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

Empirical Methods, Lecture 4

Tunga Kantarcı, FEB, Groningen University, Spring 2025

Sampling distribution

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

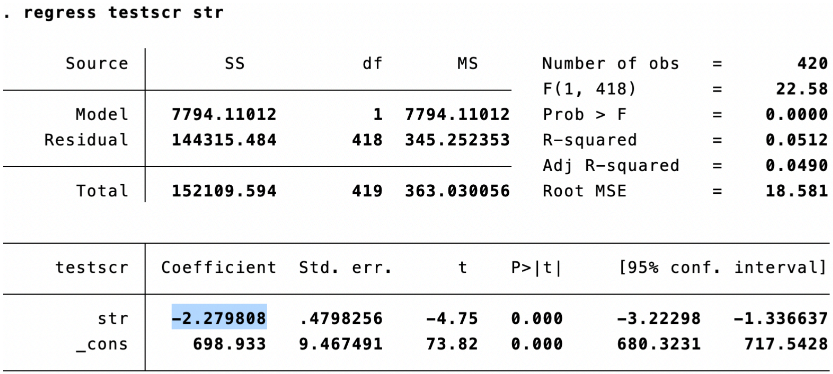
Sampling distribution

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

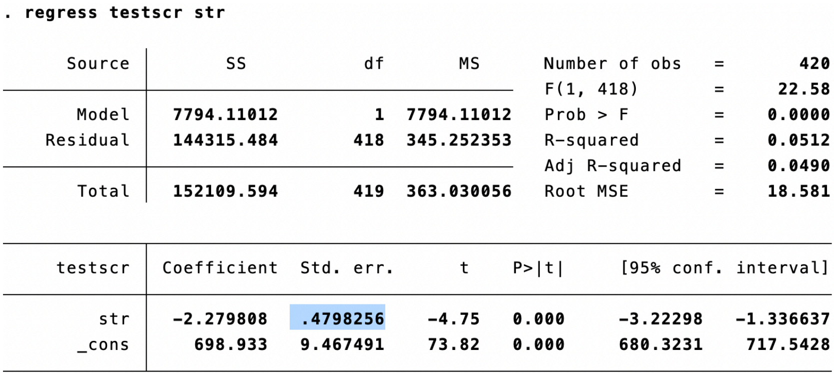


```
. regress testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11012         1  7794.11012   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

Sampling distribution

Consider now the standard error of $\hat{\beta}$.

Sampling distribution



```
. regress testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11012         1  7794.11012   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

Sampling distribution

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

Sampling distribution

The distribution of $\hat{\beta}$ 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.

However, in contrast to standard teaching practice, we will demonstrate them in an applied manner. That is, 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 $\hat{\boldsymbol{\beta}}$.

Simulating the sampling distribution

Assume a value for the true $\boldsymbol{\beta}$.

Take n random draws for $\boldsymbol{X}$ from a probability distribution.

Take n random draws for $\boldsymbol{\varepsilon}$ from a probability distribution.

Given $\boldsymbol{\beta}$ and the random draws for $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$, generate n observations for $\boldsymbol{y}$.

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

Then take n new random draws for $\boldsymbol{\varepsilon}$. Given $\boldsymbol{\beta}$ and the original random draws for $\boldsymbol{X}$, generate n new observations for $\boldsymbol{y}$. Using the newly generated $\boldsymbol{y}$ and original $\boldsymbol{X}$, obtain a new $\hat{\boldsymbol{\beta}}$.

Repeat the procedure to obtain many $\hat{\boldsymbol{\beta}}$. This gives the simulated sampling distribution of $\hat{\boldsymbol{\beta}}$.

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 other than 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.

