Lecture 4
Multiple Linear Regression: Asymptotics

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

February 10, 2022

Lecture Outline

1. Multiple Linear Regression: Asymptotics
   - Limitation of Classical Assumptions
   - Review of LLN and CLT
   - Consistency
   - Asymptotic Normality
   - Estimation of Asymptotic Covariance Matrix
   - Large Sample Tests
   - Tests of Conditional Homoskedasticity
2. Instrumental Variable Estimation
   - Endogeneity
   - Instrumental Variable Estimation
   - Two-Stage Least Squares Estimation

Are Classical Assumptions Reasonable?

- Non-random x_j: If y and x_j are all economic variables, it is not reasonable to assume that, while y is random, x_j are all non-random. When x_j are random variables, it would be difficult to infer the statistical properties of the OLS estimators.
  - $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]$ cannot be expressed as $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}(\boldsymbol{y})$.
  - $\text{var}(\hat{\boldsymbol{\beta}}) = \text{var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})$ cannot be expressed as $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\text{var}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
- When x_j are random, we need to consider the conditional variance $\text{var}(y_i|x_{i1},\ldots,x_{ik})$ and the effect of conditional heteroskedasticity, i.e., $\text{var}(y_i|x_{i1},\ldots,x_{ik})$ change with some variables x_{ij}, for deriving $\text{var}(\hat{\boldsymbol{\beta}})$.

- Normality: There is no guarantee that the data we consider are normally distributed; for example, binary variables (taking values zero and one) and count data (taking finitely many values: 1, 2, . . .) are definitely non-normal. Then, the statistics derived earlier no longer have the t or F distribution.
- If $\boldsymbol{X}$ is random, the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ need not be normally distributed even when $\boldsymbol{y}$ is. In fact, $\hat{\boldsymbol{\beta}}$ has an unknown complex distribution when $\boldsymbol{X}$ is random. This renders hypothesis testing difficult.
- Question: Is it possible to draw statistical inference without Classical Assumptions?

  Ans: Yes, we can derive asymptotic properties of the OLS estimator and related statistics under weaker and reasonable conditions.

Law of Large Numbers

Khinchine's Weak Law of Large Numbers (WLLN)

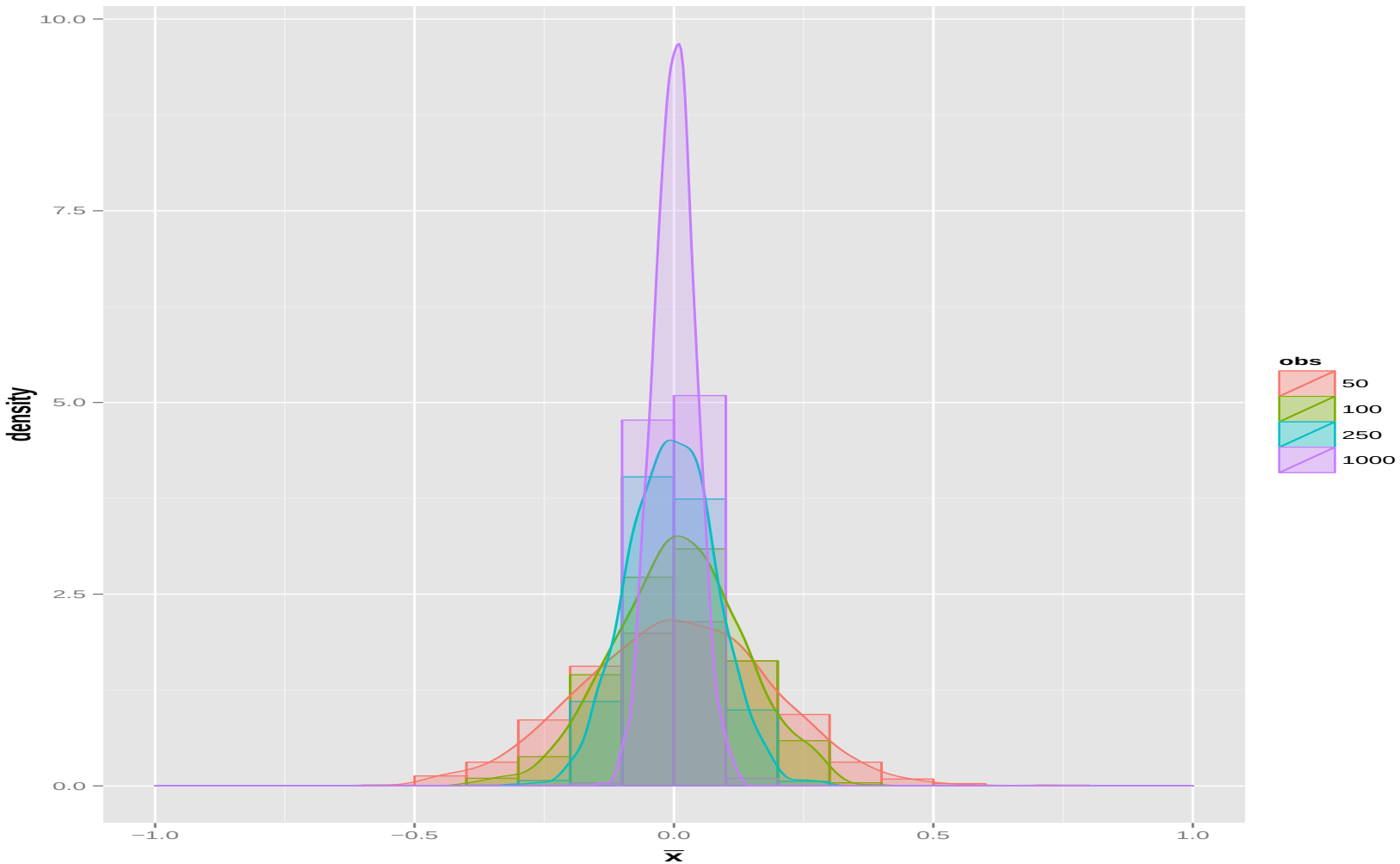
Let $z_1, \ldots, z_n$ be i.i.d. with mean μ_o. Then, $\bar{z}_n = n^{-1} \sum_{i=1}^{n} z_i$ converges in probability to μ_o, in the sense that, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\{|\bar{z}_n - \mu_o| > \varepsilon\} = 0.$$

This is denoted as $\bar{z}_n \xrightarrow{\mathbb{P}} \mu_o$.

- A WLLN ensures that $\bar{z}_n$, the sample average of z_i, is essentially close to μ_o; the probability that $\bar{z}_n$ deviates from the true mean by a certain amount approaches zero when the sample size becomes large.
- Note that i.i.d. random variables without a finite mean (e.g. Cauchy variables) do not obey a WLLN.

Simulations of LLN: $t(5)$ Variable



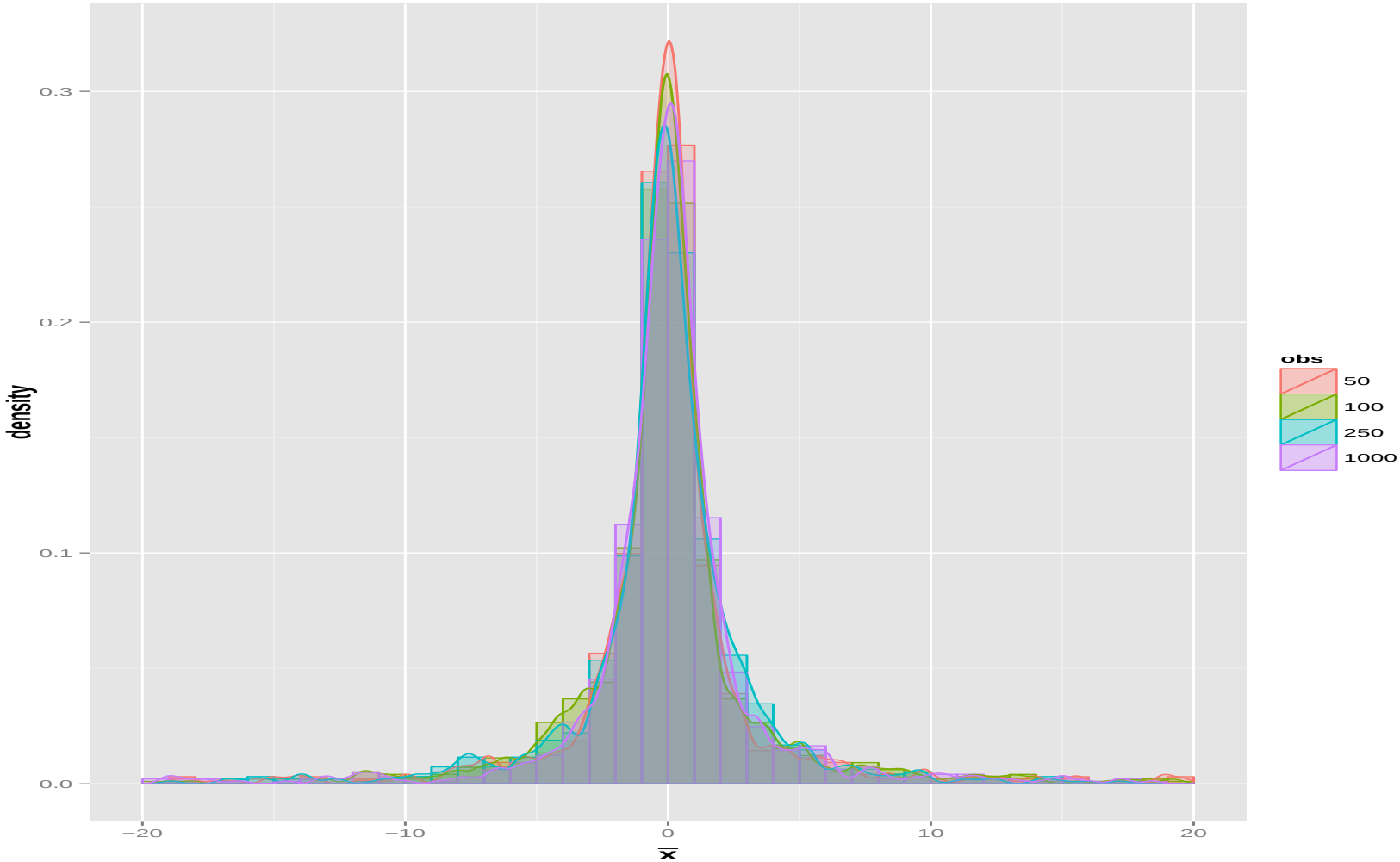
Note: i.i.d. $t(5)$ variables with 1000 replications.

