Asymptotic properties

We now look at properties of the OLS estimator that hold in **large samples**, or more precisely, in the limit as the sample size n tends to infinity, keeping the number of explanatory variables K fixed

The idea is that, at least for models estimated using large finite samples, these asymptotic results will provide a useful approximation to the behaviour of our estimators and test statistics

The power of this approach is that we can establish useful asymptotic results under **weaker assumptions** than those needed to show that the OLS estimator is unbiased and efficient in the classical linear regression model

And those needed to establish the exact finite sample distributions of the OLS estimator and the t- and F-statistics in the classical linear regression model with normally distributed errors

Consistency

We first consider a sufficient set of assumptions for the OLS estimator to be consistent in the linear model

i) $y_i = x_i'\beta + u_i$ for $i = 1, ..., n$ or $y = X\beta + u$

ii) The data on (y_i, x_i') are independently and identically distributed, with $\mathrm{E}(x_i u_i) = 0$ for all $i = 1, ..., n$

iii) The $K \times K$ matrix $M_{XX} = \mathrm{E}(x_i x_i')$ exists and is non-singular

Then $\mathrm{p}\lim_{n\to\infty} \widehat{\beta}_{\mathrm{OLS}} = \beta$, or $\widehat{\beta}_{\mathrm{OLS}} \xrightarrow{\mathrm{P}} \beta$

Recall that a sequence of random variables $\{x_n : n = 1, 2, ...\}$ **converges in probability** to the constant c if, for all $\delta > 0$ and $\varepsilon > 0$, we have

$$\mathrm{P}(|x_n - c| < \delta) > 1 - \varepsilon \quad \text{for all } n > n^*$$

or equivalently

$$\lim_{n \to \infty} \mathrm{P}(|x_n - c| < \delta) = 1$$

$$\lim_{n \to \infty} \mathrm{P}(|x_n - c| \geqslant \delta) = 0$$

This is denoted by $\mathrm{p}\lim_{n \to \infty} x_n = c$, or $x_n \xrightarrow{\mathrm{P}} c$

This definition is applied element-by-element to random vectors or matrices

For example, if we have scalar random variables $x_n \xrightarrow{\text{P}} c$ and $y_n \xrightarrow{\text{P}} d$, the random vector $(x_n, y_n) \xrightarrow{\text{P}} (c, d)$

An important property of probability limits is known as **Slutsky's theorem**, stated here for random vectors (or matrices)

Let the random vector (or matrix) x_n have a probability limit given by the vector (or matrix) c, and let $g(a)$ be a function which is continuous at the point $a = c$

Then $g(x_n) \xrightarrow{\text{P}} g(c)$

For example, let $x_n = (x_{1n}, x_{2n}) \xrightarrow{\mathrm{P}} (c_1, c_2)$ and let $g(a_1, a_2) = a_1 a_2$

Then we have $g(x_{1n}, x_{2n}) = x_{1n}x_{2n} \xrightarrow{\mathrm{P}} c_1 c_2$,

or $\mathrm{p}\lim_{n\to\infty}(x_{1n}x_{2n}) = \mathrm{p}\lim_{n\to\infty} x_{1n} \mathrm{p}\lim_{n\to\infty} x_{2n}$

The probability limit of the product of two random variables x_{1n} and x_{2n} is the product of their probability limits

Note that this property does not hold in general for expected values

We use this property to prove the consistency of $\widehat{\beta}_{\mathrm{OLS}}$ under the stated assumptions

With iid data, the Law of Large Numbers implies that the $K \times 1$ vector of sample means

$$\frac{1}{n}\sum_{i=1}^{n} x_i u_i \xrightarrow{\text{P}} \text{E}(x_i u_i)$$

with $\text{E}(x_i u_i) = 0$ from assumption (ii)

We can also write the $K \times 1$ vector $\frac{1}{n}\sum_{i=1}^{n} x_i u_i$ as $\frac{1}{n}(X'u)$, so we have

$$\left(\frac{X'u}{n}\right) \xrightarrow{\text{P}} 0$$

Similarly, with iid data, the Law of Large Numbers implies that the $K \times K$ matrix of sample means

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{\text{P}} \text{E}(x_i x_i') = M_{XX}$$

from assumption (iii), with M_{XX} non-singular so that M_{XX}^{-1} exists

We can also write the $K \times K$ matrix $\frac{1}{n}\sum_{i=1}^{n} x_i x_i'$ as $\frac{1}{n}(X'X)$, so we have

$$\left(\frac{X'X}{n}\right) \xrightarrow{\text{P}} M_{XX}$$

And using Slutsky's theorem, we have

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{\text{P}} M_{XX}^{-1}$$

Now write the OLS estimator as

$$\begin{aligned}
\widehat{\beta}_{\text{OLS}} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta + u) \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\
&= \beta + (X'X)^{-1}X'u
\end{aligned}$$

Equivalently

$$\begin{aligned}
\widehat{\beta}_{\text{OLS}} &= \beta + n(X'X)^{-1}\left(\frac{1}{n}\right)X'u \\
&= \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{n}\right)
\end{aligned}$$

$$\widehat{\beta}_{\text{OLS}} = \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{n}\right)$$

