Colin Cameron: Asymptotic Theory for OLS

1. OLS Estimator Properties and Sampling Schemes

1.1. A Roadmap

Consider the OLS model with just one regressor

$$y_i = \beta x_i + u_i.$$

The OLS estimator $\widehat{\beta} = \left(\sum_{i=1}^N x_i^2\right)^{-1} \sum_{i=1}^N x_i y_i$ can be written as

$$\widehat{\beta} = \beta + \frac{\frac{1}{N}\sum_{i=1}^N x_i u_i}{\frac{1}{N}\sum_{i=1}^N x_i^2}.$$

Then under assumptions given below (including $\mathrm{E}[u_i|x_i] = 0$)

$$\widehat{\beta} \xrightarrow{p} \beta + \frac{\operatorname{plim} \frac{1}{N}\sum_{i=1}^N x_i u_i}{\operatorname{plim} \frac{1}{N}\sum_{i=1}^N x_i^2} = \beta + \frac{0}{\operatorname{plim} \frac{1}{N}\sum_{i=1}^N x_i^2} = \beta.$$

It follows that $\widehat{\beta}$ is **consistent** for β.

And under assumptions given below (including $\mathrm{E}[u_i|x_i] = 0$ and $\mathrm{V}[u_i|x_i] = \sigma^2$)

$$\sqrt{N}(\widehat{\beta}-\beta) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N x_i u_i}{\frac{1}{N}\sum_{i=1}^N x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0, \sigma^2\left(\operatorname{plim} \frac{1}{N}\sum_{i=1}^N x_i^2\right)\right]}{\operatorname{plim} \frac{1}{N}\sum_{i=1}^N x_i^2} \xrightarrow{d} \mathcal{N}\left[0, \sigma^2\left(\operatorname{plim} \frac{1}{N}\sum_{i=1}^N x_i^2\right)^{-1}\right].$$

It follows that $\widehat{\beta}$ is **asymptotically normal** distributed with $\widehat{\beta} \overset{a}{\sim} \mathcal{N}\left[0, \sigma^2\left(\sum_{i=1}^N x_i^2\right)^{-1}\right]$.

1.2. Sampling Schemes and Error Assumptions

The key for consistency is obtaining the probability of the two averages (of $x_i u_i$ and of x_i^2), by use of laws of large numbers (LLN). And for asymptotic normality the key is the limit distribution of the average of $x_i u_i$, obtained by a central limit theorem (CLT).

Different assumptions about the stochastic properties of x_i and u_i lead to different properties of x_i^2 and $x_i u_i$ and hence different LLN and CLT.

For the data different **sampling schemes** assumptions include:

1. Simple Random Sampling (SRS).

   SRS is when we randomly draw (y_i, x_i) from the population. Then x_i are iid. So x_i^2 are iid, and $x_i u_i$ are iid if the errors u_i iid.

2. Fixed regressors.
   This occurs in an experiment where we fix the x_i and observe the resulting random y_i. Given x_i fixed and u_i iid it follows that $x_i u_i$ are inid (even if u_i are iid), while x_i^2 are nonstochastic.

3. Exogenous Stratified Sampling
   This occurs when oversample some values of x and undersample others. Then x_i are inid, so $x_i u_i$ are inid (even if u_i are iid) and x_i^2 are inid.

The simplest results assume u_i are iid. In practice for cross-section data the errors may be inid due to conditional heteroskedasticity, with $\mathrm{V}[u_i|x_i] = \sigma_i^2$ varying with i.

2. Asymptotic Theory for Consistency

Consider the limit behavior of a **sequence of random variables** b_N as $N \to \infty$. This is a stochastic extension of a sequence of real numbers, such as $a_N = 2 + (3/N)$.

Examples include: (1) b_N is an estimator, say $\widehat{\theta}$; (2) b_N is a component of an estimator, such as $N^{-1}\sum_i x_i u_i$; (3) b_N is a test statistic.

2.1. Convergence in Probability, Consistency, Transformations

Due to sampling randomness we can never be certain that a random sequence b_N, such as an estimator $\widehat{\theta}_N$, will be within a given small distance of its limit, even if the sample is infinitely large. But we can be almost certain. Different ways of expressing this almost certainty correspond to different types of convergence of a sequence of random variables to a limit. The one most used in econometrics is convergence in probability.

Recall that a sequence of nonstochastic real numbers $\{a_N\}$ converges to a if for any $\varepsilon > 0$, there exists $N^* = N^*(\varepsilon)$ such that for all $N > N^*$,

$$|a_N - a| < \varepsilon.$$

e.g. if $a_N = 2 + 3/N$, then the limit $a = 2$ since $|a_N - a| = |2 + 3/N - 2| = |3/N| < \varepsilon$ for all $N > N^* = 3/\varepsilon$.

For a sequence of r.v.'s we cannot be certain that $|b_N - b| < \varepsilon$, even for large N, due to the randomness. Instead, we require that the probability of being within ε is arbitrarily close to one.