Simulation Procedure

1. Generate a sample of n ($n = 100$ for example) realizations z_i from a given distribution F ($t(5)$ in our example) and compute the sample average: $\bar{z}_n = n^{-1} \sum_{i=1}^{n} z_i$.
2. Replicate the step above R ($R = 1000$ for example) times and obtain 1000 sample averages $\bar{z}_n$.
3. The frequency plot of these $\bar{z}_n$ is the simulated (finite-sample) density function of $\bar{z}_n$ under F.

This frequency plot would be more concentrated around the true mean 0 when n becomes larger; that is, it is less likely that $\bar{z}_n$ would be far away from 0 when n is large.

Simulations of LLN: $t(1)$ Variable



Note: i.i.d. $t(1)$ variables with 1000 replications.

Central Limit Theorem

Lindeberg-Lévy's Central Limit Theorem (CLT)

Let $z_1, \ldots, z_n$ be i.i.d. with mean μ_o and variance $\sigma_o^2 > 0$. Then,

$$\frac{\sqrt{n}(\bar{z}_n - \mu_o)}{\sigma_o} = \frac{(\bar{z}_n - \mu_o)}{\sigma_o/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\xrightarrow{D}$ stands for convergence in distribution.

A CLT ensures that the distributions of suitably normalized sample averages are essentially close to the standard normal distribution in the limit, regardless of the original distributions of z_i. Any random irregularities that are not governed by the standard normal distribution will be wiped out in the limit.

Given that $z_1, \ldots, z_n$ are i.i.d. with mean μ_o and variance σ_o^2, $\mathbb{E}(\bar{z}_n) = \mu_o$, and $\text{var}(\bar{z}_n) = \sigma_o^2/n$. As $\sigma_o^2/n \to 0$ as $n \to \infty$, $\bar{z}_n$ ought to have a degenerate distribution at μ_o in the limit. Dividing $\bar{z}_n$ by its standard deviation $\sigma_o/\sqrt{n}$, we have

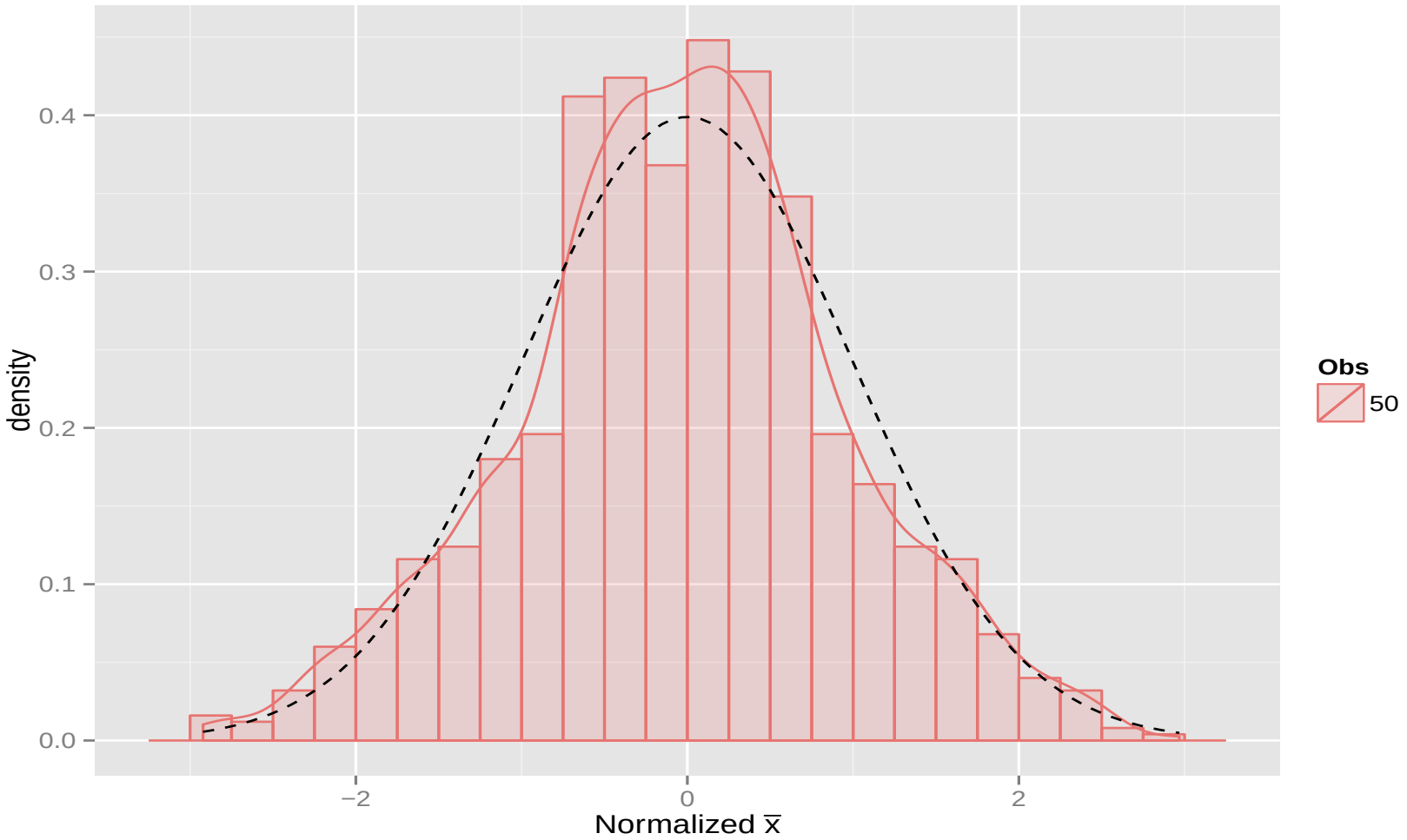
$$\frac{(\bar{z}_n - \mu_o)}{\sigma_o/\sqrt{n}} = \frac{\sqrt{n}(\bar{z}_n - \mu_o)}{\sigma_o},$$

which has mean zero and constant variance one for all n. This prevents the probability mass from shrinking towards a single point in the limit. Note that the factor $\sqrt{n}$ characterizes the rate of convergence of $\bar{z}_n$.

