Advanced Econometrics I

Jürgen Meinecke

Lecture 3 of 12

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

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

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

Example (Only Looks Like Rolling a Die)

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- $\mathcal{F} = \left\{\{1, 3, 5\}, \{2, 4, 6\}, \Omega, \emptyset\right\}$
- *Consider all* $A \in \mathcal{F}$

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

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

Definition (Random Variable—first attempt)

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

Example

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

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

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

Example

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

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 $(\Omega, \mathcal{F})$ is a function $Z : \Omega \to \mathbb{R}$ such that

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

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

$\mathcal{B}(\mathbb{R})$ 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 $\mathcal{F}$ 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 $(\Omega, \mathcal{F}, P)$, the **distribution** or **law** of the random variable is the probability measure defined by

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

We say that μ is the distribution of Z, or $\mathcal{L}(Z)$ 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 mapping 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

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

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 $\mathrm{E}\,\|Z_i\|^2 < \infty$)

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

Conventional terminology with regard to the result

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

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

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

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

But is it a good approximation?

How large does N need to be?

Illustration of CLT

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

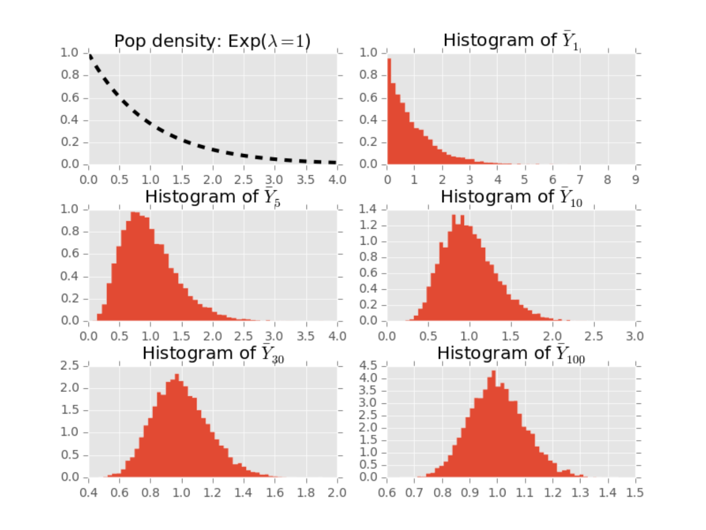
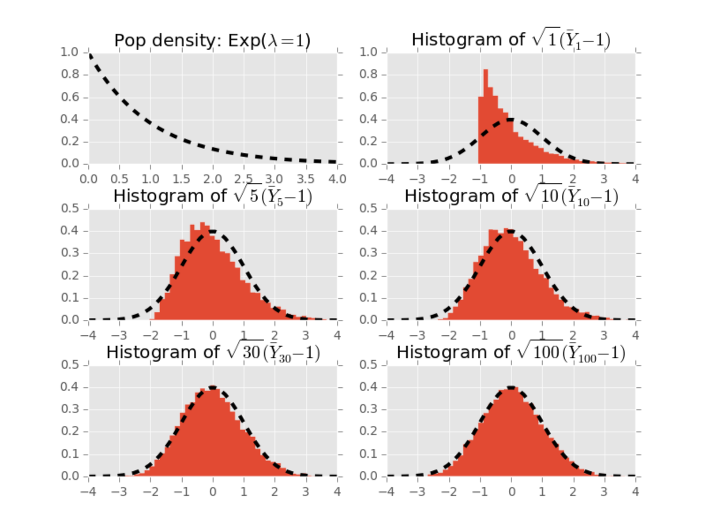


Illustration of CLT

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

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

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

Remember

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

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

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 $\hat{\beta}^{\text{OLS}} - \beta^* = o_p(1)$

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

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

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 (using Slutsky's theorem) that, given $\mathsf{E}(X_i X_i') < \infty$,

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

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)

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

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

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

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

The asymptotic variance of $\sqrt{N}(\hat{\beta}^{\mathrm{OLS}} - \beta^*)$ 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:
$\operatorname{tr} A := \sum_{i=1}^{K} a_{ii}$

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

Then

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

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

Now studying the conditional expectation

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

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)