Thus $\{b_N\}$ **converges in probability** to b if

$$\lim_{N\to\infty} \Pr[|b_N - b| < \varepsilon] = 1,$$

for any $\varepsilon > 0$. A formal definition is the following.

Definition A1: (*Convergence in Probability*) A sequence of random variables $\{b_N\}$ **converges in probability** to b if for any $\varepsilon > 0$ and $\delta > 0$, there exists $N^* = N^*(\varepsilon, \delta)$ such that for all $N > N^*$,

$$\Pr[|b_N - b| < \varepsilon] > 1 - \delta.$$

We write plim $b_N = b$, where plim is short-hand for **probability limit**, or $b_N \xrightarrow{p} b$. The limit b may be a constant or a random variable. The usual definition of convergence for a sequence of real variables is a special case of A1.

For **vector random variables** we can apply the theory for each element of $\mathbf{b}_N$. [Alternatively replace $|b_N - b|$ by the scalar $(\mathbf{b}_N - \mathbf{b})'(\mathbf{b}_N - \mathbf{b}) = (b_{1N} - b_1)^2 + \cdots + (b_{KN} - b_K)^2$ or its square root $||\mathbf{b}_N - \mathbf{b}||$.]

Now consider $\{\mathbf{b}_N\}$ to be a sequence of parameter estimates $\widehat{\boldsymbol{\theta}}$.

Definition A2: (*Consistency*) An estimator $\widehat{\boldsymbol{\theta}}$ is **consistent** for $\boldsymbol{\theta}_0$ if

$$\text{plim}\,\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0.$$

Unbiasedness $\not\Rightarrow$ consistency. Unbiasedness states $\mathrm{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}_0$. Unbiasedness permits variability around $\boldsymbol{\theta}_0$ that need not disappear as the sample size goes to infinity.

Consistency $\not\Rightarrow$ unbiasedness. e.g. add $1/N$ to an unbiased and consistent estimator - now biased but still consistent.

A useful property of plim is that it can apply to **transformations of random variables**.

Theorem A3: (*Probability Limit Continuity*). Let $\mathbf{b}_N$ be a finite-dimensional vector of random variables, and $g(\cdot)$ be a real-valued function continuous at a **constant vector point b**. Then

$$\mathbf{b}_N \xrightarrow{p} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{p} g(\mathbf{b}).$$

This theorem is often referred to as *Slutsky's Theorem*. We instead call Theorem A12 Slutsky's theorem.

Theorem A3 is one of the major reasons for the prevalence of asymptotic results versus finite sample results in econometrics. It states a very convenient property that does not hold for expectations.

For example, $\text{plim}(a_N, b_N) = (a, b)$ implies $\text{plim}(a_N b_N) = ab$, whereas $\mathrm{E}[a_N b_N]$ generally differs from $\mathrm{E}[a]\mathrm{E}[b]$.

Similarly $\text{plim}[a_N/b_N] = a/b$ provided $b \neq 0$.

2.2. Alternative Modes of Convergence

It is often easier to establish alternative modes of convergence, which in turn imply convergence in probability. [However, laws of large numbers, given in the next section, are used much more often.]

Definition A4: (*Mean Square Convergence*) A sequence of random variables $\{b_N\}$ is said to **converge in mean square** to a random variable b if

$$\lim_{N\to\infty} \mathrm{E}[(b_N - b)^2] = 0.$$

We write $b_N \xrightarrow{m} b$. Convergence in mean square is useful as $b_N \xrightarrow{m} b$ implies $b_N \xrightarrow{p} b$.

Another result that can be used to show convergence in probability is Chebychev's inequality.

Theorem A5: (*Chebyshev's Inequality*) For any random variable Z with mean μ and variance σ^2

$$\Pr[(Z-\mu)^2 > k] \leq \sigma^2/k, \quad \text{for any } k > 0.$$

A final type of convergence is almost sure convergence (denoted $\xrightarrow{as}$). This is conceptually difficult and often hard to prove. So skip. Almost sure convergence (or strong convergence) implies convergence in probability (or weak convergence).

2.3. Laws of Large Numbers

Laws of large numbers are theorems for *convergence in probability* (or *almost surely*) in the special case where the sequence $\{b_N\}$ is a **sample average**, i.e. $b_N = \bar{X}_N$ where

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Note that X_i here is general notation for a random variable, and in the regression context does not necessarily denote the regressor variables. For example $X_i = x_i u_i$.

A LLN is much easier way to get the plim than use of Definition A1 or Theorems A4 or A5. Widely used in econometrics because the estimators involve averages.

Definition A7: (*Law of Large Numbers*) A **weak law of large numbers** (LLN) specifies conditions on the individual terms X_i in $\bar{X}_N$ under which

$$(\bar{X}_N - \mathrm{E}[\bar{X}_N]) \xrightarrow{p} 0.$$

For a **strong law of large numbers** the convergence is instead almost surely.

