The sampling distribution of the OLS estimator, and the statistical properties of the OLS estimator in finite and large samples

Econometrics (35B206), Lecture 2

Tunga Kantarcı, TiSEM, Tilburg University, Spring 2019

Sampling distribution

In this lecture $\hat{\boldsymbol{\beta}}_{OLS} \equiv \hat{\boldsymbol{\beta}}$.

Sampling distribution

Consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Suppose we have n observations for $\boldsymbol{y}$ and $\boldsymbol{X}$. Using this one sample we estimate $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}$ represent this estimate. The estimation results in one $\hat{\boldsymbol{\beta}}$.

Sampling distribution

```
. regress wage educ
```

Source	SS	df	MS		
Model	7842.35455	1	7842.35455	Number of obs	= 997
Residual	31031.0745	995	31.1870095	F(1, 995)	= 251.46
				Prob > F	= 0.0000
Total	38873.429	996	39.0295472	R-squared	= 0.2017
				Adj R-squared	= 0.2009
				Root MSE	= 5.5845

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	1.135645	.0716154	15.86	0.000	.9951106	1.27618
_cons	-4.860424	.9679821	-5.02	0.000	-6.759944	-2.960903

Sampling distribution

If there is only one $\hat{\boldsymbol{\beta}}$, how can $\hat{\boldsymbol{\beta}}$ has a standard error? How can it have a distribution? How can we talk about the statistical properties of $\hat{\boldsymbol{\beta}}$?

Sampling distribution

The distribution of $\hat{\boldsymbol{\beta}}$ results from a conceptual experiment.

Sampling distribution

The experiment is about taking samples from the population repeatedly.

Consider the LRM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Take a sample of n observations for $\boldsymbol{y}$ and $\boldsymbol{X}$ from the population.

Using the $\boldsymbol{y}$ and $\boldsymbol{X}$, obtain $\hat{\boldsymbol{\beta}}$.

Sampling distribution

Take a new sample of n observations for $\boldsymbol{y}$ and $\boldsymbol{X}$ from the population.

Using the new $\boldsymbol{y}$ and $\boldsymbol{X}$, obtain a new $\hat{\boldsymbol{\beta}}$.

Repeatedly take all possible samples from the population.

Obtain many $\hat{\boldsymbol{\beta}}$.

$\hat{\boldsymbol{\beta}}$ now has a distribution. This is the sampling distribution of $\hat{\boldsymbol{\beta}}$.

The statistical properties of $\hat{\boldsymbol{\beta}}$ is about this sampling distribution of $\hat{\boldsymbol{\beta}}$.

Sampling distribution

We will study the statistical properties of $\hat{\boldsymbol{\beta}}$ in theory.

But we will also study the statistical properties in a more applied manner. E.g., we will plot the sampling distribution of $\hat{\boldsymbol{\beta}}$, and see how it behaves if we violate an assumption of the SLM.

To conduct this study, we should take samples for $\boldsymbol{y}$ and $\boldsymbol{X}$ from the population repeatedly, obtain many $\hat{\boldsymbol{\beta}}$, and plot the sampling distribution of $\hat{\boldsymbol{\beta}}$.

But taking repeated samples from the population is expensive.

What can we do?

Simulating the sampling distribution

We can simulate the sampling distribution of $\hat{\boldsymbol{\beta}}$. The simulation experiment is about obtaining $\boldsymbol{y}$ in the LRM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Take a sample of n observations for $\boldsymbol{y}$ and $\boldsymbol{X}$ from the population.

Using the $\boldsymbol{y}$ and $\boldsymbol{X}$, obtain $\hat{\boldsymbol{\beta}}$.

Simulating the sampling distribution

Draw n random numbers for ε from a probability distribution.

Given the original observations for $\boldsymbol{X}$, the initial estimate of $\hat{\boldsymbol{\beta}}$, and the random numbers for ε, generate n observations for $\boldsymbol{y}$.

Using the generated $\boldsymbol{y}$ and original $\boldsymbol{X}$, obtain a new $\hat{\boldsymbol{\beta}}$.

Repeat the procedure to generate new sets of $\boldsymbol{y}$.

Obtain many $\hat{\boldsymbol{\beta}}$.

Obtain a simulated sampling distribution for $\hat{\boldsymbol{\beta}}$.

Simulating the sampling distribution

In the experiment, we keep the n observations of $\boldsymbol{X}$ constant as we repeatedly generate new $\boldsymbol{y}$. This simplifies the experiment because then we can attribute the response of the sampling distribution to interesting counterfactual scenarios rather than to the sampling variance of $\boldsymbol{X}$. This is the same as what we do in statistical derivations. We condition on $\boldsymbol{X}$ meaning that we keep $\boldsymbol{X}$ constant. This simplifies the derivations very much. Treating $\boldsymbol{X}$ constant in repeated sampling is not realistic unless $\boldsymbol{X}$ is collected in an experimental setting where the experimenter had chosen the value for $\boldsymbol{X}$ before $\boldsymbol{y}$ was determined. We justify this treatment with the random sampling assumption.

Simulating the sampling distribution

In the experiment, we assume that $\mathrm{E}\left[\varepsilon \mid \boldsymbol{X}\right] = 0$ holds.

