Asymptotic properties

We now look at properties of the OLS estimator that hold in large samples, or more precisely, 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 data 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 much **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 small sample distributions of the OLS estimator and the t- and F-statistics in the classical linear regression model with Normally distributed errors.
(Weak) 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 \( i = 1, \ldots, N \) or \( y = X\beta + u \)

ii) The data on \((y_i, x_i)\) are independent over \( i = 1, \ldots, N \),

with \( E(u_i) = 0 \) and \( E(x_iu_i) = 0 \) for all \( i = 1, \ldots, N \)

iii) \( X \) is stochastic and full rank

iv) The \( K \times K \) matrix \( M_{XX} = p \lim_{N \to \infty} \left( \frac{X'X}{N} \right) = p \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x_i x_i' \)

exists and is non-singular

Then \( p \lim_{N \to \infty} \hat{\beta}_{OLS} = \beta \), or \( \hat{\beta}_{OLS} \xrightarrow{P} \beta \)
Recall that a sequence of random variables \( \{x_N : N = 1, 2, \ldots \} \) **converges in probability** to the constant \( a \) if, for all \( \delta > 0 \) and \( \varepsilon > 0 \), we have

\[
\Pr(|x_N - a| < \varepsilon) > 1 - \delta \quad \text{for all } N > N^*
\]

or equivalently

\[
\lim_{N \to \infty} \Pr(|x_N - a| < \varepsilon) = 1
\]

\[
\lim_{N \to \infty} \Pr(|x_N - a| \geq \varepsilon) = 0
\]

This is denoted by \( p \lim_{N \to \infty} x_N = a \), or \( x_N \xrightarrow{P} a \).
Convergence in probability (associated with weak laws of large numbers) is implied by the property of almost sure convergence (associated with strong laws of large numbers)

A random variable $x_N$ **converges almost surely** to $a$ if

$$\Pr(\lim_{N \to \infty} x_N = a) = 1$$

This is denoted by $x_N \xrightarrow{a.s.} a$
These definitions are applied element-by-element to random vectors or matrices

For example, if we have \( x_N \overset{P}{\to} a \) and \( y_N \overset{P}{\to} b \), the random vector \((x_N, y_N) \overset{P}{\to} (a, b)\)

An important property of probability limits is known as Slutsky’s Theorem

Let the random vector \( x_N \) have a probability limit given by the vector \( a \), and let \( g(.) \) be a real-valued function which is continuous at the point \( a \)

Then \( g(x_N) \overset{P}{\to} g(a) \)
For example, let $x_N = (x_{1N}, x_{2N}) \xrightarrow{P} (a_1, a_2)$ and let $g(x_1, x_2) = x_1 x_2$

Then we have $g(x_{1N}, x_{2N}) = x_{1N} x_{2N} \xrightarrow{P} a_1 a_2$,

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

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 $\hat{\beta}_{OLS}$ under the stated assumptions
\begin{align*}
\hat{\beta}_{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{align*}

Equivalently
\begin{align*}
\hat{\beta}_{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{align*}

Now
\begin{align*}
p \lim_{N \to \infty} \hat{\beta}_{OLS} &= p \lim_{N \to \infty} \beta + \left(p \lim_{N \to \infty} \left(\frac{X'X}{N}\right)\right)^{-1}\left(p \lim_{N \to \infty} \left(\frac{X'u}{N}\right)\right) \\
&= \beta + M_{XX}^{-1}p \lim_{N \to \infty} \left(\frac{X'u}{N}\right)
\end{align*}
\[ \lim_{N \to \infty} \hat{\beta}_{OLS} = \beta + M_{XX}^{-1} \lim_{N \to \infty} \left( \frac{X'u}{N} \right) \]

Consistency thus requires that

\[ \lim_{N \to \infty} \left( \frac{X'u}{N} \right) = \lim_{N \to \infty} \left( \frac{1}{N} \sum_{i=1}^{N} x_iu_i \right) = 0 \]

This in turn is implied by our assumption that \( E(x_iu_i) = 0 \)

Using a suitable Law of Large Numbers for independent observations over \( i = 1, ..., N \), we have that

\[ \lim_{N \to \infty} \left( \frac{1}{N} \sum_{i=1}^{N} x_iu_i \right) = E(x_iu_i) = 0 \]
For independent and identically distributed (iid) observations, we can use the Kolmogorov Law of Large Numbers.

