Advanced Econometrics I

Jürgen Meinecke

Week 3 Lecture

Research School of Economics, Australian National University

Roadmap

Ordinary Least Squares Estimation

- Basic Asymptotic Theory (part 2 of 2)
- Asymptotic Distribution of the OLS Estimator
- Asymptotic Variance Estimation
- Standard Errors and Confidence Intervals [next week?]
- Hypotheses Tests [next week?]

Let there be a probability space $(\Omega, \mathscr{F}, P)$

- Ω is the outcome space
- $\mathscr{F}$ collects events from Ω
- P is a probability measure on $\mathscr{F}$

Example (Only Looks Like Rolling a Die)

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathscr{F} = \{\{1, 3, 5\}, \{2, 4, 6\}, \Omega, \varnothing\}$
- *Consider all* $A \in \mathscr{F}$

$$P(A) = \begin{cases} 0 & \text{if } A = \varnothing \\ 1/2 & \text{if } A = \{1, 3, 5\} \\ 1/2 & \text{if } A = \{2, 4, 6\} \\ 1 & \text{if } A = \Omega \end{cases}$$

Notice that $P(\{2\})$ is not specified

Definition (Random Variable—first attempt)

A **random variable** on $(\Omega, \mathscr{F})$ is a function $Z : \Omega \to \mathbb{R}$.

Example

$$X(\omega) = \begin{cases} 18 & \text{if } \omega \text{ even}, \\ 24 & \text{if } \omega \text{ odd} \end{cases}$$

Induced probability $\Pr(X = 18) := P(\{2, 4, 6\}) = 1/2$

Instead of writing $\Pr(X = 18)$ I will use $P(X = 18)$

Example

$$Y(\omega) = \begin{cases} 2 & \text{if } \omega = 6, \\ 7 & \text{if } \omega = 1 \end{cases}$$

Induced probability $\Pr(Y = 2) := P(\{6\}) = ?$

The event $\{6\}$ is not assigned a probability

Of course we have a reasonable suspicion that $P(\{6\})$ should equal $1/6$, but strictly speaking this hasn't been defined two slides earlier

So we have to treat $P(\{6\})$ as unknown

To make sure that our random variable is not ill-defined like this we need to rule out such situations

Here's a more robust definition

Definition (Random Variable—second and final attempt)

A **random variable** on $(\Omega, \mathscr{F})$ is a function $Z : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : Z(w) \in B\} \in \mathscr{F} \qquad \text{for all } B \in \mathscr{B}(\mathbb{R}).$$

$\mathscr{B}(\mathbb{R})$ is the σ-algebra generated by the closed intervals $[a, b]$, for $a, b \in \mathbb{R}$

Intuition: $\mathscr{B}(\mathbb{R})$ describes all events that can be created out of all the points on the real line

$\mathscr{B}(\mathbb{R})$ is a rich set containing pretty much every subset of $\mathbb{R}$ that we will ever be dealing with (including intervals, points)

I don't need you to understand all intricacies here

Bottom line is:
The image $Z(w)$ gets pulled back to an element of $\mathscr{F}$ for which probabilities are well-defined

Using this more robust definition, Y is not a random variable

To see this, pick the subset $B = \{2\}$ from $\mathscr{B}(\mathbb{R})$

- pick $B = \{2\}$
- $\{\omega \in \Omega : Y(\omega) = 2\} = \{6\} \notin \mathscr{F}$
- same for $B = \{7\}$

The problem here is that Y is not $\mathscr{F}$-*measurable*

Definition (Distribution or Law)

Given a random variable Z on a probability space $(\Omega, \mathscr{F}, P)$, the **distribution** or **law** of the random variable is the probability measure defined by

$$\mu(B) := P(Z \in B), \qquad B \in \mathscr{B}(\mathbb{R}).$$

We say that μ is the distribution of Z, or $\mathscr{L}(Z)$ is the law of Z.

Definition (Distribution Function)

The **distribution function** of a random variable Z is defined by

$$F(z) := \mu((-\infty, z]) = P(Z \leq z), \qquad z \in \mathbb{R}.$$

F is also referred to as cumulative distribution function or cdf.

There is a one-to-one correspondence between distribution and cdfs

So we use them interchangeably

Definition (Weak Convergence)

Let F be a distribution function, and $\{F_N\}$ be a sequence of distribution functions. Then F_N **converges weakly** to F if $\lim_{N\to\infty} F_N(z) = F(z)$ for each z at which F is continuous.

We write $F_N \xrightarrow{\text{w}} F$.

Equivalently we could say $\mu_N \xrightarrow{\text{w}} \mu$ for weak convergence

Definition (Convergence in Distribution)

Let Z be a random variable, and $\{Z_N\}$ be a sequence of random variables. Then Z_N **converges in distribution or law** to Z if $F_N \xrightarrow{\text{w}} F$.

We write $Z_N \xrightarrow{\text{d}} Z$.

Now we turn to a few practical results that will help us soon when we derive the asymptotic distribution of $\hat{\beta}^{\text{OLS}}$

Theorem (Continuous Mapping Theorem)

If $Z_N \xrightarrow{d} Z$ *then* $g(Z_N) \xrightarrow{d} g(Z)$ *for continuous* g.