If a LLN can be applied then

$$\begin{array}{lll} \operatorname{plim} \bar{X}_N & = \lim \mathrm{E}[\bar{X}_N] & \text{in general} \\ & = \lim N^{-1} \sum_{i=1}^{N} \mathrm{E}[X_i] & \text{if } X_i \text{ independent over } i \\ & = \mu & \text{if } X_i \text{ iid.} \end{array}$$

Leading examples of laws of large numbers follow.

Theorem A8: (*Kolmogorov LLN*) Let $\{X_i\}$ be iid (independent and identically distributed). If and only if $\mathrm{E}[X_i] = \mu$ exists and $\mathrm{E}[|X_i|] < \infty$, then $(\bar{X}_N - \mu) \xrightarrow{as} 0$.

Theorem A9: (*Markov LLN*) Let $\{X_i\}$ be inid (independent but not identically distributed) with $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{V}[X_i] = \sigma_i^2$. If $\sum_{i=1}^{\infty} (\mathrm{E}[|X_i - \mu_i|^{1+\delta}]/i^{1+\delta}) < \infty$, for some $\delta > 0$, then $(\bar{X}_N - N^{-1} \sum_{i=1}^{N} \mathrm{E}[X_i]) \xrightarrow{as} 0$.

The Markov LLN allows nonidentical distribution, at expense of require existence of an absolute moment beyond the first. The rest of the side-condition is likely to hold with cross-section data. e.g. if set $\delta = 1$, then need variance plus $\sum_{i=1}^{\infty} (\sigma_i^2/i^2) < \infty$ which happens if σ_i^2 is bounded.

Kolmogorov LLN gives almost sure convergence. Usually convergence in probability is enough and we can use the weaker *Khinchine's Theorem.*

Theorem A8b: (*Khinchine's Theorem*) Let $\{X_i\}$ be iid (independent and identically distributed). If and only if $\mathrm{E}[X_i] = \mu$ exists, then $(\bar{X}_N - \mu) \xrightarrow{p} 0$.

Which LLN should I use in regression applications? It depends on the sampling scheme.

3. Consistency of OLS Estimator

Obtain probability limit of $\widehat{\beta} = \beta + [\frac{1}{N} \sum_{i=1}^{N} x_i u_i]/[\frac{1}{N} \sum_{i=1}^{N} x_i^2]$.

3.1. Simple Random Sampling (SRS) with iid errors

Assume x_i iid with mean μ_x and u_i iid with mean 0.

As $x_i u_i$ are iid, apply Khinchine's Theorem yielding $N^{-1} \sum_i x_i u_i \xrightarrow{p} \mathrm{E}[xu] = \mathrm{E}[x] \times \mathrm{E}[u] = 0$.

As x_i^2 are iid, apply Khinchine's Theorem yielding $N^{-1} \sum_i x_i^2 \xrightarrow{p} \mathrm{E}[x^2]$ which we assume exists.

By Theorem A3 (Probability Limit Continuity) $\text{plim}[a_N/b_N] = a/b$ if $b \neq 0$. Then

$$\text{plim}\,\widehat{\beta} = \beta + \frac{\text{plim}\,\frac{1}{N}\sum_{i=1}^{N} x_i u_i}{\text{plim}\,\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta + \frac{0}{\text{E}[x^2]} = \beta.$$

3.2. Fixed Regressors with iid errors

Assume x_i fixed and that u_i iid with mean 0 and variance σ^2.

Then $x_i u_i$ are inid with mean $\text{E}[x_i u_i] = x_i \text{E}[u_i] = 0$ and variance $\text{V}[x_i u_i] = x_i^2\sigma^2$. Apply Markov LLN yielding $N^{-1}\sum_i x_i u_i - N^{-1}\sum_i \text{E}[x_i u_i] \xrightarrow{p} 0$, so $N^{-1}\sum_i x_i u_i \xrightarrow{p} 0$. The side-condition with $\delta = 1$ is $\sum_{i=1}^{\infty} x_i^2\sigma^2/i^2$ which is satisfied if x_i is bounded.

We also assume $\lim N^{-1}\sum_i x_i^2$ exists.

By Theorem A3 (Probability Limit Continuity) $\text{plim}[a_N/b_N] = a/b$ if $b \neq 0$. Then

$$\text{plim}\,\widehat{\beta} = \beta + \frac{\text{plim}\,\frac{1}{N}\sum_{i=1}^{N} x_i u_i}{\lim\,\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta + \frac{0}{\lim\,\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \beta.$$

3.3. Exogenous Stratified Sampling with iid errors

Assume x_i inid with mean $\text{E}[x_i]$ and variance $\text{V}[x_i]$ and u_i iid with mean 0.

Now $x_i u_i$ are inid with mean $\text{E}[x_i u_i] = \text{E}[x_i]\text{E}[u_i] = 0$ and variance $\text{V}[x_i u_i] = \text{E}[x_i^2]\sigma^2$, so need Markov LLN. This yields $N^{-1}\sum_i x_i u_i \xrightarrow{p} 0$, with the side-condition satisfied if $\text{E}[x_i^2]$ is bounded.

