Lecture 4
Multiple Linear Regression: Asymptotics

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

March 1, 2020

Lecture Outline

1. Mulple 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

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's 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 $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ need not be normally distributed even when $\boldsymbol{y}$ is. In fact, $\hat{\beta}$ 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)

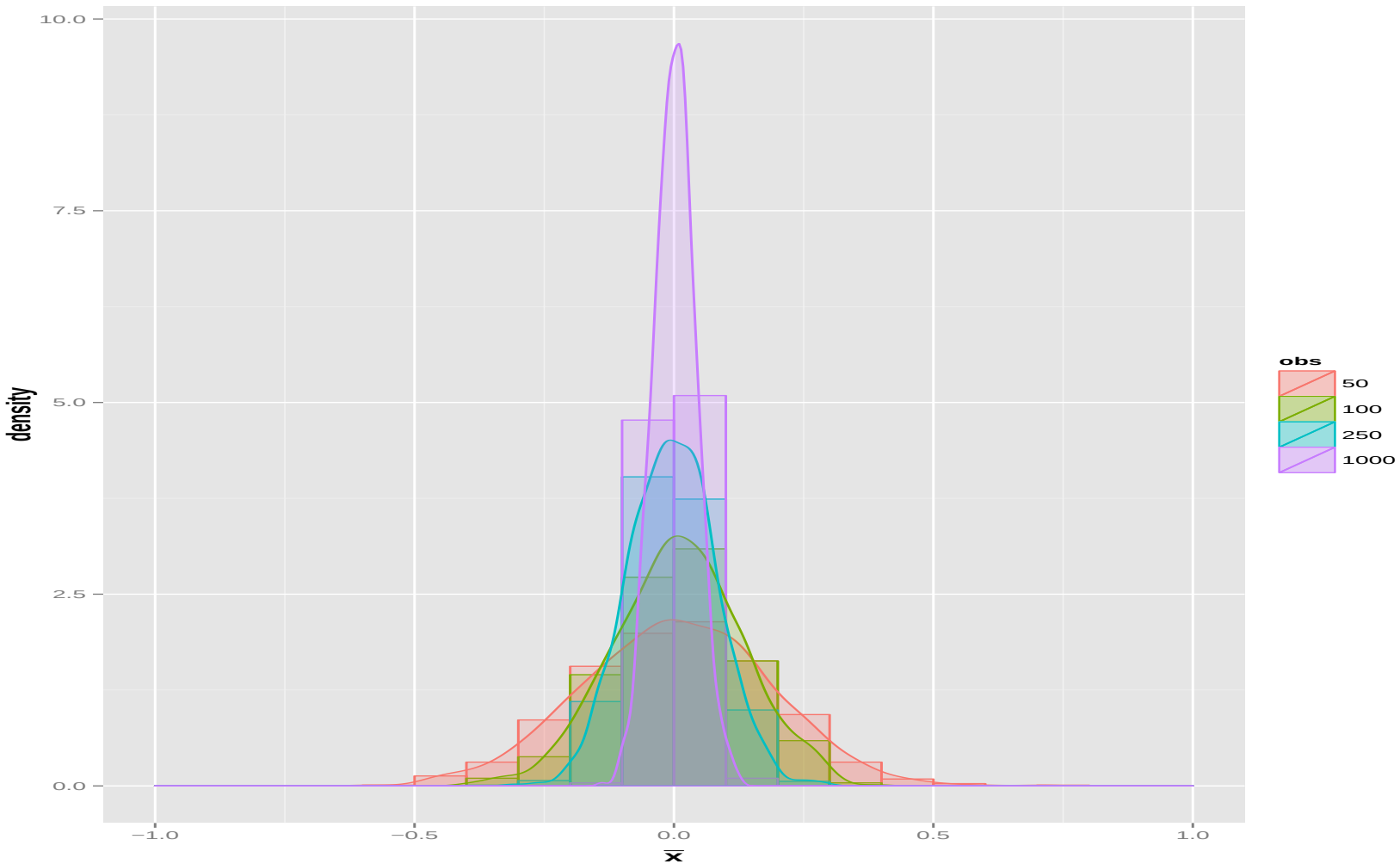
Let $z_1, \ldots, z_n$ be i.i.d. with mean μ_o. Then, $\bar{z}_n = n^{-1} \sum_{i=1}^n z_i$ converges in probability to μ_o, in the sense that, for any $\varepsilon > 0$,