Now using Slutsky's theorem we have

$$\text{p}\lim_{n\to\infty}\widehat{\beta}_{\text{OLS}} = \text{p}\lim_{n\to\infty}\beta + \left(\text{p}\lim_{n\to\infty}\left(\frac{X'X}{n}\right)\right)^{-1}\left(\text{p}\lim_{n\to\infty}\left(\frac{X'u}{n}\right)\right)$$

The consistency property $\text{p}\lim_{n\to\infty}\widehat{\beta}_{\text{OLS}} = \beta$ thus follows from assumptions (ii) and (iii), which ensure that:

$\text{p}\lim_{n\to\infty}\left(\frac{X'u}{n}\right) = 0$

$\left(\text{p}\lim_{n\to\infty}\left(\frac{X'X}{n}\right)\right)^{-1} = M_{XX}^{-1}$ exists and is finite

Then we have the result that $\text{p}\lim_{n\to\infty}\widehat{\beta}_{\text{OLS}} = \beta$, or $\widehat{\beta}_{\text{OLS}} \xrightarrow{\text{P}} \beta$

The assumption that the data on (y_i, x_i') are independently and identically distributed simplifies the proof given here, but is stronger than needed to establish consistency of the OLS estimator

The key assumption needed is the orthogonality condition $\mathrm{E}(x_i u_i) = 0$

Data on (y_i, x_i') that are assumed to be independent over $i = 1, 2, ..., n$ but not identically distributed would require a Law of Large Numbers for that setting

[This case is covered in section 4.4 of Cameron and Trivedi (2005)]

For time series models with lagged dependent variables, the data on (y_i, x_i') are not independent over $i = 1, 2, ..., n$, since x_i includes y_{i-1}

This case would require a Law of Large Numbers for stationary, ergodic processes

[This case is covered in Sections 2.2 and 2.3 of Hayashi (2000)]

In contrast, we know that $\widehat{\beta}_{\text{OLS}}$ is *not* unbiased in dynamic models with lagged dependent variables, since the strict exogeneity assumption $\text{E}(y|X) = X\beta$ was needed to establish unbiasedness

When we state that $\widehat{\beta}_{\text{OLS}}$ is consistent, or $\widehat{\beta}_{\text{OLS}} \xrightarrow{\text{P}} \beta$

- we view the sequence of OLS estimators obtained from increasing sample sizes as a sequence of random vectors
- consistency holds for each element of the vector, so that we have $\widehat{\beta}_k \xrightarrow{\text{P}} \beta_k$ for each $k = 1, ..., K$

Notice that we do *not* need to assume either conditional homoskedasticity ($\text{Var}(u_i|x_i) = \sigma^2$) or normality ($u_i|x_i \sim$ normal) to establish consistency

Nor do we require that the conditional expectation of y_i given x_i is linear ($\text{E}(y_i|x_i) = x_i'\beta$), let alone the stronger form $\text{E}(y|X) = X\beta \leftrightarrow \text{E}(u|X) = 0$

The key assumption needed is the orthogonality condition $\mathrm{E}(x_i u_i) = 0$

We normally consider linear models in which either:

a) we have an intercept term, so that $x_{1i} = 1$ for all $i = 1, 2, ..., n$, in which case the first element of the $K \times 1$ vector $\mathrm{E}(x_i u_i) = 0$ implies that we have $\mathrm{E}(u_i) = 0$; or

b) we have $\mathrm{E}(y_i) = 0$ and $\mathrm{E}(x_i) = 0$ (for example, original variables may have been expressed as deviations from their means), which also implies that we have $\mathrm{E}(u_i) = 0$

Given that we have $\mathrm{E}(u_i) = 0$, the assumption that $\mathrm{E}(x_i u_i) = 0$ implies that $\mathrm{Cov}(x_i, u_i) = \mathrm{E}(x_i u_i) - \mathrm{E}(x_i)\mathrm{E}(u_i) = 0$, and hence that $\mathrm{Corr}(x_i, u_i) = 0$

Here we are assuming that we are interested in the linear relation between y_i and the explanatory variables in x_i with the property that the additive error term u_i is uncorrelated *in the population* with each of the included explanatory variables in x_i

The statements

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

$$\text{and} \quad \mathrm{E}(x_i u_i) = 0$$

define the **linear projection** of y_i on x_i to be $\mathrm{E}^*(y_i|x_i) = x_i'\beta$

With iid data on (y_i, x_i'), the OLS estimator is a consistent estimator for the parameters of this linear projection

Note that the linear conditional expectation assumption $\mathrm{E}(y_i|x_i) = x_i'\beta \leftrightarrow \mathrm{E}(u_i|x_i) = 0$ found in the linear regression model implies the orthogonality assumption $\mathrm{E}(x_i u_i) = 0$ found in the linear projection model

Using the Law of Iterated Expectations

$$\mathrm{E}(x_i u_i) = \mathrm{E}_x[\mathrm{E}_{u|x}(x_i u_i|x_i)] = \mathrm{E}_x[x_i \mathrm{E}_{u|x}(u_i|x_i)] = \mathrm{E}_x[0] = 0$$

if we have $\mathrm{E}(u_i|x_i) = 0$

With iid data on (y_i, x_i'), the OLS estimator is therefore a consistent estimator for the parameters of the linear regression model

