampling variance: the variance associated with a sampling distribution. It shows how spread out the distribution of an estimator is.
- If $\{Y_i: i = 1,2,\ldots,n\}$ is a random sample from a population with mean μ and variance σ^2, then Y has the same mean as the population, but its sampling variance equal σ^2/n. Hence, the larger sample size, the smaller sampling variance.
- Relative efficiency: If W_1 and W_2 are two unbiased estimators, W_1 is efficient relative to W_2 when $Var(W_1) \leq Var(W_2)$ for all θ, with strict inequality for at least one value of θ

Asymptotic properties of estimators

- Consistency: Let W_n be an estimator of θ based on a sample Y_1, Y_2, …, Y_n of size n. Then W_n is a consistent estimator of θ if for every $\varepsilon > 0$, $P(|W_n - \theta| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$
- Probability limit: $\text{plim}(W_n) = \theta$
- If W_n is an unbiased estimator of θ and $\text{Var}(W_n) \rightarrow 0$ as $n \rightarrow \infty$, then $\text{plim}(W_n) = \theta$
- Law of large number: Let Y_1, Y_2, …, Y_n be independent, identically distributed random variables with mean μ. Then, $\text{plim}(\bar{Y}_n) = \mu$
- Asymptotic Normality implies that $P(Z_n < z) \rightarrow \Phi(z)$ as $n \rightarrow \infty$, or $P(Z_n < z) \approx \Phi(z)$
  - The cumulative distribution function of Z_n gets closer and closer to the cdf of the standard normal distribution as the sample size n gets large.

Sampling Distributions as $n \uparrow$



$$n_1 < n_2 < n_3$$

Estimators that are consistent, meaning as n → ∞, the distribution of the estimator collapses to the parameter value

Asymptotic properties of estimators

- Central limit theorem: the standardized average of any population with mean μ and variance σ^2 is asymptotically ~N(0,1), or

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

- CLT basically says that for non-normal data, the distribution of the sample means has an approximate normal distribution, no matter what the distribution of the original data looks like, as long as the sample size is large enough (usually at least 30) and all samples have the same size.
- It's important because it means that we can approximate the distribution of certain statistics, even if we know very little about the underlying sampling distribution.

General approaches to parameter estimation

- We want estimators to be unbiasedness, consistency, and efficiency
- Method of Moments

The parameter θ is shown to be related to some expected value in the distribution of Y, usually *E(Y) or E(Y²)*.

Suppose θ is related to the population mean as $\theta = g(\mu)$ for some function g. Because the sample average $\bar{Y}$ is an unbiased and consistent estimator of μ, it gives us the estimator $g(\bar{Y})$ of μ

We replace the population moment, μ, with its sample counter part, $\bar{Y}$ when we know that $g(\bar{Y})$ is consistent for θ, and if g(μ) is a linear function of μ, then $g(\bar{Y})$ is unbiased as well.

- sample covariance
- sample correlation coefficient

General approaches to parameter estimation

- Maximum Likelihood

Let $\{Y_1, Y_2, ..., Y_n\}$ be a random sample from the population distribution $f(y;\theta)$. Because of the random sampling assumption, the joint distribution of $\{Y_1, Y_2, ..., Y_n\}$ is simply the product of the densities: $f(y_1;\theta)f(y_2;\theta)\cdots f(y_n;\theta)$. In the discrete case, this is $P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n)$

Define the likelihood function as

$$L(\theta; Y_1, ..., Y_n) = f(y_1;\theta)f(y_2;\theta)\cdots f(y_n;\theta)$$

The maximum likelihood estimator of θ, W, is the value of θ that maximizes the likelihood function. This value depends on the random sample. Out of all the possible values for θ, the value that makes the likelihood of the observed data largest should be chosen.