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

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



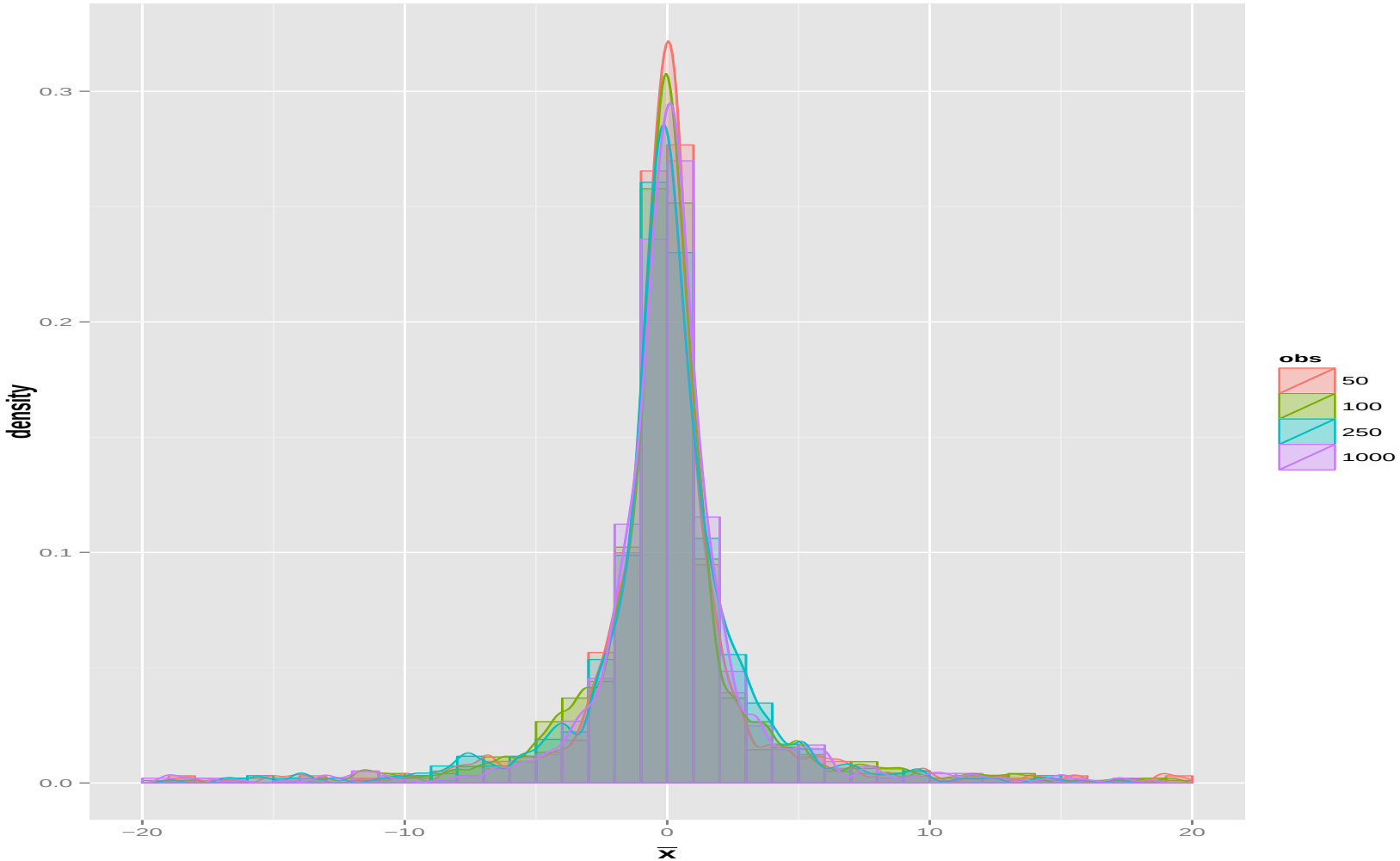
Note: i.i.d. $t(5)$ variables with 1000 replications.

Simulation Procedure

1. Generate a sample of n ($n = 100$ for example) realizations z_i from $t(5)$ distribution 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 resulting frequency plot of these $\bar{z}_n$ is the simulated (finite-sample) density function of $\bar{z}_n$ under $t(5)$.