In general, the converse is not true: $\mathrm{E}(x_i u_i) = 0$ does not imply $\mathrm{E}(u_i|x_i) = 0$

In this sense, the linear projection model makes a weaker assumption: we can specify $\mathrm{E}(x_i u_i) = 0$ without requiring the conditional expectation $\mathrm{E}(y_i|x_i)$ to be a linear function

Similarly the orthogonality property (for any bounded function $g(x_i)$)

$$\mathrm{E}\{[y_i - \mathrm{E}(y_i|x_i)]g(x_i)\} = \mathrm{E}\{[y_i - x_i'\beta]g(x_i)\} = 0$$

in the linear regression model implies the orthogonality condition (setting $g(x_i) = x_i$)

$$\mathrm{E}\{[y_i - x_i'\beta]x_i\} = \mathrm{E}(u_i x_i) = \mathrm{E}(x_i u_i) = 0$$

in the linear projection model, while the converse does not hold

$\mathrm{E}(u_i|x_i) = 0$ also implies that $\mathrm{E}(u_i) = 0$, since

$$\mathrm{E}(u_i) = \mathrm{E}_x[\mathrm{E}_{u|x}(u_i|x_i)] = \mathrm{E}_x[0] = 0$$

And $\mathrm{E}(u_i|x_i) = 0$ thus implies that $\mathrm{cov}(x_i, u_i) = 0$, since

$$\mathrm{cov}(x_i, u_i) = \mathrm{E}(x_i u_i) - \mathrm{E}(x_i)\mathrm{E}(u_i) = 0 - 0 = 0$$

So $\mathrm{E}(u_i|x_i) = 0$ implies $\mathrm{cov}(x_i, u_i) = 0$, and hence $\mathrm{corr}(x_i, u_i) = 0$

In general the converse is not true: $\mathrm{corr}(x_i, u_i) = 0$ does not imply that $\mathrm{E}(u_i|x_i) = 0$

[Although we know that if x_i and u_i are jointly normally, then zero covariance does imply independence, in which case we have $\mathrm{E}(u_i|x_i) = 0$]

The consistency property $\widehat{\beta}_{\mathrm{OLS}} \xrightarrow{\mathrm{P}} \beta$ is sometimes called 'weak consistency', to emphasise that we have only established convergence in probability (and not stronger concepts of convergence for sequences of random vectors)

Consistency is a large sample or asymptotic property, i.e. a property that holds in the limit as the sample size n increases to infinity

Unbiasedness is a small sample or finite sample property

An unbiased estimator (with $\mathrm{E}(\widehat{\beta}) = \beta$) will usually be consistent, but need not be

Consistency further requires that $\mathrm{Var}(\widehat{\beta})$ falls to zero as the sample size increases

A consistent estimator need not be unbiased

For example, if $\widehat{\beta}$ is an unbiased and consistent estimator of β, then $\widetilde{\beta} = \widehat{\beta} + \frac{1}{n}$ is a consistent but biased estimator of β

$$\mathrm{E}(\widetilde{\beta}) = \mathrm{E}(\widehat{\beta}) + \tfrac{1}{n} = \beta + \tfrac{1}{n} \neq \beta \leftrightarrow \text{biased}$$

$$\mathrm{p}\lim_{n\to\infty}(\widetilde{\beta}) = \mathrm{p}\lim_{n\to\infty}(\widehat{\beta}) + \mathrm{p}\lim_{n\to\infty}(\tfrac{1}{n}) = \beta + 0 = \beta \leftrightarrow \text{consistent}$$

Asymptotic normality

Consistency is a desirable property for an estimator to have, particularly if we have the luxury of working with large datasets

But consistency is not sufficient for (asymptotic) inference

To obtain (approximate) confidence intervals and to conduct (asymptotically valid) hypothesis tests, we require an asymptotic distribution result, giving the limit distribution for a suitably scaled transformation of the estimator

We start by considering a sufficient set of assumptions for the OLS estimator, suitably scaled, to have a limit distribution that is normal

Assumptions

i) $y_i = x_i'\beta + u_i$ for $i = 1, ..., n$ or $y = X\beta + u$

ii) The data on (y_i, x_i') are independently and identically distributed,

with $\mathrm{E}(x_i u_i) = 0$ for all $i = 1, ..., n$

iii) The $K \times K$ matrix $M_{XX} = \mathrm{E}(x_i x_i')$ exists and is non-singular

iv) The $K \times K$ matrix $M_{X\Omega X} = \mathrm{E}(u_i^2 x_i x_i')$ exists and is non-singular

Then $\sqrt{n}(\widehat{\beta}_{\mathrm{OLS}} - \beta) \xrightarrow{\mathrm{D}} \mathrm{N}(0, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1})$

Recall that a sequence of random variables $\{x_n : n = 1, 2, ...\}$ **converges in distribution** to the random variable x if

$$F_{x_n} \to F_x \text{ as } n \to \infty \quad \text{or} \quad \lim_{n\to\infty} F_{x_n} = F_x$$

at all values of a where $F_x(a)$ is continuous, where F_{x_n} is the distribution function of x_n and F_x is the distribution function of x