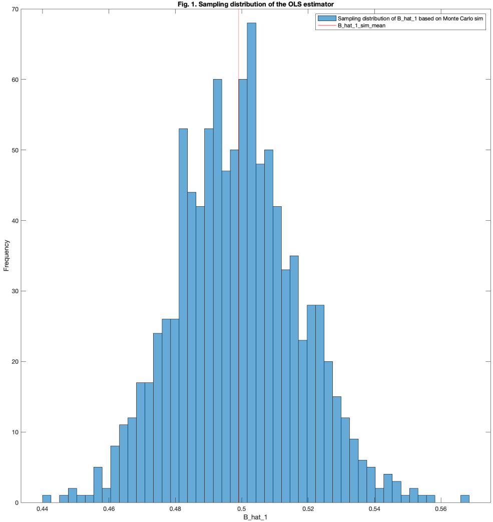
Simulating the sampling distribution

```
# Simulate the sampling distribution of the OLS estimator
N_sim = 1000;
N_obs = 9000;
B_true = 0.5;
x = random('Uniform',-1,1,[N_obs],1);
B_hat_sim = NaN(1,N_sim);
for i = 1:N_sim
    e = random('Normal',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

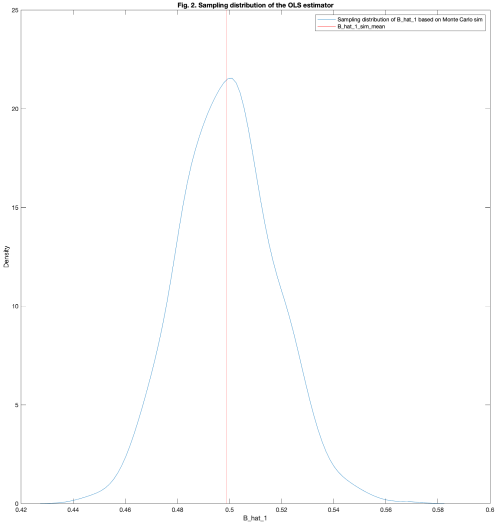
```
histogram(B_hat_sim)
```

Fig. 1. Sampling distribution of the OLS estimator



Simulating the sampling distribution

```
kdesnity(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 need to make a distinction between a small and large sample because statistical properties of any statistic depend on the sample size.

We want the sampling distribution of $\hat{\beta}$ to behave in certain ways in a small, or finite, sample. We want the sampling distribution of $\hat{\beta}$ 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 the slides.

Statistical properties in finite samples, unbiasedness

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

Statistical properties in finite samples, unbiasedness

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

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

We break down this statement, to understand what it says.

Statistical properties in finite samples, unbiasedness

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

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

The sampling distribution of coefficient estimates is created by repeatedly taking all possible samples of a same size from the population. Here we consider the average of all these estimates. Recall that the mean in the population is the expected value. That is, expected value is a population term. It is not a sample term.

Statistical properties in finite samples, unbiasedness

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

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

$\boldsymbol{X}$ is the sample data at hand for the explanatory variables. Conditioning on $\boldsymbol{X}$ means that we fix $\boldsymbol{X}$, take many samples of $\boldsymbol{Y}$, and use samples of $\boldsymbol{Y}$ and the single sample of $\boldsymbol{X}$ to produce the sampling distribution of $\hat{\boldsymbol{\beta}}$. Fixing $\boldsymbol{X}$ allows us to focus on the variability in $\hat{\boldsymbol{\beta}}$ due to the random variation in $\boldsymbol{Y}$.

Why do we focus on the variation in $\boldsymbol{Y}$? The error of the regression, ε, captures the random variation in $\boldsymbol{Y}$ that is not explained by the observed explanatory variables $\boldsymbol{X}$. By focusing on the random variation in $\boldsymbol{Y}$, we can better understand and model the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Statistical properties in finite samples, unbiasedness

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

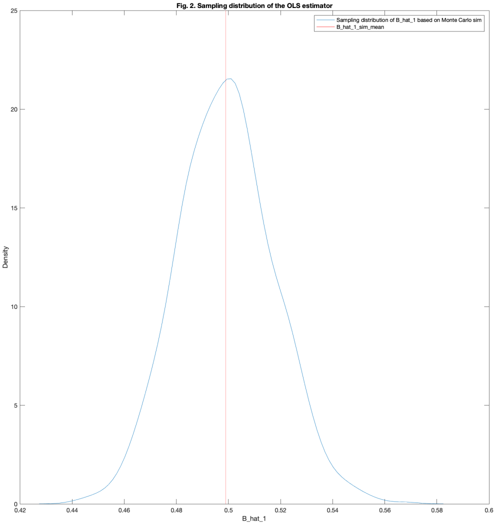
$$\mathrm{E}\left[\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right] = \boldsymbol{\beta}.$$

It says that the mean of the sampling distribution of $\hat{\boldsymbol{\beta}}$, given the sample data at hand on $\boldsymbol{X}$, 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 $\boldsymbol{\beta}$.