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

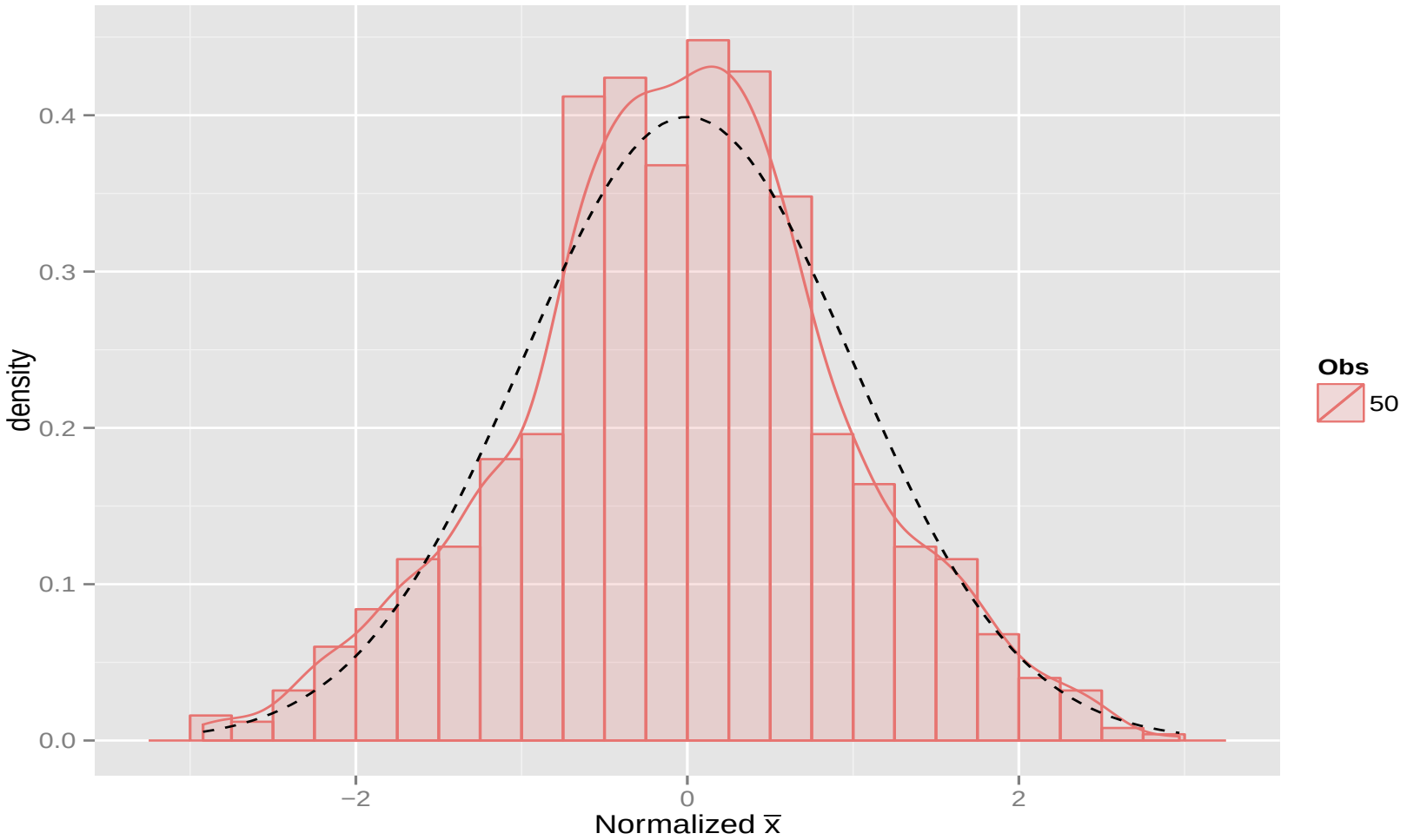
$$\frac{(\bar{z}_n - \mu_o)}{\sigma_o/\sqrt{n}} = \frac{\sqrt{n}(\bar{z}_n - \mu_o)}{\sigma_o},$$

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

Remarks:

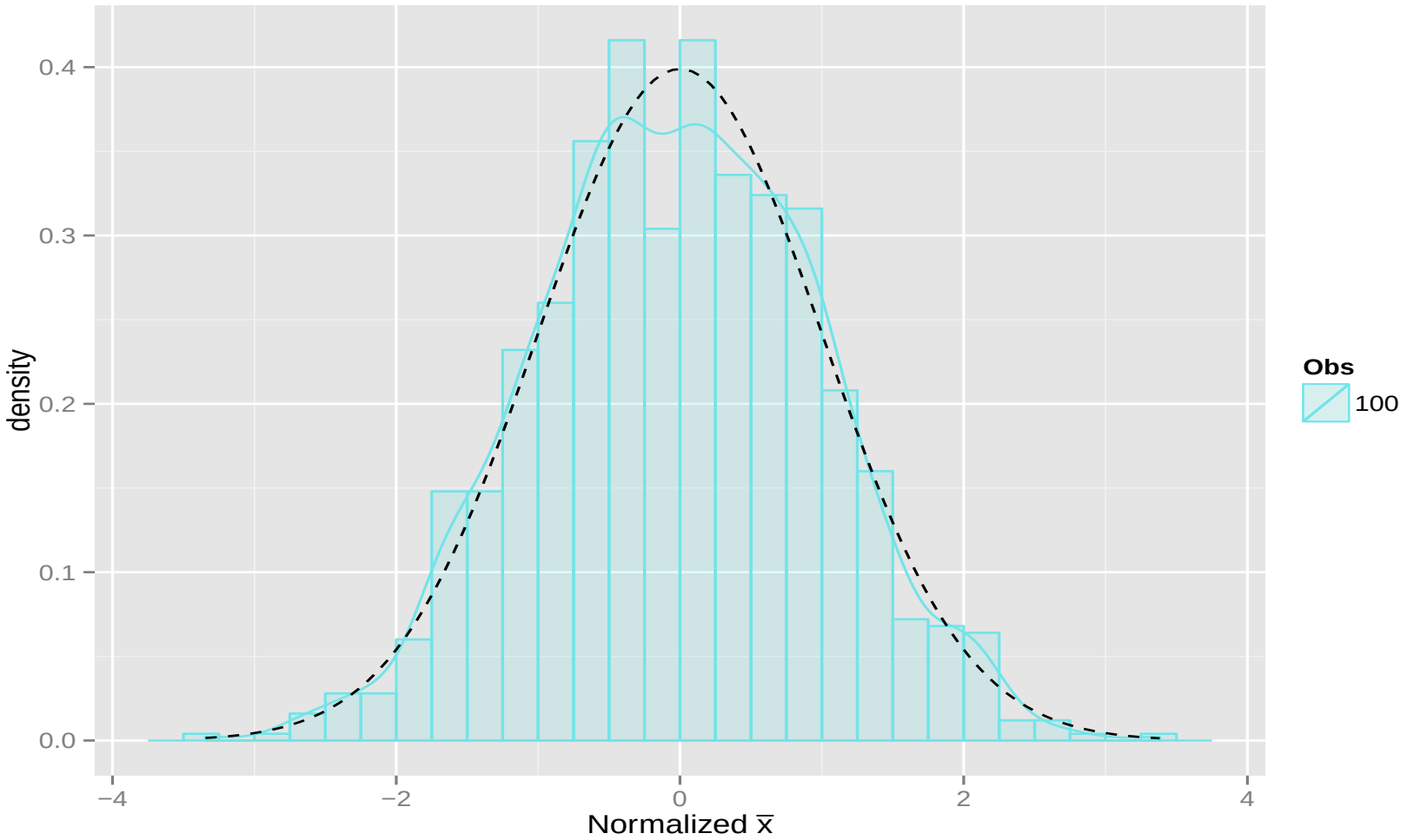
- i.i.d. random variables need not obey this CLT if they do not have a finite variance, e.g., $t(2)$ variables.
- Both LLN and CLT may hold for non-i.i.d. random variables under stronger conditions (e.g., higher order moments exist).