For independent but not identically distributed observations, we can use the Markov Law of Large Numbers.

For time series models with lagged dependent variables, the observations \((y_i, x_i)\) are not independent over \(i = 1, \ldots, N\), since \(x_i\) includes \(y_{i-1}\), violating assumption (ii).

In this case, we can still use Laws of Large Numbers for stationary, ergodic processes to establish conditions under which \(\hat{\beta}_{OLS}\) remains consistent.
[Sections 2.2 and 2.3 of Hayashi (2000) provide the details]

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

When we state that \( \hat{\beta}_{OLS} \) is consistent, or \( \hat{\beta}_{OLS} \xrightarrow{P} \beta \),

\[ \bullet \] we view the sequence of OLS estimates from increasing sample sizes as a sequence of random vectors.

\[ \bullet \] consistency holds for each element of the vector, so that \( \hat{\beta}_k \xrightarrow{P} \beta_k \) for \( k = 1, \ldots, K \).
Notice that we do not need to assume either conditional homoskedasticity 
\( V(u_i|x_i) = \sigma^2 \) or Normality \( u_i|x_i \) has a Normal distribution \) to establish 
consistency of the OLS estimator.

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

The key assumption for the OLS estimator to be consistent is that the 
error term \( u_i \) is uncorrelated with (orthogonal to) each of the 
explanatory variables in the vector \( x_i \), as implied by the statement 
\( E(x_iu_i) = 0 \) (here a \( K \times 1 \) vector with each element equal to zero).
Note that the conditional independence assumption \( E(u_i|x_i) = 0 \) implies the orthogonality assumption \( E(x_iu_i) = 0 \)

Using the Law of Iterated Expectations

\[
E(x_iu_i) = E_x[E_{u|x}(x_iu_i|x_i)] = E_x[x_iE_{u|x}(u_i|x_i)] = E_x[0] = 0
\]

if \( E(u_i|x_i) = 0 \)

\( E(u_i|x_i) = 0 \) also implies that \( E(u_i) = 0 \), since

\[
E(u_i) = E_x[E_{u|x}(u_i|x_i)] = E_x[0] = 0
\]
And $E(u_i|x_i) = 0$ implies that $cov(x_i, u_i) = 0$, since

$$cov(x_i, u_i) = E(x_iu_i) - E(x_i)E(u_i) = 0 - 0 = 0$$

So we have that $E(u_i|x_i) = 0$ implies $cov(x_i, u_i) = 0$

In general the converse is not true.

Although we know that if $x_i$ and $u_i$ are both Normally distributed, then zero covariance does imply independence, in which case we have $E(u_i|x_i) = 0$
The statements

\[ y_i = x_i' \beta + u_i \]

and

\[ E(u_i) = 0 \quad \text{and} \quad E(x_i u_i) = 0 \]

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

With independent observations on \((y_i, x_i)\) over \(i = 1, ..., N\), the OLS estimator provides a consistent estimator for the parameters of this linear projection.
More accurately, this property should be called ‘weak consistency’, since we have only established convergence in probability

But this is commonly abbreviated to the term ‘consistency’

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 $E(\hat{\beta}) = \beta$) will usually be consistent, but need not be consistent.

Consistency further requires that $V(\hat{\beta})$ falls as the sample size increases.

A consistent estimator need not be unbiased.

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

$$E(\widetilde{\beta}) = E(\hat{\beta}) + \frac{1}{N} = \beta + \frac{1}{N} \neq \beta \leftrightarrow \text{biased}$$

$$p \lim_{N \to \infty}(\widetilde{\beta}) = p \lim_{N \to \infty}(\hat{\beta}) + p \lim_{N \to \infty}(\frac{1}{N}) = \beta + 0 = \beta \leftrightarrow \text{consistent}$$
Asymptotic Normality

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

But this is not sufficient for inference.