If $x \sim \mathrm{N}(\mu, \sigma^2)$, we write $x_n \xrightarrow{\mathrm{D}} \mathrm{N}(\mu, \sigma^2)$

This definition extends to random vectors, with F_{x_n} and F_x interpreted as the distribution functions of the random vectors x_n and x

If the vector $x \sim \mathrm{N}(\mu, \Sigma)$, we write $x_n \xrightarrow{\mathrm{D}} \mathrm{N}(\mu, \Sigma)$

The distribution to which the random vector x_n converges in the limit as $n \to \infty$ is referred to as the **limit distribution** or the limiting distribution

An important property of limit distributions is known as the **continuous mapping theorem**, stated here for random vectors

Let the random vector $x_n \xrightarrow{\text{D}} x$, where x is a random vector, and let $g(a)$ be a continuous function

Then $g(x_n) \xrightarrow{\text{D}} g(x)$

Again if $x_n = (x_{1n}, x_{2n}) \xrightarrow{\text{D}} (x_1, x_2)$ and $g(a_1, a_2) = a_1 a_2$, we have that $g(x_{1n}, x_{2n}) = x_{1n} x_{2n} \xrightarrow{\text{D}} x_1 x_2$

Another useful result is the **generalised Slutsky's theorem**, stated here for random vectors (or matrices)

Let the random vector $x_n \xrightarrow{\text{D}} x$, where x is a random vector, and the random vector (or matrix) $y_n \xrightarrow{\text{P}} c$, where c is non-stochastic, and let $g(a, b)$ be a function that is continuous at (a, c) for all values of a

Then $g(x_n, y) \xrightarrow{\text{D}} g(x, c)$

Implications:

If x_n is a $K \times 1$ random vector and y_n is a $K \times 1$ random vector, with $x_n \xrightarrow{\text{D}} x$ and $y_n \xrightarrow{\text{P}} c$, we have

i) $x_n + y_n \xrightarrow{\text{D}} x + c$

ii) $y_n' x_n \xrightarrow{\text{D}} c'x$

If x_n is a $K \times 1$ random vector and A_n is a $K \times K$ random matrix, with $x_n \xrightarrow{\text{D}} x$ and $A_n \xrightarrow{\text{P}} A$, we also have

iii) $A_n x_n \xrightarrow{\text{D}} Ax$

iv) $x_n' A_n^{-1} x_n \xrightarrow{\text{D}} x' A^{-1} x$ provided A^{-1} exists

The matrix version of the product rule $A_n x_n \xrightarrow{\text{D}} Ax$ implies that if $x_n \xrightarrow{\text{D}} \text{N}(0, \Sigma)$ and $A_n \xrightarrow{\text{P}} A$ then $A_n x_n \xrightarrow{\text{D}} \text{N}(0, A\Sigma A')$

We use this result to show the asymptotic normality of $\widehat{\beta}_{\text{OLS}}$ under the stated assumptions

We also require a **Central Limit Theorem** for random vectors

Suppose that $z_1, z_2, ..., z_n$ are independently and identically distributed random vectors, with $\text{E}(z_i) = \mu$ and $\text{Var}(z_i) = \Sigma$ finite

Let $\overline{z}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$ denote the vector of sample means for a sample of size n, and let $y_n = \sqrt{n}(\overline{z}_n - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_i - \mu)$

Then $y_n \xrightarrow{\text{D}} \text{N}(0, \Sigma)$

Now consider the $K \times 1$ random vectors $(x_i u_i)$ for $i = 1, 2, ..., n$

These are independently and identically distributed from assumption (ii)

We have $\mathrm{E}(x_i u_i) = 0$ directly from assumption (ii)

From assumption (iv), we obtain

$$\mathrm{Var}(x_i u_i) = \mathrm{E}[(x_i u_i)(x_i u_i)'] = \mathrm{E}(u_i^2 x_i x_i') = M_{X\Omega X} \text{ finite}$$

Using the Central Limit Theorem, for the $K \times 1$ vector of sample means $\frac{1}{n}\sum_{i=1}^{n} x_i u_i$, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i u_i\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \xrightarrow{\text{D}} \text{N}(0, M_{X\Omega X})$$

We can also write the $K \times 1$ vector $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i$ as $\frac{1}{\sqrt{n}}(X'u)$, so we have

$$\left(\frac{X'u}{\sqrt{n}}\right) \xrightarrow{\text{D}} \text{N}(0, M_{X\Omega X})$$

Previously we showed that assumption (iii) implies that

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{\text{P}} M_{XX}^{-1}$$

As before, we write the OLS estimator in the form

$$\widehat{\beta}_{\text{OLS}} = \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{n}\right)$$

or equivalently

$$\widehat{\beta}_{\text{OLS}} - \beta = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{n}\right)$$

Scaled by $\sqrt{n}$, we have

$$\sqrt{n}\left(\widehat{\beta}_{\text{OLS}} - \beta\right) = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{\sqrt{n}}\right)$$

$$\sqrt{n}\left(\widehat{\beta}_{\text{OLS}}-\beta\right)=\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{\sqrt{n}}\right)$$