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



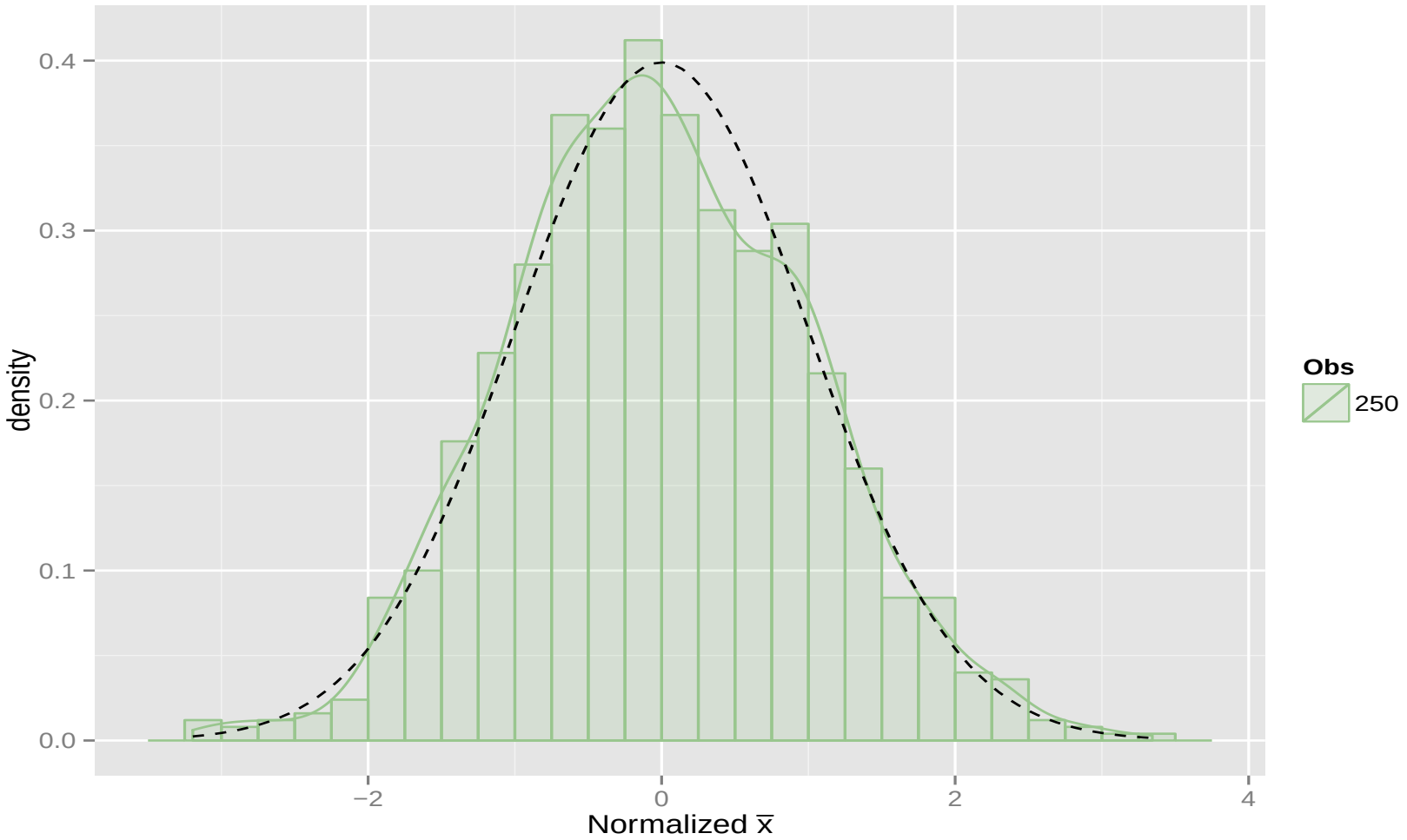
Note: i.i.d. $t(5)$ variables with 1000 replications.