And x_i^2 are inid, so need Markov LLN with side-condition that requires e.g. existence and boundedness of $\text{E}[x_i^4]$.

Combining again get $\text{plim}\,\widehat{\beta} = \beta$.

4. Asymptotic Theory for Asymptotic Normality

Given consistency, the estimator $\widehat{\theta}$ has a *degenerate distribution* that collapses on θ_0 as $N \to \infty$. So cannot do statistical inference. [Indeed there is no reason to do it if $N \to \infty$.] Need to magnify or *rescale* $\widehat{\theta}$ to obtain a random variable with *nondegenerate distribution* as $N \to \infty$.

4.1. Convergence in Distribution, Transformation

Often the appropriate scale factor is $\sqrt{N}$, so consider $b_N = \sqrt{N}(\widehat{\theta} - \theta_0)$. b_N has an extremely complicated cumulative distribution function (cdf) F_N. But like any other function F_N it may have a limit function, where convergence is in the usual (nonstochastic) mathematical sense.

Definition A10: (*Convergence in Distribution*) A sequence of random variables $\{b_N\}$ is said to **converge in distribution** to a random variable b if

$$\lim_{N\to\infty} F_N = F,$$

at every continuity point of F, where F_N is the distribution of b_N, F is the distribution of b, and convergence is in the usual mathematical sense.

We write $b_N \xrightarrow{d} b$, and call F the **limit distribution** of $\{b_N\}$.
$b_N \xrightarrow{p} b$ implies $b_N \xrightarrow{d} b$.
In general, the reverse is not true. But if b is a constant then $b_N \xrightarrow{d} b$ implies $b_N \xrightarrow{p} b$.

To extend limit distribution to **vector random variables** simply define F_N and F to be the respective cdf's of vectors $\mathbf{b}_N$ and $\mathbf{b}$.

A useful property of convergence in distribution is that it can apply to **transformations of random variables**.

Theorem A11: (*Limit Distribution Continuity).* Let $\mathbf{b}_N$ be a finite-dimensional vector of random variables, and $g(\cdot)$ be a continuous real-valued function. Then

$$\mathbf{b}_N \xrightarrow{d} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{d} g(\mathbf{b}). \tag{4.1}$$

This result is also called the *Continuous Mapping Theorem.*

Theorem A12: (*Slutsky's Theorem*) If $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$, where a is a random variable and b is a constant, then

$$\begin{array}{ll} (i) & a_N + b_N \xrightarrow{d} a + b \\ (ii) & a_N b_N \xrightarrow{d} ab \\ (iii) & a_N/b_N \xrightarrow{d} a/b, \quad \text{provided } \Pr[b=0] = 0. \end{array} \tag{4.2}$$

Theorem A12 (also called *Cramer's Theorem*) permits one to separately find the limit distribution of a_N and the probability limit of b_N, rather than having to consider the joint behavior of a_N and b_N. Result (ii) is especially useful and is sometimes called the **Product Rule**.

4.2. Central Limit Theorems

Central limit theorems give *convergence in distribution* when the sequence $\{b_N\}$ is a *sample average.* A CLT is much easier way to get the plim than e.g. use of Definition A10.

By a LLN $\bar{X}_N$ has a degenerate distribution as it converges to a constant, $\lim \mathrm{E}[\bar{X}_N]$. So scale $(\bar{X}_N - \mathrm{E}[\bar{X}_N])$ by its standard deviation to construct a random variable with unit variance that will have a nondegenerate distribution.

Definition A13: (*Central Limit Theorem*) Let

$$Z_N = \frac{\bar{X}_N - \mathrm{E}[\bar{X}_N]}{\sqrt{\mathrm{V}[\bar{X}_N]}},$$

where $\bar{X}_N$ is a sample average. A **central limit theorem** (CLT) specifies the conditions on the individual terms X_i in $\bar{X}_N$ under which

$$Z_N \xrightarrow{d} \mathcal{N}[0,1],$$

i.e. Z_N converges in distribution to a standard normal random variable.

Note that

$$\begin{array}{lll} Z_N & = (\bar{X}_N - \mathrm{E}[\bar{X}_N])/\sqrt{\mathrm{V}[\bar{X}_N]} & \text{in general} \\ & = \sum_{i=1}^N (X_i - \mathrm{E}[X_i])/\sqrt{\sum_{i=1}^N \mathrm{V}[X_i]} & \text{if } X_i \text{ independent over } i \\ & = \sqrt{N}(\bar{X}_N - \mu)/\sigma & \text{if } X_i \text{ iid.} \end{array}$$

If $\bar{X}_N$ satisfies a central limit theorem, then so too does $h(N)\bar{X}_N$ for functions $h(\cdot)$ such as $h(N) = \sqrt{N}$, since

$$Z_N = \frac{h(N)\bar{X}_N - \mathrm{E}[h(N)\bar{X}_N]}{\sqrt{\mathrm{V}[h(N)\bar{X}_N]}}.$$

Often apply the CLT to the normalization $\sqrt{N}\bar{X}_N = N^{-1/2}\sum_{i=1}^N X_i$, since $\mathrm{V}[\sqrt{N}\bar{X}_N]$ is finite.