Now the $K\times K$ matrix $\left(\frac{X'X}{n}\right)^{-1}\xrightarrow{\text{P}}M_{XX}^{-1}$

And the $K\times 1$ vector $\left(\frac{X'u}{\sqrt{n}}\right)\xrightarrow{\text{D}}\text{N}(0,M_{X\Omega X})$

Using the generalised Slutsky's theorem, and the symmetry of M_{XX}^{-1}, we obtain the limit distribution of the product

$$\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{\sqrt{n}}\right)\xrightarrow{\text{D}}\text{N}(0,M_{XX}^{-1}M_{X\Omega X}M_{XX}^{-1})$$

Then we have the result that

$$\sqrt{n}\left(\widehat{\beta}_{\text{OLS}}-\beta\right)\xrightarrow{\text{D}}\text{N}(0,M_{XX}^{-1}M_{X\Omega X}M_{XX}^{-1})$$

An estimator $\widehat{\theta}$ of the parameter vector θ with the property that

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) \xrightarrow{\mathrm{D}} \mathrm{N}(0,V)$$

for some finite variance matrix V is said to be **asymptotically normal**

Appropriately scaled, the estimator has a non-degenerate limit distribution which is normal

The variance matrix V in this limit distribution is sometimes called the asymptotic variance matrix of the estimator $\widehat{\theta}$, and sometimes denoted by $\mathrm{avar}(\widehat{\theta})$ or $\mathrm{Avar}(\widehat{\theta})$ (for example, Hayashi (2000))

NB. This terminology is not universal; some authors use the term asymptotic variance in a different sense; see Hayashi (2000), p95

As was the case for our consistency result, the assumption that the data on (y_i, x_i') are independently and identically distributed simplifies the proof of asymptotic normality given here, but is stronger than needed to establish asymptotic normality for the OLS estimator

Data on (y_i, x_i') that are assumed to be independent over $i = 1, 2, ..., n$ but not identically distributed would require a Central Limit Theorem for that setting

[This case is covered in section 4.4 of Cameron and Trivedi (2005)]

For time series models with lagged dependent variables, we cannot assume independent observations

This case would require a Central Limit Theorem for stationary, ergodic Martingale difference sequences

[This case is covered in sections 2.2 and 2.3 of Hayashi (2000)]

Notice that we do *not* need to assume conditional homoskedasticity ($\text{Var}(u_i|x_i) = \sigma^2$) to establish asymptotic normality of the OLS estimator

Nor do we require that the conditional expectation of y_i given x_i is linear ($\text{E}(y_i|x_i) = x_i'\beta$), or the stronger form $\text{E}(y|X) = X\beta$

Nor do we require that the conditional distribution $y|X \sim \text{N}(X\beta, \sigma^2 I)$

However, to explore the implications of our asymptotic normality result, we first consider the *special case* of a classical linear regression model in which we have the linear conditional expectation assumption $\mathrm{E}(u_i|x_i) = 0$, and the conditional homoskedasticity assumption $\mathrm{Var}(u_i|x_i) = \mathrm{E}(u_i^2|x_i) = \sigma^2$

We assume that the data on (y_i, x_i') are independently and identically distributed, and that $M_{XX} = \mathrm{E}(x_i x_i')$ exists and is non-singular

We do *not* assume that the conditional distribution $u_i|x_i \sim \mathrm{N}(0, \sigma^2)$

With $\mathrm{E}(u_i^2|x_i) = \sigma^2$ for $i = 1, 2, ..., n$, we can derive

$$\begin{aligned} M_{X\Omega X} &= \mathrm{E}(u_i^2 x_i x_i') = \mathrm{E}_x[\mathrm{E}_{u|x}(u_i^2 x_i x_i'|x_i)] \\ &= \mathrm{E}_x[x_i x_i' \mathrm{E}_{u|x}(u_i^2|x_i)] = \mathrm{E}_x[x_i x_i' \sigma^2] \\ &= \sigma^2 \mathrm{E}(x_i x_i') = \sigma^2 M_{XX} \end{aligned}$$

Our limit distribution result

$$\sqrt{n}\left(\widehat{\beta}_{\mathrm{OLS}} - \beta\right) \xrightarrow{\mathrm{D}} \mathrm{N}(0, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1})$$

then simplifies in this special case to give

$$\sqrt{n}\left(\widehat{\beta}_{\mathrm{OLS}} - \beta\right) \xrightarrow{\mathrm{D}} \mathrm{N}(0, \sigma^2 M_{XX}^{-1})$$

Asymptotic inference

Consistency $\widehat{\beta}_{\text{OLS}} \xrightarrow{\text{P}} \beta$

The parameter vector β is not stochastic, so the limit distribution of $\widehat{\beta}_{\text{OLS}}$ itself is degenerate

In practice, we do not get to estimate parameter vectors using infinitely large samples

To form confidence intervals and to test hypotheses, we need to approximate the distribution of the estimator

Asymptotic distribution results allow us to approximate this distribution accurately in large finite samples

