Asymptotic Theory

Greene Ch. 4, App. D, Kennedy App. C, ***R*** scripts: *module3s1a*, *module3s1b*, *module3s1c*, *module3s1d*

Overview of key concepts

Consider a random variable x_n that itself is a function of other random variables and sample size "*n*" (e.g the sample mean, or the sample variance). Such a random variable is called a "*sequence random variable*", since its outcome depends also on "*n*".

We're interested in the properties of x_n as $n \to \infty$. There are two possible scenarios: (i) x_n can "converge in probability" to some constant *c*, or (ii) "converge in distribution" to another random variable "*x*". Most estimators used in practice converge in probability.

If x_n converges to *x*, we call the distribution of *x* the "limiting distribution" of x_n. However, for most estimators of interest we don't know or cannot derive this limiting distribution. This also includes estimators that converge in probability – a single spike over a constant c is not a very exciting or useful distribution! In that case, we derive instead the limiting distribution of a *function* involving x_n (i.e. a popular one is $\sqrt{n}\left(x_n - c\right)$, where x_n is a sample statistic and $c = \text{plim}\left(x_n\right)$). This function does not collapse to a spike.

This limiting distribution is exactly correct as $n \to \infty$. We use it to derive the "approximate" distribution (or "asymptotic" distribution) of x_n itself that holds approximately for finite "*n*". Then from that, we can get the "asymptotic expectation" (for example to assess "asymptotic unbiasedness" or "consistency"), and the "asymptotic variance" (for example to assess asymptotic "efficiency").

Finally, we can then use these properties to compare different estimators, and to perform hypothesis tests.

Convergence in Probability

A sequence random variable x_n converges in probability to constant c if

$$\lim_{n\to\infty} prob\left(\left|x_n - c\right| > \varepsilon\right) = 0 \qquad \forall \varepsilon > 0 \tag{1}$$

In words, as n increases it becomes increasingly unlikely for x_n to be "far" from c (ε here is just some arbitrary constant, say 0.001, not an error term). If (1) holds, we say "***plim*** $\boldsymbol{x_n = c}$**",** or "x_n is consistent for c".

Example: Consider the following binary mixture distribution:

$$prob\left(x_n = 0\right) = 1 - \frac{1}{n} \qquad prob\left(x_n = n\right) = \frac{1}{n}$$

Clearly, as $n \to \infty$ it is less and less likely that $x_n = n$. Since there are only two possible outcomes for x_n, it must be that $\text{plim } x_n = 0$.

"Plims" are generally difficult to derive directly, but the following sufficient condition helps in most practical cases:

If x_n has mean μ_n and variance σ_n^2 such that the ordinary limits of μ_n and σ_n^2 are c and 0 respectively we say that x_n "converges in mean square error" (or "converges in quadratic mean") to *c* and

$$\text{plim } x_n = c$$

So convergence in MSE implies convergence in probability (but the reverse is not necessarily true).

Example: For a sample mean of values drawn from *any* distribution with mean μ and variance σ^2, we have:

$$x_n = \tfrac{1}{n}\sum_{i=1}^{n} x_i \qquad E(x_n) = \mu \qquad V(x_n) = \frac{\sigma^2}{n}$$

$$\lim_{n\to\infty} E(x_n) = \mu \quad \lim_{n\to\infty} V(x_n) = 0$$

So by convergence in MSE, the sample mean x_n is a consistent estimator of the population mean. The sampling distribution of the sample mean converges to a spike at μ as the sample size goes to infinity -> see ***R*** script *module3s1a*.

This notion extends to the mean of any function of a given random variable. Specifically, for any function $g(x)$, where x is some random variable (not necessarily a sequence RV), if $E(g(x))$ and $V(g(x))$ are finite constants, then $\text{plim } \tfrac{1}{n}\sum_{i=1}^{n} g(x_i) = E(g(x))$.

Example: Consider a normally distributed RV *x* with mean μ and variance 1. It is known that

$$E(\exp(x)) = \exp\left(\mu + \tfrac{1}{2}\right), \quad V(\exp(x)) = \exp(2\mu + 2) - \exp(2\mu + 1)$$

Thus: $\text{plim} \tfrac{1}{\text{n}}\sum_{i=1}^{n}(\exp(x_i)) = \exp\left(\mu + \tfrac{1}{2}\right)$

Note: The above results are specific examples of "*Laws of Large Numbers" (LLNs)*. These laws make statements about the behavior of *averages* of samples when sample sizes become very large.