To obtain confidence intervals and to conduct hypothesis tests, we require an asymptotic distribution result, relating the distribution of the estimator in large samples to a known (and ideally a convenient) distribution.

We start by considering a sufficient set of assumptions for the OLS estimator, suitably scaled, to have a limit distribution that is Normal.
To link this asymptotic distribution result to the finite sample distribution result obtained previously for the classical Normal linear regression model, we first assume that the conditional variance of the error term \( E(u_i^2|x_i) \) satisfies the conditional homoskedasticity assumption \( E(u_i^2|x_i) = \sigma^2 \) for all \( i = 1, ..., N \).

This assumption is needed for the OLS estimator to have the particular asymptotic variance derived here, although not to establish that the asymptotic distribution is Normal.
Assumptions

i) \( y_i = x_i' \beta + u_i \) for \( i = 1, \ldots, N \) \ or \ \( y = X \beta + u \)

ii) The data on \((y_i, x_i)\) are independent over \( i = 1, \ldots, N \),

with \( E(u_i) = 0 \) and \( E(x_iu_i) = 0 \) and \( E(u_i^2| x_i) = \sigma^2 \) for all \( i = 1, \ldots, N \)

iii) \( X \) is stochastic and full rank

iv) The \( K \times K \) matrix \( M_{XX} = p \lim_{N \to \infty} \left( \frac{X'X}{N} \right) = p \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x_i x_i' \)

exists and is non-singular

v) The \( K \times 1 \) vector \( \left( \frac{X'u}{\sqrt{N}} \right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i u_i \overset{D}{\to} N(0, \sigma^2 M_{XX}) \)

Then \( \sqrt{N}(\hat{\beta}_{OLS} - \beta) \overset{D}{\to} N(0, \sigma^2 M_{XX}^{-1}) \)
Recall that a sequence of random variables \( \{x_N : N = 1, 2, \ldots\} \) converges \textbf{in distribution} to the random variable \( x \) if and only if

\[
F_N \to F \quad \text{as} \quad N \to \infty \quad \text{or} \quad \lim_{N \to \infty} F_N = F
\]

where \( F_N \) is the cumulative distribution function of \( x_N \) and \( F \) is the cumulative distribution function of \( x \)

If \( x \sim N(\mu, \sigma^2) \), we write \( x_N \overset{D}{\to} N(\mu, \sigma^2) \)

This definition extends to random vectors, with \( F_N \) and \( F \) interpreted as the cumulative distribution functions of the random vectors \( x_N \) and \( x \)

If the vector \( x \sim N(\mu, \Sigma) \), we write \( x_N \overset{D}{\to} N(\mu, \Sigma) \)
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.

Let the random vector $x_N \overset{D}{\to} x$ and let $g(.)$ be a continuous, real-valued function.

Then $g(x_N) \overset{D}{\to} g(x)$

Again if $x_N = (x_{1N}, x_{2N}) \overset{D}{\to} (x_1, x_2)$ and $g(x_1, x_2) = x_1x_2$, we have that $g(x_{1N}, x_{2N}) = x_{1N}x_{2N} \overset{D}{\to} x_1x_2$
Other useful results are known variously as the transformation theorem, Cramer’s theorem or (again) as Slutsky’s theorem.