Asymptotic normality $\sqrt{n}(\widehat{\beta}_{\text{OLS}} - \beta) \xrightarrow{\text{D}} \text{N}(0, V)$

where $V = \sigma^2 M_{XX}^{-1}$ in the classical linear regression case

Appropriately scaled, we have a non-degenerate limit distribution

To use this limit distribution to approximate the distribution of $\widehat{\beta}_{\text{OLS}}$ in large finite samples, we multiply by the scalar $\frac{1}{\sqrt{n}}$ and add the non-stochastic vector β to both sides, giving

$$\widehat{\beta}_{\text{OLS}} \overset{\text{a}}{\sim} \text{N}\left(\beta, V/n\right)$$

This gives us the approximation to the distribution of the OLS estimator based on our asymptotic distribution theory, sometimes called the 'asymptotic distribution' of $\widehat{\beta}_{\text{OLS}}$

$$\widehat{\beta}_{\mathrm{OLS}} \overset{\mathrm{a}}{\sim} \mathrm{N}\left(\beta, V/n\right)$$

The variance matrix V/n in this approximate (asymptotic) distribution is *also* sometimes called the asymptotic variance matrix of the estimator $\widehat{\beta}_{\mathrm{OLS}}$, and *also* sometimes denoted by $\mathrm{avar}(\widehat{\beta}_{\mathrm{OLS}})$ or $\mathrm{Avar}(\widehat{\beta}_{\mathrm{OLS}})$ (for example, Cameron and Trivedi (2005))

NB. As previously noted, this terminology and notation is not universal (see Hayashi (2000), p95); to avoid (further) confusion, we will refrain from using this notation at all

To make use of this approximation, we require a consistent estimator of the $K \times K$ variance matrix V

In the classical linear regression case, we can easily obtain a consistent estimator of $V = \sigma^2 M_{XX}^{-1}$

Our earlier result that

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{\text{P}} M_{XX}^{-1}$$

gives us a consistent estimator of the $K \times K$ matrix M_{XX}^{-1}

Both

$$\widehat{\sigma}^2_{\text{OLS}} = \frac{1}{n-K}\sum_{i=1}^{n} \widehat{u}_i^2 = \frac{\widehat{u}'\widehat{u}}{n-K}$$

and

$$\widehat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{\widehat{u}'\widehat{u}}{n}$$

are consistent estimators of the scalar σ^2 in this setting (although only $\widehat{\sigma}^2_{\mathrm{OLS}}$ is unbiased)

Using Slutsky's theorem, the estimator

$$\widehat{V} = \widehat{\sigma}^2_{\mathrm{OLS}}\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{\mathrm{P}} V$$

Since $\widehat{V} \xrightarrow{\mathrm{P}} V$, the difference between V and $\widehat{V}$ becomes negligible in the limit as $n \to \infty$, and we can replace the unknown V by the estimator $\widehat{V}$ without changing the form of our limit distribution result

This gives the approximation to the distribution of $\widehat{\beta}_{\text{OLS}}$ as

$$\widehat{\beta}_{\text{OLS}} \overset{\text{a}}{\sim} \text{N}\left(\beta, \widehat{V}/n\right)$$

which is the form used to construct (approximate) confidence intervals and (asymptotically valid) test statistics

Now consider

$$\begin{aligned}
\widehat{V}/n &= \left(\frac{1}{n}\right) \widehat{\sigma}^2_{\text{OLS}} \left(\frac{X'X}{n}\right)^{-1} \\
&= \left(\frac{1}{n}\right) \widehat{\sigma}^2_{\text{OLS}} n (X'X)^{-1} \\
&= \widehat{\sigma}^2_{\text{OLS}} (X'X)^{-1}
\end{aligned}$$

This expression coincides with the exact finite sample conditional variance $\text{Var}(\widehat{\beta}_{\text{OLS}}|X)$ that we obtained in the classical linear regression model with normally distributed errors (i.e. with the additional assumption that $u_i|x_i \sim \text{N}(0, \sigma^2)$)

And the approximate distribution

$$\widehat{\beta}_{\text{OLS}} \overset{\text{a}}{\sim} \text{N}\left(\beta, \widehat{V}/n\right) \quad \text{or} \quad \widehat{\beta}_{\text{OLS}} \overset{\text{a}}{\sim} \text{N}\left(\beta, \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1}\right)$$

coincides with the exact finite sample conditional distribution of $\widehat{\beta}_{\text{OLS}}|X$ in the classical linear regression model with normally distributed errors

There is a special case of the classical linear regression model for which we know that this approximation will be accurate

In large finite samples, the asymptotic distribution result provides a rationale for using this expression to estimate the variance of $\widehat{\beta}_{\text{OLS}}$, and for treating $\widehat{\beta}_{\text{OLS}}$ as a random vector with a normal distribution, more generally, in linear models where we maintain the conditional homoskedasticity assumption $\text{E}(u_i^2|x_i) = \sigma^2$, even if

- we have no reason to assume that the error terms u_i are normally distributed, conditional on x_i
- or to maintain the linear conditional expectation (strict exogeneity) assumption of the classical linear regression model