Simulating the sampling distribution

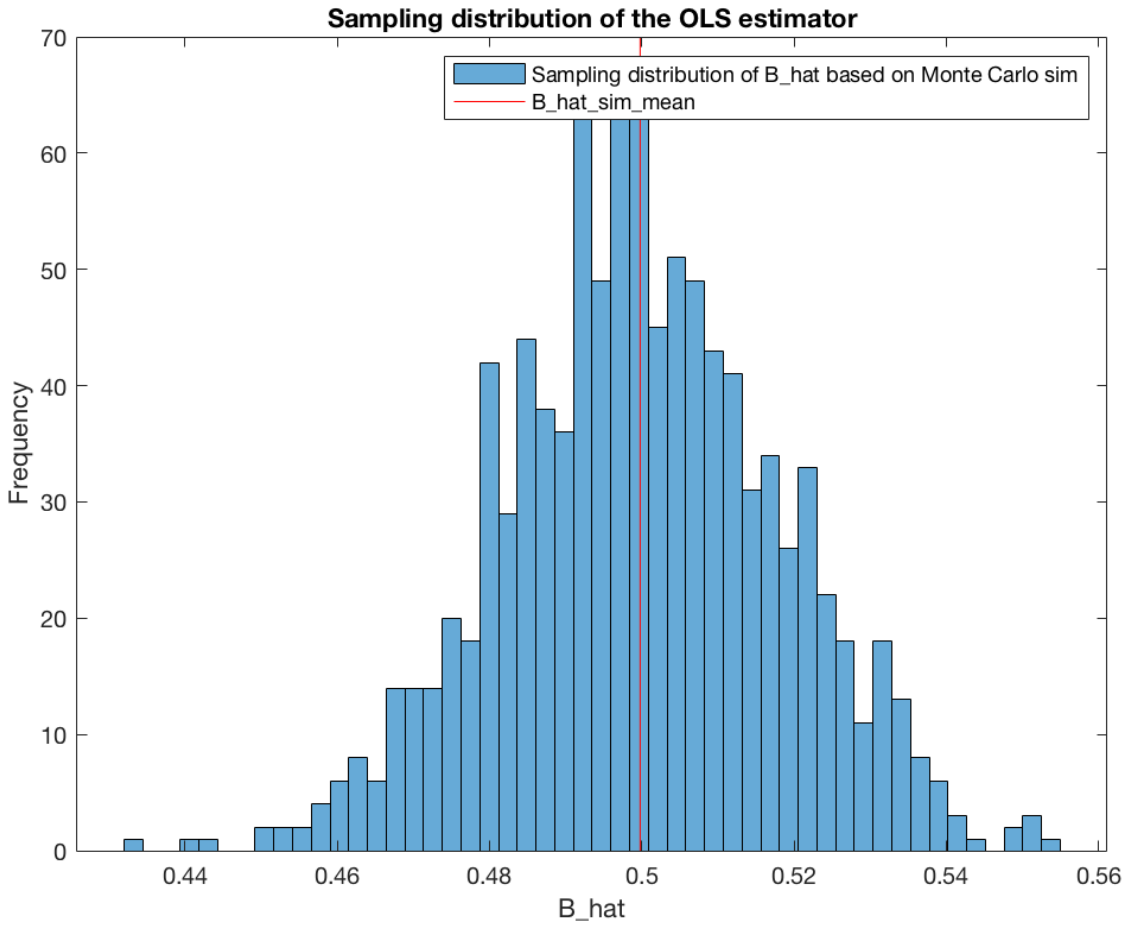
If there is no initial sample data for $\boldsymbol{X}$, we can keep the experiment purely hypothetical. Take random draws for $\boldsymbol{X}$ from a distribution. Do the same for ε. Assume a value for β. Generate $\boldsymbol{y}$. Obtain a first $\hat{\beta}$. Generate new ε. Using the original $\boldsymbol{X}$ and the assumed β, generate new $\boldsymbol{y}$. Obtain a new $\hat{\beta}$. Repeat the procedure to obtain many $\hat{\beta}$. This is what we do next.

Simulating the sampling distribution

```
# Simulate the sampling distribution of the OLS estimator
N_sim = 1000
N_obs = 9000
B_true = 0.5
x = unifrnd(-1,1,N_obs,1)
B_hat_sim = NaN(1,N_sim)
for i = 1:N_sim
    e = normrnd(0,1,N_obs,1)
    y = x*B_true+e
    B_hat_sim(1,i) = inv(x'*x)*x'*y
end
```

Simulating the sampling distribution

```
histogram(B_hat_sim)
```



Statistical properties

How do we want the sampling distribution of $\hat{\beta}$ to behave? How do we want the mean and the variance of this distribution to behave?

Statistical properties

We make a distinction between a small and large sample. We want the sampling distribution of $\hat{\boldsymbol{\beta}}$ to behave in certain ways in a finite sample. We want the sampling distribution of $\hat{\boldsymbol{\beta}}$ to behave in certain ways in a large sample.

Statistical properties

Behaviour in a finite sample means that the behaviour does not depend on n. We fix n, and study the behaviour. If we change n, the behaviour is not affected. Behaviour in a large sample means that the behaviour depends on n. We increase n, and study how the sampling distribution behaves. Why we differentiate between a small and large sample will become clear later in these slides.

Statistical properties in finite samples, unbiasedness

Suppose that we calculate a $\hat{\beta}$ with the sample at hand. $\hat{\beta}$ is an estimate of the true β. We want to believe that this particular $\hat{\beta}$ is close to β in some criterion.

Statistical properties in finite samples, unbiasedness

Consider the mean of the sampling distribution of $\hat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right].$$

Note that sampling distribution is created by repeatedly taking all possible samples from the population. The mean in the population is the expected value!