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



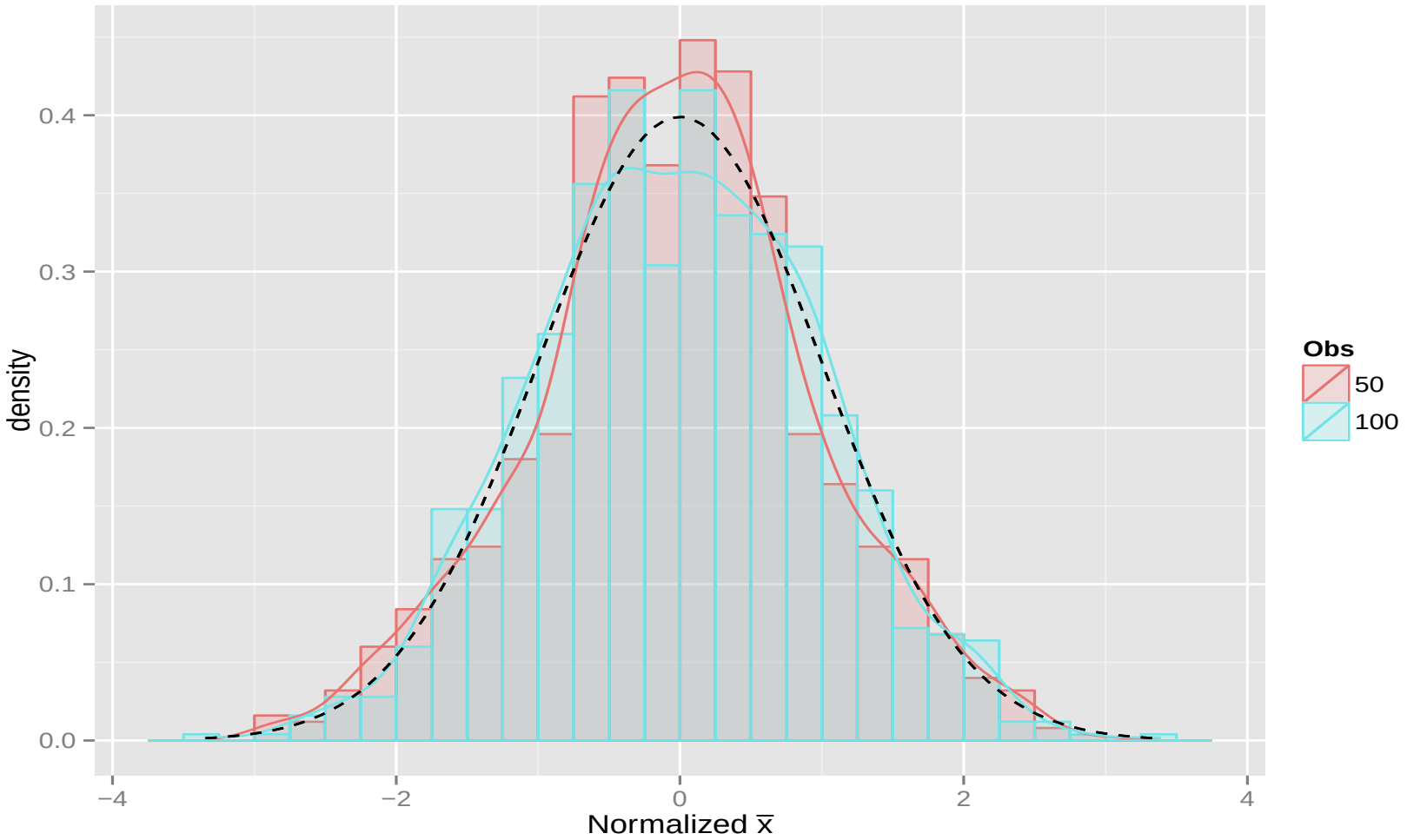
Note: i.i.d. $t(5)$ variables with 1000 replications.