Corollary

If $Z_N \xrightarrow{d} N(0, \Omega)$ *then*

$$AZ_N \xrightarrow{d} N(0, A\Omega A')$$

$$(A + o_p(1))Z_N \xrightarrow{d} N(0, A\Omega A'),$$

and since $Z \sim N(0, \Omega) \Rightarrow Z'\Omega^{-1}Z \sim \chi^2(dim(Z))$,

$$Z_N'\Omega^{-1}Z_N \xrightarrow{d} \chi^2(dim(Z_N))$$

$$Z_N'(\Omega + o_p(1))^{-1}Z_N \xrightarrow{d} \chi^2(dim(Z_N)).$$

Another important result for the sample average $\bar{Z}_N := \sum_{i=1}^{N} Z_i/N$.

Theorem (Central Limit Theorem (CLT))

Let $Z_1, Z_2, \ldots$ be a sequence of independent and identically distributed random vectors with $E\|Z_i\|^2 < \infty$. Then

$$\sqrt{N}(\bar{Z}_N - \mu_Z) \xrightarrow{d} N\Big(0, E\left((Z_i - \mu_Z)(Z_i - \mu_Z)'\right)\Big),$$

where $\mu_z := EZ_i$.

Notice:

- $\|z\| := \sqrt{z'z}$ is the Euclidian norm here
- $E\|Z_i\|^2 < \infty$ is an economical way of saying that all components of Z_i have finite means, variances, and covariances

The CLT is a remarkable result

From the WLLN we know that $(\bar{Z}_N - \mu_Z) \xrightarrow{p} 0$

At the same time $\sqrt{N} \to \infty$

Yet their product converges to a normal distribution!

The restrictions imposed in it don't seem very strong

For example, it does not matter what distribution the Z_i come from (as long as $\mathsf{E}\,\|Z_i\|^2 < \infty$)

The sample average multiplied by $\sqrt{N}$ converges to a normal distribution

Conventional terminology with regard to the result

$$\sqrt{N}(\bar{Z}_N - \mu_Z) \xrightarrow{\mathsf{d}} \mathsf{N}(0, \Omega)$$

where $\Omega := \mathsf{E}\left((Z_i - \mu_Z)(Z_i - \mu_Z)'\right)$

- $\bar{Z}_N$ is *asymptotically normally distributed*
- The *large sample distribution* of $\bar{Z}_N$ is normal
- Ω is the asymptotic variance of $\sqrt{N}(\bar{Z}_N - \mu_Z)$
- Ω/N is the asymptotic variance of $\bar{Z}_N$

Primitive usage

- when the sample size N is large yet finite
- the sample average $\bar{Z}_N$ *almost* has a normal distribution
- around the population mean μ_Z
- with variance Ω/N
- irrespective of the underlying distribution of the $Z_1, Z_2, \ldots$

Practical meaning of CLT: for large sample sizes

$$\bar{Z}_N \overset{approx}{\sim} N(\mu_Z, \Omega/N)$$

Let's sketch the proof for a scalar-version of the CLT, where $\mathrm{E}Z_i = \mu_Z$ and $\mathrm{Var}\ Z_i = \sigma_Z^2$

We know from undergrad that $\mathrm{E}\bar{Z}_N = \mu_Z$ and $\mathrm{Var}\ \bar{Z}_N = \sigma_Z^2/N$, therefore CLT says that

$$\sqrt{N}(\bar{Z}_N - \mu_Z) \xrightarrow{\mathsf{d}} \mathsf{N}(0, \sigma_Z^2)$$

or, equivalently,

$$\frac{\sqrt{N}(\bar{Z}_N - \mu_Z)}{\sigma_Z} \xrightarrow{\mathsf{d}} \mathsf{N}(0, 1)$$

To prove this, we need a new concept

Definition (Moment Generating Function)

Let Z be a random variable, the **moment generating function (mgf)** of Z is given by $M_Z(t) = \mathrm{E}\left(e^{tZ}\right)$, where $t \in \mathbb{R}$.

Fun facts about the mgf

- The curvature of the mgf at zero describes all moments:
  $\frac{d^k M_Z}{dt^k}(0) = \mathrm{E}Z^k$
  kth derivative evaluated at zero is equal to kth moment (hence that name)
- not every random variable has a well-defined mgf
  (there exists a generalization, called *characteristic function* that overcomes this problem, mgf is a slightly less general version but easier to work with)
- for random variables whose mgf exist:
  two random variables have identical distributions if and only if their mgf are the same

Mgf can be a useful device for establishing limiting distributions

Lemma (Curtiss' Continuity Theorem)

Let $M_Z(t)$ be the mgf of Z and let $M_{Z_N}(t)$ be the mgf of Z_N.
If $\lim_{N\to\infty} M_{Z_N}(t) = M_Z(t)$ for every t then $Z_N \xrightarrow{d} Z$.

This is based on Lévy's Continuity Theorem (1937)

We're interested in showing $\frac{\sqrt{N}(\bar{Z}_N - \mu_Z)}{\sigma_Z} \xrightarrow{\mathsf{d}} \mathsf{N}(0,1)$

Let's consider the mgf of $\widetilde{Z}_N := \frac{\sqrt{N}(\bar{Z}_N - \mu_Z)}{\sigma_Z}$
and show that its limit is equal to the mgf of a $N(0,1)$

Wait! What is the mgf of the standard normal distribution?

Lemma

The mgf of the standard normal distribution is $t \mapsto e^{t^2/2}$.