Statistical properties in finite samples, unbiasedness

The criterion we want $\hat{\boldsymbol{\beta}}$ to satisfy is

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] = \boldsymbol{\beta}.$$

It says that the mean of the sampling distribution of $\hat{\boldsymbol{\beta}}$ is equal to $\boldsymbol{\beta}$. It says that on average $\hat{\boldsymbol{\beta}}$ will correctly estimate $\boldsymbol{\beta}$. If this is true, we say that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Statistical properties in finite samples, unbiasedness

In the sumlation above, `mean(B_hat_sim)` gives 0.4998. `B_true` was 0.5.

Statistical properties in finite samples, unbiasedness

In practice, what does unbiasedness imply? Suppose that you draw an unlucky sample from the population, and obtain a bad $\hat{\boldsymbol{\beta}}$. Or think of our simulation experiment. Suppose that you take n draws for ε from its assumed distribution which turn out to be extreme. $\hat{\boldsymbol{\beta}}$ will be far from its population mean $\mathrm{E}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right]$ which is equal to $\boldsymbol{\beta}$. Hence, in practice, to satisfy unbiasedness as much as possible, the sample we draw should be typical.

Statistical properties in finite samples, unbiasedness, proof

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}.
\end{aligned}
$$

Taking the expectation conditional on $\boldsymbol{X}$,

$$
\begin{aligned}
\mathrm{E}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] &= \mathrm{E}\left[\boldsymbol{\beta} \mid \boldsymbol{X}\right] + \mathrm{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] \\
&= \boldsymbol{\beta}
\end{aligned}
$$

if $\mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] = 0$.

Hence, $\mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] = 0$ is another requirement for unbiasedness.

Statistical properties in finite samples, unbiasedness

By the LIE, it is also true that

$$\begin{aligned} \mathsf{E}\left[\hat{\boldsymbol{\beta}}\right] &= \mathsf{E}_{\boldsymbol{X}}\left[\mathsf{E}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right]\right] \\ &= \mathsf{E}_{\boldsymbol{X}}\left[\boldsymbol{\beta}\right] \\ &= \boldsymbol{\beta}. \end{aligned}$$

Statistical properties in finite samples, unbiasedness

Is the OLS estimator the only unbiased estimator? Consider a competitor estimator

$$\hat{\boldsymbol{\beta}}_0 = \boldsymbol{C}\boldsymbol{y}.$$

$\boldsymbol{C}$ is some K×n matrix that depends on $\boldsymbol{X}$. Taking the expectation conditional on $\boldsymbol{X}$,

$$\begin{aligned}
\mathrm{E}\left[\hat{\boldsymbol{\beta}}_0 \mid \boldsymbol{X}\right] &= \mathrm{E}\left[\boldsymbol{C}\boldsymbol{y} \mid \boldsymbol{X}\right] \\
&= \mathrm{E}\left[\boldsymbol{C}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \mid \boldsymbol{X}\right] \\
&= \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] \\
&= \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

if $\mathrm{E}\left[\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right] = 0$, and if $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$.

The OLS estimator is not the only unbiased estimator!

Statistical properties in finite samples, unbiasedness

The OLS estimator is not the only unbiased estimator. But recall that we wanted to believe in the OLS estimator in some criterion, and we have considered unbiasedness as a criterion. But now we have more than one estimator that is unbiased. Why should we still believe in the OLS estimator $\hat{\beta}$?

Statistical properties in finite samples, unbiasedness

Another objection to the unbiasedness criterion is summarised nicely by the story of three econometricians who go duck hunting. The first shoots about a foot in front of the duck, the second about a foot behind. The third yells: "We got him!"

Statistical properties in finite samples, efficiency

We need to judge $\hat{\boldsymbol{\beta}}$ on an additional criterion than unbiasedness to preserve our belief in $\hat{\boldsymbol{\beta}}$. This new criterion is

$$\text{Var}\left[\hat{\boldsymbol{\beta}}_0 \mid \boldsymbol{X}\right] \geq \text{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right].$$

It says that the variance of the sampling distribution of the OLS estimator $\hat{\boldsymbol{\beta}}$ is the smallest when compared to the variance of the sampling distribution of any other competing unbiased estimator $\hat{\boldsymbol{\beta}}_0$.

If this is true, we say that $\hat{\boldsymbol{\beta}}$ is the best unbiased estimator.

Statistical properties in finite samples, efficiency, proof

Derive the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon},$$

and hence

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}.$$

Using the general variance formula, conditional on $\boldsymbol{X}$,

$$\begin{aligned}
\operatorname{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] &= \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right] \\
&= \mathrm{E}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \mid \boldsymbol{X}\right] \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right]\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}
\end{aligned}$$

if $\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] = \sigma^2\boldsymbol{I}$.

Statistical properties in finite samples, efficiency, proof

Derive the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_0$, conditional on $\boldsymbol{X}$.

$$\hat{\boldsymbol{\beta}}_0 = \boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\varepsilon}$$

where $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$ by unbiasedness of $\hat{\boldsymbol{\beta}}_0$. Hence,

$$\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\varepsilon}.$$

Using the general variance formula, conditional on $\boldsymbol{X}$,

$$\begin{aligned}
\operatorname{Var}\left[\hat{\boldsymbol{\beta}}_0 \mid \boldsymbol{X}\right] &= \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right] \\
&= \mathrm{E}\left[\boldsymbol{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{C}' \mid \boldsymbol{X}\right] \\
&= \boldsymbol{C}\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right]\boldsymbol{C}' \\
&= \boldsymbol{C}\sigma^2\boldsymbol{I}\boldsymbol{C}' \\
&= \sigma^2\boldsymbol{C}\boldsymbol{C}'
\end{aligned}$$

since $\boldsymbol{C}$ depends on $\boldsymbol{X}$, and if $\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}\right] = \sigma^2\boldsymbol{I}$.