Slutsky Theorem

For any continuous function $g(x_n)$ that is itself not a function of *n*, we have $\text{plim } g(x_n) = g(\text{plim } x_n)$.

This is an important LLN-type result, which allows us to find plims for highly nonlinear *x_n*'s. In fact, often times we can't even derive analytical results for the expectation of such an *x_n*. Thanks to the Slutsky theorem, we can at least say something about the consistency of *x_n*.

Example:

Let $\overline{x} = \tfrac{1}{\text{n}}\sum_{i=1}^{n} x_i$ and $s^2 = \tfrac{1}{\text{n-1}}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Then: $\text{plim}\left(\frac{\overline{x}^2}{s^2}\right) = \frac{\text{plim } \overline{x}^2}{\text{plim } s^2} = \frac{(\text{plim } \overline{x})^2}{\text{plim } s^2} = \frac{\mu^2}{\sigma^2}$

Convergence in Distribution / Limiting Distribution

Assume we draw a sample of *n* x_i's from some distribution, and compute the sequence RV x_n (mean, median, etc.). Then we repeat this process many, many times to get a sampling distribution for $\mathrm{x_n}$ (just like we do in script *module3s1a*). Assume this series of $\mathrm{x_n}$'s has a cumulative distribution function $F_n(x)$, which is likely unknown. Consider another (non-sequence) RV *x* with cdf $F(x)$.

We say that x_n *converges in distribution (CID)* to *x* if $\lim_{n\to\infty}\left|F_n(x_n)-F(x)\right|=0$ at all continuity points of $F(x)$, and $F(x)$ is the *limiting distribution* of $\mathrm{x_n}$. Symbolically: $x_n \xrightarrow{d} x$.
The *limiting mean* and *limiting variance* of x_n are the mean and variance of the limiting distribution, assuming these moments exist.

Often times (but not always..) the moments of the limiting distribution of x_n are the ordinary limits of the moments of the finite sample distribution of x_n.

<u>Example</u>: Limiting distribution of $\mathrm{t_{n-1}}$ (Greene p. 1048)

Consider a sample size *n* from a standard normal distribution. The popular t-statistic for a test of the hypothesis that the population mean is zero is given by

$$t_{n-1}=\frac{\overline{x}_n}{\left(\frac{\sqrt{s_n^2}}{\sqrt{n}}\right)} \quad \text{where} \quad s_n^2=\tfrac{1}{\text{n-1}}\sum_{i=1}^{n}\left(\mathrm{x_i}-\overline{x}_n\right)^2,$$

Where (n-1) are called "degrees of freedom". Clearly *t* is a sequence RV as it depends on *n*. In this case we actually know the exact (finite sample) distribution of t_{n-1}:

$$f\left(t_{n-1}\right)=\frac{\Gamma(n/2)}{\Gamma\left((n-1)/2\right)}\left((n-1)\pi\right)^{-1/2}\left(1+\frac{t_{n-1}^2}{n-1}\right)^{-n/2} \tag{2}$$

This distribution has a mean of zero and a variance of (n-1)/(n-3). The cdf is generically given by $F_{n-1}(t)=\int_{-\infty}^{t} f_{n-1}(x)\,dx$. As n goes to infinity, t_{n-1} converges to the standard normal, i.e. $t_{n-1}\xrightarrow{d} n(0,1)$.

Note that $\lim_{n\to\infty} 0=0,$ and $\lim_{n\to\infty}\frac{n-1}{n-3}=1$, so here we have a case where the moments of the limiting distribution (i.e. the standard normal) are identical to the limit of the moments of the finite sampling distribution.

See script *module3s1b* for a graphical illustration of this example.

Greene p. 1049 (theorem D. 16) shows some important rules for limiting distributions. Here is perhaps the most important, sort of the analog to the Slutsky Theorem for Convergence in Probability:

If $x_n\xrightarrow{d} x$ and $g(x_n)$ is a continuous function then $g(x_n)\xrightarrow{d} g(x)$.

Example:

We now know that $t_{n-1} \xrightarrow{d} n(0,1)$. How about the limiting distribution of t_{n-1}^2? Based on the rule above the limiting distribution will be that of the square of a standard normal, which is the *chi²* distribution with one DOF, i.e. $t_{n-1}^2 \xrightarrow{d} chi^2(1)$. See ***R*** script `module3s1b`.