To interpret this estimator for the variance of $\widehat{\beta}_{\text{OLS}}$ (under conditional homoskedasticity), $\widehat{V}/n = \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1}$, it may be helpful to consider the **special case** with 1 explanatory variable (denoted x_i) and no intercept

$$y_i = x_i\beta + u_i$$

where we assume that both y_i and x_i have zero mean, perhaps because original variables have been transformed into deviations from their sample means

In this case $X'X$ is simply $\sum_{i=1}^{n} x_i^2$

Since all the x_i^2 terms are non-negative, $X'X = \sum_{i=1}^{n} x_i^2$ increases as we add more observations to the sample

$\widehat{\sigma}^2_{\text{OLS}} \xrightarrow{\text{P}} \sigma^2$, so $\widehat{\sigma}^2_{\text{OLS}}$ does not increase with the sample size n

In this example, it is clear that

$$\widehat{V}/n = \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1} = \frac{\widehat{\sigma}^2_{\text{OLS}}}{\sum_{i=1}^{n} x_i^2}$$

decreases with the sample size n

Our estimate of the variance of $\widehat{\beta}_{\text{OLS}}$ will tend to fall as we add more observations to the sample

The square root of this (scalar) estimated variance, sometimes termed the asymptotic standard error, also falls as n increases, indicating that we can estimate the (scalar) parameter β more precisely as we add more observations to the sample

In the limit as $n \to \infty$, we have

$$\widehat{V}/n = \frac{\widehat{\sigma}^2_{\text{OLS}}}{\sum_{i=1}^{n} x_i^2} \to 0$$

consistent with our result that the limit distribution of $\widehat{\beta}_{\text{OLS}}$ is degenerate

The observation that $X'X = \sum_{i=1}^{n} x_i^2$ increases with n is perfectly consistent with our result that $\left(\frac{X'X}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \xrightarrow{\text{P}} \text{E}(x_i^2) = M_{XX} < \infty$

In the scalar case, with x_i having zero mean, $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ is an estimator of the variance of the random variable x_i

With iid observations, $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ is a consistent estimator of the variance

In this case, our assumption simply requires that the variance of x_i exists, and is finite

With cross-section data, this assumption is usually innocuous

With stationary time series data, this assumption is also satisfied

Although this suggests that a different approach will be required for non-stationary time series (e.g. random walks, unit roots, ...)

Now let the scalar $\widehat{v}_{kk}$ denote the k^{th} element on the main diagonal of the $K \times K$ variance matrix estimator $\widehat{V}/n = \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1}$

Since

$$\widehat{\beta}_k \overset{\text{a}}{\sim} \text{N}(\beta_k, \widehat{v}_{kk})$$

we also have

$$t_k = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{se_k} \overset{\text{a}}{\sim} \text{N}(0,1)$$

Simple hypothesis tests of the form $H_0 : \beta_k = \beta_k^0$ can then be conducted in large finite samples by comparing the test statistic $t_k = \frac{\widehat{\beta}_k - \beta_k^0}{se_k}$ computed under the null hypothesis to critical values of the standard normal $\text{N}(0,1)$ distribution at the desired significance level

For example, to test the null $H_0 : \beta_k = \beta_k^0$ against the two-sided alternative $H_1 : \beta_k \neq \beta_k^0$ at (approximately) the 5% level of significance, we can compare the absolute value of $t_k = \frac{\widehat{\beta}_k - \beta_k^0}{se_k}$ that we obtain to the critical value 1.96

In this context, the standard error $se_k = \sqrt{\widehat{v}_{kk}}$ is sometimes referred to as the asymptotic standard error (sometimes denoted ase_k)

Similarly the test is sometimes referred to as an asymptotic t-test, to emphasise that its validity is based on an asymptotic approximation to the distribution of the test statistic

Why can we treat t_k here as a draw from the standard normal distribution, even though we have estimated the unknown σ^2?

The limit distribution result

$$\sqrt{n}(\widehat{\beta}_{\text{OLS}} - \beta) \xrightarrow{\text{D}} \text{N}(0, \sigma^2 M_{XX}^{-1})$$

implies that

$$z_k = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} \xrightarrow{\text{D}} \text{N}(0, 1)$$

where v_{kk} is the (unknown) k^{th} element on the main diagonal of $\left(\frac{1}{n}\right) \sigma^2 M_{XX}^{-1}$

Using $se_k = \sqrt{\widehat{v}_{kk}}$ to estimate the unknown $\sqrt{v_{kk}}$ gives

$$t_k = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} = \left(\frac{\sqrt{v_{kk}}}{\sqrt{\widehat{v}_{kk}}}\right) \left(\frac{\widehat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) = \left(\frac{\sqrt{v_{kk}}}{\sqrt{\widehat{v}_{kk}}}\right) z_k$$

Since $\sqrt{\widehat{v}_{kk}}$ is a consistent estimator of $\sqrt{v_{kk}}$, we have $\sqrt{\widehat{v}_{kk}} \xrightarrow{\mathrm{P}} \sqrt{v_{kk}}$, or

$$\left(\frac{\sqrt{v_{kk}}}{\sqrt{\widehat{v}_{kk}}}\right) \xrightarrow{\mathrm{P}} 1$$