Statistical properties in finite samples, efficiency, proof

$$
\begin{aligned}
\text{Var}\left[\hat{\beta}_0 \mid \boldsymbol{X}\right] - \text{Var}\left[\hat{\beta} \mid \boldsymbol{X}\right] &= \sigma^2 \boldsymbol{C}\boldsymbol{C}' - \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2 \left[\boldsymbol{C}\boldsymbol{C}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \\
&= \sigma^2 \left[\boldsymbol{C}\boldsymbol{C}' - \boldsymbol{I}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{I}'\right] \\
&= \sigma^2 \left[\boldsymbol{C}\boldsymbol{C}' - \boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}'\right] \\
&= \sigma^2 \boldsymbol{C} \left[\boldsymbol{C}' - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}'\right] \\
&= \sigma^2 \boldsymbol{C} \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right] \boldsymbol{C}' \\
&= \sigma^2 \boldsymbol{C} \left[\boldsymbol{I} - \boldsymbol{P}\right] \boldsymbol{C}' \\
&= \sigma^2 \boldsymbol{C}\boldsymbol{M}\boldsymbol{C}' \\
&= \sigma^2 \boldsymbol{C}\boldsymbol{M}\boldsymbol{M}'\boldsymbol{C}' \\
&= \sigma^2 \boldsymbol{C}\boldsymbol{M}(\boldsymbol{C}\boldsymbol{M})' \\
&\geq 0
\end{aligned}
$$

since $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}$, $\boldsymbol{M}$ is symmetric and idempotent, and $\boldsymbol{C}\boldsymbol{M}(\boldsymbol{C}\boldsymbol{M})'$ is positive semidefinite (or nonnegative definite) (Greene, A-114).

Statistical properties in finite samples, efficiency

Let $\boldsymbol{q}$ be a $K \times 1$ vector of constants. Use $\boldsymbol{q}$ to obtain a linear combination of the true coefficients of a LRM such that

$$\boldsymbol{q}'\boldsymbol{\beta} = c_1\beta_1 + c_2\beta_2 + \ldots + c_K\beta_K.$$

It is easy to verify that $\boldsymbol{q}'\hat{\boldsymbol{\beta}}_0$ and $\boldsymbol{q}'\hat{\boldsymbol{\beta}}$ are unbiased estimators of $\boldsymbol{q}'\boldsymbol{\beta}$. The difference between the variances of these estimators is

$$\begin{aligned}
\text{Var}\left[\boldsymbol{q}'\hat{\boldsymbol{\beta}}_0 \mid \boldsymbol{X}\right] - \text{Var}\left[\boldsymbol{q}'\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] &= \boldsymbol{q}'\text{Var}\left[\hat{\boldsymbol{\beta}}_0 \mid \boldsymbol{X}\right]\boldsymbol{q} - \boldsymbol{q}'\text{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right]\boldsymbol{q} \\
&= \boldsymbol{q}'\left[\text{Var}\left[\hat{\boldsymbol{\beta}}_0 \mid \boldsymbol{X}\right] - \text{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right]\right]\boldsymbol{q} \\
&= \boldsymbol{q}'\sigma^2\boldsymbol{CM}(\boldsymbol{CM})'\boldsymbol{q} \\
&= \sigma^2\boldsymbol{q}'\boldsymbol{CM}(\boldsymbol{CM})'\boldsymbol{q} \\
&\geq 0
\end{aligned}$$

since $\text{Var}[\boldsymbol{aX}] = \boldsymbol{a}\text{Var}[\boldsymbol{X}]\boldsymbol{a}'$ for the constant and random vectors $\boldsymbol{a}$ and $\boldsymbol{X}$, and $\boldsymbol{q}'\boldsymbol{CM}(\boldsymbol{CM})'\boldsymbol{q}$ is positive semidefinite.

Statistical properties in finite samples, efficiency

This shows that the OLS estimator $\hat{\boldsymbol{\beta}}$ yields the smallest variance when it is used to estimate any linear combination of $\boldsymbol{\beta}$.

This result allows us to carry out a joint hypothesis test on $\boldsymbol{\beta}$ using a suitable test statistic which will need to employ the variance of $\boldsymbol{q}'\hat{\boldsymbol{\beta}}$. E.g., suppose we want to test if

$$\boldsymbol{q}'\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \beta_2 - \beta_3$$

is 0. In a suitable test statistic, we now know that we can use

$$\text{Var}\left[\boldsymbol{q}'\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right]$$

because we know that

$$\text{Var}\left[\boldsymbol{q}'\hat{\boldsymbol{\beta}}_0 \mid \boldsymbol{X}\right] \geq \text{Var}\left[\boldsymbol{q}'\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right].$$

Is the OLS estimator a linear estimator?

If an estimator is a linear function of the dependent variable, it is a linear estimator. Is $\hat{\boldsymbol{\beta}}$ a linear estimator?

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

$\hat{\boldsymbol{\beta}}$ is a linear function of the values of $\boldsymbol{y}$. The values of $\boldsymbol{y}$ are linearly combined using weights that are a non-linear function of the values of $\boldsymbol{X}$. Hence, $\hat{\boldsymbol{\beta}}$ is a linear estimator with respect to how it uses the values of the dependent variable only, irrespective of how it uses the values of the regressors.