Remarks:

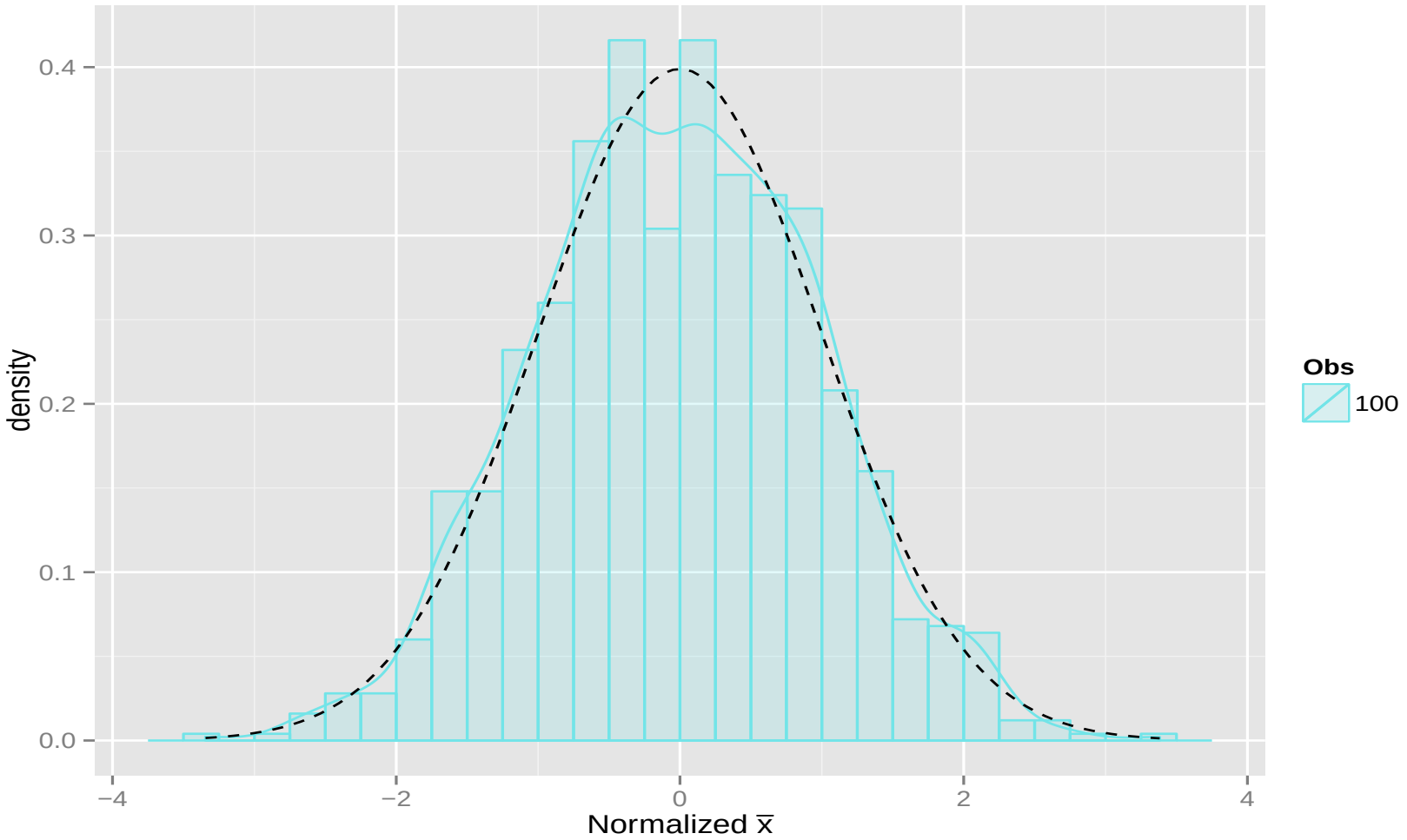
- i.i.d. random variables need not obey this CLT if they do not have a finite variance, e.g., $t(2)$ variables.
- Both LLN and CLT may hold for non-i.i.d. random variables under stronger conditions (e.g., higher order moments exist).

Simulations of CLT: $t(5)$ Variable; Sample 50



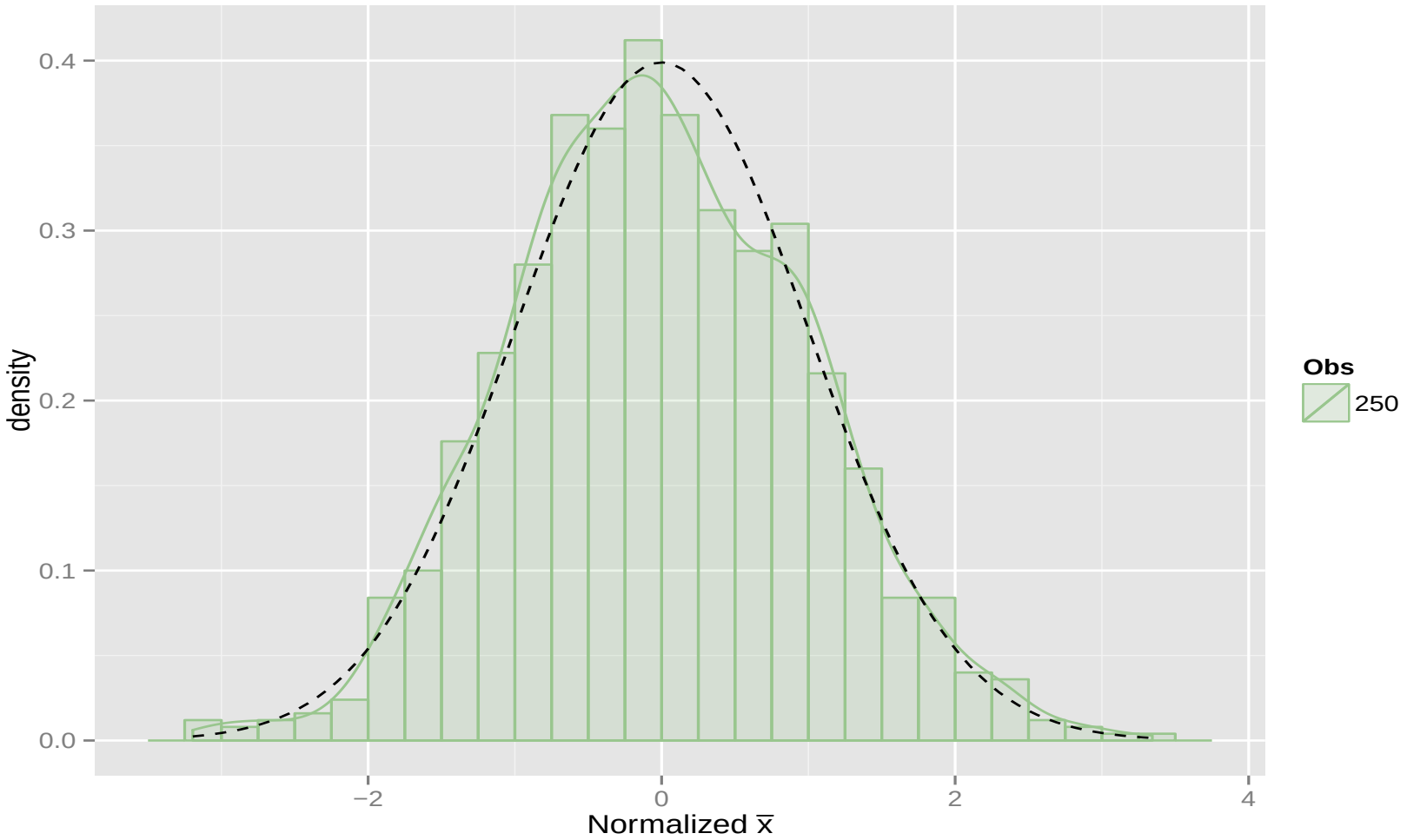
Note: i.i.d. $t(5)$ variables with 1000 replications.

Simulations of CLT: $t(5)$ Variable; Sample 100



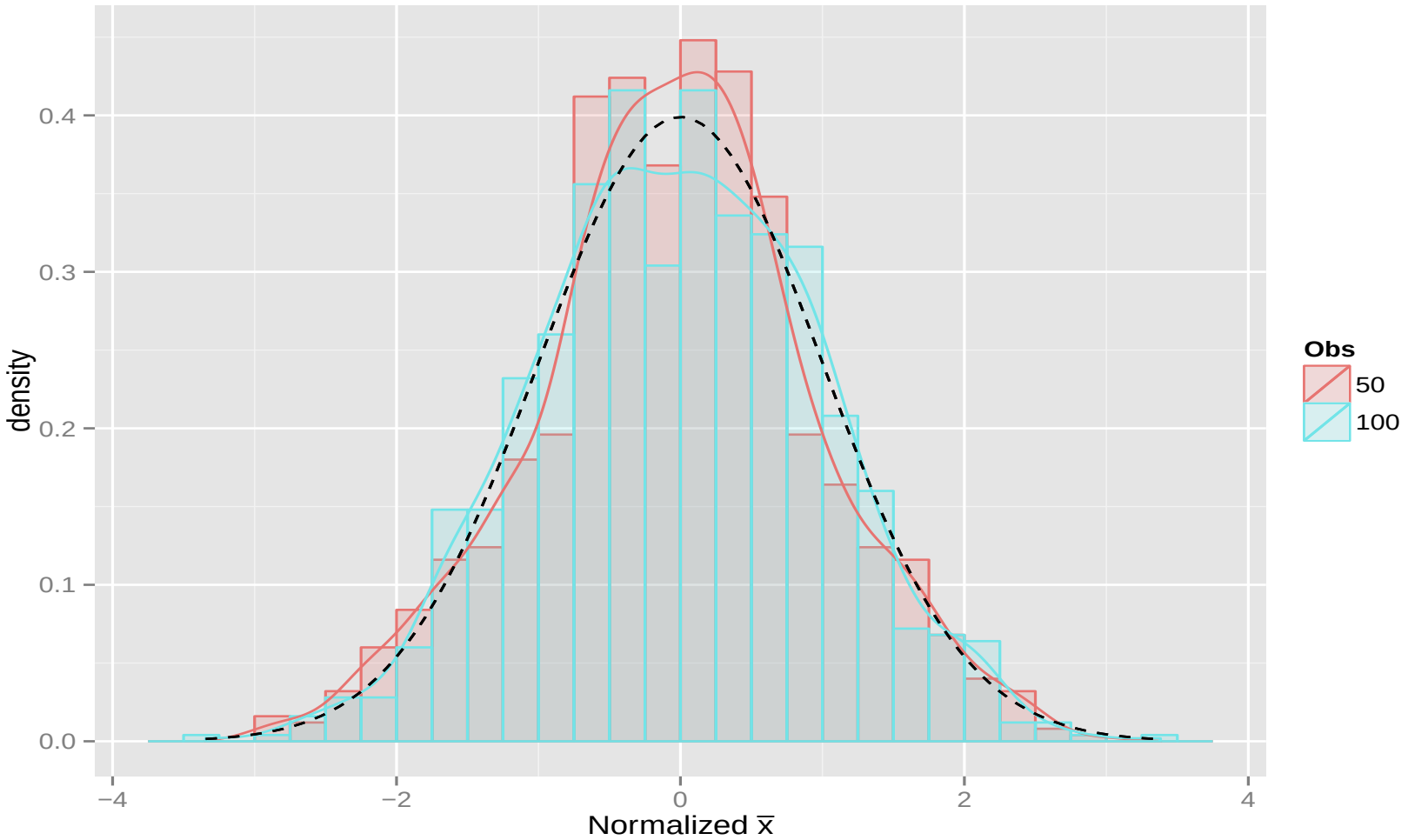
Note: i.i.d. $t(5)$ variables with 1000 replications.