Examples of central limit theorems include the following.

Theorem A14: (*Lindeberg-Levy CLT*) Let $\{X_i\}$ be iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{V}[X_i] = \sigma^2$. Then $Z_N = \sqrt{N}(\bar{X}_N - \mu)/\sigma \xrightarrow{d} \mathcal{N}[0,1]$.

Theorem A15: (*Liapounov CLT*) Let $\{X_i\}$ be independent with $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{V}[X_i] = \sigma_i^2$. If $\lim \left(\sum_{i=1}^N \mathrm{E}[|X_i - \mu_i|^{2+\delta}]\right) / \left(\sum_{i=1}^N \sigma_i^2\right)^{(2+\delta)/2} = 0$, for some choice of $\delta > 0$, then $Z_N = \sum_{i=1}^N (X_i - \mu_i)/\sqrt{\sum_{i=1}^N \sigma_i^2} \xrightarrow{d} \mathcal{N}[0,1]$.

Lindberg-Levy is the CLT in introductory statistics. For the iid case the LLN required μ exists, while CLT also requires σ^2 exists.

For inid data the Liapounov CLT additionally requires existence of an absolute moment of higher order than two.

Which CLT should I use in regression applications? It depends on the sampling scheme.

5. Limit Distribution of OLS Estimator

Obtain limit distribution of $\sqrt{N}(\widehat{\beta}-\beta) = [\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i]/[\frac{1}{N}\sum_{i=1}^{N} x_i^2]$.

5.1. Simple Random Sampling (SRS) with iid errors

Assume x_i iid with mean μ_x and second moment $\mathrm{E}[x^2]$, and assume u_i iid with mean 0 and variance σ^2.

Then $x_i u_i$ are iid, with mean 0 and variance $\sigma^2\mathrm{E}[x^2]$. [Proof for variance: $\mathrm{V}_{x,u}[xu] = \mathrm{E}_x[\mathrm{V}[xu|x]]+\mathrm{V}_x[\mathrm{E}[xu|x]] = \mathrm{E}_x[x^2\sigma^2]+0 = \sigma^2\mathrm{E}_x[x^2]$. Apply Lindeberg-Levy CLT yielding

$$\sqrt{N}\left(\frac{N^{-1}\sum_{i=1}^{N} x_i u_i - 0}{\sqrt{\sigma^2\mathrm{E}[x^2]}}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\sqrt{\sigma^2\mathrm{E}[x^2]}} \xrightarrow{d} \mathcal{N}[0,1].$$

Using Slutsky's theorem that $a_N \times b_N \xrightarrow{d} a \times b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$), this implies

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i \xrightarrow{d} \mathcal{N}[0, \sigma^2\mathrm{E}[x^2]].$$

Then using Slutsky's theorem that $a_N/b_N \xrightarrow{d} a/b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$)

$$\sqrt{N}(\widehat{\beta}-\beta) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0, \sigma^2\mathrm{E}[x^2]\right]}{\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N} x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0, \sigma^2\mathrm{E}[x^2]\right]}{\mathrm{E}[x^2]} \xrightarrow{d} \mathcal{N}\left[0, \sigma^2\left(\mathrm{E}[x^2]\right)^{-1}\right],$$

where we use result from consistency proof that $\operatorname{plim} N^{-1}\sum_{i=1}^{N} x_i^2 = \mathrm{E}[x^2]$.

5.2. Fixed Regressors with iid errors

Assume x_i fixed and and u_i iid with mean 0 and variance σ^2.

Then $x_i u_i$ are inid with mean 0 and variance $\mathrm{V}[x_i u_i] = x_i^2\sigma^2$. Apply Liapounov LLN yielding

$$\sqrt{N}\left(\frac{N^{-1}\sum_{i=1}^{N} x_i u_i - 0}{\sqrt{\lim N^{-1}\sum_{i=1}^{N} x_i^2\sigma^2}}\right) = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i}{\sqrt{\sigma^2 \lim N^{-1}\sum_{i=1}^{N} x_i^2}} \xrightarrow{d} \mathcal{N}[0,1].$$

Using Slutsky's theorem that $a_N \times b_N \xrightarrow{d} a \times b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$), this implies

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i u_i \xrightarrow{d} \mathcal{N}[0, \sigma^2 \lim \frac{1}{N} \sum_{i=1}^{N} x_i^2].$$

Then using Slutsky's theorem that $a_N/b_N \xrightarrow{d} a/b$ (for $a_N \xrightarrow{d} a$ and $b_N \xrightarrow{p} b$)

$$\sqrt{N}(\widehat{\beta}-\beta) = \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i u_i}{\frac{1}{N} \sum_{i=1}^{N} x_i^2} \xrightarrow{d} \frac{\mathcal{N}\left[0, \sigma^2 \lim \frac{1}{N} \sum_{i=1}^{N} x_i^2\right]}{\lim \frac{1}{N} \sum_{i=1}^{N} x_i^2} \xrightarrow{d} \mathcal{N}\left[0, \sigma^2 \left(\lim \frac{1}{N} \sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

5.3. Exogenous Stratified Sampling with iid errors

Assume x_i inid with mean $\mathrm{E}[x_i]$ and variance $\mathrm{V}[x_i]$ and u_i iid with mean 0.