Is the OLS estimator a linear estimator?

Consider the bivariate LRM

$$\boldsymbol{y} = \boldsymbol{x}_0\beta_0 + \boldsymbol{x}_1\beta_1 + \boldsymbol{\varepsilon}.$$

$\boldsymbol{x}_0$ is a column of ones. In this model

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) y_i - \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \bar{y}.$$

$\hat{\beta}_1$ is a linear function of the values of $\boldsymbol{y}$.

Statistical properties in finite samples

Gauss-Markov Theorem. In the LRM with regressor matrix $\boldsymbol{X}$, the OLS estimator, $\hat{\boldsymbol{\beta}}$, is the minimum variance, linear, unbiased estimator of $\boldsymbol{\beta}$. For any vector of constants $\boldsymbol{q}$, the minimum variance linear unbiased estimator of $\boldsymbol{q}'\boldsymbol{\beta}$ in the regression model is $\boldsymbol{q}'\hat{\boldsymbol{\beta}}$.

Notes on the variance of the OLS estimator

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}.$$

$\boldsymbol{X}$ contains the column of ones $\boldsymbol{x}_0$ and other k variables.

$$\text{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

and it can be shown that

$$\text{Var}\left[\hat{\beta}_j \mid \boldsymbol{X}\right] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2\left(1 - R_j^2\right)}, \; j = 0, 1, \ldots, k.$$

where R_j^2 is the coefficient of determination from regressing $\boldsymbol{x}_j$ on all the other regressors. The expression shows that the variance of the OLS estimator is (i) increasing with the variance of ε, (ii) decreasing with the sample size, (iii) decreasing with the sample variance of $\boldsymbol{X}$, and (iv) increasing with R_j^2.

Notes on the variance of the OLS estimator

By the LIE, it is also true that

$$\begin{aligned}
\text{Var}\left[\hat{\boldsymbol{\beta}}\right] &= \text{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right] \\
&= \text{E}_{\boldsymbol{X}}\left[\text{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \boldsymbol{X}\right]\right] \\
&= \text{E}_{\boldsymbol{X}}\left[\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \\
&= \sigma^2 \text{E}_{\boldsymbol{X}}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right].
\end{aligned}$$

Notes on the variance of the OLS estimator

$$\text{Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

The problem is that σ^2 is unobserved. An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-K}.$$

s is the standard error of the regression.

The estimator of the variance-covaraince matrix of $\hat{\boldsymbol{\beta}}$ is then given by

$$\text{Est. Var}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-K}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Notes on the variance of the OLS estimator

The estimator of the standard error of $\hat{\boldsymbol{\beta}}$ is

$$\text{Est. S.E.}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] = \sqrt{\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}}.$$

What is a standard error? Recall the notion of the sampling distribution of $\hat{\boldsymbol{\beta}}$. The standard deviation of the sampling distribution of $\hat{\boldsymbol{\beta}}$ is the standard error of $\hat{\boldsymbol{\beta}}$. This is important! Hence it is an exercise in the lab.

Statistical properties in finite samples, normality

We have shown that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}.$$

Assume for the first time that $\boldsymbol{\varepsilon}$ is multivariate normal (A6). That is,

$$\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N\left[\mathbf{0}, \sigma^2 \boldsymbol{I}\right].$$

Is $\hat{\boldsymbol{\beta}}$ multivariate normal?

Statistical properties in finite samples, normality

We condition on $\boldsymbol{X}$ and hence treat it is as given. The matrix

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

is $K \times n$. Recast it as a $K \times n$ matrix

$$\begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \dots & \boldsymbol{w}_n \end{bmatrix}.$$

ε is $n \times 1$. $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon$ becomes

$$\boldsymbol{w}_1\varepsilon_1 + \boldsymbol{w}_2\varepsilon_2 + \ldots + \boldsymbol{w}_n\varepsilon_n.$$

Hence, $\hat{\boldsymbol{\beta}}$ is a linear combination of the elements of ε. A linear combination of normal random variables is normal. Hence, $\hat{\boldsymbol{\beta}}$ is multivariate normal.