Who would want a biased estimator, right?

Statistical properties in finite samples, unbiasedness

Recall the sampling distribution of $\hat{\beta}$, created using simulation:



Notes on the variance of the OLS estimator, S.E. estimator

The aim is to calculate the mean of the sampling distribution of the OLS coefficient estimates. We consider an unrealistic setup where we are able to do repeated sampling, via a simulation exercise, and hence observe the sampling distribution. The second paragraph considers the realistic case where we cannot do repeated sampling to have a sampling distribution. We have only one sample data at hand.

Statistical properties in finite samples, unbiasedness

`mean(B_hat_sim)` gives 0.4998! `B_hat_sim` contains OLS coefficient estimates from repeated sampling. Here we consider all `B_hat_sim` that are created in the simulation exercise. That is, `B_hat_sim` is a vector containing all simulated OLS coefficient estimates from repeated sampling, and we take the mean of it.

Notes on the variance of the OLS estimator, S.E. estimator

In the simulation exercise `B_true` was set to 0.5. Are you surprised that 0.4998 is so close to 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 needs to be typical.

Statistical properties in finite samples, unbiasedness

In practice, if it was difficult to draw a sample that is typical, the unbiasedness criterion would have been criticized. This 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!" In practice, the random sample econometricians take is usually representative of the population, so is typical, so to get close to $\hat{\beta}$, so that econometricians are not mocked by the joke.

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 a 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 $\boldsymbol{X}$,

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

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

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

It says that the variance of the sampling distribution of the unbiased 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.

We skip the proof.

Is the OLS estimator a linear estimator?

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

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

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

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

$\hat{\beta}_1$ 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 (best), 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, formula

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.

Notes on the variance of the OLS estimator, formula

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

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

and hence

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

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

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

if $\operatorname{E}\left[\varepsilon\varepsilon' \mid \boldsymbol{X}\right] = \sigma^2\boldsymbol{I}$. Hence, homoskedasticity is a requirement.

Notes on the variance of the OLS estimator, formula

By the LIE, it is also true that

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

Skip.

Notes on the variance of the OLS estimator, formula

It can be shown that element j of

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

is

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

Notes on the variance of the OLS estimator, Var. estimator

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

We cannot use this formula because σ^2 is unobserved. An unbiased estimator of σ is

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

Or, since the vector multiplication implies sum of squared residuals, we have

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-K}}.$$

This is the standard error of the regression.

Notes on the variance of the OLS estimator, Var. estimator

The estimator of the variance-covariance 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, S.E. estimator

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

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

Or in level form it is

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

Note that the current model is the multiple linear regression model. In the simple linear regression model, we have

$$\text{Est. S.E.}\left[\hat{\beta}_j \mid \boldsymbol{X}\right] = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}},\ \ j = 0, 1.$$

Notes on the variance of the OLS estimator, S.E. estimator

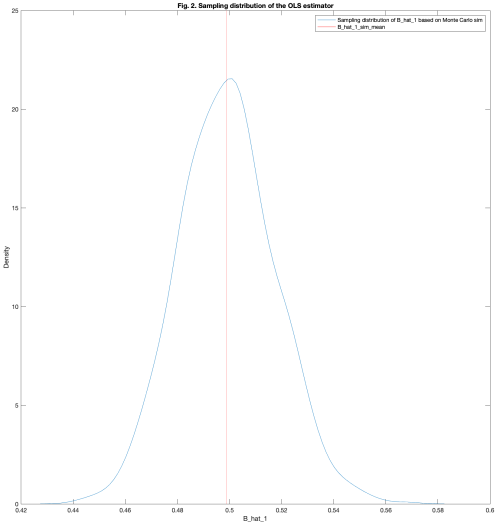
What does the standard error estimate of the OLS estimate really represent?