Most estimators used in practice are sequence random variables that converge in probability. This implies that their limiting distribution collapses to a spike, which causes a dilemma, since we need a nonzero asymptotic variance to derive key properties of the estimator and perform hypothesis tests.

The way around this is to first perform a "*stabilizing transformation*" (ST) of the estimator to a random variable with a "well defined" limiting distribution. The most common ST is as follows:

$$ST\left(\hat{\boldsymbol{\theta}}\right) = z_n = \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \text{plim}\hat{\boldsymbol{\theta}}\right) = \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \tag{3}$$

This construct usually converges to a well defined limiting density, i.e. $z_n \xrightarrow{d} f(z)$. An estimator (i.e. our $\hat{\boldsymbol{\theta}}$ from above) with this property is called "root n consistent".

To get from there to the "approximate limiting distribution" or "*asymptotic distribution*" of $\hat{\boldsymbol{\theta}}$ itself we need to apply the (or, better, one of the many versions of the) ***Central Limit Theorem*** (CLT, according to Greene the "single most important theorem in econometrics"..). For a scalar random variable, the *Lindberg-Levy CLT* states:

If $x_1, x_2, \cdots, x_n$ constitute a random sample from *any pdf* with a finite mean μ and variance σ^2 and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ then $\sqrt{n}\left(\bar{x} - \mu\right) \xrightarrow{d} normal\left(0, \sigma^2\right)$ and $\sqrt{n}\left(\frac{\bar{x} - \mu}{\sigma}\right) \xrightarrow{d} normal\left(0,1\right)$.

Similar CLTs for scalars and vectors are given in Greene's Appendix D.

(So remember the following: Laws of Large Numbers apply to Convergence in Probability, Central Limit Theorems apply to Convergence in Distribution.)

See ***R*** script `module3s1c` for examples of stabilizing transformations.

Asymptotic Distributions

An asymptotic distribution of some estimator $\hat{\boldsymbol{\theta}}$ is the distribution used to approximate the true (unknown) finite sample distribution. In econometrics, we generally derive asymptotic distributions by first deriving the limiting distribution of ST($\hat{\boldsymbol{\theta}}$), then applying a CLT. Specifically:

Scalar case:

If $\sqrt{n}\left(\dfrac{\bar{x}-\mu}{\sigma}\right)\xrightarrow{d} normal\left(0{,}1\right)$, then “approximately” or “asymptotically” $\bar{x}\xrightarrow{d} normal\left(\mu,\sigma^2/n\right)$ or, equivalently, $\bar{x}\overset{a}{\sim} n\left(\mu,\sigma^2/n\right)$.

Vector case:

If $\hat{\boldsymbol{\theta}}$ is an estimator for parameter vector $\boldsymbol{\theta}$, and $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d} n\left(\mathbf{0},\mathbf{V}\right)$ then $\hat{\boldsymbol{\theta}}\overset{a}{\sim} n\left(\boldsymbol{\theta},\tfrac{1}{n}\mathbf{V}\right)$.

The term $\tfrac{1}{n}\mathbf{V}$ is called the “a*symptotic variance-covariance matrix*” (or, sloppily, “asymptotic variance”), which I will generically label as $V_a\left(\hat{\boldsymbol{\theta}}\right)$. If the above holds we say that $\hat{\boldsymbol{\theta}}$ is “asymptotically normally distributed” or “*asymptotically normal*”. Also, if the asymptotic variance of any other consistent, asymptotically normal estimator (call it $\tilde{\boldsymbol{\theta}}$) exceeds $V_a\left(\hat{\boldsymbol{\theta}}\right)$by a non-negative definite matrix, $\hat{\boldsymbol{\theta}}$ is said to be asymptotically efficient.