Simulations of CLT: $t(5)$ Variable; Sample 250



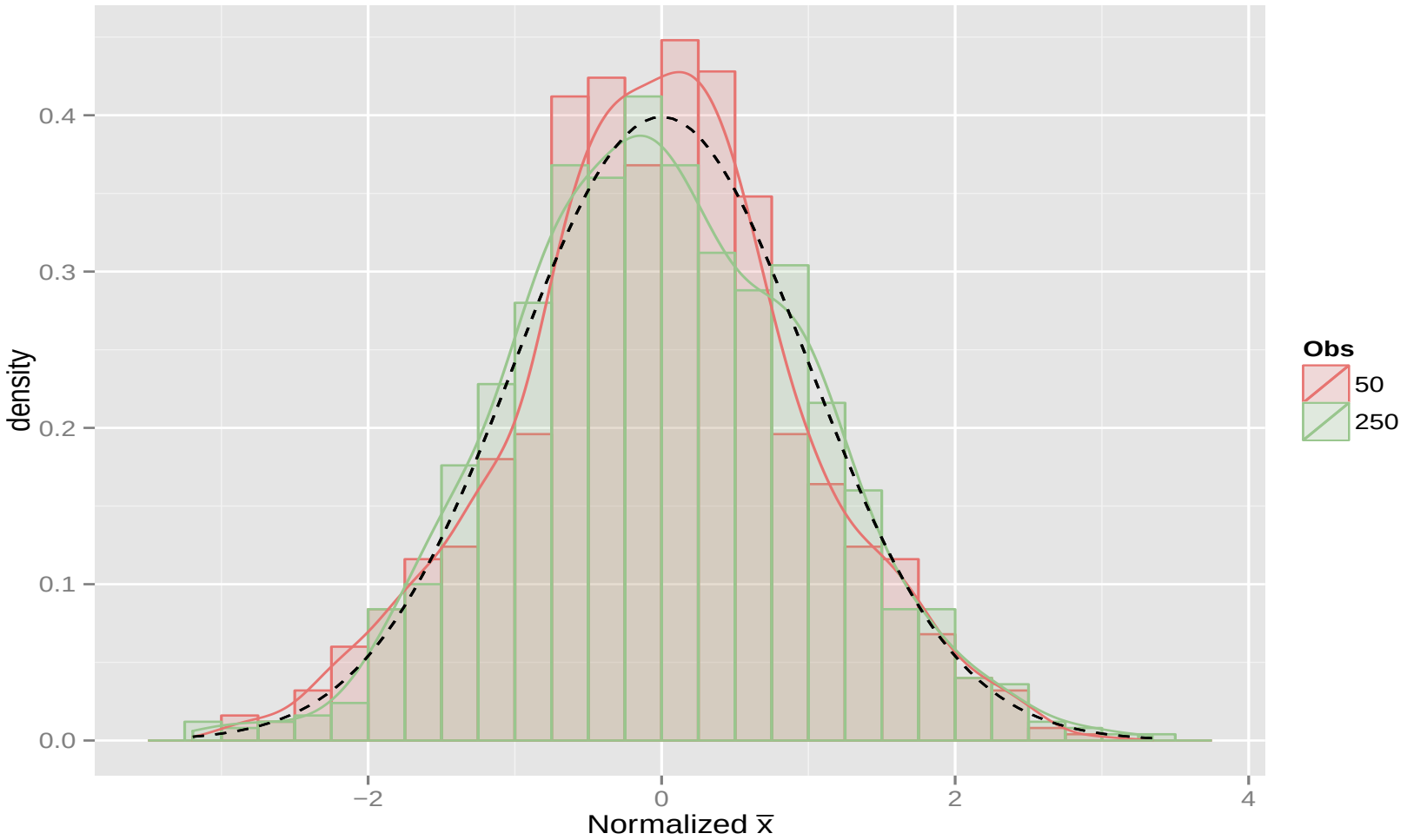
Note: i.i.d. $t(5)$ variables with 1000 replications.

Comparison of CLT: $t(5)$ Variable; Samples 50 & 100



Note: i.i.d. $t(5)$ variables with 1000 replications.

Comparison of CLT: $t(5)$ Variable; Samples 50 & 250



Note: i.i.d. $t(5)$ variables with 1000 replications.

Multivariate Versions of LLN and CLT

- Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be i.i.d. random vectors with mean $\boldsymbol{\mu}_o$. We say $\bar{\boldsymbol{z}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\mu}_o$ if each element of $\bar{\boldsymbol{z}}_n$ converges in probability to the corresponding element in $\boldsymbol{\mu}_o$.
- Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be i.i.d. with mean $\boldsymbol{\mu}_o$ and positive definite covariance matrix $\boldsymbol{\Sigma}_o$. The multivariate CLT reads:

$$\boldsymbol{\Sigma}_o^{-1/2}\sqrt{n}(\bar{\boldsymbol{z}}_n - \boldsymbol{\mu}_o) = \boldsymbol{\Sigma}_o^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{z}_i - \boldsymbol{\mu}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

- Cramér-Wold Device: Let $\boldsymbol{\zeta}_n = \boldsymbol{\Sigma}_o^{-1/2}\sqrt{n}(\bar{\boldsymbol{z}}_n - \boldsymbol{\mu}_o)$ and $\boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Then, $\boldsymbol{\zeta}_n \xrightarrow{D} \boldsymbol{\nu}$ if, and only if, $\boldsymbol{a}'\boldsymbol{\zeta}_n \xrightarrow{D} \boldsymbol{a}'\boldsymbol{\nu}$ for all $\boldsymbol{a}$ with $\|\boldsymbol{a}\| = 1$. That is, $\boldsymbol{\zeta}_n$ obeys a multivariate CLT if, and only if, any linear combination of $\boldsymbol{\zeta}_n$ obeys a univariate CLT.

OLS Consistency

Let $\boldsymbol{x}_i = (1,\ x_{i1},\ \ldots,\ x_{ik})'$ denote the i^{th} column of $\boldsymbol{X}'$, we can write the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ as:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, n.$$

It follows that the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$ can be expressed as:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right).$$

Modern Assumption I

The random variables y_i are $y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + \varepsilon_i$, $i = 1, \ldots, n$, such that

(i) $\boldsymbol{x}_i$, $i = 1, \ldots, n$, are i.i.d. random vectors with finite second moment;

(ii) ε_i, $i = 1, \ldots, n$, are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\boldsymbol{x}_i\varepsilon_i) = \boldsymbol{0}$.

Given Modern Assumption I,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i (\boldsymbol{x}_i' \boldsymbol{b}_o + \varepsilon_i) \right)$$
$$= \boldsymbol{b}_o + \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i \right).$$

Clearly, $\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{b}_o$ when the second term converges to zero. Writing

$$\hat{\boldsymbol{\beta}} = \boldsymbol{b}_o + \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i \right),$$

we have from the WLLN that $n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{x}_i \varepsilon_i) = \boldsymbol{0}$, and

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i') =: \boldsymbol{M}_{xx}.$$

It follows that

$$\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{b}_o + \boldsymbol{M}_{xx}^{-1}\mathbf{0} = \boldsymbol{b}_o.$$

OLS Consistency

Given the specification $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$, suppose that Modern Assumption I holds. Then, $\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{b}_o$; that is, $\hat{\boldsymbol{\beta}}$ is weakly consistent for $\boldsymbol{b}_o$.