Recall the sampling distribution of $\hat{\boldsymbol{\beta}}$. The standard deviation of the sampling distribution of $\hat{\boldsymbol{\beta}}$ is the standard error of $\hat{\boldsymbol{\beta}}$!

Est. S.E. $\left[\hat{\beta}_j \mid \boldsymbol{X}\right]$ is just a mathematical formula to estimate this standard deviation using the sample data at hand because we cannot take repeated samples from the population and draw the sampling distribution of the OLS estimates and calculate the standard deviation of them.

Notes on the variance of the OLS estimator, S.E. estimator

The sampling distribution of $\hat{\beta}$ we created using simulation:



Notes on the variance of the OLS estimator, S.E. estimator

The aim is to calculate the S.D. of the sampling distribution of the OLS coefficient estimates. The first paragraph below considers an unrealistic setup where we are able to do repeated sampling, via a simulation exercise, and hence observe the sampling distribution. The second paragraph considers the realistic case where we cannot do repeated sampling to have a sampling distribution. We have only one sample data at hand.

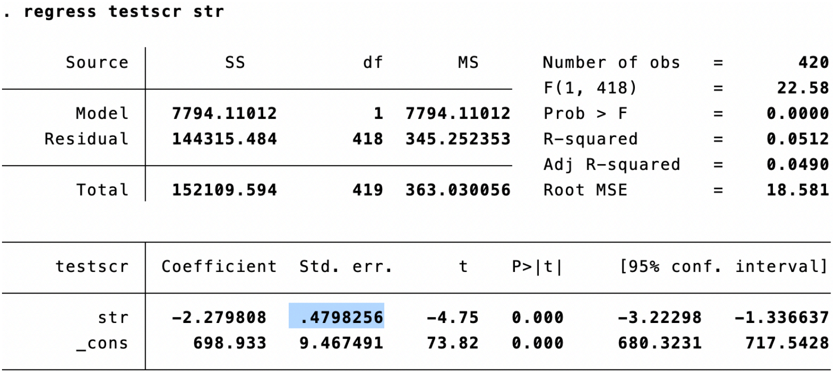
Notes on the variance of the OLS estimator, S.E. estimator

`std(B_hat_sim)` gives 0.0183. `B_hat_sim` contains OLS coefficient estimates from repeated sampling. Here we consider all `B_hat_sim` that are created in the simulation exercise. That is, `B_hat_sim` is a vector containing all simulated OLS coefficient estimates from repeated sampling, and we take the S.D. of it.

Notes on the variance of the OLS estimator, S.E. estimator

Est. S.E. $\left[\hat{\beta}_j \mid \boldsymbol{X}\right]$ gives 0.0182! Here among all `B_hat_sim` that are created in the simulation exercise, we pick one, and use that in the formula for Est. S.E. $\left[\hat{\beta}_j \mid \boldsymbol{X}\right]$ to estimate the S.D. of the sampling distribution of the OLS coefficient estimates.