For scalar random variables $x_N$ and $y_N$, these can be stated as

If $x_N \overset{D}{\rightarrow} x$ and $y_N \overset{P}{\rightarrow} y$, where $x$ is a random variable and $y$ is a constant, then

i) $x_N + y_N \overset{D}{\rightarrow} x + y$

ii) $x_N y_N \overset{D}{\rightarrow} xy$ (product rule)

iii) $x_N / y_N \overset{D}{\rightarrow} x/y$ provided $y \neq 0$
Results (i) and (ii) extend to (conformable) random vectors

If $x_N$ is a $K \times 1$ random vector, $x$ is a $K \times 1$ random vector, $y_N$ is a $K \times 1$ random vector, and $y$ is a $K \times 1$ non-stochastic vector, we have

i) $x_N + y_N \xrightarrow{D} x + y$

ii) $y_N x_N \xrightarrow{D} y'x$

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

iii) $A_N x_N \xrightarrow{D} Ax$

iv) $x_N A_N^{-1} x_N \xrightarrow{D} x'A^{-1}x$ provided $A^{-1}$ exists
The matrix version of the product rule $A_N x_N \xrightarrow{D} Ax$ implies that if $x_N \xrightarrow{D} N(0, \Sigma)$ and $A_N \xrightarrow{P} A$ then $A_N x_N \xrightarrow{D} N(0, A\Sigma A')$

We use this result to show the asymptotic Normality of $\hat{\beta}_{OLS}$ under the stated assumptions

Using

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

we have

$$\hat{\beta}_{OLS} - \beta = \left( \frac{X'X}{N} \right)^{-1} \left( \frac{X'u}{N} \right)$$
and

\[
\sqrt{N} \left( \hat{\beta}_{OLS} - \beta \right) = \left( \frac{X'X}{N} \right)^{-1} \left( \frac{X'u}{\sqrt{N}} \right)
\]

Now the \( K \times K \) matrix \( \left( \frac{X'X}{N} \right)^{-1} \xrightarrow{P} M_{XX}^{-1} \) from assumption (iv)

And the \( K \times 1 \) vector \( \left( \frac{X'u}{\sqrt{N}} \right) \xrightarrow{D} N(0, \sigma^2 M_{XX}) \) from assumption (v)

The product \( \left( \frac{X'X}{N} \right)^{-1} \left( \frac{X'u}{\sqrt{N}} \right) \) thus has the same limit distribution as \( M_{XX}^{-1} \left( \frac{X'u}{\sqrt{N}} \right) \),

which gives us that

\[
\left( \frac{X'X}{N} \right)^{-1} \left( \frac{X'u}{\sqrt{N}} \right) \xrightarrow{D} N(0, \sigma^2 M_{XX}^{-1} M_{XX} M_{XX}^{-1})
\]

using the symmetry of \( M_{XX}^{-1} \).
Hence

\[ \sqrt{N} \left( \hat{\beta}_{OLS} - \beta \right) \xrightarrow{D} N(0, \sigma^2 M_{XX}^{-1}) \]

More primitive assumptions can be used to derive assumption (v):

\[ \left( \frac{X'u}{\sqrt{N}} \right) \xrightarrow{D} N(0, \sigma^2 M_{XX}) \]

Given assumption (ii), with independence and conditional homoskedasticity, we can use the **Lindeberg-Levy Central Limit Theorem** for independent and identically distributed random vectors.
If \( \{z_i : i = 1, 2, \ldots\} \) is a sequence of iid \( K \times 1 \) random vectors with \( E(z_i) = 0 \) and \( V(z_i) = E(z_i z_i') = \Sigma \) finite, then

\[
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_i \xrightarrow{D} N(0, \Sigma)
\]

Letting \( z_i = x_i u_i \), we have iid \( K \times 1 \) random vectors with \( E(x_i u_i) = 0 \)

Assuming finite variance, we obtain

\[
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} x_i u_i = \left( \frac{X' u}{\sqrt{N}} \right) \xrightarrow{D} N(0, V(x_i u_i))
\]
Moreover

\[ V(x_i u_i) = E[(x_i u_i)(x_i u_i)'] = E(x_i u_i u_i' x'_i) = E(u_i^2 x_i x'_i) \]

since \( u_i \) is a scalar

But

\[
E(u_i^2 x_i x'_i) = E_x \left[ E_{u|x}(u_i^2 x_i x'_i | x_i) \right] = E_x \left[ x_i x'_i E_{u|x}(u_i^2 | x_i) \right] \\
= E_x \left[ x_i x'_i \sigma^2 \right] = \sigma^2 E_x(x_i x'_i) = \sigma^2 M_{XX}
\]

using the conditional homoskedasticity assumption \( E(u_i^2 | x_i) = \sigma^2 \)