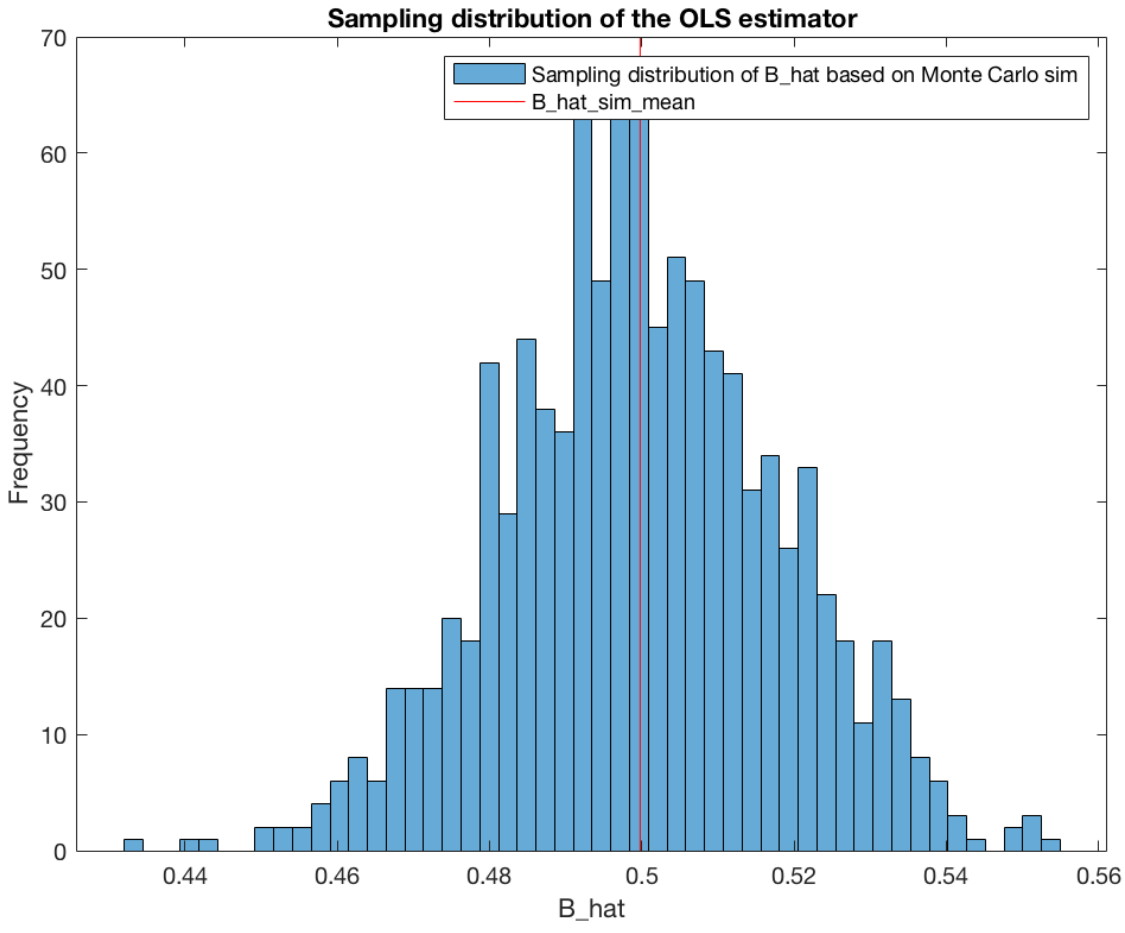
Statistical properties in finite samples, normality

Using the mean and variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ derived above,

$$\hat{\boldsymbol{\beta}} \mid \boldsymbol{X} \sim N\left[\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right].$$

Statistical properties in finite samples, normality

```
histogram(B_hat_sim)
```



Recall that in our simulation we assumed that the error is normal.

Statistical properties in finite samples, normality

In finite sample analysis, the normal distribution of $\hat{\boldsymbol{\beta}}$ is a consequence of the assumption that $\boldsymbol{\varepsilon}$ is normal, and that $\boldsymbol{X}$ is constant. In large sample analysis, we will obtain an approximate normal distribution for $\hat{\boldsymbol{\beta}}$, without assuming that $\boldsymbol{\varepsilon}$ is normal, and $\boldsymbol{X}$ is constant.

Statistical properties in large samples

We used unbiasedness and efficiency as criteria to judge if $\hat{\beta}$ is a good estimator. These criteria do not depend on n. We would consider new criteria to judge $\hat{\beta}$ based on n. But why?

Statistical properties in large samples

First reason. $\hat{\beta}$ should come closer to the population parameter β if we increase n and come closer to the population N. Who wants an estimator that does not satisfy this?

Statistical properties in large samples

Second reason. Estimators developed to estimate true parameters in complicated models are usually biased. We could check if a biased estimator in a small sample becomes an unbiased estimator in a large sample. E.g., the OLS estimator is biased if the lagged dependent variable is an explanatory variable in the model. However, it is consistent.

Statistical properties in large samples

Third reason. Often the derivation of a property of an estimator is not tractable in a small sample but in a large sample. This is because the expected value of a non-linear function of a statistic is not the non-linear function of the expected value of that statistic. But the plim of a non-linear function of a statistic is the non-linear function of the plim of that statistic.

Statistical properties in large samples

Recall the sampling distribution of $\hat{\boldsymbol{\beta}}$ obtained in the simulation experiment. Imagine creating a sequence of sampling distributions of $\hat{\boldsymbol{\beta}}$ with successively larger n. If the distributions in this sequence become more and more similar in form to some specific distribution as n becomes extremely large, this specific distribution is called the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.

Statistical properties in large samples, consistency

If the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ becomes concentrated on the particular value $\boldsymbol{\beta}$ as n approaches infinity, $\boldsymbol{\beta}$ is said to be the probability limit of $\hat{\boldsymbol{\beta}}$. We then write