(Proof: see assignment)

Notice $\widetilde{Z}_N := \frac{\sqrt{N}(\bar{Z}_N - \mu_Z)}{\sigma_Z} = \frac{(\sum Z_i - N\mu_Z)}{\sigma_Z\sqrt{N}}$

$$
\begin{aligned}
M_{\widetilde{Z}_N}(t) &= \mathsf{E}\left(e^{t\widetilde{Z}_N}\right) = \mathsf{E}\left(\exp\left(t\frac{\sum(Z_i - N\mu_Z)}{\sigma_Z\sqrt{N}}\right)\right) \\
&= \mathsf{E}\left(\exp\left(t\frac{(Z_1 - \mu_Z)}{\sigma_Z\sqrt{N}}\right) \cdot \exp\left(t\frac{(Z_2 - \mu_Z)}{\sigma_Z\sqrt{N}}\right) \cdots \exp\left(t\frac{(Z_N - \mu_Z)}{\sigma_Z\sqrt{N}}\right)\right) \\
&= \mathsf{E}\left(\exp\left(t\frac{(Z_1 - \mu_Z)}{\sigma_Z\sqrt{N}}\right)\right) \cdots \mathsf{E}\left(\exp\left(t\frac{(Z_N - \mu_Z)}{\sigma_Z\sqrt{N}}\right)\right) \\
&= \mathsf{E}\left(\left(\exp\left(t\frac{(Z_1 - \mu_Z)}{\sigma_Z\sqrt{N}}\right)\right)\right)^N \\
&= m_{Z_1}\left(\frac{t}{\sigma_z\sqrt{N}}\right)^N
\end{aligned}
$$

where we define $m_{Z_1}(t) := \mathsf{E}\left(e^{t(Z_1 - \mu_Z)}\right)$

Copy and paste last line: $m_{Z_1}(t) := \mathsf{E}\left(e^{t(Z_1-\mu_Z)}\right)$

Notice that

- $m_{Z_1}(0) = 1$
- $m'_{Z_1}(0) = \mathsf{E}(Z_1 - \mu_Z) = 0$
- $m''_{Z_1}(0) = \mathsf{E}(Z_1 - \mu_Z)^2 = \sigma_Z^2$

Applying a second order Taylor approximation (at zero):

$$\begin{aligned} m_{Z_1}(t) &\approx m_{Z_1}(0) + m'_{Z_1}(0) \cdot t + (1/2) m''_{Z_1}(0) \cdot t^2 \\ &= 1 + (1/2)\sigma_Z^2 \cdot t^2 \end{aligned}$$

and therefore,

$$\begin{aligned} m_{Z_1}\left(\frac{t}{\sigma_z\sqrt{N}}\right) &= 1 + \sigma_Z^2 \cdot \frac{t^2}{2 \cdot \sigma_Z^2 N} \\ &= 1 + \frac{t^2/2}{N} \end{aligned}$$

Connecting the dots

$$M_{\widetilde{Z}_N}(t) = m_{Z_1}\left(\frac{t}{\sigma_z\sqrt{N}}\right)^N = \left(1 + \frac{t^2/2}{N}\right)^N$$

And finally, to evaluate the limit use this result:

Lemma

$\lim_{N\to\infty}\left(1 + \frac{c}{N}\right)^N = e^c.$

It follows that

$$\lim_{N\to\infty} M_{\widetilde{Z}_N}(t) = \lim_{N\to\infty}\left(1 + \frac{t^2/2}{N}\right)^N = e^{t^2/2},$$