Then \( V(x_i u_i) = \sigma^2 M_{XX} \) and \( \left( \frac{X'u}{\sqrt{N}} \right) \xrightarrow{D} N(0, \sigma^2 M_{XX}) \), as stated in assumption (v)
The conditional homoskedasticity assumption is only required to obtain this particular expression for variance of the limit distribution of \( \left( \frac{X'u}{\sqrt{N}} \right) \), and hence the implied variance of the limit distribution of \( \hat{\beta}_{OLS} \).

With independence and conditional heteroskedasticity (i.e. \( E(u_i^2|x_i) \neq \sigma^2 \) for all \( i = 1, \ldots, N \)), we can use the Liapounov Central Limit Theorem for independent but not identically distributed random vectors, to obtain that the limit distribution of \( \left( \frac{X'u}{\sqrt{N}} \right) \) is Normal, but with a different expression for the variance.
For time series models with lagged dependent variables, we cannot assume independent observations.

In this case, we can still use Central Limit Theorems for stationary, ergodic Martingale difference sequences to establish conditions under which $\hat{\beta}_{OLS}$ remains asymptotically Normal.

[Sections 2.2 and 2.3 of Hayashi (2000) provide the details]
Asymptotic inference

(Weak) Consistency \( \hat{\beta}_{OLS} \xrightarrow{P} \beta \)

The parameter vector \( \beta \) is not stochastic, so the limit distribution of \( \hat{\beta}_{OLS} \) 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 in finite samples.

Asymptotic distribution results allow us to approximate this distribution accurately if the (finite) sample size \( N \) is sufficiently large.
Asymptotic Normality \[ \sqrt{N}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N(0, \sigma^2 M_{XX}^{-1}) \]

 Appropriately scaled, we have a non-degenerate limit distribution

 To use this limit distribution to approximate the distribution of \( \hat{\beta}_{OLS} \) in large samples, we multiply by the scalar \( \frac{1}{\sqrt{N}} \) and add the non-stochastic vector \( \beta \) to both sides

 This gives us the approximation to the distribution of the OLS estimator based on our asymptotic distribution theory, or more simply the ‘asymptotic distribution’ of \( \hat{\beta}_{OLS} \), as
\[ \hat{\beta}_{OLS} \sim N \left( \beta, \left( \frac{1}{N} \right) \sigma^2 M_{XX}^{-1} \right) \]

The variance matrix \( \left( \frac{1}{N} \right) \sigma^2 M_{XX}^{-1} \) is the approximation to the variance of the OLS estimator based on our asymptotic distribution theory, or more simply the ‘asymptotic variance’ of \( \hat{\beta}_{OLS} \)

We denote this by \( avar(\hat{\beta}_{OLS}) = \left( \frac{1}{N} \right) \sigma^2 M_{XX}^{-1} \) or (if it is clear from the context that we are referring to the asymptotic variance) more simply by \( V(\hat{\beta}_{OLS}) = \left( \frac{1}{N} \right) \sigma^2 M_{XX}^{-1} \)

*** Note that some authors use the term ‘asymptotic variance’ and the notation \( avar() \) differently ***
Warning: some authors, including Hayashi (2000), use the term ‘asymptotic variance’ and the notation \(a\text{var}(\hat{\beta}_{OLS})\) to denote the variance of the limit distribution of \(\sqrt{N}(\hat{\beta}_{OLS} - \beta)\), here \(\sigma^2M_{XX}^{-1}\), rather than the approximation to the variance of \(\hat{\beta}_{OLS}\) that we obtain from this limit distribution, here \((\frac{1}{N})\sigma^2M_{XX}^{-1}\), as we use the term here.