Similar to fixed regressors will need to use Liapounov CLT. We will get

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left[0, \sigma^2 \left(\operatorname{plim} \frac{1}{N} \sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

6. Asymptotic Distribution of OLS Estimator

From consistency we have that $\widehat{\beta}$ has a degenerate distribution with all mass at β, while $\sqrt{N}(\widehat{\beta} - \beta)$ has a limit normal distribution. For formal asymptotic theory, such as deriving hypothesis tests, we work with this limit distribution. But for exposition it is convenient to think of the distribution of $\widehat{\beta}$ rather than $\sqrt{N}(\widehat{\beta} - \beta)$. We do this by introducing the artifice of "asymptotic distribution".

Specifically we consider N large but not infinite, and drop the probability limit in the preceding result, so that

$$\sqrt{N}(\widehat{\beta} - \beta) \sim \mathcal{N}\left[0, \sigma^2 \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

It follows that the **asymptotic distribution** of $\widehat{\beta}$ is

$$\widehat{\beta} \overset{a}{\sim} \mathcal{N}\left[\beta, \sigma^2 \left(\sum_{i=1}^{N} x_i^2\right)^{-1}\right].$$

Note that this is exactly the same result as we would have got if $y_i = \beta x_i + u_i$ with $u_i \sim \mathcal{N}[0, \sigma^2]$.

7. Multivariate Normal Limit Theorems

The preceding CLTs were for scalar random variables.

Definition A16a: (*Multivariate Central Limit Theorem*) Let $\boldsymbol{\mu}_N = \mathrm{E}[\bar{\mathbf{X}}_N]$ and $\mathbf{V}_N = \mathbf{V}[\bar{\mathbf{X}}_N]$. A **multivariate central limit theorem** (CLT) specifies the conditions on the individual terms X_i in $\bar{X}_N$ under which

$$\mathbf{V}_N^{-1/2}(\mathbf{b}_N - \boldsymbol{\mu}_N) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{I}].$$

This is formally established using the following result.

Theorem A16: (*Cramer-Wold Device*) Let $\{\mathbf{b}_N\}$ be a sequence of random $k \times 1$ vectors. If $\boldsymbol{\lambda}'\mathbf{b}_N$ converges to a normal random variable for every $k \times 1$ constant non-zero vector $\boldsymbol{\lambda}$, then $\mathbf{b}_N$ converges to a multivariate normal random variable.

The advantage of this result is that if $\mathbf{b}_N = \bar{\mathbf{X}}_N$, then $\boldsymbol{\lambda}'\mathbf{b}_N = \lambda_1\bar{X}_{1N} + \cdots + \lambda_k\bar{X}_{kN}$ will be a scalar average and we can apply a scalar CLT, yielding

$$\frac{\boldsymbol{\lambda}'\bar{\mathbf{X}}_N - \boldsymbol{\lambda}'\boldsymbol{\mu}_N}{\sqrt{\boldsymbol{\lambda}'\mathbf{V}_N\boldsymbol{\lambda}}} \xrightarrow{d} \mathcal{N}[0,1], \quad \text{and hence} \quad \mathbf{V}_N^{-1/2}(\mathbf{b}_N - \boldsymbol{\mu}_N) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{I}].$$

Microeconometric estimators can often be expressed as

$$\sqrt{N}(\widehat{\theta} - \boldsymbol{\theta}_0) = \mathbf{H}_N\mathbf{a}_N,$$

where $\operatorname{plim}\mathbf{H}_N$ exists and $\mathbf{a}_N$ has a limit normal distribution. The distribution of this product can be obtained directly from part (ii) of Theorem A12 (Slutsky's theorem). We restate it in a form that arises for many estimators.

Theorem A17: (*Limit Normal Product Rule*) If a vector $\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\boldsymbol{\mu}, \mathbf{A}]$ and a matrix $\mathbf{H}_N \xrightarrow{p} \mathbf{H}$, where $\mathbf{H}$ is positive definite, then

$$\mathbf{H}_N\mathbf{a}_N \xrightarrow{d} \mathcal{N}[\mathbf{H}\boldsymbol{\mu}, \mathbf{H}\mathbf{A}\mathbf{H}'].$$

For example, the OLS estimator

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{\sqrt{N}}\mathbf{X}'\mathbf{u},$$

is $\mathbf{H}_N = (N^{-1}\mathbf{X}'\mathbf{X})^{-1}$ times $\mathbf{a}_N = N^{-1/2}\mathbf{X}'\mathbf{u}$ and we find the plim of $\mathbf{H}_N$ and the limit distribution of $\mathbf{a}_N$.