Then t_k has the same limit distribution as z_k, so that we also have

$$t_k = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\widehat{v}_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{se_k} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$$

However, since this test is only justified by an asymptotic approximation, and since we know that a $t(n-K)$ random variable converges in distribution to a $\mathrm{N}(0,1)$ random variable as $n \to \infty$ with K fixed, there is *also* an asymptotic justification for comparing the test statistic $t_k = \frac{\widehat{\beta}_k - \beta_k^0}{se_k}$ computed under the null hypothesis to critical values of the $t(n-K)$ distribution

Some econometric packages use the $t(n-K)$ approximation to the distribution of $t_k = \frac{\widehat{\beta}_k - \beta_k^0}{se_k}$ for reporting p-values and confidence intervals

You may notice that Stata's regress command does this

In general, the exact distribution of the test statistic is unknown, and depends on features of the data generating process

Except in the classical linear regression model with normally distributed errors, both the $\mathrm{N}(0,1)$ approximation and the $t(n-K)$ approximation only have an asymptotic justification, and they coincide in the limit as $n \to \infty$ for given K

In large samples, this difference becomes negligibly small

Testing linear restrictions

Using the approximation to the distribution of $\widehat{\beta}_{\mathrm{OLS}}$

$$\widehat{\beta}_{\mathrm{OLS}} \overset{\mathrm{a}}{\sim} \mathrm{N}\left(\beta, \widehat{V}/n\right)$$

the approximate (asymptotic) distribution of the $p \times 1$ random vector $\widehat{\theta}_{\mathrm{OLS}} = H\widehat{\beta}_{\mathrm{OLS}}$, where H is a non-stochastic $p \times K$ matrix with $\mathrm{rank}(H) = p$ (full row rank), is obtained as

$$\widehat{\theta}_{\mathrm{OLS}} \overset{\mathrm{a}}{\sim} \mathrm{N}(H\beta, H(\widehat{V}/n)H')$$

We also have $\mathrm{p}\lim_{n\to\infty} \widehat{\theta}_{\mathrm{OLS}} = H\,\mathrm{p}\lim_{n\to\infty} \widehat{\beta}_{\mathrm{OLS}} = H\beta = \theta$, or $\widehat{\theta}_{\mathrm{OLS}} \xrightarrow{\mathrm{P}} \theta$

The approximate (asymptotic) distribution of the scalar random variable w given by the quadratic form

$$w = (\widehat{\theta}_{\text{OLS}} - H\beta)' \left[H(\widehat{V}/n)H' \right]^{-1} (\widehat{\theta}_{\text{OLS}} - H\beta)$$

is obtained as

$$w \overset{\text{a}}{\sim} \chi^2(p)$$

In the special case of the classical linear regression model, we have $\widehat{V}/n = \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1}$ and $H(\widehat{V}/n)H' = \widehat{\sigma}^2_{\text{OLS}} H(X'X)^{-1}H'$

In this case we have

$$w = (\widehat{\theta}_{\text{OLS}} - H\beta)' \left[\widehat{\sigma}^2_{\text{OLS}} H(X'X)^{-1}H' \right]^{-1} (\widehat{\theta}_{\text{OLS}} - H\beta) \overset{\text{a}}{\sim} \chi^2(p)$$

Joint hypothesis tests of the form $H_0 : H\beta = \theta^0$ can then be conducted by computing the test statistic

$$w = (\widehat{\theta}_{\text{OLS}} - \theta^0)' \left[\widehat{\sigma}^2_{\text{OLS}} H(X'X)^{-1}H'\right]^{-1} (\widehat{\theta}_{\text{OLS}} - \theta^0)$$

which has a distribution that is well approximated in large finite samples by the $\chi^2(p)$ distribution, if the null hypothesis is true

We can compare the value of the test statistic we obtain to critical values of the $\chi^2(p)$ distribution at the desired level of significance

For example, if $p = 2$ and we want a test at (approximately) the 5% significance level, we can compare the computed value of the test statistic w to 5.99, the upper 95% percentile of the $\chi^2(2)$ distribution

Alternatively, since this test only has an asymptotic justification, and we know that p times a $F(p, n-K)$ random variable converges in distribution to a $\chi^2(p)$ random variable as $n \to \infty$ with K fixed, there is *also* an asymptotic justification for comparing the test statistic w/p computed under the null hypothesis to critical values of the $F(p, n-K)$ distribution

Some econometric packages use the $F(p, n-K)$ approximation to the distribution of w/p for reporting p-values

You may notice that Stata does this

In large samples, this difference also becomes negligibly small

The test statistic w is called an (asymptotic) Wald statistic, and the corresponding test is called an (asymptotic) Wald test

The Wald test is one of three main approaches used for testing restrictions on the parameter vector in econometric models

The other two approaches

- Lagrange multiplier or score tests
- criterion difference or likelihood ratio tests

will be covered later in the course

The (asymptotic) Wald test can also be used to test non-linear restrictions (for example, $H_0 : \beta_2\beta_3 = 0$), using an approximation called the delta method

This will also be covered later in the course