Expressions in Hayashi (2000) which involve the term \(a\text{var}()\) or \(\hat{a}\text{var}()\) should be divided by the sample size \(N\), or multiplied by \((\frac{1}{N})\), to make them consistent with the notation on the slides.

Neither $\sigma^2$ nor $M_{XX}$ are known

The obvious choice to estimate $M_{XX} = p \lim_{N \to \infty} \left( \frac{X'X}{N} \right)$ is

$$\widehat{M}_{XX} = \left( \frac{X'X}{N} \right)$$

Since $p \lim_{N \to \infty} \widehat{M}_{XX} = p \lim_{N \to \infty} \left( \frac{X'X}{N} \right) = M_{XX}$, this gives us a consistent estimator of $M_{XX}$

Both the OLS estimator

$$\hat{\sigma}^2 = \frac{1}{N - K} \sum_{i=1}^{N} \hat{u}_i^2 = \frac{\hat{u}'\hat{u}}{N - K}$$
and the Maximum Likelihood estimator

\[ \hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 = \frac{\hat{u}'\hat{u}}{N} \]

can be shown to be consistent estimators of \( \sigma^2 \) under our assumptions

Using the OLS estimator of \( \sigma^2 \) gives us a consistent estimator of \( avar(\hat{\beta}_{OLS}) \) as

\[
avar(\hat{\beta}_{OLS}) = \left( \frac{1}{N} \right) \hat{\sigma}^2 \hat{M}_{XX}^{-1} \\
= \left( \frac{1}{N} \right) \hat{\sigma}^2 \left( \frac{X'X}{N} \right)^{-1} \\
= \left( \frac{1}{N} \right) \hat{\sigma}^2 N(X'X)^{-1} \\
= \hat{\sigma}^2 (X'X)^{-1}
\]
Since $\text{avar}(\hat{\beta}_{OLS}) \xrightarrow{P} \text{avar}(\hat{\beta}_{OLS})$, we also have

$$\hat{\beta}_{OLS} \xrightarrow{a} N \left( \beta, \text{avar}(\hat{\beta}_{OLS}) \right)$$

or

$$\hat{\beta}_{OLS} \xrightarrow{a} N \left( \beta, \hat{\sigma}^2 (X'X)^{-1} \right)$$

Note that this asymptotic approximation to the distribution of $\hat{\beta}_{OLS}$ (in the model with conditional homoskedasticity) coincides with the exact result we obtained for the distribution of $\hat{\beta}_{OLS}$ in the classical linear regression model with Normally distributed errors.

There is at least one version of the linear model for which we know this approximation based on the asymptotic distribution theory will be accurate
In large samples, this asymptotic distribution result provides a motivation for using this expression to estimate the variance of $\hat{\beta}_{OLS}$, and for treating $\hat{\beta}_{OLS}$ as a random vector with a Normal distribution, even if:

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

This will be useful in practice in sample sizes that are large enough for the asymptotic distribution to provide a good approximation to the distribution of $\hat{\beta}_{OLS}$, but not so large for the variance of $\hat{\beta}_{OLS}$ to become negligibly small
To interpret this expression for the asymptotic variance of $\hat{\beta}_{OLS}$ (under conditional homoskedasticity), $\text{avar}(\hat{\beta}_{OLS}) = \hat{\sigma}^2 (X'X)^{-1}$, 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

Then $X'X$ 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 with the sample size $N$

$\hat{\sigma}^2 \xrightarrow{P} \sigma^2$, so $\hat{\sigma}^2$ does not increase with the sample size $N$

In this example, it is clear that

$$\text{avar}(\hat{\beta}_{OLS}) = \hat{\sigma}^2 (X'X)^{-1} = \frac{\hat{\sigma}^2}{N \sum_{i=1}^{N} x_i^2}$$

decreases with the sample size $N$
So although the expression $\widehat{avar}(\beta_{OLS}) = \hat{\sigma}^2(X'X)^{-1}$ does not appear to depend on the sample size $N$, our estimate of the asymptotic variance will tend to fall as we add more observations to the sample.