- Aside from the WLLN effect, consistency hinges on the condition $\mathbb{E}(\boldsymbol{x}_i\varepsilon_i) = \mathbf{0}$, i.e., ε_i and $\boldsymbol{x}_i$ are uncorrelated. Thus, the linear specification suffices to capture the systematic part of y_i; no other information could be of help in a linear fashion.
- Consistency ensures that, when more information (a larger sample) become available, it is more likely that the OLS estimates will be close to the true parameter values. Note that consistency is a property that depends on sample size n, but unbiasedness is not.

Consistency and Finite Samples

Q1: Why is consistency relevant if we only have finite samples in practice?

Ans: It is true that we could never have an infinite sample, but consistency ensures that an estimator can approximate the true parameters arbitrarily well when enough information are available.

Q2: For what sample size should we expect estimates to well approximate the true parameters?

Ans: It depends on the stochastic properties of the data. If the data are independent, it typically requires a larger sample to achieve the same degree of accuracy when the data distributions have fatter tails than does the normal distribution. If the data are dependent (correlated), an even larger sample may be needed. We must emphasize that there is no "magic number" (such as 30, 50 or 100) for the desired sample size.

Exclusion of Important Variables

Given the specification $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$, suppose y_i actually depend on more variables $\boldsymbol{z}_i$ such that $y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + \boldsymbol{z}_i'\boldsymbol{d}_o + \varepsilon_i$, where ε_i satisfy Modern Assumption I(ii) with $\mathbb{E}(\boldsymbol{x}_i\varepsilon_i) = \boldsymbol{0}$ and $\mathbb{E}(\boldsymbol{z}_i\varepsilon_i) = \boldsymbol{0}$. That is, our specification excludes the important variables $\boldsymbol{z}$ that should have been included in the model. In this case,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i(\boldsymbol{x}_i'\boldsymbol{b}_o + \boldsymbol{z}_i'\boldsymbol{d}_o + \varepsilon_i)\right) \\
&= \boldsymbol{b}_o + \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left[\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{z}_i'\right)\boldsymbol{d}_o + \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\varepsilon_i)\right)\right].
\end{aligned}$$

By the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{z}_i' \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{x}_i \boldsymbol{z}_i') =: \boldsymbol{M}_{xz}, \quad \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{x}_i \varepsilon_i) = \boldsymbol{0}.$$

It follows that $\hat{\boldsymbol{\beta}}$ is inconsistent for $\boldsymbol{b}_o$:

$$\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{b}_o + \boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xz}\boldsymbol{d}_o + \boldsymbol{M}_{xx}^{-1} \cdot \boldsymbol{0} = \boldsymbol{b}_o + \boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xz}\boldsymbol{d}_o.$$

Note that consistency would still obtain if the vector of the omitted variables $\boldsymbol{z}$ is orthogonal to $\boldsymbol{x}$, i.e., $\boldsymbol{M}_{xz} = \mathbb{E}(\boldsymbol{x}_i \boldsymbol{z}_i') = \boldsymbol{0}$. This shows that inconsistency arises because the included variables in $\boldsymbol{x}$ are correlated with the omitted variables in $\boldsymbol{z}$. Given the specification $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$, the correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$ can also be understood as $\mathbb{E}(\boldsymbol{x}_i u_i) \neq \boldsymbol{0}$, because u_i now includes the omitted variables $\boldsymbol{z}_i$.

Recall that the dummy variable model for estimating the average treatment effect on the outcome variable Y is:

$$y_i = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + D_i\gamma + u_i,$$

where D is the treatment indicator. D is endogenous when D and the outcome Y depend on some unobserved factors. For example, in the study of "return to college education", D is the dummy for college attendance, and Y is the earning. Here, D and Y may depend on the factors that characterize "ability", such that people with better "ability" are more likely to go to college and earn a higher pay. Thus, y_i actually follow:

$$y_i = \alpha_o + \boldsymbol{x}_i'\boldsymbol{\beta}_o + D_i\gamma_o + (A_i\delta_o + \varepsilon_i),$$

where A_i denote the unobserved "ability" index. The regression error u_i now includes $A_i\delta_o$ and ε_i, and $\mathbb{E}(D_i u_i) \neq \mathbf{0}$ because D_i also depend on A_i. As a result, the OLS estimators are inconsistent.

Inclusion of Irrelevant Variables

Suppose we specify multiple linear regression: $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta} + u_i$, but y_i depend only on $\boldsymbol{x}_i$: $y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + \varepsilon_i$, where $\mathbb{E}(\boldsymbol{x}_i\varepsilon_i) = \mathbf{0}$. That is, we include the irrelevant variables $\boldsymbol{z}$ in our specification. As discussed before, y_i can be written as

$$y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + \boldsymbol{z}_i' \cdot \mathbf{0} + \varepsilon_i,$$

with $\mathbf{0}$ as the true parameter associated with $\boldsymbol{z}_i$, and u_i agree with ε_i. Then, as long as $\mathbb{E}(\boldsymbol{z}_i\varepsilon_i) = 0$, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{b}_o$, and $\hat{\boldsymbol{\delta}}$ is consistent for $\mathbf{0}$. This shows that a more general model specification does not affect consistency.

Asymptotic Normality

In addition to Modern Assumption I(i) and (ii), we also postulate the following condition.

Modern Assumption I(iii)

The random variables y_i are $y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + \varepsilon_i$, $i = 1, \ldots, n$, such that

(iii) $\boldsymbol{x}_i\varepsilon_i$ obey a multivariate CLT:

$$\boldsymbol{V}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i\varepsilon_i \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),$$

where for $\boldsymbol{V} = \mathrm{var}(\boldsymbol{x}_i\varepsilon_i)$ is a positive definite matrix.

Recall

$$\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i \right).$$

We can then write

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o) = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \underbrace{\boldsymbol{V}^{1/2} \boldsymbol{V}^{-1/2}}_{=\boldsymbol{I}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i \right).$$

Modern Assumption I(iii) and the fact that $n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{\mathbb{P}} \boldsymbol{M}_{xx}$ lead to:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o) \xrightarrow{D} \boldsymbol{M}_{xx}^{-1} \boldsymbol{V}^{1/2} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),$$

where the limit is simply a linear transformation of $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Note that $\boldsymbol{x}_i\boldsymbol{x}_i'$ is symmetric, and so is $\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i') = \boldsymbol{M}_{xx}$. Also, $\boldsymbol{V}$ is a covariance matrix and must also be symmetric. It follows that the limit is a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix:

$$(\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}^{1/2})\boldsymbol{I}(\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}^{1/2})' = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2}\boldsymbol{M}_{xx}^{-1} = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}\boldsymbol{M}_{xx}^{-1}.$$

We have established the following result.

OLS Asymptotic Normality

Given the specification $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$, suppose that Modern Assumption I(i), (ii) and (iii) hold. Then,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}\boldsymbol{M}_{xx}^{-1}),$$

where $\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}\boldsymbol{M}_{xx}^{-1}$ is known as the asymptotic covariance matrix.

When ε_i are conditionally homoskedastic such that $\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i) = \sigma_o^2$,

$$\boldsymbol{V} = \mathbb{E}(\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i') = \mathbb{E}[\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i)\boldsymbol{x}_i \boldsymbol{x}_i')] = \sigma_o^2 \mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i'),$$

where the 2nd equality follows from the law of iterated expectations. In this case, the asymptotic covariance matrix simplifies to

$$\boldsymbol{M}_{xx}^{-1} \boldsymbol{V} \boldsymbol{M}_{xx}^{-1} = \boldsymbol{M}_{xx}^{-1} (\sigma_o^2 \boldsymbol{M}_{xx}) \boldsymbol{M}_{xx}^{-1} = \sigma_o^2 \boldsymbol{M}_{xx}^{-1},$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \sigma_o^2 \boldsymbol{M}_{xx}^{-1}).$$

When $\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i)$ depend on $\boldsymbol{x}_i$, i.e., ε_i are conditionally heteroskedastic, the asymptotic covariance matrix has the *sandwich* form and cannot be simplified.

Some Remarks

- OLS consistency and asymptotic normality would hold provided that data satisfy proper WLLN and CLT. We do not require $\boldsymbol{x}$ to be non-random, nor do we impose normality and (conditional) homoskedasticity on y_i .
- Under these weaker conditions, we cannot derive the exact distribution of $\hat{\boldsymbol{\beta}}$, but are able to derive the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o)$. The limiting normal distribution serves as an approximation to the unknown, exact distribution.
- For a given sample size, whether the approximation to normality is good depends on the stochastic properties of the data. A larger sample typically yields better approximation to normality, but, again, there is no "magic number" for the sample size (such as 30, 50 or 100) that assures good approximation.