which is the mgf of a standard normal distribution

It follows that $\widetilde{Z}_N \xrightarrow{\mathsf{d}} N(0,1)$

Illustration of CLT

The underlying distribution of $Z_1, \ldots, Z_N$ is exponential

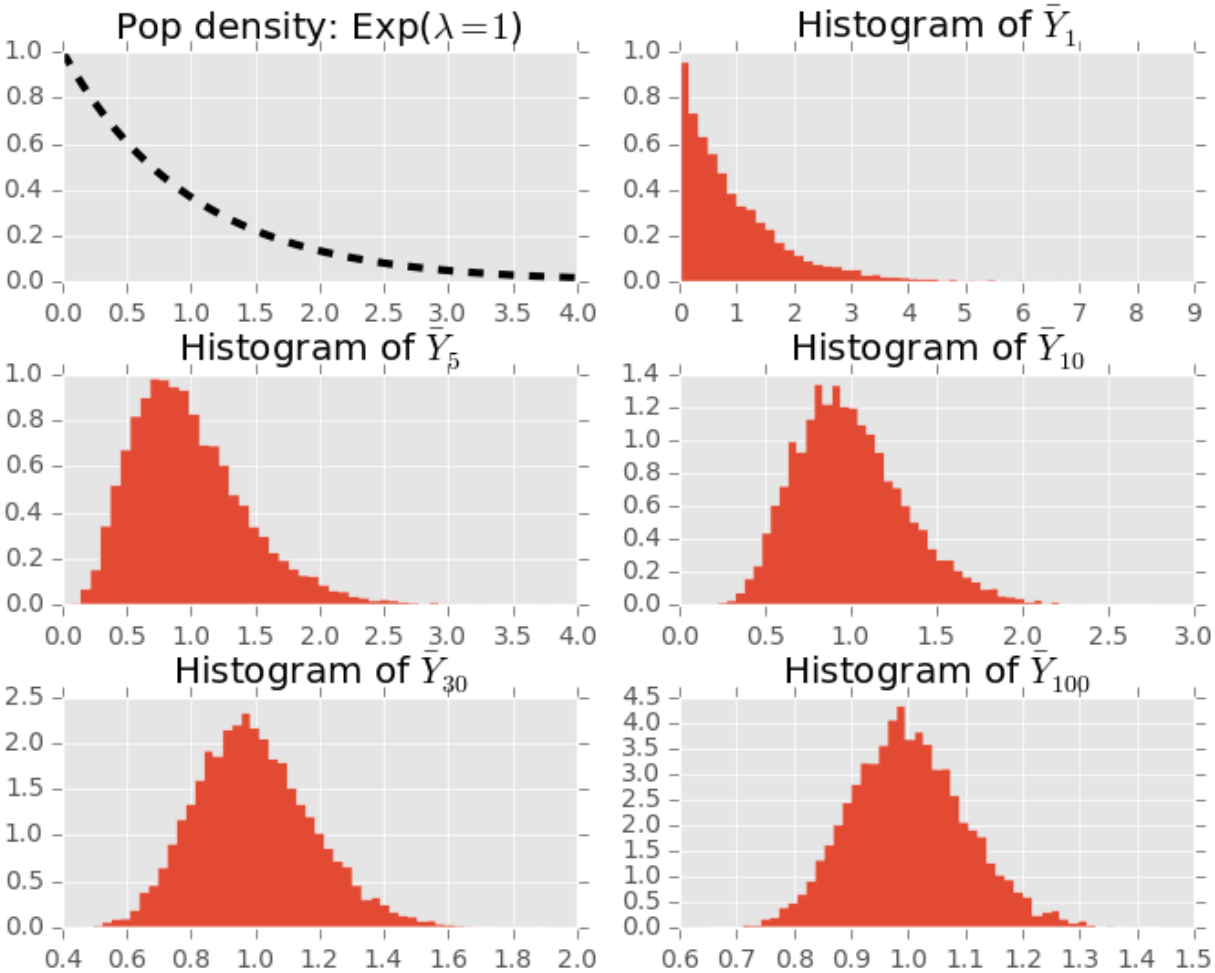
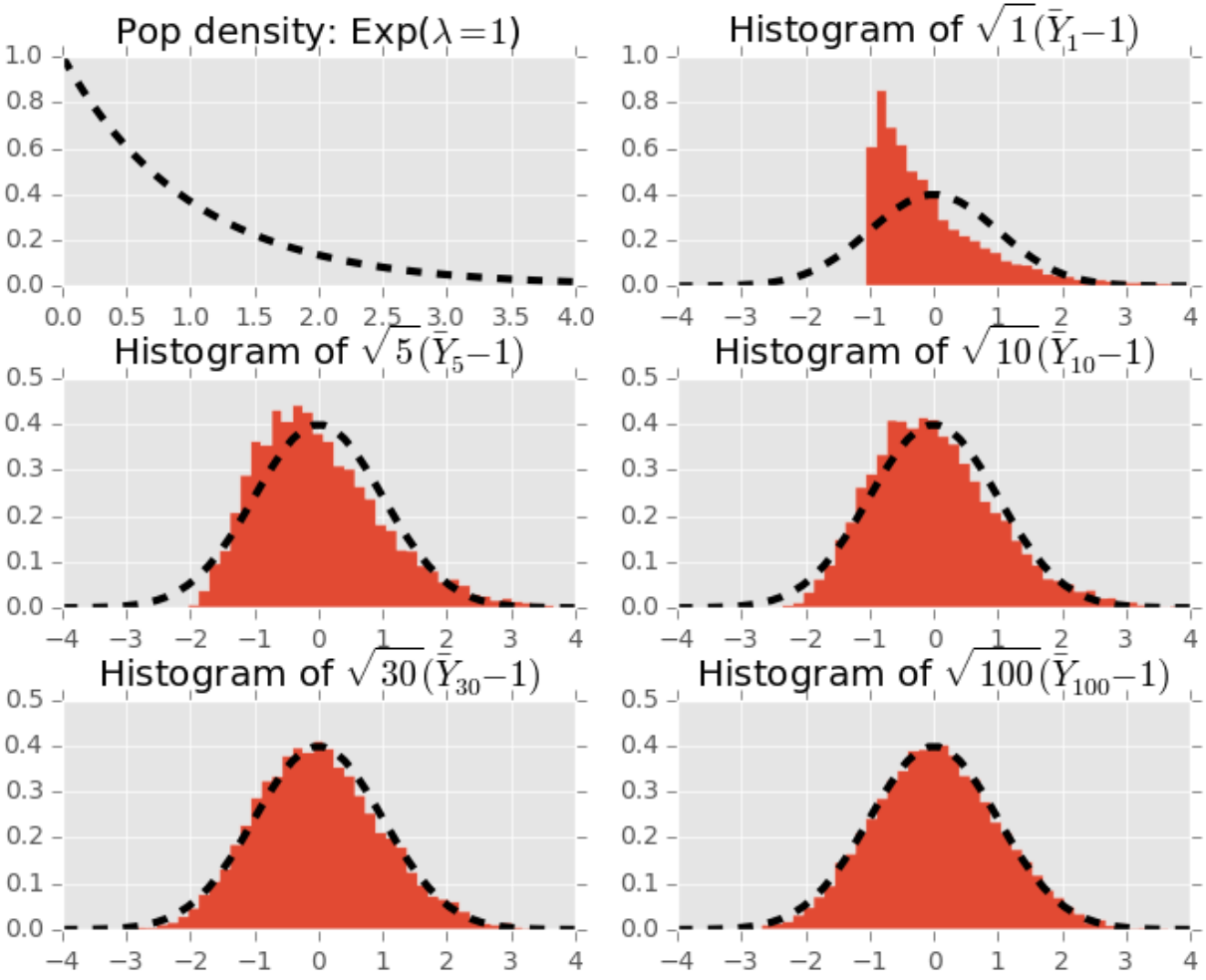


Illustration of CLT

The underlying distribution of $Z_1, \ldots, Z_N$ is exponential



Roadmap

Ordinary Least Squares Estimation

- Basic Asymptotic Theory (part 2 of 2)
- Asymptotic Distribution of the OLS Estimator
- Asymptotic Variance Estimation
- Standard Errors and Confidence Intervals [next week?]
- Hypotheses Tests [next week?]