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta},$$

or

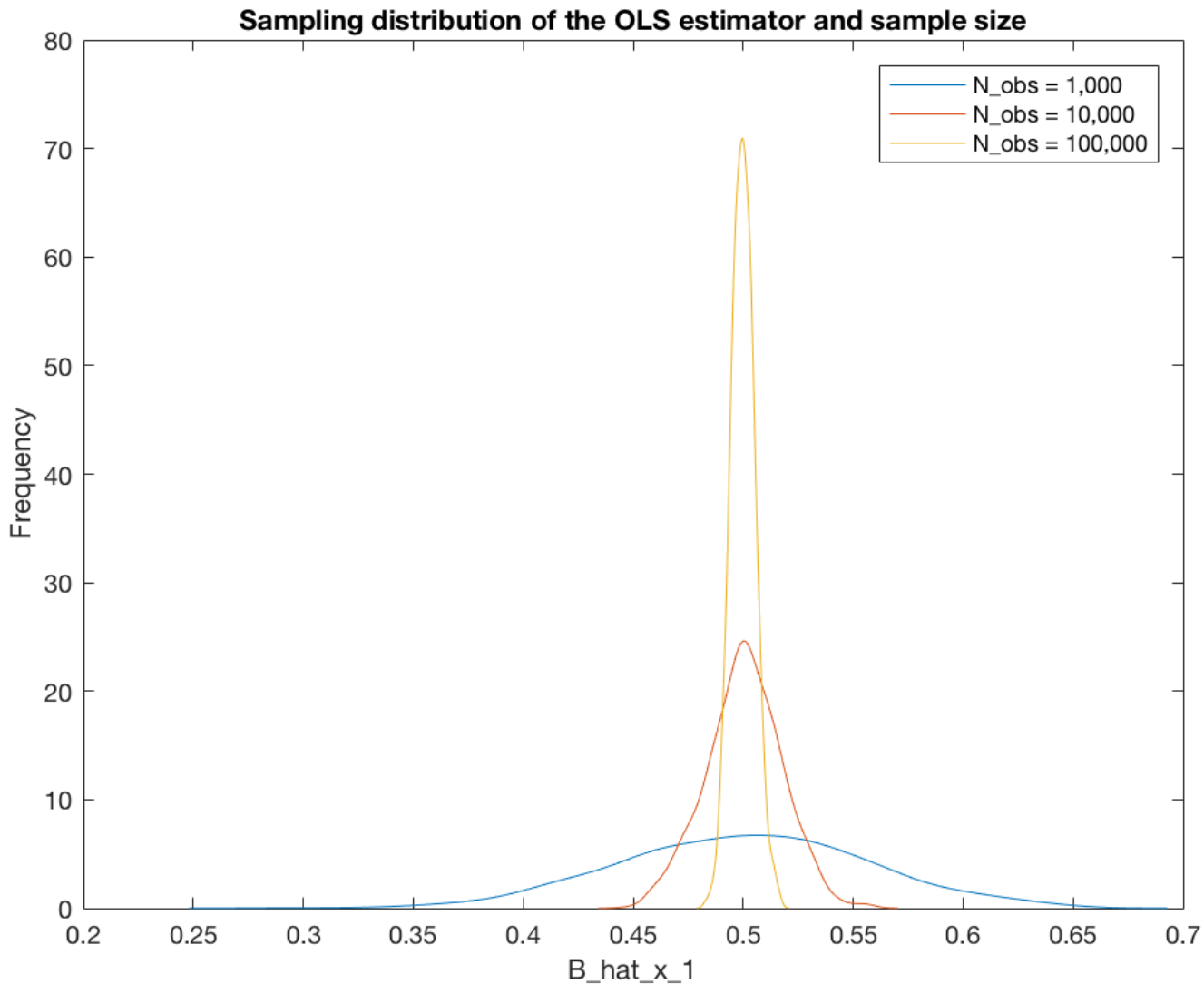
$$\text{plim}\, \hat{\boldsymbol{\beta}} = \boldsymbol{\beta},$$

or

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1).$$

We then say that $\hat{\boldsymbol{\beta}}$ is consistent. This is our first large sample criterion.

Statistical properties in large samples, consistency



Statistical properties in large samples, consistency

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}.\end{aligned}$$

Statistical properties in large samples, consistency

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \dots & 1 & \dots & 1 \\ x_{21} & x_{22} & \dots & x_{2i} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \dots & x_{ji} & \dots & x_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{ki} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} 1 & x_{21} & \dots & x_{j1} & \dots & x_{k1} \\ 1 & x_{22} & \dots & x_{j2} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{2i} & \dots & x_{ji} & \dots & x_{ki} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & \dots & x_{jn} & \dots & x_{kn} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_i' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix}$$

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

Statistical properties in large samples, consistency

$$\boldsymbol{X}'\boldsymbol{\varepsilon} = \begin{bmatrix} 1 & 1 & \dots & 1 & \dots & 1 \\ x_{21} & x_{22} & \dots & x_{2i} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \dots & x_{ji} & \dots & x_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{ki} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i .$$

Statistical properties in large samples, consistency

Hence,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

becomes

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

Take the plim of both sides of the equation. When taking the plim, we do not condition on $\boldsymbol{X}$. To derive asymptotic results, we do not need the technical simplification brought by fixing $\boldsymbol{X}$ in repeated samples.

Statistical properties in large samples, consistency

Using the sum rule of plim (Greene, Theorem D.14),

$$\operatorname{plim} \hat{\boldsymbol{\beta}} = \operatorname{plim} \boldsymbol{\beta} + \operatorname{plim} \left[\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i \right].$$

Using the product rule of plim (Greene, Theorem D.14),

$$\operatorname{plim} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \operatorname{plim} \left[\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \right] \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i.$$

Statistical properties in large samples, consistency

Assuming that x_i is i.i.d. (A5), and using the WLLN (Greene, Theorem D.5),

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{p} \mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right].$$

Statistical properties in large samples, consistency

Using the ratio rule of plim (Greene, Theorem D.14), and assuming that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists (A2),

$$\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \xrightarrow{p} \left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}.$$

Statistical properties in large samples, consistency

Assuming that ε_i is i.i.d. (A5), assuming that $\mathrm{E}[\boldsymbol{x}_i\varepsilon_i] = \mathbf{0}$ (A3), and using the WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\varepsilon_i \xrightarrow{p} \mathrm{E}[\boldsymbol{x}_i\varepsilon_i] = \mathbf{0}.$$

Statistical properties in large samples, consistency

$$\operatorname{plim} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \underbrace{\operatorname{plim} \left[\left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \right]}_{\left(\mathsf{E}[\boldsymbol{x}_i \boldsymbol{x}_i']\right)^{-1}} \underbrace{\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i}_{\mathbf{0}}$$

Hence,

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}.$$

Statistical properties in large samples, consistency

The probability limit can be seen as the large-sample equivalent of the expected value. Hence, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ can be seen as the large-sample equivalent of unbiasedness.