Notes on the variance of the OLS estimator, S.E. estimator



```
. regress testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11012         1  7794.11012   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

The Root mean squared error is the standard error of the regression. Std. err. is the standard error estimate of the OLS estimate.

Notes on the variance of the OLS estimator, determinants

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

The expression shows that the S.E. of the OLS estimator is

i. higher if the variance of $\hat{\varepsilon}$ is higher,

ii. lower if the sample size is larger,

iiii. lower if the sample variance of x_{ij} is larger, and

iv. higher if R_j^2 is higher, where R_j^2 is the coefficient of determination from regressing $\boldsymbol{x}_j$ on all other regressors.

Notes on the variance of the OLS estimator, determinants

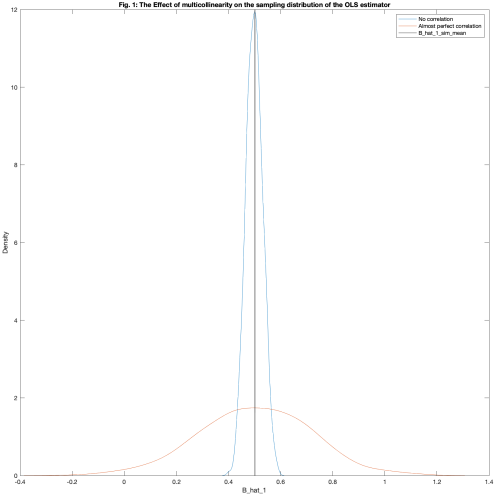
Let's pay more attention to the effect of R_j^2 on Est. S.E. $\left[\hat{\beta}_j \mid \boldsymbol{X}\right]$.

R_j^2 will always be positive because explanatory variables will always be correlated to some extent, even spuriously. So some correlation between explanatory variables is not a problem.

What is a problem is if they are strongly correlated, a phenomenon known as multicollinearity. When explanatory variables are strongly correlated, they share a significant amount of common information, which makes it challenging to estimate their individual effects accurately. This leads to inflated S.E.s for the estimated regression coefficients, resulting in less reliable estimates.

Notes on the variance of the OLS estimator, determinants

Sampling distribution of an OLS estimator under almost perfect correlation and no correlation:



Notes on the variance of the OLS estimator, determinants

S.D. of the sampling distribution of an OLS estimator under different correlation levels:

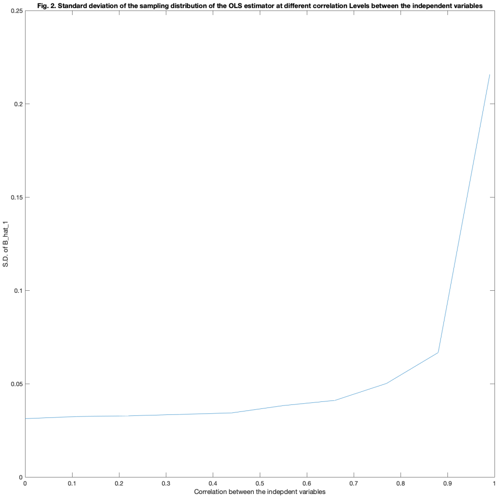


Fig. 2. Standard deviation of the sampling distribution of the OLS estimator at different correlation Levels between the independent variables

Notes on the variance of the OLS estimator, determinants

Note that this is about multicollinearity, not perfect multicollinearity. The latter was already discussed under assumption A2.

Statistical properties in finite samples, normality

We have shown that

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

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

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

$\boldsymbol{\varepsilon}$ is $n \times 1$. $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$ 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 $\boldsymbol{\varepsilon}$. A linear combination of normal random variables is normal. Hence, $\hat{\boldsymbol{\beta}}$ is multivariate normal.