Estimation of Asymptotic Covariance Matrix

Let $\boldsymbol{D}_o := \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}\boldsymbol{M}_{xx}^{-1}$. Then, $\boldsymbol{D}_o^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. By the WLLN, $\boldsymbol{M}_{xx} = \mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')$ can be consistently estimated by its sample counterpart:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i',$$

and $\boldsymbol{V} = \mathbb{E}(\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i')$ can be consistently estimated by (proof omitted):

$$\widehat{\boldsymbol{V}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i',$$

where $\hat{u}_i$ are the OLS residuals. This $\widehat{\boldsymbol{V}}$ is known as a heteroskedasticity-consistent covariance matrix estimator because $\boldsymbol{V}$ permits conditional heteroskedasticity of an unknown form.

A weakly consistent estimator of $\boldsymbol{D}_o$ is thus

$$\widehat{\boldsymbol{D}} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1};$$

this is known as the Eicker-White estimator. Replacing $\boldsymbol{D}_o$ by its consistent estimator $\widehat{\boldsymbol{D}}$ we obtain:

$$\widehat{\boldsymbol{D}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{b}_o) = \underbrace{\widehat{\boldsymbol{D}}^{-1/2}\boldsymbol{D}_o^{1/2}}_{\xrightarrow{\mathrm{P}}\boldsymbol{I}}\underbrace{\boldsymbol{D}_o^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{b}_o)}_{\xrightarrow{D}\mathcal{N}(\boldsymbol{0},\boldsymbol{I})} \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\boldsymbol{I}).$$

In practice, we use the following covariance matrix estimator:

$$\tilde{\boldsymbol{D}} = \widehat{\boldsymbol{D}}/n = \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1},$$

such that

$$\tilde{\boldsymbol{D}}^{-1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{b}_o) = \widehat{\boldsymbol{D}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{b}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\boldsymbol{I}).$$

Let $\tilde{d}_{jj}$ denote the j^{th} diagonal element of $\tilde{\boldsymbol{D}}$; its square root, $\sqrt{\tilde{d}_{jj}}$, is also referred to as the Eicker-White standard error for $\hat{\beta}_{j-1}$.

Special case: When there is conditional homoskedasticity: $\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i) = \sigma_o^2$, $\boldsymbol{D}_o = \sigma_o^2 \boldsymbol{M}_{xx}^{-1}$ can be consistently estimated by

$$\widehat{\boldsymbol{D}} = \hat{\sigma}^2 \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1}, \quad \text{with} \quad \hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2.$$

In this case, $\tilde{\boldsymbol{D}} = \widehat{\boldsymbol{D}}/n = \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$, which is exactly the estimator $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ obtained earlier under Classical Assumption.

Remark: It is important to note that the estimator $\hat{\sigma}^2 (n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1}$ would not be consistent for D_o when conditional heteroskedasticity is present. Consequently, we would lose asymptotic normality if $\hat{\boldsymbol{\beta}}$ is normalized using this inconsistent covariance matrix estimator.

Testing A Single Hypothesis

Suppose we are testing only one parameter: $\boldsymbol{R}\boldsymbol{b}_o = b_2 = c$, with

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Recall that by asymptotic normality,

$$\tilde{\boldsymbol{D}}^{-1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),$$

where $\tilde{\boldsymbol{D}} = (\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}(\sum_{i=1}^n \hat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i')(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}$. It is easy to see that

$$(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1/2}[\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o)] \xrightarrow{D} \mathcal{N}(0, 1).$$

In this case, $\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}' = \tilde{d}_{33}$, the third diagonal element of $\tilde{\boldsymbol{D}}$.

Then under the null hypothesis, the statistic

$$(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c) = \frac{\hat{\beta}_2 - c}{\text{EW-se}(\hat{\beta}_2)} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\text{EW-se}(\hat{\beta}_2) = \sqrt{\tilde{d}_{33}}$ is the Eicker-White standard error of $\hat{\beta}_2$. Note that this is the t statistic standardized by the Eicker-White standard error and has the limiting distribution $\mathcal{N}(0,1)$.

Suppose the hypothesis involves two parameters: $\boldsymbol{R}\boldsymbol{b}_o = b_2 - b_3 = c$, with

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 1 & -1 & 0 & \cdots & 0 \end{pmatrix}.$$

In this case, one can verify $\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}' = \tilde{d}_{33} + \tilde{d}_{44} - 2\tilde{d}_{34}$, where $\tilde{d}_{ij}$ the (i,j) th term of $\tilde{\boldsymbol{D}}$. Then,

$$(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - c) = \frac{\hat{\beta}_2 - \hat{\beta}_3 - c}{\text{EW-se}(\hat{\beta}_2 - \hat{\beta}_3)} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\text{EW-se}(\hat{\beta}_2 - \hat{\beta}_3) = \sqrt{\tilde{d}_{33} + \tilde{d}_{44} - 2\tilde{d}_{34}}$.

Limiting Distribution of the t Statistic

Under the null hypothesis $\boldsymbol{R}\boldsymbol{b}_o = c$, where $\boldsymbol{R}$ is $1 \times (k+1)$, we have the t statistic:

$$\frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - c}{\text{EW-se}(\boldsymbol{R}\hat{\boldsymbol{\beta}})} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\text{EW-se}(\boldsymbol{R}\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{1/2}$ denotes the Eicker-White standard error.

Remark: As $\widehat{\boldsymbol{D}}$ is consistent for $\boldsymbol{D}_o$ when conditional heteroskedasticity is present, the t statistic based on the standard errors from $\tilde{\boldsymbol{D}}$ is said to be robust to conditional heteroskedasticity. Such t statistic uses the robust, Eicker-White standard error and is valid even when conditional heteroskedasticity is present with an unknown form. This is the statistic we should use in empirical studies with cross-section data.

A Special Case

When there is conditional homoskedasticity: $\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i) = \sigma_o^2$, $\boldsymbol{D}_o = \sigma_o^2 \boldsymbol{M}_{xx}^{-1}$. In this case, the Eicker-White covariance matrix estimator remains consistent but is not needed. Instead, we may use $\tilde{\boldsymbol{D}} = \hat{\sigma}^2 (\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i')^{-1}$, so that the t statistic is:

$$\frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - c}{\text{se}(\boldsymbol{R}\hat{\boldsymbol{\beta}})} = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - c}{\hat{\sigma}\sqrt{\boldsymbol{R}(\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i')^{-1}\boldsymbol{R}'}} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\text{se}(\boldsymbol{R}\hat{\boldsymbol{\beta}})$ is the conventional OLS standard error of $\boldsymbol{R}\hat{\boldsymbol{\beta}}$.

Remark: This t statistic is valid only when conditional homoskedasticity is present. This result is analogous to that under Classical Assumption.

Testing Multiple Hypotheses

Consider now the joint hypothesis $\boldsymbol{R}\boldsymbol{b}_o = \boldsymbol{c}$, where $\boldsymbol{R}$ is $q \times (k+1)$. Then under the null hypothesis,

$$(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q),$$

so that the limit contains q independent $\mathcal{N}(0,1)$ variables. The Wald statistic is the inner product of the left-hand side:

$$\begin{aligned}\mathcal{W} &= (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1/2}(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \\ &= (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}),\end{aligned}$$

and its limiting distribution is therefore the sum of q independent, squared $\mathcal{N}(0,1)$ variables, i.e., the $\chi^2(q)$ distribution.

We have established the following result.

Limiting Distribution of the Wald Statistic

Under the null hypothesis $\boldsymbol{R}\boldsymbol{b}_o = \boldsymbol{c}$, where $\boldsymbol{R}$ is $q \times (k+1)$, we have the Wald statistic:

$$(\boldsymbol{R}\hat{\beta} - \boldsymbol{c})'(\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\beta} - \boldsymbol{c}) \xrightarrow{D} \chi^2(q),$$

where $\tilde{\boldsymbol{D}} = (\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}(\sum_{i=1}^n \hat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i')(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}$.

Remark: The Wald statistic is robust to conditional heteroskedasticity when the Eicker-White covariance matrix estimator $\tilde{\boldsymbol{D}}$ is used in the statistic.

Example: Consider the hypothesis of 3 parameters being zero:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}, \quad \boldsymbol{R}\boldsymbol{b}_o = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

The Wald statistic is then

$$\mathcal{W} = \begin{pmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 \end{pmatrix} (\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}')^{-1} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} \xrightarrow{D} \chi^2(3).$$

You can easily check $\boldsymbol{R}\tilde{\boldsymbol{D}}\boldsymbol{R}'$ is a 3×3 matrix but not a diagonal matrix in this case; write down this matrix using the notations d_{ij} (homework).