Theorem A17 also justifies *replacement* of a limit distribution variance matrix *by a consistent estimate* without changing the limit distribution. Given

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{B}],$$

then it follows by Theorem A17 that

$$\mathbf{B}_N^{-1/2} \times \sqrt{N}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{I}]$$

for any $\mathbf{B}_N$ that is a consistent estimate for $\mathbf{B}$ and is positive definite.

A formal multivariate CLT yields $\mathbf{V}_N^{-1/2}(\mathbf{b}_N-\boldsymbol{\mu}_N) \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{I}]$. Premultiply by $\mathbf{V}_N^{1/2}$ and apply Theorem A17, giving simpler form

$$\mathbf{b}_N-\boldsymbol{\mu}_N \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{V}],$$

where $\mathbf{V}=\operatorname{plim}\mathbf{V}_N$ and we assume $\mathbf{b}_N$ and $\mathbf{V}_N$ are appropriately scaled so that $\mathbf{V}$ exists and is positive definite.

Different authors express the **limit variance matrix V** in different ways.

1. General form: $\mathbf{V}=\operatorname{plim}\mathbf{V}_N$. With fixed regressors $\mathbf{V}=\lim\mathbf{V}_N$.
2. Stratified sampling or fixed regressors: Often $\mathbf{V}_N$ is a matrix average, say $\mathbf{V}_N = N^{-1}\sum_{i=1}^N \mathbf{S}_i$,where $\mathbf{S}_i$ is a square matrix. A LLN gives $\mathbf{V}_N-\mathrm{E}[\mathbf{V}_N] \xrightarrow{p} \mathbf{0}$. Then $\mathbf{V}=\lim\mathrm{E}[\mathbf{V}_N]=\lim N^{-1}\sum_{i=1}^N\mathrm{E}[\mathbf{S}_i]$.
3. Simple random sampling: $\mathbf{S}_i$ are iid, $\mathrm{E}[\mathbf{S}_i]=\mathrm{E}[\mathbf{S}]$, so $\mathbf{V}=\mathrm{E}[\mathbf{S}]$.

As an example, $\operatorname{plim} N^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i' = \lim N^{-1}\sum_i\mathrm{E}[\mathbf{x}_i\mathbf{x}_i']$ if LLN applies and $=\mathrm{E}[\mathbf{x}\mathbf{x}']$ under simple random sampling.

8. Asymptotic Normality

It can be convenient to re-express results in terms of $\widehat{\boldsymbol{\theta}}$ rather than $\sqrt{N}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$.

Definition A18: (*Asymptotic Distribution of* $\widehat{\boldsymbol{\theta}}$) If

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0},\mathbf{B}],$$

then we say that *in large samples* $\widehat{\boldsymbol{\theta}}$ is **asymptotically normally distributed** with

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\theta}_0, N^{-1}\mathbf{B}\right],$$

where the term "in large samples" means that N is large enough for good approximation but not so large that the variance $N^{-1}\mathbf{B}$ goes to zero.

A more shorthand notation is to implicitly presume asymptotic normality and use the following terminology.

Definition A19: (*Asymptotic Variance of* $\widehat{\boldsymbol{\theta}}$) If (**??**) holds then we say that the **asymptotic variance matrix** of $\widehat{\boldsymbol{\theta}}$ is

$$\mathrm{V}[\widehat{\boldsymbol{\theta}}] = N^{-1}\mathbf{B}. \tag{8.1}$$

Definition A20: (*Estimated Asymptotic Variance of* $\widehat{\boldsymbol{\theta}}$) If (**??**) holds then we say that the **estimated asymptotic variance matrix** of $\widehat{\theta}$ is

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = N^{-1}\widehat{\mathbf{B}}. \tag{8.2}$$

where $\widehat{\mathbf{B}}$ is a consistent estimate of $\mathbf{B}$.

Some authors use the $\mathrm{Avar}[\widehat{\boldsymbol{\theta}}]$ and $\widehat{\mathrm{Avar}}[\widehat{\boldsymbol{\theta}}]$ in definitions A19 and A20 to avoid potential confusion with the variance operator $\mathrm{V}[\cdot]$. It should be clear that here $\mathrm{V}[\widehat{\boldsymbol{\theta}}]$ means asymptotic variance of an estimator since few estimators have closed form expressions for the finite sample variance.

As an example of definitions 18-20, if $\{X_i\}$ are iid $[\mu, \sigma^2]$ then $\sqrt{N}(\bar{X}_N - \mu)/\sigma \xrightarrow{d} \mathcal{N}[0,1]$, or equivalently that $\sqrt{N}\bar{X}_N \xrightarrow{d} \mathcal{N}[\mu, \sigma^2]$. Then $\bar{X}_N \overset{a}{\sim} \mathcal{N}\left[\mu,\sigma^2/N\right]$; the asymptotic variance of $\bar{X}_N$ is σ^2/N; and the estimated asymptotic variance of $\bar{X}_N$ is s^2/N, where s^2 is a consistent estimator of σ^2 such as $s^2 = \sum_i \left(X_i - \bar{X}_N\right)^2/(N-1)$.