We know that $\hat{\beta}^{\text{OLS}} \in L_2$

We would like to know the exact distribution of $\hat{\beta}^{\text{OLS}}$ for finite samples (so-called small sample distribution)

Remember

$$\hat{\beta}^{\text{OLS}} = \beta^* + \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1} \sum_{i=1}^{N} X_i u_i$$
$$\beta^* = \mathsf{E}(X_i X_i')^{-1} \mathsf{E}(X_i Y_i)$$

We suspect that $\hat{\beta}^{\text{OLS}} | X_i \sim \mathsf{N}(\cdot, \cdot)$ if $u_i \sim \mathsf{N}(\cdot, \cdot)$

In the absence of such a restrictive assumption, we are unable to determine the exact distribution of $\hat{\beta}^{\text{OLS}}$

We approximate exact distribution by asymptotic distribution

Our hope is that the asymptotic (aka large sample) distribution is a good approximation

The CLT will be our main tool in deriving the asymptotic distribution of $\hat{\beta}^{\text{OLS}}$

Big picture: we already know that $\hat{\beta}^{\text{OLS}} - \beta^* = \text{o}_p(1)$

From what I said earlier, we may suspect that $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*)$ could converge to a normal distribution

To derive this result, let's recall the following representation of the OLS estimator from last week:

$$\hat{\beta}^{\text{OLS}} = \beta^* + \left(\frac{1}{N} \sum_{i=1}^{N} X_i X_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i u_i \right)$$

Let's re-arrange terms ...

Copy and past, for convenience:

$$\hat{\beta}^{\text{OLS}} = \beta^* + \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} X_i u_i\right)$$

Then isolating $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*)$:

$$\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*) = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_i u_i\right)\right)$$

Can you see how the CLT can now be applied to the second factor on the rhs?

Let's break the rhs up again into its bits and pieces

We've already shown last week that, given $\mathsf{E}(X_i X_i') < \infty$,

$$\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} = \mathsf{E}(X_i X_i')^{-1} + \mathsf{o}_p(1) = \mathsf{O}_p(1)$$

For the second factor on the rhs, we know that $\mathsf{E}\left(\sum X_i u_i / N\right) = 0$, then applying the CLT is easy:

$$\left(\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_i u_i\right)\right) \xrightarrow{\mathsf{d}} \mathsf{N}(0, \mathsf{E}(u_i^2 X_i X_i'))$$

Using our tools from basic asymptotic theory (part 2)

Proposition (Asymptotic Distribution of OLS Estimator)

$$\sqrt{N}(\hat{\beta}^{OLS} - \beta^*) = \left(N^{-1}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(N^{-1/2}\sum_{i=1}^{N} X_i u_i\right)$$
$$\xrightarrow{d} N(0, \Omega)$$

where $\Omega := E(X_i X_i')^{-1} E(u_i^2 X_i X_i') E(X_i X_i')^{-1}$.

Ω is the asymptotic variance of $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*)$

Ω/N is the asymptotic variance of $\hat{\beta}^{\text{OLS}}$

We take this to mean that $\hat{\beta}^{\text{OLS}}$ has an *approximate* normal distribution with mean β^* and variance Ω/N

Roadmap

- Ordinary Least Squares Estimation
  - Basic Asymptotic Theory (part 2 of 2)
  - Asymptotic Distribution of the OLS Estimator
  - Asymptotic Variance Estimation
  - Standard Errors and Confidence Intervals [next week?]
  - Hypotheses Tests [next week?]

The asymptotic variance of $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*)$ is

$$\Omega := \mathsf{E}(X_i X_i')^{-1} \mathsf{E}(u_i^2 X_i X_i') \mathsf{E}(X_i X_i')^{-1}$$

The rhs is a function of unobserved population moments

How would we estimate Ω?

Clearly, we estimate $\mathsf{E}(X_i X_i')$ by $(1/N) \sum_{i=1}^{N} X_i X_i'$

But what about $\mathsf{E}(u_i^2 X_i X_i')$?

We don't know u_i

If we observed u_i then we would surely use $(1/N)\sum_{i=1}^{N} u_i^2 X_i X_i'$

That would be an unbiased variance estimator

But we don't observe the errors u_i, instead we "observe" the residuals $\hat{u}_i := Y_i - X_i'\hat{\beta}^{\text{OLS}}$

So how about using $(1/N)\sum_{i=1}^{N} \hat{u}_i^2 X_i X_i'$ to estimate the middle piece?