A Special Case

When there is conditional homoskedasticity: $\mathbb{E}(\varepsilon^2|\boldsymbol{x}_i) = \sigma_o^2$, $\boldsymbol{D}_o = \sigma_o^2 \boldsymbol{M}_{xx}^{-1}$. In this case, $\tilde{\boldsymbol{D}} = \hat{\sigma}^2(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}$, so that

$$\mathcal{W} = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c})' \left[\boldsymbol{R}\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\boldsymbol{R}'\right]^{-1} \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{c}}{\hat{\sigma}^2} \xrightarrow{D} \chi^2(q).$$

- This Wald statistic is not robust to conditional heteroskedasticity, because $\hat{\sigma}^2(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}$ is not a consistent estimator when conditional heteroskedasticity is present.
- It can be seen that $\mathcal{W}/q$ is nothing but the conventional F statistic. Thus, F statistics are not robust to conditional heteroskedasticity. In practice, we may replace the F tests in Lecture 3 with the corresponding, robust Wald tests.

Example: Wage Regressions

The estimated wage model based on Taiwan's 2010 male data (11561 obs):

	3.8939	+ 0.0800 educ	+ 0.0166 exper	
OLS-se	(0.0198)	(0.0012)	(0.0003)	
t	(197.05)	(65.41)	(50.45)	
EW-se	(0.0208)	(0.0013)	(0.0004)	
t	(186.94)	(61.63)	(43.71)	
	$\bar{R}^2 = 0.2893$	$\hat{\sigma} = 0.3595$	Reg $F = 1992.4$	
	3.7902	+ 0.0779 educ	+ 0.0365 exper	$-$ 0.0005 exper^2
OLS-se	(0.0199)	(0.0012)	(0.0009)	(0.00002)
t	(190.60)	(64.77)	(38.72)	($-$22.47)
EW-se	(0.0207)	(0.0013)	(0.0010)	(0.00003)
t	(183.53)	(61.04)	(35.05)	($-$18.63)
	$\bar{R}^2 = 0.319$	$\hat{\sigma} = 0.3519$	Reg $F = 1600.2$	

Regression F statistics are based on White's standard errors.

Tests of Conditional Homoskedasticity

When there is no conditional heteroskedasticity, the asymptotic covariance matrix of the OLS estimator has a simpler form and can be consistently estimated by the conventional OLS covariance matrix estimator. It is thus desirable to first check if the data exhibit conditional homoskedasticity. Rejecting this hypothesis suggests that the robust, Eicker-White standard errors should be employed.

The well known Breusch-Pagan (BP) test is to test the null hypothesis of conditional homoskedasticity: $\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i, \boldsymbol{w}_i) = \sigma_o^2$, where the variables in $\boldsymbol{w}_i$ are from the information set but different from those of $\boldsymbol{x}_i$. The BP test is designed to test against the alternative that $\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i, \boldsymbol{w}_i)$ is an unknown function of $\boldsymbol{z}_i$ and hence changes with i, where $\boldsymbol{z}_i$ contains the elements of $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ that may affect the conditional variance of ε_i.

Breusch-Pagan Test

The Breusch-Pagan statistic is computed as follows:

1. Regress y_i on $\boldsymbol{x}_i$ to obtain the OLS residuals $\hat{u}_i$.
2. Regress $\hat{u}_i^2$ on constant one and $\boldsymbol{z}_i$ and obtain R^2_{aux}.
3. The BP statistic is $n\,R^2_{\text{aux}} \xrightarrow{D} \chi^2(m)$, where m is the number of elements in $\boldsymbol{z}_i$.

Remark:

- Intuitively, the residuals $\hat{u}_i$ approximate the true errors. When there is homoskedasticity, we expect that $\hat{u}_i^2$ do not depend on $\boldsymbol{z}_i$, so that the auxiliary regression in Step 2 above would have a small R^2_{aux} (even after multiplied by n). Otherwise, R^2_{aux} would be large, suggesting the null hypothesis is false.
- A drawback of the BP test is that the choice of $\boldsymbol{z}_i$ is subjective.

White Test

As for the choice of $\boldsymbol{z}_i$, the White test of White (1980) proposes to choose all non-constant regressors of $\boldsymbol{x}_i$, and their respective squares and pairwise products. The resulting test is also known as a test of conditional heteroskedasticity of unknown form. For example, when $\boldsymbol{x}_i$ contains the constant one and 3 variables x_{i1}, x_{i2}, x_{i3},

$$\boldsymbol{z}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i1}^2, x_{i2}^2, x_{i3}^2, x_{i1}x_{i2}, x_{i1}x_{i3}, x_{i2}x_{i3})'.$$

Then $n\,R^2_{\text{aux}}$ is distributed as $\chi^2(9)$ in the limit. A difficulty of the White test is that the number of variables in $\boldsymbol{z}_i$ may be too large; this may happen even when the original regression contains only a moderate number of regressors. Note that the White test was originally derived as an information matrix test.

Endogeneity

In linear regressions, a regressor is said to be endogenous if it is correlated with the regression error.

- Treatment effect model: The treatment indicator and outcome variable may both depend on some unobserved factors and hence are correlated. In this case, the OLS estimator is inconsistent
- Measurement error problem: For $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the regressor x_i can only be observed with random error: $w_i = x_i + v_i$. The operational model is:

$$y_i = \beta_0 + \beta_1 w_i + (\varepsilon_i - \beta_1 v_i),$$

with $\varepsilon_i - \beta_1 v_i$ the regression error. Clearly, w_i must be correlated with the regression error. Here, the random error v_i plays the same role as the unobserved factor in the case above.

- Simultaneous equation system: The following examples are taken from Chapter 3 of Hayashi (2000).
  1. Consider the following demand-supply system:

$$q_i = \alpha_0 + \alpha_1 p_i + u_i, \quad \text{(demand)}$$
$$q_i = \beta_0 + \beta_1 p_i + v_i. \quad \text{(supply)}$$

     Solving this system for p_i and q_i, we find p_i depend on both u_i and v_i and hence are endogenous in both equations.
  2. A simple economic system with the GNP identity: $y_i = c_i + g_i$, where c is consumption and g is investment, and the consumption function:

$$c_i = \alpha_0 + \alpha_1 y_i + u_i, \quad 0 < \alpha_1 < 1.$$

     It is easily calculated that

$$y_i = \frac{\alpha_0}{1-\alpha_1} + \frac{1}{1-\alpha_1} g_i + \frac{1}{1-\alpha_1} u_i,$$

     so that y_i are correlated with u_i.

Instrumental Variable Estimation

Suppose that $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_o + u_i$, where $\boldsymbol{x}_i$ is $(k+1) \times 1$ and contains some endogenous regressors such that $\mathbb{E}(\boldsymbol{x}_i u_i) \neq \boldsymbol{0}$. Let $\boldsymbol{z}_i$ be the $(k+1) \times 1$ vector that is correlated with $\boldsymbol{x}_i$ but uncorrelated with u_i, i.e., $\mathbb{E}(\boldsymbol{z}_i u_i) = \boldsymbol{0}$. The variables in $\boldsymbol{z}_i$ are referred to as the instrumental variables (IV) or simply the instruments. Instead of solving the normal equations $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}$ for $\boldsymbol{\beta}$, the IV method solves:

$$\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}.$$

This system is just identified in the sense that it contains $k+1$ equations with $k+1$ unknowns. Given that $\boldsymbol{Z}'\boldsymbol{X}$ is of full rank (so that it is invertible), the solution is the IV estimator:

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \left(\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{z}_i y_i\right).$$

It is easy to see that the IV estimator is consistent for $\boldsymbol{\beta}_o$:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{IV}} &= \left(\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{z}_i (\boldsymbol{x}_i' \boldsymbol{\beta}_o + u_i)\right) \\ &= \boldsymbol{\beta}_o + \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{x}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i u_i\right) \\ &\xrightarrow{\mathbb{P}} \boldsymbol{\beta}_o + [\mathbb{E}(\boldsymbol{z}_i \boldsymbol{x}_i')]^{-1} \mathbb{E}(\boldsymbol{z}_i u_i) = \boldsymbol{\beta}_o,\end{aligned}$$

because $\mathbb{E}(\boldsymbol{z}_i u_i) = \boldsymbol{0}$. As for the OLS asymptotics, we assume the CLT:

$$n^{-1/2} \sum_{i=1}^n \boldsymbol{z}_i u_i \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}), \quad \boldsymbol{V} = \mathbb{E}(u_i^2 \boldsymbol{z}_i \boldsymbol{z}_i').$$

The asymptotic distribution result for $\hat{\boldsymbol{\beta}}_{\text{IV}}$ is immediate:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{M}_{zx}^{-1} \boldsymbol{V} \boldsymbol{M}_{xz}^{-1}),$$

where $\boldsymbol{M}_{zx} := \mathbb{E}(\boldsymbol{z}_i \boldsymbol{x}_i') = \boldsymbol{M}_{xz}'$.