The square root of this (scalar) asymptotic variance, termed the asymptotic standard error, also falls as $N$ increases, indicating that we can estimate the parameter $\beta$ more precisely as we add more observations to the sample.

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

$$\widehat{avar}(\beta_{OLS}) = \frac{\hat{\sigma}^2}{N} \sum_{i=1}^{N} x_i^2 \to 0$$

consistent with our result that the limit distribution of $\hat{\beta}_{OLS}$ is degenerate.
The observation that \( X'X = \sum_{i=1}^{N} x_i^2 \) increases with \( N \) is perfectly consistent with our assumption that 
\[
\left( \frac{X'X}{N} \right) = \frac{1}{N} \sum_{i=1}^{N} x_i^2 \xrightarrow{P} M_{XX}
\]

In the scalar case, with \( x_i \) having zero mean, \( \frac{1}{N} \sum_{i=1}^{N} x_i^2 \) is the sample variance of the random variable \( x \)

With independent observations, \( \frac{1}{N} \sum_{i=1}^{N} x_i^2 \) is a consistent estimator of the population variance

In this case, our assumption simply requires that the population variance of \( x \) 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 $\hat{v}_{kk}$ again denote the $k^{th}$ element on the main diagonal of the $K \times K$ (asymptotic) variance matrix $\hat{\sigma}^2(X'X)^{-1} = \text{avar}(\hat{\beta}_{OLS})$

Since

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

we also have

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

**Simple hypothesis tests** of the form $H_0 : \beta_k = \beta_k^0$ can be conducted in large samples by comparing the test statistic $t_k = \frac{\hat{\beta}_k - \beta_k^0}{se_k}$ computed under the null hypothesis to critical values of the standard Normal $N(0, 1)$ 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 the 5% level of significance, we can compare the absolute value of $t_k = \frac{\hat{\beta}_k - \beta_k^0}{se_k}$ that we obtain to the critical value 1.96.

Strictly speaking, the standard error $se_k = \sqrt{v_{kk}}$ should be referred to here as the asymptotic standard error, sometimes denoted $ase_k$.

This test is sometimes referred to as an asymptotic t-test.
Why can we treat $t_k$ here as a draw from the standard Normal distribution, even though we have estimated the unknown $\sigma^2$?

The limit distribution result

$$\sqrt{N}(\hat{\beta}_{OLS} - \beta) \overset{D}{\rightarrow} N(0, \sigma^2 M_{XX}^{-1})$$

implies that

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\nu}_{kk}}} \overset{D}{\rightarrow} N(0, 1)$$

where $\nu_{kk}$ is the (unknown) $k^{th}$ element on the main diagonal of $(\frac{1}{N}) \sigma^2 M_{XX}^{-1}$.

Using $se_k = \sqrt{\hat{\nu}_{kk}}$ to estimate the unknown $\sqrt{\nu_{kk}}$ gives

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\nu}_{kk}}} = \left( \frac{\sqrt{\nu_{kk}}}{\sqrt{\hat{\nu}_{kk}}} \right) \left( \frac{\beta_k - \beta_k}{\sqrt{\nu_{kk}}} \right) = \left( \frac{\sqrt{\nu_{kk}}}{\sqrt{\hat{\nu}_{kk}}} \right) z_k$$
Since $\sqrt{\hat{v}_{kk}}$ is a consistent estimator of $\sqrt{v_{kk}}$, we have $\sqrt{\hat{v}_{kk}} \xrightarrow{P} \sqrt{v_{kk}}$, or

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

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

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

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 $N(0, 1)$ random variable as $N \to \infty$ with $K$ fixed, there is also an asymptotic justification for comparing the test statistic $t_k = \frac{\hat{\beta}_k - \beta_k^0}{se_k}$ 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 $t_k = \frac{\hat{\beta}_k - \beta^0_k}{se_k}$ for reporting p-values and confidence intervals.

You may have noticed that Stata does this.