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



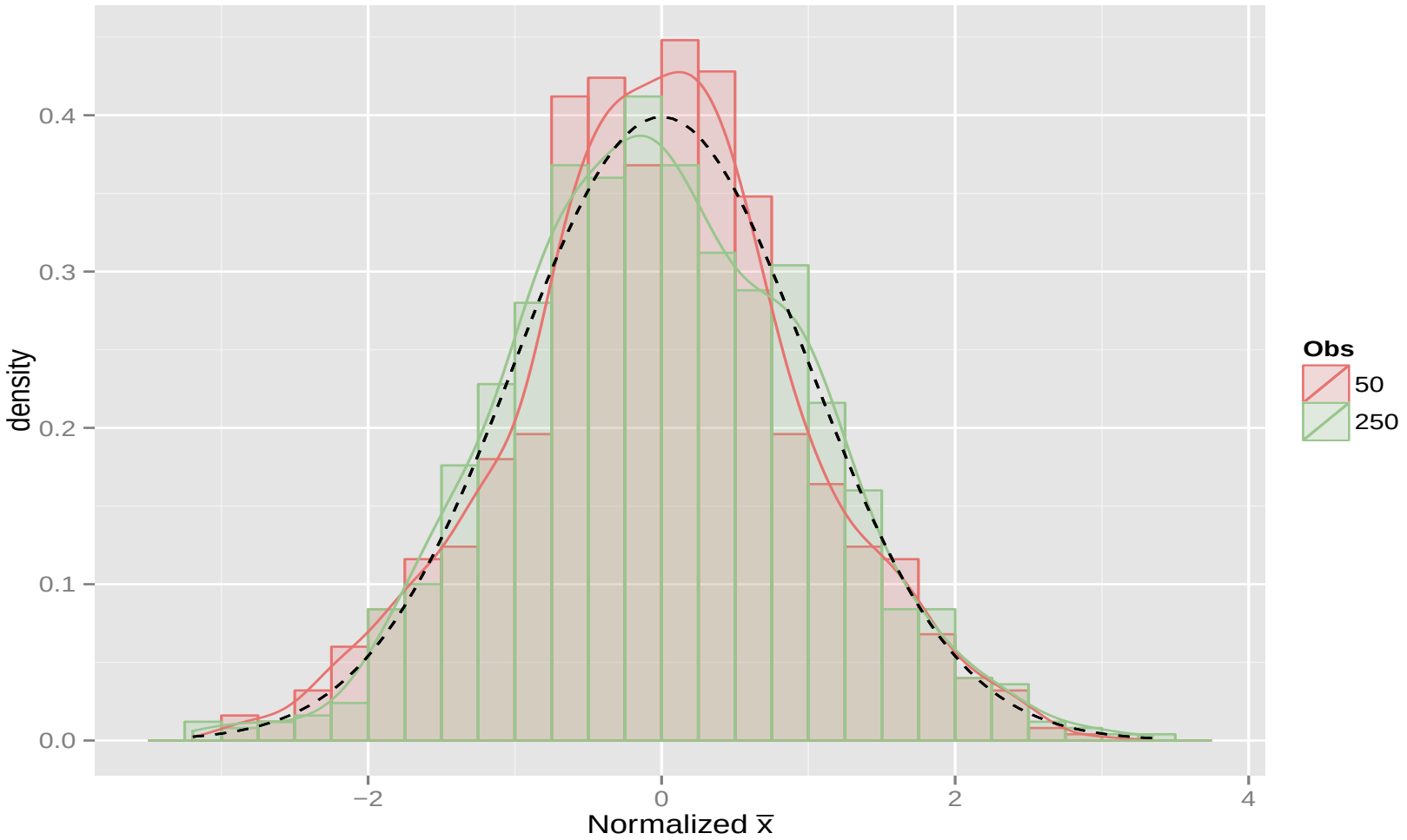
Note: i.i.d. $t(5)$ variables with 1000 replications.

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.

Consistency

Modern Assumption 0

The random variables y_i, $i = 1, \ldots, n$, follow the population model:

$$y_i = b_0 + b_1 x_i + u_i,$$

for some numbers b_0, b_1, where (i) $(x_i, y_i)'$, $i = 1, \ldots, n$, are i.i.d. such that x_i has mean μ_x and variance σ_x^2, (ii) $\mathbb{E}(u_i) = 0$, and $\mathbb{E}(x_i u_i) = 0$.

For the simple regression specification: $y_i = \beta_0 + \beta_1 x_i + u_i$,

$$\hat{\beta}_1 = b_1 + \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) u_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Then, $\hat{\beta}_1 \xrightarrow{\mathbb{P}} b_1$ when the second term on the right converges to zero.

Assumption (i) ensures that x_i obey WLLN, so that the sample average converges in probability to the true mean: $\bar{x} \xrightarrow{\mathbb{P}} \mu_x$. Also, the sample variance converges in probability to the true variance:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x+\mu_x-\bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2+(\mu_x-\bar{x})^2 \\ &\quad + 2(\mu_x-\bar{x})\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x),\end{aligned}$$

where the 2nd and 3rd terms all converge to zero in probability, and the first term converges to $\mathbb{E}(x_i-\mu_x)^2=\sigma_x^2$. Assumption (ii) and WLLN ensure that $n^{-1}\sum_{i=1}^{n} x_i u_i \xrightarrow{\mathbb{P}} \mathbb{E}(x_i u_i)=0$ and

$$\frac{1}{n}\sum_{i=1}^{n} u_i = \frac{1}{n}\sum_{i=1}^{n}(y_i-b_0-b_1 x_i) \xrightarrow{\mathbb{P}} \mathbb{E}(u_i)=0.$$

Combining the results above we have

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})u_i = \frac{1}{n}\sum_{i=1}^{n} x_i u_i - \bar{x}\left(\frac{1}{n}\sum_{i=1}^{n} u_i\right) \xrightarrow{\mathbb{P}} 0.$$

Consequently,

$$\hat{\beta}_1 \xrightarrow{\mathbb{P}} b_1 + 0/\sigma_x^2 = b_1.$$

It is also easy to show $\hat{\beta}_0 \xrightarrow{\mathbb{P}} b_0$ (homework). We have proved the following result.

OLS Consistency

Given the simple regression, $y_i = \beta_0 + \beta_1 x_i + u_i$, suppose that Modern Assumption 0(i) and (ii) hold. Then, $\hat{\beta}_0 \xrightarrow{\mathbb{P}} b_0$ and $\hat{\beta}_1 \xrightarrow{\mathbb{P}} b_1$. That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are weakly consistent for b_0 and b_1, respectively.