While this is in principal the right idea, it results in a biased variance estimator

Let's try understand the source of this bias

First some new tools

Let $M_X := I_N - P_X$ with $P_X := X(X'X)^{-1}X'$

Then $\hat{u} = M_X u$

Cool facts about M_X:
$M_X = M_X'$ (symmetric) and $M_X M_X = M_X$ (idempotent)

The **trace of a $K \times K$ matrix** is the sum of its diagonal elements:
$\text{tr}\, A := \sum_{i=1}^{K} a_{ii}$

Savvy tricks: $\text{tr}\,(AB) = \text{tr}\,(BA)$ and $\text{tr}\,(A + B) = \text{tr}\, A + \text{tr}\, B$

Then

$$\hat{\sigma}_u^2 := \sum_{i=1}^{N} \hat{u}_i^2 / N = \frac{\text{tr}\,(\hat{u}\hat{u}')}{N} = \frac{\text{tr}\,(\hat{u}'\hat{u})}{N} = \frac{\text{tr}\,((M_X u)'(M_X u))}{N}$$
$$= \frac{\text{tr}\,(u' M_X' M_X u)}{N} = \frac{\text{tr}\,(u' M_X u)}{N} = \frac{\text{tr}\,(M_X u u')}{N}$$

Aside: $\dim M_X = N \times N$ and $\dim(uu') = N \times N$

Now studying the conditional expectation

$$\begin{aligned}
\mathsf{E}\left(\hat{\sigma}_u^2|X\right) &= \mathsf{E}\left(\text{tr}\left(M_X uu'\right)|X\right)/N \\
&= \text{tr}\ \left(E\left(M_X uu'|X\right)\right)/N \\
&= \text{tr}\ \left(M_X \mathsf{E}\left(uu'|X\right)\right)/N \\
&= \sigma_u^2 \cdot \text{tr}\ \left(M_X\right)/N \\
&= \sigma_u^2\left(\frac{N-K}{N}\right) \\
&< \sigma_u^2,
\end{aligned}$$

where in the fourth equality we simplified our lives by setting $\mathsf{E}(uu'|X) = \sigma_u^2 I_N$ (conditional homoskedasticity)

(The fifth equality will be justified in Assignment 3)

Big picture: $\hat{\sigma}_u^2$ is downwards biased which is not good

Confidence intervals based on $\hat{\sigma}_u^2$ would be too narrow

Statistical inference based on $\hat{\sigma}_u^2$ would be too optimistic

There is an easy fix!

Use $s_u^2 := \frac{N}{N-K}\hat{\sigma}_u^2 = \frac{1}{N-K}\sum_{i=1}^{N}\hat{u}_i^2$ instead

Obviously s_u^2 will be unbiased

I'm not particularly concerned about this bias

That's because N should be a much larger number than K

The whole idea of using asymptotic approximations to finite sample distributions is to let $N \to \infty$ while K is fixed

In other words $\lim_{N\to\infty} \hat{\sigma}_u^2 = \lim_{N\to\infty} s_u^2$
(asymptotic bias is the same)

Combining things, we propose the following asymptotic variance estimator

Definition (Asymptotic Variance Estimator)

$$\hat{\Omega} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\frac{1}{N-K}\sum_{i=1}^{N} \hat{u}_i^2 X_i X_i'\right)\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}$$

Stata calculates $\hat{\Omega}$ when you type something like

```
regress lwage schooling experience, robust
```

Textbooks call $\hat{\Omega}$ the *heteroskedasticity robust variance estimator*

The *standard errors* derived from $\hat{\Omega}$ are sometimes referred to as **Eicker-Huber-White** standard errors
(or some subset permutation of these names)

Notice: Wooldridge, on page 61, proposes this version

Definition (Asymptotic Variance Estimator)

$$\hat{\Omega}^{\text{Wool}}_{\text{dridge}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i^2 X_i X_i'\right) \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}$$

This is NOT what Stata implements
(to the best of my knowledge)

But from what I said earlier, it merely creates rounding error

Asymptotically they are all identical
(because K is a finite number)

Roadmap

- Ordinary Least Squares Estimation
  - Basic Asymptotic Theory (part 2 of 2)
  - Asymptotic Distribution of the OLS Estimator
  - Asymptotic Variance Estimation
  - Standard Errors and Confidence Intervals `[next week?]`
  - Hypotheses Tests `[next week?]`

Why do we care about the distribution of $\hat{\beta}^{\text{OLS}}$?

Knowing the distribution helps us understand *precision* of the estimate

In addition, people use the distribution to construct statistical tests

I prefer to focus on precision and ignore statistical testing

For the sake of illustration, let's tentatively assume that $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*) \sim \text{N}(0, \Omega)$

The point here is that we assume that the normal distribution is *exact*, not just an asymptotic approximation

Proposition

Let r be some K dimensional nonstochastic vector. Then

$$\sqrt{N}(r'\hat{\beta}^{OLS} - r'\beta^*) \sim N(0, r'\Omega r)$$

Corollary

$$\frac{r'\hat{\beta}^{OLS} - r'\beta^*}{\sqrt{r'\Omega r/N}} \sim N(0, 1)$$

You can pick r to consider any linear combination of the elements of β^* that you are interested in

Most times people use $r = e_k$ where e_k is the k-th unit vector taking the value 1 in position k and the value zero elsewhere

That way you are grabbing the kth element of a vector, or the (k, k) element of a matrix

- $\beta_k^* = e_k' \beta^* = \beta^{*\prime} e_k$
- $\omega_{kk} = e_k' \Omega e_k$

Therefore

$$\frac{\hat{\beta}_k^{\text{OLS}} - \beta_k^*}{\sqrt{\omega_{kk}/N}} = \frac{e_k' \hat{\beta}^{\text{OLS}} - e_k' \beta^*}{\sqrt{e_k' \Omega e_k / N}} \sim \text{N}(0, 1)$$

The OLS estimator is a point estimator

It is unlikely that $\hat{\beta}^{\text{OLS}} = \beta^*$
(in fact, that event has probability zero)

Instead of a point estimator, should we consider an interval estimator?