General approaches to parameter estimation

- More convenient to work with the log-likelihood function:

$$\log[L(\theta;Y_1,...,Y_n)] = \sum_{i=1}^{n} \log[f(y_i;\theta)]$$

- Maximum likelihood estimation (MLE) is usually consistent and sometimes unbiased, generally the most asymptotically efficient estimator when the population model f(y) is correctly specified.
- The MLE is sometimes the minimum variance unbiased estimator.
- If the distribution of Y conditional on a set of explanatory variables, X_1, X_2,..., X_k, then we replace the density with $f(Y_i \mid X_{i1},...,X_{ik};\theta_1,...,\theta_p)$, where this density is allowed to depend on *p* parameters.

General approaches to parameter estimation

- Least Squares

  Find the m that makes the sum of squared deviations $\sum_{i=1}^{n}(Y_i - m)^2$ as small as possible.

- The sample mean, $\bar{Y}$, is a least squares estimator of the population mean.

- For some important distributions, including the normal and the Bernoulli, the sample average $\bar{Y}$ is also the maximum likelihood estimator of the population mean. All three approached often result in the same estimator. In other cases, the estimators are similar but not identical.

Interval Estimation and Confidence Intervals

- Suppose the population has a normal(μ,1) distribution and let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample from this population. Then, we have $\bar{Y}$ ~ Normal(μ, 1/n).
- Standardize $\bar{Y}$, which has a standard normal distribution, we have

$$P(\bar{Y} - 1.96/\sqrt{n} < \mu < \bar{Y} + 1.96/\sqrt{n}) = 0.95$$

  - The probability that the random interval $[\bar{Y} - 1.96/\sqrt{n}, \bar{Y} + 1.96/\sqrt{n}]$ contains the population mean μ is 95% >> interval estimate (w/ small y)
  - It is also called a 95% confidence interval, meaning that this random interval (w/ big Y) contains μ with probability 0.95.
  - This is NOT "the probability that μ is in the above interval is 0.95".
- If unknown variance, we need the sample standard deviation which is constructed

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{1/2}$$

Interval Estimation and Confidence Intervals

- We must rely on the t distribution instead of the standard normal distribution.

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

- So now we need c – the value such that 95% of the area in the t_{n-1} is between c and –c: $P(-c < t_{n-1} < c) = 0.95$. Once c has been properly chosen, we now have the random interval

  $[\bar{Y} - c \cdot S/\sqrt{n}, \bar{Y} + c \cdot S/\sqrt{n}]$ contains μ with prob .95.

- More generally, let c_α denote the 100(1-α) percentile in the t_{n-1} distribution. Hence, a 100(1-α)% confidence interval is obtained as $[\bar{y} - c_{\alpha/2} \cdot s/\sqrt{n}, \bar{y} + c_{\alpha/2} \cdot s/\sqrt{n}]$

- The t distribution approaches the standard normal distribution as the df gets large: for α = .05, $c_{\alpha/2} \rightarrow 1.96$ as $n \rightarrow \infty$.