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

Both the $N(0, 1)$ 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

The asymptotic distribution of the \( p \times 1 \) random vector \( \hat{\theta}_{OLS} = H \hat{\beta}_{OLS} \), where \( H \) is a non-stochastic \( p \times K \) matrix with \( \text{rank}(H) = p \) (full row rank), is given by

\[
\hat{\theta}_{OLS} \overset{a}{\sim} N(H\beta, H\var{\hat{\beta}_{OLS}}H')
\]

where \( p \lim_{N \to \infty} \hat{\theta}_{OLS} = H p \lim_{N \to \infty} \hat{\beta}_{OLS} = H \beta \)

and \( \var{\hat{\theta}_{OLS}} = H\var{\hat{\beta}_{OLS}}H' = \hat{\sigma}^2 H(X'X)^{-1}H' \)

Then the quadratic form

\[
w = (\hat{\theta}_{OLS} - H\beta)' \left[ \var{\hat{\theta}_{OLS}} \right]^{-1} (\hat{\theta}_{OLS} - H\beta) \overset{a}{\sim} \chi^2(p)
\]
Joint hypothesis tests of the form $H_0 : H \beta = \theta^0$ can thus be conducted by computing the test statistic

$$w = (\hat{\theta}_{OLS} - \theta^0)' \left[ \widehat{avar}(\hat{\theta}_{OLS}) \right]^{-1} (\hat{\theta}_{OLS} - \theta^0)$$

which has a $\chi^2(p)$ asymptotic 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 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 also have noticed that Stata does this.

In large samples, this difference also becomes negligibly small.
Testing non-linear restrictions

These asymptotic Wald tests of restrictions on the parameter vector $\beta$ can be extended to cover non-linear restrictions

For example, in the model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

we may be interested in testing the non-linear restriction that $\beta_2 \beta_3 = 1 \leftrightarrow \beta_2 \beta_3 - 1 = 0$
This hypothesis can be formulated as

\[
h \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \beta_2 \beta_3 - 1 = 0
\]

or \( h(\beta) = \beta_2 \beta_3 - 1 = 0 \) for a vector-valued function \( h(\beta) \) with continuous first derivatives

Let \( H(\beta) \) be the \( p \times K \) matrix of first derivatives \( H(\beta) = \partial h(\beta)/\partial \beta' \), so that in our example we have

\[
H(\beta) = \partial h(\beta)/\partial \beta' = \begin{pmatrix} 0 & \beta_3 & \beta_2 \end{pmatrix}
\]

with full row rank (i.e. \( \text{rank}(H(\beta)) = 1 \) in our example)
We can then use the **delta method** to show that

\[
    w = h(\hat{\beta}_{OLS})' \left[ H(\hat{\beta}_{OLS})\hat{avar}(\hat{\beta}_{OLS})H(\hat{\beta}_{OLS})' \right]^{-1} h(\hat{\beta}_{OLS}) \sim \chi^2(p)
\]

under the null hypothesis that \( h(\beta) = 0 \)

We compute the test statistic \( w \) using our estimate of the parameter vector \( \hat{\beta}_{OLS} \) and its asymptotic variance \( \hat{avar}(\hat{\beta}_{OLS}) \), and compare the result to critical values of the \( \chi^2(p) \) distribution.

The Wald test statistic for linear restrictions, stated previously, can be obtained as a special case, setting \( h(\beta) = H\beta - \theta^0 = 0 \)

Note that the representation \( h(\beta) = 0 \) for non-linear restrictions is not unique.
For example, to test the restriction that $\beta_2\beta_3 = 1$, we may consider the representations

$$\beta_2\beta_3 - 1 = 0$$

or

$$\beta_2 - \frac{1}{\beta_3} = 0$$

or

$$\beta_3 - \frac{1}{\beta_2} = 0$$

Although Wald test statistics based on different representations of non-linear restrictions are asymptotically equivalent, they can take different values in finite samples, and there is no guarantee that different Wald tests of the same non-linear restriction will give the same outcome in finite samples.