EC508: Econometrics
Sampling Distribution of the OLS Estimator

Jean-Jacques Forneron

Spring, 2023

Boston University

The Sampling Distribution of the OLS Estimator (SW 4.5)

- The OLS estimator is computed from a sample of data. A different sample yields a different value of $\hat{\beta}_1$. This is the source of the "sampling uncertainty" of $\hat{\beta}_1$. We want to:
    - quantify the sampling uncertainty associated with use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$
    - construct a confidence interval for β_1
    - All these require figuring out the sampling distribution of the OLS estimator. Two steps to get there:
        - Probability framework for linear regression
        - Distribution of the OLS estimator

Probability Framework for Linear Regression

The probability framework for linear regression is summarized by the three least squares assumptions.

- **Population**:
  The group of interest (ex: all possible school districts)
- **Random variables**: Y, X
  Ex: (Test Score, STR)
- **Joint distribution of** (Y, X). We assume:
  - The population regression function is linear
  - $\mathbb{E}(u|X) = 0$ (1st Least Squares Assumption)
  - X, Y have nonzero finite fourth moments (3rd L.S.A.)
- **Data Collection by simple random sampling** implies:
  $\{(X_i, Y_i)\}, i = 1, \ldots, n$, are i.i.d. (2nd L.S.A.)

Mean Independence

- Mean Independence: $\mathbb{E}(u_i|X_i) = 0$
- Implies $cov(u_i, X_i) = 0$: no correlation between X and u
- Implied by X and u independent + u mean zero

The Sampling Distribution of $\hat{\beta}_1$

- Like $\bar{Y}$, $\hat{\beta}_1$ has a sampling distribution.
- What is $\mathbb{E}(\hat{\beta}_1)$?
  - If $\mathbb{E}(\hat{\beta}_1) = \beta_1$, then OLS is unbiased – a good thing!
- What is *var*$(\hat{\beta}_1)$? (measure of sampling uncertainty)
  - We need to derive a formula so we can compute the standard error of $\hat{\beta}_1$.
- What is the distribution of $\hat{\beta}_1$ in small samples?
  - It is very complicated in general
- What is the distribution of $\hat{\beta}_1$ in large samples?
  - In large samples, $\hat{\beta}_1$ is normally distributed.

The mean and variance of the sampling distribution of $\hat{\beta}_1$

Some preliminary algebra:

$$
\begin{aligned}
&Y_i = \beta_0 + \beta_1 X_i + u_i \\
&\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u} \\
&\Rightarrow Y_i - \bar{Y} = \beta_1 [X_i - \bar{X}] + u_i - \bar{u}
\end{aligned}
$$

This implies that:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \\
&= \frac{\sum_{i=1}^{n} \beta_1 (X_i - \bar{X}_n)^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} + \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)(u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}
\end{aligned}
$$

The mean and variance of the sampling distribution of $\hat{\beta}_1$

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} \beta_1 (X_i - \bar{X}_n)^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} + \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)(u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)(u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \\
&= \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n) u_i}{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}
\end{aligned}
$$

Now we can calculate $\mathbb{E}(\hat{\beta}_1)$ and $var(\hat{\beta}_1)$:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \beta_1 + \mathbb{E}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}\right) \\
&= \beta_1 + \mathbb{E}\left(\mathbb{E}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}\middle| X_1, \dots, X_n\right)\right) \\
&= \beta_1 + \mathbb{E}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)\mathbb{E}\left(u_i \middle| X_1, \dots, X_n\right)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}\right)
\end{aligned}$$

LSA 2: (Y_i, X_i) iid implies $\mathbb{E}\left(u_i \middle| X_1, \dots, X_n\right) = \mathbb{E}\left(u_i \middle| X_i\right)$

LSA 1: $\mathbb{E}\left(u_i \middle| X_i\right) = 0$

Together: $\mathbb{E}(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1$ is an **unbiased** estimator of β_1

Weak Law of Large Numbers

- Let $Z_1, \ldots, Z_n$ be iid with $\mathbb{E}(|Z_i|^2) < \infty$
- Then:

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} \mathbb{E}(Z_i)$$

- $\xrightarrow{p}$ is the convergence in probability:

$$\mathbb{P}(|\bar{Z}_n - \mathbb{E}(Z_i)| > \varepsilon) \to 0, \text{ as } n \to \infty$$

- The WLLN can be proved using Chebyshev's inequality:

$$\mathbb{P}(|\bar{Z}_n - \mathbb{E}(Z_i)| > \varepsilon) \leq \frac{\mathbb{E}(|\bar{Z}_n - \mathbb{E}(Z_i)|^2)}{\varepsilon^2}$$

Consistency of OLS

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

- Let $Z_i = X_i u_i$. Z has mean zero, finite variance: $\bar{Z}_n \xrightarrow{p} 0$
- $\bar{X}_n \bar{u}_n \xrightarrow{p} \mathbb{E}(X_i)\mathbb{E}(u_i) = 0$
- $$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \\ &\xrightarrow{p} \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 \\ &= var(X_i) > 0 \end{aligned}$$
- Together these imply:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} \xrightarrow{p} \beta_1 + 0$$