Definition A21: (*Asymptotic Efficiency*) A consistent asymptotically normal estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is said to be **asymptotically efficient** if it has an asymptotic variance-covariance matrix equal to the Cramer-Rao lower bound

$$-\left(\mathrm{E}\left[\left.\frac{\partial^2 \ln L_N}{/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_0}\right]\right)^{-1}.$$

9. OLS Estimator with Matrix Algebra

Now consider $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, so

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

Note that the $k \times k$ matrix $\mathbf{X}'\mathbf{X} = \sum_i \mathbf{x}_i\mathbf{x}_i'$ where $\mathbf{x}_i$ is a $k \times 1$ vector of regressors for the i^{th} observation.

9.1. Consistency of OLS

To prove consistency we rewrite this as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} N^{-1}\mathbf{X}'\mathbf{u}.$$

The reason for renormalization in the right-hand side is that $N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i'$ is an average that converges in probability to a finite nonzero matrix if $\mathbf{x}_i$ satisfies assumptions that permit a LLN to be applied to $\mathbf{x}_i\mathbf{x}_i'$.

Then

$$\operatorname{plim}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{u}\right),$$

using Slutsky's Theorem (Theorem A.3). The OLS estimator is therefore **consistent** for $\boldsymbol{\beta}$ (i.e., $\operatorname{plim}\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \boldsymbol{\beta}$) if

$$\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}.$$

If a law of LLN can be applied to the average $N^{-1}\mathbf{X}'\mathbf{u} = N^{-1}\sum_i \mathbf{x}_i u_i$ then a necessary condition for this to hold is that $\mathrm{E}[\mathbf{x}_i u_i] = \mathbf{0}$. The fundamental condition for consistency of OLS is that $\mathrm{E}[u_i|\mathbf{x}_i] = \mathbf{0}$ so that $\mathrm{E}[\mathbf{x}_i u_i] = \mathbf{0}$.

9.2. Limit Distribution of OLS

Given consistency, the limit distribution of $\widehat{\boldsymbol{\beta}}$ is degenerate with all the mass at $\boldsymbol{\beta}$. To obtain a limit distribution we multiply $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ by $\sqrt{N}$, so

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} N^{-1/2}\mathbf{X}'\mathbf{u}.$$

We know $\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}$ exists and is finite and nonzero from the proof of consistency. For iid errors, $\mathrm{E}[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{I}$ and $\mathrm{V}[\mathbf{X}'\mathbf{u}|\mathbf{X}] = \mathrm{E}[\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}'|\mathbf{X}] = \boldsymbol{\sigma}^2\mathbf{X}'\mathbf{X}$ we assume that a CLT can be applied to yield

$$N^{-1/2}\mathbf{X}'\mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \sigma^2\left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)].$$

Then by Theorem A17: (Limit Normal Product Rule)

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\xrightarrow{d} \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} \times \mathcal{N}[\mathbf{0}, \sigma^2\left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)] \\ &\xrightarrow{d} \mathcal{N}[\mathbf{0}, \sigma^2\left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}].\end{aligned}$$

9.3. Asymptotic Distribution of OLS

Then dropping the limits

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}[\mathbf{0}, \sigma^2\left(N^{-1}\mathbf{X}'\mathbf{X}\right)],$$

so

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

The asymptotic variance matrix is

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

and is consistently estimated by the estimated variance matrix

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = s^2(\mathbf{X}'\mathbf{X})^{-1},$$

where s^2 is consistent for σ^2. For example, $s^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}/(N-k)$ or $s^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}/N$.

9.4. OLS with Heteroskedatic Errors

What if the errors are heteroskedastic? If $\mathrm{E}[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \boldsymbol{\Sigma} = \mathrm{Diag}[\sigma_i^2]$ then $\mathrm{V}[\mathbf{X}'\mathbf{u}|\mathbf{X}] = \mathrm{E}[\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}'|\mathbf{X}] = \mathbf{X}'\boldsymbol{\Sigma}\mathbf{X} = \sum_{i=1}^N \sigma_i^2 \mathbf{x}_i\mathbf{x}_i'$. A CLT gives

$$N^{-1/2}\mathbf{X}'\mathbf{u} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \operatorname{plim} N^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}],$$

leading to

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\xrightarrow{d} \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} \times \mathcal{N}[\mathbf{0}, \operatorname{plim} N^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}] \\ &\xrightarrow{d} \mathcal{N}[\mathbf{0}, \sigma^2 \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} \times \operatorname{plim} N^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X} \times \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}].\end{aligned}$$

Then dropping the limits etcetera

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}].$$

The asymptotic variance matrix is

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

White (1980) showed that this can be consistently estimated by

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}(\textstyle\sum_{i=1}^N \widehat{u}_i{}^2\mathbf{x}_i\mathbf{x}_i')(\mathbf{X}'\mathbf{X})^{-1},$$

even though $\widehat{u}_i{}^2$ is not consistent for σ_i^2.