- Consistency hinges on the condition $\mathbb{E}(x_i u_i) = 0$, i.e., u_i and x_i are uncorrelated. That is, 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 the sample size becomes larger (i.e., when more information becomes available), it is more likely that the OLS estimates will be close to the true parameter values. Noted that, while unbiasedness does not depend on sample size n, consistency is a property that holds when n becomes infinitely large.

Consistency in Matrix Notations

Let $\boldsymbol{\xi}_i = (x_{i1} \ \ldots \ x_{ik})'$ be the i th observation of k regressors and $\boldsymbol{x}_i = (1 \ \boldsymbol{\xi}_i')' = (1 \ x_{i1} \ \ldots \ x_{ik})'$ be the column vector containing constant one and $\boldsymbol{\xi}_i$. That is, $\boldsymbol{x}_i$ is the i th column of $\boldsymbol{X}'$, so that the linear specification $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$ can be written as:

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

Modern Assumption I

The random variables y_i, $i = 1, \ldots, n$, follow the population model: $y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + u_i$, $\boldsymbol{b}_o = (b_0 \ b_1 \ \ldots \ b_k)'$, where (i) $(\boldsymbol{\xi}_i' \ y_i)'$, $i = 1, \ldots, n$, are i.i.d. random vectors such that $\boldsymbol{\xi}_i$ has mean $\boldsymbol{\mu}_x$ and variance $\boldsymbol{\Sigma}_x$, and (ii) $\mathbb{E}(\boldsymbol{x}_i u_i) = \boldsymbol{0}$.

Recall that the OLS estimator is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$, where $\boldsymbol{X}'\boldsymbol{X}$ is a $(k+1) \times (k+1)$ matrix and can be expressed as:

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

and $\boldsymbol{X}'\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{x}_i y_i$. Given Modern Assumption I,

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

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

To show this, we write

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

By WLLN, $n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i u_i \xrightarrow{\mathbb{P}} \mathbb{E}(\boldsymbol{x}_i u_i) = \boldsymbol{0}$, 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

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

OLS Consistency

Given the multiple linear regression, $y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + u_i$, suppose that Modern Assumption I(i) and (ii) hold. Then, $\hat{\boldsymbol{\beta}} \xrightarrow{\mathbb{P}} \boldsymbol{b}_o$; that is $\hat{\boldsymbol{\beta}}$ is weakly consistent for $\boldsymbol{b}_o$.

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 follow the population model: $y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + \boldsymbol{z}_i'\boldsymbol{d}_o + u_i$, where u_i satisfy Modern Assumption I(ii) such that $\mathbb{E}(\boldsymbol{x}_i u_i) = \boldsymbol{0}$. That is, we exclude ℓ important variables $\boldsymbol{z}$ from the specification. In this case, we can write:

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

Under WLLN,

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

It follows that

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

This shows that $\hat{\boldsymbol{\beta}}$ is inconsistent for $\boldsymbol{b}_o$ when important variables in $\boldsymbol{z}$ are omitted. There is an exception. When $\boldsymbol{z}$ is orthogonal to the vector of included variables $\boldsymbol{x}$, i.e., $\mathbb{E}(\boldsymbol{x}_i \boldsymbol{z}_i') = \mathbf{0}$ (this rarely happens), consistency still obtains:

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

Thus, when $\boldsymbol{x}$ and $\boldsymbol{z}$ are orthogonal, the probability limit of $\hat{\boldsymbol{\beta}}$ is not affected by the presence or absence of $\boldsymbol{z}$ in the linear specification.

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 follow the population model: $y_i = \boldsymbol{x}_i'\boldsymbol{b}_o + u_i$, where u_i satisfy Modern Assumption I(ii) such that $\mathbb{E}(\boldsymbol{x}_i u_i) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{z}_i u_i) = 0$. That is, we include ℓ irrelevant variables $\boldsymbol{z}$ in our specification. As discussed before, the population model can be written as

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

with $\mathbf{0}$ as the true parameter associated with $\boldsymbol{z}_i$. Consequently, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{b}_o$, and $\hat{\boldsymbol{\delta}}$ is consistent for $\mathbf{0}$.

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, $i = 1, \ldots, n$, follow the population model: $y_i = \boldsymbol{x}_i' \boldsymbol{b}_o + u_i$, where $\boldsymbol{b}_o = (b_0 \ b_1 \ \ldots \ b_k)'$, where (iii) for $\boldsymbol{V} = \mathrm{var}(\boldsymbol{x}_i u_i)$,

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

- Under Modern Assumption I, $\boldsymbol{x}_i u_i$ are i.i.d. random vectors with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}$. Thus, Modern Assumption I(iii) above ensures that $\boldsymbol{x}_i u_i$ satisfy a multivariate CLT.