Example:

We already know that the asymptotic variance of the MLE estimator $\hat{\boldsymbol{\theta}}$ is the inverse of the information matrix, i.e. $V_a\left(\hat{\boldsymbol{\theta}}\right)=\left(I\left(\boldsymbol{\theta}\right)\right)^{-1}$. It is also true that $\hat{\boldsymbol{\theta}}$ is consistent (as we will show below) and asymptotically normal. So together: $\hat{\boldsymbol{\theta}}\overset{a}{\sim} n\left(\boldsymbol{\theta},\left(I\left(\boldsymbol{\theta}\right)\right)^{-1}\right)$. It can be shown that no other consistent, asymptotically normal estimator has a “smaller” asymptotic variance, so $\hat{\boldsymbol{\theta}}$ is also asymptotically efficient. The smallest possible asymptotic variance is often referred to as the *Cramer-Rao Bound (CRB).* So we can say that the MLE estimator achieves the CRB.

(Note: A “consistent” estimator is often also called “*asymptotically unbiased*”).

Asymptotic Properties of the Least Squares Estimator

Even so we already know the finite sample properties of **b**, it’s still important to also understand its asymptotic qualities (since we often need to compare it to other estimators with unknown finite properties).

*Consistency of **b***

Assume $\text{plim}\left(\tfrac{1}{\text{n}}\mathbf{X}'\mathbf{X}\right)=\mathbf{Q}$, a positive definite matrix. We can write

$$\mathbf{b}=\boldsymbol{\beta}+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\varepsilon}=\boldsymbol{\beta}+\left(\tfrac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\tfrac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} \tag{4}$$

Note that the 1/n terms cancel out, so mathematically the last expression is truly equivalent to the first and second. Now, using the Slutsky Theorem:

$$\text{plim}\,\mathbf{b}=\boldsymbol{\beta}+\left(\text{plim}\left(\tfrac{1}{\text{n}}\mathbf{X}'\mathbf{X}\right)\right)^{-1}\text{plim}\left(\tfrac{1}{\text{n}}\mathbf{X}'\boldsymbol{\varepsilon}\right)=\beta+\mathbf{Q}^{\mathbf{-1}}\text{plim}\left(\tfrac{1}{\text{n}}\mathbf{X}'\boldsymbol{\varepsilon}\right) \tag{5}$$

As shown in detail in Greene, Ch. 4, the plim of the last term is zero, thus $\operatorname{plim}\mathbf{b}=\boldsymbol{\beta}$, i.e. **b** is consistent for $\boldsymbol{\beta}$.

*Asymptotic Normality of **b** and the "Delta Method"*

To derive the full asymptotic distribution of **b** we first perform a stabilizing transformation.....

$$\sqrt{n}\left(\mathbf{b}-\boldsymbol{\beta}\right)=\left(\tfrac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\tfrac{1}{\sqrt{n}}\mathbf{X}'\boldsymbol{\varepsilon} \tag{6}$$

... and then apply a multivariate version of the CLT (details see textbook) to derive

$$\mathbf{b}\overset{a}{\sim}n\left(\boldsymbol{\beta},\tfrac{\sigma^2}{n}\mathbf{Q}^{\mathbf{-1}}\right) \tag{7}$$

In practice the asymptotic variance is estimated as $\hat{V}_a\left(\mathbf{b}\right)=s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$, which of course is the same procedure we used for the finite variance. It can be shown that $s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ is a consistent estimator for $\tfrac{\sigma^2}{n}\mathbf{Q}^{\mathbf{-1}}$.

We can also note that **b** will be asymptotically efficient is we add the normality assumption for the error term. This result flows form the fact that under normal errors $\mathbf{b}=\hat{\boldsymbol{\beta}}_{\mathbf{MLE}}$, and MLE estimators are always asymptotically efficient (i.e. achieve the CRB).

Estimating the Variance for a function of original estimates: The Delta Method and the Krinsky-Robb procedure

Often times our ultimate construct of interest is a (potentially nonlinear) function of all or some of the elements in $\boldsymbol{\beta}$, i.e. $f\left(\boldsymbol{\beta}\right)$. Using $f\left(\mathbf{b}\right)$ as the estimator for $f\left(\boldsymbol{\beta}\right)$, we can use the following asymptotic result:

$$f\left(\mathbf{b}\right)\overset{a}{\sim}n\left(f\left(\boldsymbol{\beta}\right),\boldsymbol{\Gamma}\left(V_a\left(\mathbf{b}\right)\right)\boldsymbol{\Gamma}'\right)\quad\text{where}\quad\boldsymbol{\Gamma}=\frac{\partial f\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}'} \tag{8}$$

In practice, the asymptotic variance is estimated by

$$\hat{V}_a\left(f\left(\mathbf{b}\right)\right)=\mathbf{C}\hat{V}_a\left(\mathbf{b}\right)\mathbf{C}'\quad\text{where}\quad\mathbf{C}=\frac{\partial f\left(\mathbf{b}\right)}{\partial\mathbf{b}'} \tag{9}$$

This derivation is commonly referred to as the "*Delta Method*". It can be used for any asymptotically normally distributed estimator, not just the OLS estimator.