Considerations:

- the smallest interval we would consider is $\hat{\beta}^{\text{OLS}}$ itself
- by having a proper interval, we can make sure that β^* is covered with a probability larger than zero (unlike for point estimates)
- the largest interval covers the whole real line and guarantees a 100% coverage probability (not very informative though)
- there's a tension between two goals:
  high coverage probability vs narrow (informative) interval

Idea: accept a coverage probability that is a little less than 100%, say 95%, and hope to obtain an informative interval

Because $\frac{\hat{\beta}_k^{\text{OLS}} - \beta_k^*}{\sqrt{\omega_{kk}/N}} \sim \text{N}(0,1)$,

the obvious interval that comes to mind is $\left[\hat{\beta}_k^{\text{OLS}} \pm c \cdot \sqrt{\omega_{kk}/N}\right]$

This is symmetric around the point estimate because of the symmetry of the normal distribution

A clever choice of c will ensure a 95% coverage probability:

$$P\left(\beta_k^* \in \left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\omega_{kk}/N}\right]\right) = 0.95$$

Careful! Don't read this literally as
"the probability that β_k^* is in interval"

That's incorrect! It makes it sound as if β_k^* is a random variable

The random object is the interval $\left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\omega_{kk}/N}\right]$

So the way to read the above statement is
"the probability that the interval covers β_k^*"

Many people do not understand what a confidence interval can tell them and what it cannot tell them

It means:
Prior to repeatedly estimating $\hat{\beta}_k^{\text{OLS}}$ in separate random experiments, the probability is 95% that the random interval $\left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\omega_{kk}/N}\right]$ contains β_k^*

A frequentist's thought experiment: if I were given 100 random samples of size N then about 95 of them will yield confidence intervals that contain β_k^* (but I don't know which ones)

Common misconceptions regarding confidence intervals

The following statements are all false

- The specific 95% confidence interval presented by a study has a 95% chance of containing the coefficient
- The true coefficient β_k^* has a 95% probability of falling inside the confidence interval
- A coefficient outside the 95% confidence interval is refuted by the data

The first two in particular are believed by many people

Google: *Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations*, by Greenland at al, a worthwhile read

A few slides ago we tentatively assumed $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*) \sim \text{N}(0, \Omega)$

Now let's generalize by going back to $\sqrt{N}(\hat{\beta}^{\text{OLS}} - \beta^*) \xrightarrow{\text{d}} \text{N}(0, \Omega)$

It's easy to adjust earlier results accordingly, basically by replacing '$\sim$' with '$\xrightarrow{\text{d}}$'

Proposition

Let r be some K dimensional nonstochastic vector. Then

$$\sqrt{N}(r'\hat{\beta}^{OLS} - r'\beta^*) \xrightarrow{d} N(0, r'\Omega r)$$

Corollary

$$\frac{r'\hat{\beta}^{OLS} - r'\beta^*}{\sqrt{r'\Omega r/N}} \xrightarrow{d} N(0, 1)$$

You may replace Ω by $\hat{\Omega} = \Omega + o_p(1)$:

Proposition

$$\frac{r'\hat{\beta}^{OLS} - r'\beta^*}{\sqrt{r'\hat{\Omega}r/N}} \xrightarrow{d} N(0,1)$$

Grabbing one element from that vector: $\frac{\hat{\beta}_k^{\text{OLS}} - \beta_k^*}{\sqrt{\hat{\omega}_{kk}/N}} \xrightarrow{\text{d}} \text{N}(0,1)$

where $\hat{\omega}_{kk} := e_k'\hat{\Omega}e_k$
(this is a number that we can compute from the sample data)

The confidence interval for β_k^* therefore is $\left[\hat{\beta}_k^{\text{OLS}} \pm 1.96 \cdot \sqrt{\hat{\omega}_{kk}/N}\right]$

Terminology:
the term $\sqrt{\hat{\omega}_{kk}/N}$ is also called the **asymptotic standard error** of $\hat{\beta}_k^{\text{OLS}}$

Aside: by convention, an estimator of the standard deviation of an estimator is called a *standard error*

Definition (Asymptotic Standard Error of $\hat{\beta}^{OLS}$)

Let Ω/N be the asymptotic variance of $\hat{\beta}^{\text{OLS}}$. The **asymptotic standard errors** of the OLS estimator $\hat{\beta}^{\text{OLS}}$ and $\hat{\beta}_k^{\text{OLS}}$ are

$$\text{se}(\hat{\beta}^{\text{OLS}}) = \sqrt{\text{diag } \hat{\Omega}/N}$$
$$\text{se}(\hat{\beta}_k^{\text{OLS}}) = e_k' \cdot \text{se}(\hat{\beta}^{\text{OLS}}),$$

where $\hat{\Omega}/N$ is the estimator of the asymptotic variance of $\hat{\beta}^{\text{OLS}}$

We obtain this result regarding the asymptotic coverage probability:

Proposition

$$\lim_{N\to\infty} P\left(\beta_k^* \in [\hat{\beta}_k^{OLS} \pm 1.96 \cdot se(\hat{\beta}_k^{OLS})]\right) = 0.95$$

Roadmap

Ordinary Least Squares Estimation

- Basic Asymptotic Theory (part 2 of 2)
- Asymptotic Distribution of the OLS Estimator
- Asymptotic Variance Estimation
- Standard Errors and Confidence Intervals [next week?]
- Hypotheses Tests [next week?]