Recall

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

We can then write

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

While $n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{\mathbb{P}} \boldsymbol{M}_{xx}$, Modern Assumption I(iii) gives:

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

It follows that

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

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. Thus, $\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}^{1/2}\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is a linear transformation of $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and has a multivariate normal distribution with mean $\boldsymbol{0}$ 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 multiple linear regression, $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$, suppose that Modern Assumption I(i), (ii) and (iii) hold. Then,

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

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

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

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

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 $\mathbb{E}(u_i^2|\boldsymbol{x}_i)$ depend on $\boldsymbol{x}_i$, i.e., u_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 still able to show that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o)$ is normally distributed in the limit. This 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 $\boldsymbol{D} := \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}\boldsymbol{M}_{xx}^{-1}$. By WLLN, $\boldsymbol{M}_{xx}$ can be consistently estimated by

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

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

$$\widehat{\boldsymbol{V}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\boldsymbol{x}_i\boldsymbol{x}_i', \qquad \hat{u}_i \text{ the OLS residuals.}$$

$\widehat{\boldsymbol{V}}$ is known as the heteroskedasticity-consistent covariance matrix estimator because $\boldsymbol{V}$ permits conditional heteroskedasticity. A weakly consistent estimator of $\boldsymbol{D}$ is the following Eicker-White estimator:

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

As $\boldsymbol{D}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}_o) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, we can replace $\boldsymbol{D}$ by its consistent estimator $\widehat{\boldsymbol{D}}$ and obtain:

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

Letting

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

we have $\tilde{\boldsymbol{D}}^{-1/2} = \widehat{\boldsymbol{D}}^{-1/2}\sqrt{n}$, so that

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

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: $\mathbb{E}(u_i^2|\boldsymbol{x}_i) = \sigma_o^2$, $\boldsymbol{D} = \sigma_o^2 \boldsymbol{M}_{xx}^{-1}$ 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,

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

which is exactly the estimator $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ obtained earlier under Classical Assumption. This should not be surprising because $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ was derived under the assumption of homoskedasticity (and non-random $\boldsymbol{X}$).

Under Modern Assumption with conditional homoskedasticity, we have

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

The standard error for $\hat{\beta}_{j-1}$ is $\hat{\sigma}$ times $\sqrt{m^{jj}}$, the square root of the j^{th} diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Remark: When the estimator $\hat{\sigma}^2(n^{-1}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i')^{-1}$ is used but conditional heteroskedasticity is present, this estimator would not be consistent for $\boldsymbol{D}$. Consequently, we lose asymptotic normality because $\hat{\boldsymbol{\beta}}$ is normalized by an incorrect (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 a 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 $\boldsymbol{D}$ 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 differs from the conventional t statistic in that the former uses the robust, Eicker-White standard error. This is the statistic we should use in empirical studies with cross-section data.

A Special Case

When there is conditional homoskedasticity: $\mathbb{E}(u_i^2|\boldsymbol{x}_i) = \sigma_o^2$, $\boldsymbol{D} = \sigma_o^2 \boldsymbol{M}_{xx}^{-1}$. In this case, the Eicker-White covariance matrix estimator, though still consistent, 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. It would not have the limiting normal distribution if there is conditional heteroskedasticity.

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:

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

because the limit is the sum of q independent, squared $\mathcal{N}(0,1)$ variables. The Wald statistic is robust to conditional heteroskedasticity when the Eicker-White covariance matrix estimator is used.

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: $\mathbb{E}(u_i^2|\boldsymbol{x}_i) = \sigma_o^2$, $\boldsymbol{D} = \sigma_o^2 \boldsymbol{M}_{xx}^{-1}$. 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 $\mathcal{W}/q$ here is nothing but the conventional F statistic. Thus, F statistics are also 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: $\mathbb{E}(u_i^2|\boldsymbol{x}_i, \boldsymbol{w}_i) = \sigma_o^2$, against the alternative that $\mathbb{E}(u_i^2|\boldsymbol{x}_i, \boldsymbol{w}_i)$ is a function of $\boldsymbol{z}_i$, where $\boldsymbol{w}_i$ is a vector of variables from the information set but different from $\boldsymbol{x}_i$, $\boldsymbol{z}_i$ contains the elements of $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ that may affect the conditional variance of u_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 u_i. When there is homoskedasticity, we expect that $\hat{u}_i^2$ do not depend on $\boldsymbol{z}_i$ and hence yield a very small R^2_{aux} (even after multiplied by n). Otherwise, R^2_{aux} ought to be larger, suggesting the null hypothesis is false.
- In practice, the choice of the elements of $\boldsymbol{z}_i$, i.e., the variables that may affect the conditional variance of u_i, is usually subjective.

White Test

One way to determine $\boldsymbol{z}_i$ is to choose all (or some) variables of $\boldsymbol{\xi}_i$ as the elements of $\boldsymbol{z}_i$. Another well known choice is to set, for example,

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

when $\boldsymbol{\xi}_i$ contains 3 variables x_{i1}, x_{i2}, x_{i3}. Then $n\, R^2_{\text{aux}}$ is distributed as $\chi^2(9)$ in the limit. That is, the elements of $\boldsymbol{z}_i$ include all variables of $\boldsymbol{\xi}_i$, their squares and their pairwise products. The resulting BP test is also known as the White test of conditional heteroskedasticity. A difficulty of the White test is that the number of variables in $\boldsymbol{z}_i$ may be too large even when the original regression contains only a moderate number of regressors.