The asymptotic covariance matrix $\boldsymbol{D}_o = \boldsymbol{M}_{zx}^{-1}\boldsymbol{V}\boldsymbol{M}_{xz}^{-1}$ can be estimated using the Eicker-White-type estimator:

$$\widehat{\boldsymbol{D}} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{z}_i\boldsymbol{z}_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{z}_i'\right)^{-1},$$

where $\hat{u}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\text{IV}}$ are the IV residuals. As before,

$$\tilde{\boldsymbol{D}} = \widehat{\boldsymbol{D}}/n = \left(\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{z}_i\boldsymbol{z}_i'\right)\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{z}_i'\right)^{-1},$$

so that

$$\tilde{\boldsymbol{D}}^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}_o) = \widehat{\boldsymbol{D}}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

The Eicker-White standard errors for $\hat{\boldsymbol{\beta}}_{\text{IV}}$ are taken form $\tilde{\boldsymbol{D}}$ and robust to potential conditional heteroskedasticity. The Wald tests can be conducted as in the OLS context.

Remarks:

- In practice, $\boldsymbol{x}_i$ may contain only one endogenous variable. For example, write $\boldsymbol{x}_i = (\boldsymbol{\xi}_i' \; x_{i1})'$, where $\boldsymbol{\xi}_i = (1 \; x_{i2} \; \ldots \; x_{ik})'$ contains all exogenous regressors such that $\mathbb{E}(\boldsymbol{\xi}_i u_i) = \mathbf{0}$, and x_{i1} is endogenous such that $\mathbb{E}(x_{i1} u_i) \neq 0$. We can set $\boldsymbol{z}_i = (\boldsymbol{\xi}_i' \; \zeta_i)$, where ζ_i is an IV with $\mathbb{E}(\zeta_i u_i) = 0$. That is, only one IV is needed for this endogenous variable. Finding a proper IV is practically challenging, however.
- The condition that $\boldsymbol{Z}'\boldsymbol{X}$ is of full rank requires that the variables in $\boldsymbol{Z}$ must be correlated with the variables in $\boldsymbol{X}$. In other words, we cannot use an arbitrary but irrelevant variable as the instrument.
- There may exist multiple instruments for a single endogenous variable. In that case, $\boldsymbol{Z}$ is $n \times \ell$ with $\ell > k+1$, and the system $\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\beta) = \mathbf{0}$ is over-identified, and the IV estimator cannot be computed. We will discuss this estimation problem later.

An Alternative Expression

Consider the simple linear regression: $y_i = \beta_0 + \beta_1 x_i + u_i$, with x_i endogenous. It is readily verified that the IV estimator for β_1 is (check):

$$\hat{\beta}_{1,\text{IV}} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}.$$

The IV estimator can be expressed as the ratio of two regression coefficients:

$$\hat{\beta}_{1,\text{IV}} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y}) / \sum_{i=1}^n (z_i - \bar{z})^2}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})' / \sum_{i=1}^n (z_i - \bar{z})^2},$$

where the numerator is the slope of the regression of y on 1 and z, and the denominator is the slope of the regression of x on 1 and z.

To see why it is the case, write the regression for the denominator term of $\hat{\beta}_{1,\text{IV}}$ as:

$$x_i = \gamma_0 + \gamma_1 z_i + v_i.$$

Plugging this into the original model: $y_i = \beta_0 + \beta_1 x_i + u_i$, we obtain the so-called "reduced form" (model without endogenous regressor):

$$y_i = (\beta_0 + \beta_1\gamma_0) + \beta_1\gamma_1 z_i + (\beta_1 v_i + u_i).$$

which is the regression for the numerator term of $\hat{\beta}_{1,\text{IV}}$. While the numerator of the IV estimator provides an estimate of $\beta_1\gamma_1$, the denominator yields an estimate of γ_1. Their ratio is thus an estimate of β_1, as it ought to be.

Two-Stage Least Squares Estimation

Again consider the case $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_o + u_i$, where $\boldsymbol{x}_i = (\boldsymbol{\xi}_i' \; x_{i1})'$ with x_{i1} endogenous and $\boldsymbol{\xi}_i$ exogenous, and $\boldsymbol{\beta}_o = (\boldsymbol{b}' \; \beta_1)'$. We may first regress x_{i1} on $\boldsymbol{\xi}_i$ and z_i to obtain the fitted values $\hat{x}_{i1}$; here, z_i may contain multiple instruments. Plugging the fitted values into the original model we have

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_o + u_i = \boldsymbol{\xi}_i'\boldsymbol{b} + \beta_1\hat{x}_{i1} + [\beta_1(x_{i1} - \hat{x}_{i1}) + u_i].$$

This suggests the second-stage regression of y_i on $\boldsymbol{\xi}_i$ and $\hat{x}_{i1}$; note that $\boldsymbol{\xi}_i$ are uncorrelated with u_i, $\sum_{i=1}^n \boldsymbol{\xi}_i'(x_{i1} - \hat{x}_{i1}) = 0$, and $\sum_{i=1}^n \hat{x}_{i1}(x_{i1} - \hat{x}_{i1}) = 0$. The resulting estimators of $\boldsymbol{b}$ and β_1 are known as the two-stage least squares (2SLS) estimators.

We now focus on the case that $\boldsymbol{Z}$ is the $n \times \ell$ matrix of instruments with $\ell > k+1$, i.e., the number of instruments is greater than the number of parameters. The system, $\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\beta) = \boldsymbol{0}$, is over-identified and cannot be solved. Instead, we solve the following just-identified system:

$$\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\beta) = \boldsymbol{0},$$

and obtain the 2SLS estimator:

$$\hat{\beta}_{\text{2SLS}} = [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}].$$

Let $\boldsymbol{P}_Z = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, an orthogonal projection matrix. The first-stage regression of $\boldsymbol{X}$ on $\boldsymbol{Z}$ leads to the matrix of fitted values: $\widehat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{P}_Z\boldsymbol{X}$. Then, the second-stage regression of $\boldsymbol{y}$ on $\widehat{\boldsymbol{X}}$ yields the estimator:

$$(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}})^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{y},$$

which is exactly the 2SLS estimator $\hat{\beta}_{\text{2SLS}}$.

Asymptotic Properties of the 2SLS Estimator

Given $\boldsymbol{y} = \boldsymbol{X}\beta_o + \boldsymbol{u}$, we have

$$\begin{aligned}\hat{\beta}_{\text{2SLS}} &= [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}] \\ &= [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{X}\beta_o + \boldsymbol{u})] \\ &= \beta_o + [(\boldsymbol{X}'\boldsymbol{Z}/n)(\boldsymbol{Z}'\boldsymbol{Z}/n)^{-1}(\boldsymbol{Z}'\boldsymbol{X}/n)]^{-1} \\ &\qquad [(\boldsymbol{X}'\boldsymbol{Z}/n)(\boldsymbol{Z}'\boldsymbol{Z}/n)^{-1}(\boldsymbol{Z}'\boldsymbol{u}/n)].\end{aligned}$$

By suitable WLLN, we have seen that $\boldsymbol{X}'\boldsymbol{Z}/n \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{x}_i\boldsymbol{z}_i') =: \boldsymbol{M}_{xz}$, $\boldsymbol{Z}'\boldsymbol{X}/n \xrightarrow{\mathbb{P}} \boldsymbol{M}_{xz}' =: \boldsymbol{M}_{zx}$, $\boldsymbol{Z}'\boldsymbol{Z}/n \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{z}_i\boldsymbol{z}_i') =: \boldsymbol{M}_{zz}$, and

$$\boldsymbol{Z}'\boldsymbol{u}/n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i u_i \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{z}_i u_i) = \boldsymbol{0}.$$

It follows that $\hat{\beta}_{\text{2SLS}} \xrightarrow{\mathbb{P}} \beta_o$.

Assuming the CLT: $n^{-1/2}\sum_{i=1}^{n} \boldsymbol{z}_i u_i \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$, it is readily verified that (check):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{2SLS}} - \boldsymbol{\beta}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}_o),$$

with $\boldsymbol{D}_o = (\boldsymbol{M}_{xz}\boldsymbol{M}_{zz}^{-1}\boldsymbol{M}_{zx})^{-1}(\boldsymbol{M}_{xz}\boldsymbol{M}_{zz}^{-1}\boldsymbol{V}\boldsymbol{M}_{zz}^{-1}\boldsymbol{M}_{zx})(\boldsymbol{M}_{xz}\boldsymbol{M}_{zz}^{-1}\boldsymbol{M}_{zx})^{-1}$.
The asymptotic covariance matrix can be estimated by its sample counterpart $\widehat{\boldsymbol{D}}$. The robust standard errors for $\hat{\boldsymbol{\beta}}_{\text{2SLS}}$ are taken from:

$$\tilde{\boldsymbol{D}} = \widehat{\boldsymbol{D}}/n = [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1} \\ [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\widehat{\boldsymbol{V}}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}][\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1},$$

where $\widehat{\boldsymbol{V}} = \sum_{i=1}^{n} \hat{u}_i^2 \boldsymbol{z}_i \boldsymbol{z}_i'$, with $\hat{u}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\text{2SLS}}$.

Remark: The asymptotic covariance matrix must be estimated directly from the formula above; the covariance matrix estimator from the second-stage regression is not a correct estimator for $\boldsymbol{D}_o$.