Hypothesis Testing

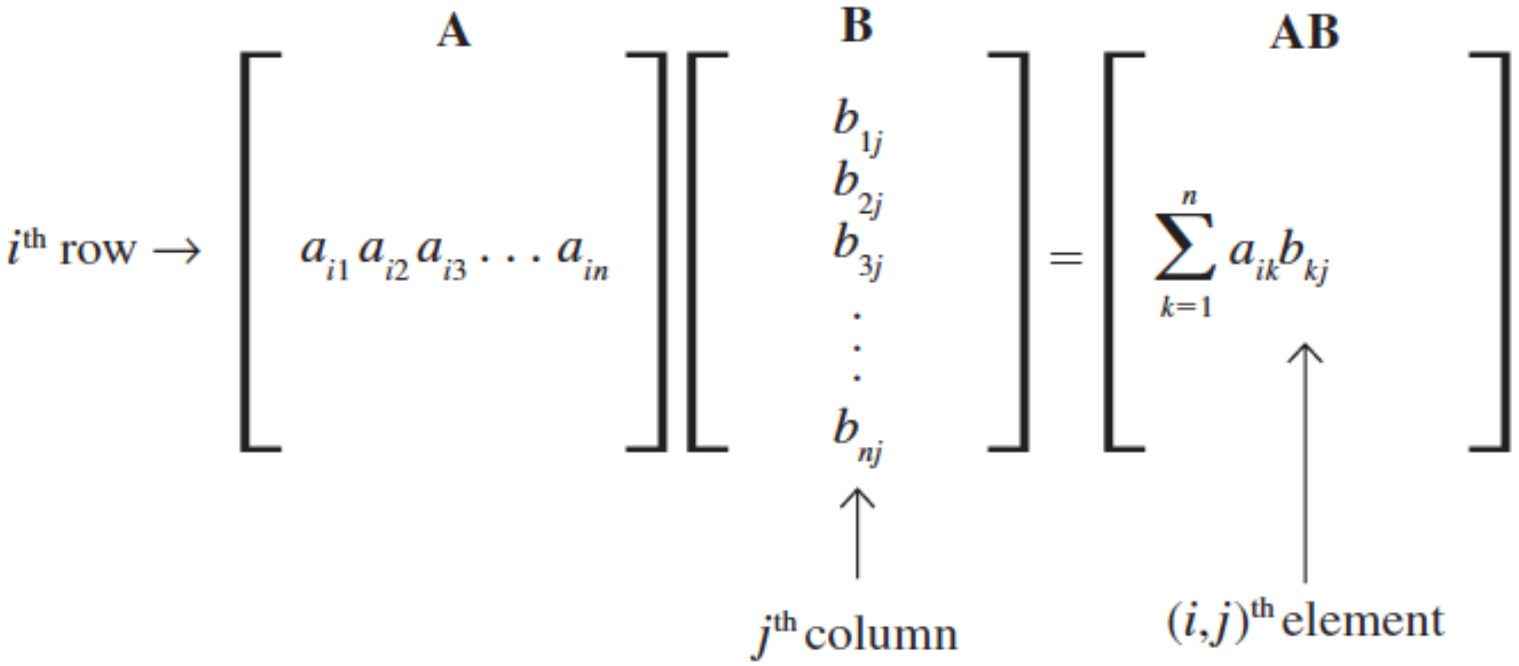
- Mostly, we do a test about the mean in a normal population.
- Null hypothesis H_0: $\mu = \mu_0$
- Alternative hypothesis H_a: $\mu > \mu_0$, $\mu < \mu_0$, or $\mu \neq \mu_0$
- Again, variance is replaced with the sample standard deviation. We need t-distribution to evaluate. Let call a test statistic as *T*. Under the null hypothesis, the random variable $T = \sqrt{n}(\bar{Y} - \mu_0)/S$ has a t_{n-1} distribution.
- Suppose we set 5% significance level. (Type I error: we reject H_0 when it is in fact true. We need fairly small Type I error.) The critical value c is chosen so that $P(T > c \mid H_0) = .05$ (one-tailed test). The rejection rule is $t > c$, where c is the $100(1-\alpha)$ percentile in a t_{n-1} distribution.
- P-value: the largest significance level we could carry out the test and still fail to reject H_0

Matrix (basic)

EE426 Econometrics 2

Matrix Operation

- Matrix multiplication: $\mathbf{AB} = \left[\sum_{k=1}^{n} a_{ik} b_{kj}\right].$



- same result with summation between 2 variables

- Transpose: if $A = [a_{ij}]_{m \times n}$, $A' = [a_{ji}]_{n \times m}$
- Symmetric matrix: $A' = A$

Matrix Operation

- Properties of Matrix Operation

1. $(a + b)A = aA + bA$
2. $a(A + B) = aA + aB$
3. $(ab)A = a(bA)$
4. $a(AB) = (aA)B$
5. $A + B = B + A$
6. $(A + B) + C = A + (B + C)$
7. $(AB)C = A(BC)$
8. $A(B + C) = AB + AC$
9. $(A + B)C = AC + BC$
10. $IA = AI = A$
11. $A + 0 = 0 + A = A$
12. $A – A = 0$
13. $A0 = 0A = 0$
14. $AB \neq BA$