You study $Y_i = X_i\beta^* + u_i$ where $\mathsf{E}(u_i X_i) = 0$

For simplicity X_i is a scalar

For some reason you are interested in the value of β^*

In particular, you want to know $\beta^* \overset{?}{=} 0$

You remember that OLS delivers a consistent estimator

You obtain $\hat{\beta}^{\mathsf{OLS}} = 0.18$

What do you do?

You consider two states of nature:

- $\beta^* = 0$
- $\beta^* \neq 0$

These are mutually exclusive and exhaustive

You can look at them as *hypotheses*

Definition (Statistical Hypothesis)

A **statistical hypothesis** is a statement about a population parameter.

One is the null hypothesis, and one the alternative hypothesis (of course denoted by H_0 and H_1)

You would like to know which one is *true* (if there is such a thing)

To determine which hypothesis is true, you propose:

$$\textit{if } \hat{\beta}^{OLS} = 0 \textit{ then } \beta^* = 0\textit{, else } \beta^* \neq 0$$

According to this decision rule, you decide that $\beta^* \neq 0$
(because $0.18 \neq 0$)

You have just conducted a *hypothesis test*

Definition

A **statistical hypothesis test** is a decision rule that specifies

(i) for which sample values H_0 is considered true;

(ii) for which sample values H_1 is considered true.

The hypothesis test

$$\textit{if } \hat{\beta}^{OLS} = 0 \textit{ then } \beta^* = 0\textit{, else } \beta^* \neq 0$$

is not good because you will almost certainly conclude that $\beta^* \neq 0$

This test is extremely conservative

You understand that $\hat{\beta}^{\text{OLS}}$ could be nonzero even if $\beta^* = 0$

The estimator $\hat{\beta}^{\text{OLS}}$ is subject to sampling error

As sensible test should reflect this possibility of sampling error, and therefore the variance of $\hat{\beta}^{\text{OLS}}$ should play a role too

If we are unable to quantify the exact variance of $\hat{\beta}^{\text{OLS}}$, the asymptotic variance will be good enough

The most common statistic to combine information of the point estimate and its variance is the t-statistic

Definition (t-Statistic)

Let $\hat{\theta}$ be an estimator and $\text{se}(\hat{\theta})$ be its asymptotic standard error. Then

$$t_{\hat{\theta}}(\theta) := \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})}$$

is the **t-statistic** or **t-ratio** for θ.

It has the shape of the *standardized* estimator $\hat{\theta}$

Let's say we have two competing estimators, labelled $\hat{\beta}^{\text{OLS}}$ and $\hat{\beta}^{\text{IV}}$ and we want to test if $\beta^* = 24$

Then we would look at $t_{\hat{\beta}^{\text{OLS}}}(24)$ and $t_{\hat{\beta}^{\text{IV}}}(24)$

It should be clear that because $\hat{\beta}^{\text{OLS}} = \beta^* + \text{o}_p(1)$

$$t_{\hat{\beta}^{\text{OLS}}}(\beta^*) \xrightarrow{\text{d}} \text{N}(0,1)$$

Software packages such as Stata have the terrible habit of reporting $t_{\hat{\beta}^{\text{OLS}}}(0)$ as part of a standard regression output

$t_{\hat{\beta}^{\text{OLS}}}(0)$ facilitates a hypothesis test of the null $\beta^* = 0$ against the alternative $\beta^* \neq 0$, the critical value is simply ± 1.96

It is not clear that the null $\beta^* = 0$ is interesting at all

There is an awful practice in applied econometrics to focus on the value of t-statistics, or, equivalently, on *significance stars*

The vast majority of researchers present their estimation tables with *STATA significance stars*

- $|t| > 1.64$ receives one *star*
- $|t| > 1.96$ receives two *stars*
- $|t| > 2.58$ receives three *stars*

It's like the Michelin restaurant guide: the more *stars*, the better!

For example, if the return to schooling is estimated to equal 0.14 and it is statistically significant at the 95% level, then the table will say 0.14^{**}

Many applied papers limit the discussion of their results only to those coefficient estimates with *stars* attached, that is, only to those who are *statistically significant*

Results that don't have any *stars* are often ignored

Our average Monday seminar follows this pattern

Sadly, PhD students copy this terrible practice

I have had countless conversations with PhD students whose goal it is to obtain *stars* in their tables

Because the opinion is: NO STARS, NO PAPER!

The research objective becomes: obtain *stars*

But often times stars are out of reach

Try do your estimations without *stars* or t-statistics

They are simplistic or reductionist

They seem to apply a binary world:
results are either statistically significant or irrelevant

(Also, they encourage *star*-hacking:
the strong incentive to obtain stars)

So what should you be doing?

What ought to be best practice?
(But admittedly and unfortunately isn't)

Report standard errors and confidence intervals

They offer a notion of *precision of estimates*

Also, never ever say this:
"*The estimate is highly significance*"
(or variations thereof)
It demonstrates that you don't understand what you are doing

(Also: don't use STATA)