Statistical properties in large samples, consistency

$\hat{\boldsymbol{\beta}}$ is also consistent under weaker versions of A5 and A3. The former is an exercise in the tutorial. The latter is as follows. To show consistency, we assume

$$\mathrm{E}\left[\boldsymbol{x}_i \varepsilon_i\right] = \mathbf{0}$$

which is weak exogeneity, and not

$$\mathrm{E}\left[\boldsymbol{X} \varepsilon_i\right] = \mathbf{0}$$

which is strict exogeneity. But remember that in finite sample analysis, for unbiasedness, we have assumed strict exogeneity. This shows that as n increases, we enjoy a weaker model assumption.

Statistical properties in large samples: asy. efficiency

The variance of the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is called the asymptotic variance of $\hat{\boldsymbol{\beta}}$. Among the consistent estimators, if the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is smaller than the asymptotic variance of any other estimator, $\hat{\boldsymbol{\beta}}$ is said to be asymptotically efficient. This is our second large sample criterion.

Statistical properties in large samples: asy. efficiency

Suppose that the asymptotic variance of a competitor estimator $\hat{\boldsymbol{\beta}}_0$ is $\boldsymbol{\Omega}$. Then, under assumptions A1, A2, A3, A4, and A5, it can be shown that

$$\begin{aligned} \text{Asy. Var}(\hat{\boldsymbol{\beta}}_0) - \text{Asy. Var}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\Omega} - \frac{\sigma^2}{n} \left(\text{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i' \right] \right)^{-1} \\ &\geq 0. \end{aligned}$$

This is the large-sample equivalent of the efficiency criteria we have considered in finite samples. We do not consider the proof.

Statistical properties in large samples: asy. normality

Consider our earlier result

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

Rearrange the terms, and multiply both sides of the equation with $\sqrt{n}$ to obtain

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

Statistical properties in large samples: asy. normality

We already know that

$$\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \xrightarrow{p} \left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}.$$

Convergence in probability implies convergence in distribution. Hence,

$$\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \xrightarrow{d} \left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}.$$

Note that $\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]$ is a constant matrix and has no variance.

Statistical properties in large samples: asy. normality

Assuming that $\mathrm{E}[\boldsymbol{x}_i \varepsilon_i] = 0$ (A3); assuming that $\boldsymbol{x}_i$ and ε_i are both i.i.d. (A5), and applying the CLT (Greene, Theorem D.19A),

$$\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\varepsilon_i \xrightarrow{d} N\left[\mathbf{0}, \sigma^2 \mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right].$$

We did not assume that ε_i is normal (A6). We are enjoying the CLT!

Statistical properties in large samples: asy. normality

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}}_{\xrightarrow{d}\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}}\underbrace{\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\varepsilon_i}_{\xrightarrow{d}N\left[\mathbf{0},\sigma^2\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right]}}\ .$$

Using the product rule of limiting distributions (Greene, Theorem D.16),

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}N\left[\mathbf{0},\sigma^2\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right].$$

Statistical properties in large samples: asy. normality

$$
\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &\xrightarrow{d} \left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1} N\left[\mathbf{0}, \sigma^2 \mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right] \\
&\xrightarrow{d} N\left[\mathbf{0}, \sigma^2\left(\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\right)\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\left(\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\right)'\right] \\
&\xrightarrow{d} N\left[\mathbf{0}, \sigma^2\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\right] \\
&\xrightarrow{d} N\left[\mathbf{0}, \sigma^2\left(\mathrm{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'\right]\right)^{-1}\right],
\end{aligned}
$$

using the property that the transpose and inverse operations commute from the second to the third line.

Statistical properties in large samples: asy. normality

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{d} N\left[\mathbf{0}, \sigma^2\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}\right]$$

is a limiting distribution.

This means that when n is infinite, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ has an exact distribution, which is the limiting distribution.

This implies that when n is finite, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ has an approximate distribution, which is close to the limiting distribution. At least this is what we want to believe in, and therefore we assume this. This approximate distribution is also called the asymptotic distribution. The point is that we know that if we let n increase, the approximation is better. We know that if we let n go to infinity, we will get the exact distribution (Greene, Section D.3).

Statistical properties in large samples: asy. normality

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{d} N\left[\mathbf{0}, \sigma^2\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}\right]$$

Assuming that this limiting distribution holds approximately for finite n,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{a} N\left[\mathbf{0}, \sigma^2\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}\right].$$

$$\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{a} N\left[\mathbf{0}, \sigma^2 \frac{1}{n}\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}\right].$$

$$\hat{\boldsymbol{\beta}} \xrightarrow{a} N\left[\boldsymbol{\beta}, \sigma^2 \frac{1}{n}\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}\right],$$

or

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \sigma^2 \frac{1}{n}\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}\right].$$

Statistical properties in large samples: asy. normality

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \sigma^2 \frac{1}{n}\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}\right].$$

σ^2 and $1/n\left(\mathrm{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i'\right]\right)^{-1}$ are population terms of the limiting distribution. They are unobserved. In practice, they are estimated with $\hat{\sigma}^2$ and $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ given sample data, respectively.