The square roots of the diagonal of $\hat{V}_a\left(f\left(\mathbf{b}\right)\right)$ can then be interpreted as *standard errors* for $f\left(\mathbf{b}\right)$. This is usually the ultimate objective of applying the Delta Method.

An alternative and asymptotically equivalent procedure to derive these standard errors is via simulation, e.g. as suggested by Krinsky and Robb (1986). The *Krinsky-Robb (KR) method* works as follows:

1. Take r=1..R (a large number, say 10,000) draws of coefficient estimates from their asymptotic distribution, i.e. from $\mathbf{b_r} \sim n\left(\mathbf{b}, \hat{V}_a(\mathbf{b})\right)$.
2. For each case, compute $f(\mathbf{b_r})$.
3. Examine the empirical variance-covariance matrix for these draws. Use the diagonal to construct standard errors.

For an empirical example of the Delta method & KR see ***R*** script `module3s1d`.
In that example, we have original estimated the error standard deviation $\hat{\sigma}$, along with its estimated variance $V(\hat{\sigma})$ (the last row, last column element of the inverted negative Hessian). Assume you're instead interested in an estimate of the error variance, $\hat{\sigma}^2$. You can use the Delta Method to obtain standard errors for this variance:

In this (single-variable) case we have

$$f(\hat{\sigma}) = \hat{\sigma}^2$$
$$C = \frac{\partial \hat{\sigma}^2}{\partial \hat{\sigma}} = 2\hat{\sigma}$$
$$\hat{V}\left(\hat{\sigma}^2\right) = 2\hat{\sigma} * \hat{V}(\hat{\sigma}) * 2\hat{\sigma} = 4(\hat{\sigma})^2 \hat{V}(\hat{\sigma})$$

<u>Analytical Example for the Delta method:</u>

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \qquad f(\boldsymbol{\beta}) = \begin{bmatrix} \beta_1/(1-\beta_2) - \beta_3 \\ \beta_1\beta_2{}^2 \end{bmatrix}$$

$$f(\mathbf{b}) = \begin{bmatrix} f_1(\mathbf{b}) \\ f_2(\mathbf{b}) \end{bmatrix} = \begin{bmatrix} b_1/(1-b_2) - b_3 \\ b_1 b_2{}^2 \end{bmatrix} \qquad \hat{V}_a(\mathbf{b}) = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_{33} \end{bmatrix}$$

$$\mathbf{C} = \left(\frac{\partial f(\mathbf{b})}{\partial \mathbf{b}'}\right) = \begin{bmatrix} \frac{\partial f_1(\mathbf{b})}{\partial b_1} & \frac{\partial f_1(\mathbf{b})}{\partial b_2} & \frac{\partial f_1(\mathbf{b})}{\partial b_3} \\ \frac{\partial f_2(\mathbf{b})}{\partial b_1} & \frac{\partial f_2(\mathbf{b})}{\partial b_2} & \frac{\partial f_2(\mathbf{b})}{\partial b_3} \end{bmatrix} = \begin{bmatrix} \frac{1}{1-b_2} & \frac{b_1}{(1-b_2)^2} & -1 \\ b_2^2 & 2b_1b_2 & 0 \end{bmatrix}$$

$$\hat{V}_a\left(f(\mathbf{b})\right) = \mathbf{C}\hat{V}_a(\mathbf{b})\mathbf{C}'$$

We would usually let the computer solve the final expression. Note the dimensions: If **b** is *k* by 1 and *f*(**b**) has *J* elements (i.e. is *J* by 1) , **C** will be J by k, and $\hat{V}_a\left(f(\mathbf{b})\right) = \mathbf{C}\hat{V}_a(\mathbf{b})\mathbf{C}'$ will be *J* by *J.*

<u>**Caveat:**</u>

Keep in mind that the Delta Method and KR procedure are asymptotic concepts – these methods can be highly unreliable in a small-sample context!