Note: AB is possible if #rows A = #columns B

- Properties of Transpose

1. $(A')' = A$
2. $(aA)' = aA'$
3. $(A + B)' = A' + B'$
4. $(AB)' = B'A'$
5. $x'x = \Sigma x^2$ when x is n × 1 vector

Matrix Operation

- Partitioned Matrix Multiplication:

$A_{n\times k}$ and $B_{n\times m}$

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}$$

$$\mathbf{A}'\mathbf{B} = \sum_{i=1}^{n} \mathbf{a}_i'\mathbf{b}_i,$$

Each i, $\mathbf{a}_i'\mathbf{b}_i$ is k×m matrix >> A'B can be written as the sum of n matrices

As a special case, we have

$$\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n} \mathbf{a}_i'\mathbf{a}_i,$$

Trace and Inverse

- Trace: for any n×n matrix A, $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$, is the sum of its diagonal elements.
- Properties of Trace

1. $\mathrm{tr}(I_n) = n$
2. $\mathrm{tr}(A') = \mathrm{tr}(A)$
3. $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$
4. $\mathrm{tr}(aA) = a\mathrm{tr}(A)$

- Inverse: for any n×n matrix A has an inverse, A^{-1}, provided that $A^{-1}A = I_n$ and $AA^{-1} = I_n$. A is called "invertible or nonsingular matrix"
- Properties Inverse

1. If an inverse exists, it is unique.
2. $(aA)^{-1} = (1/a)A^{-1}$, $a \neq 0$ และ A is invertible
3. $(AB)^{-1} = B^{-1}A^{-1}$
4. $(A')^{-1} = (A^{-1})'$

Linear Independence and Rank of Matrix

- Linear Independence: Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ be a set of n×1 vector. These are linearly independent vectors if, and only if,

  $$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_r\mathbf{x}_r = 0$$

  implies that $a_1 = a_2 = \ldots = a_r = 0$
- If the above has a not equal to 0, it means we can write at least one vector x as a linear combination of the others. Call $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ as 'linearly dependent'.

- Rank(A) = maximum number of of linearly independent columns of A.
  - If A (n×m) has rank(A) = m, then A has full column rank
  - If A is n×m, its rank can be at most m

1. rank(A') = rank(A)
2. A (n×k), rank(A) ≤ min(n,k)
3. If A (k×k) and rank(A) = k, A is invertible.

Differentiation

- For a given n×1 vector a, define the linear function as

  $f(\mathbf{x}) = \mathbf{a}'\mathbf{x}$

- Partial derivatives: $\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{a}'$

- A quadratic form in a symmetric matrix A (n×n):

  $g(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$

- Partial derivatives: $\partial g(\mathbf{x})/\partial \mathbf{x} = 2\mathbf{x}'\mathbf{A}$

  (1×n vector)

Note: We can write a quadratic form in the summation as follows:

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>1} a_{ij}x_i x_j$$

Random Vectors

- Expected value

1. Let y (n×1) be a random vector,

   $E(y) = [E(y_1), E(y_2), ..., E(y_n)]'$

2. Let Z (n×m) be a random matrix,

   $E(Z) = [E(z_{ij})]$, (size n×m)

- Properties

1. $E(Ay + b) = AE(y) + b$, where A and b are nonrandom
2. $E(AZB) = AE(Z)B$, where A(p×n) and B(m×k) are nonrandom

Random Vectors

- Variance-Covariance Matrix

Let y (n×1) be a random vector. Then,

$$\mathrm{Var}(\mathbf{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} \qquad \sigma_j^2 = \mathrm{Var}(y_j) \qquad \sigma_{ij} = \mathrm{Cov}(y_i, y_j)$$

is a symmetric matrix

1. Var(a'y) = a'[Var(y)]a ≥ 0
2. Var(y) = E[(y – μ)(y-μ')], μ = E(y)
3. If y_i y_j are not correlated, Var(y) is a diagonal matrix, Var(y_i) = σ^2 for *i* = 1, 2,…, n and can be written as Var(y) = $\sigma^2 \mathrm{I}_n$
4. Var(Ay + b) = A[Var(y)]A'