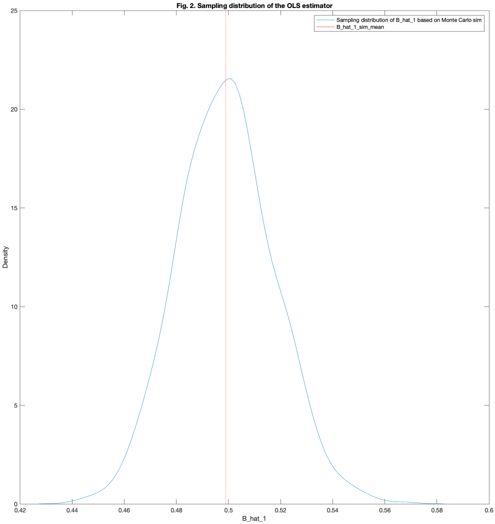
Statistical properties in finite samples, normality

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

Simulated sampling distribution of $\hat{\boldsymbol{\beta}}$, where errors are 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 later in these slides, 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 finite samples, normality

What happens if the assumption of normality of ε does not hold in finite samples?

Statistical properties in finite samples, normality

Error when it has a normal and and when it has a t distribution:

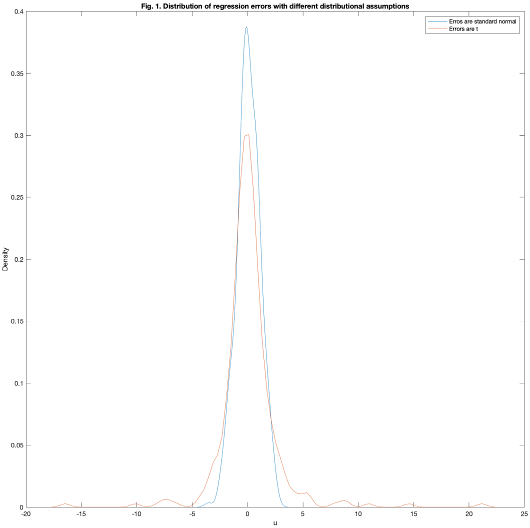
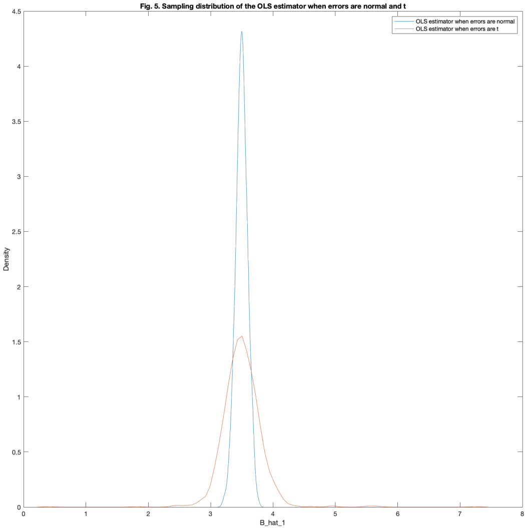


Fig. 1. Distribution of regression errors with different distributional assumptions

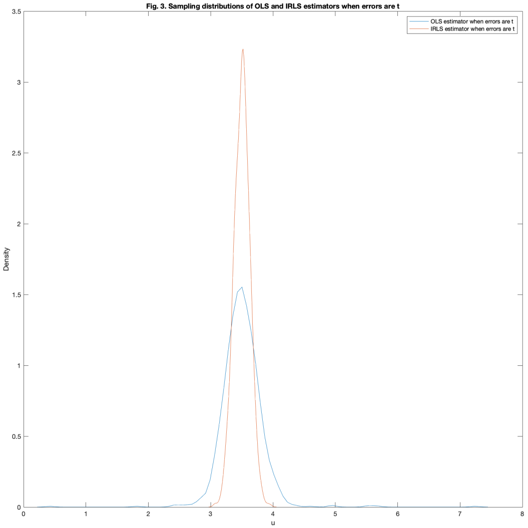
Statistical properties in finite samples, normality

OLS estimator when the error has a normal and t distribution:



Statistical properties in finite samples, normality

OLS and IRLS estimators when the error has a t distribution:



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 will consider new criteria to judge $\hat{\beta}$ based on n. But why do we want to do that?

Statistical properties in large samples

First reason. $\hat{\boldsymbol{\beta}}$ should come closer to the population $\boldsymbol{\beta}$ 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 population 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 unbiased asymptotically, that is when n is very large, which makes the OLS estimator as we say, 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 probability limit of a non-linear function of a statistic is the non-linear function of the probability limit of that statistic.

Statistical properties in large samples

Recall the sampling distribution of $\hat{\beta}$ obtained in the simulation experiment. Imagine creating a sequence of sampling distributions of $\hat{\beta}$ 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 $\hat{\beta}$.

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

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

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, proof

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

Statistical properties in large samples, consistency, proof

$$
\begin{aligned}
\boldsymbol{X}'\boldsymbol{X} &= \begin{bmatrix} 1 & 1 & \dots & 1 & \dots & 1 \\ x_{21} & x_{22} & \dots & x_{2i} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \dots & x_{ji} & \dots & x_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{ki} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} 1 & x_{21} & \dots & x_{j1} & \dots & x_{k1} \\ 1 & x_{22} & \dots & x_{j2} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{2i} & \dots & x_{ji} & \dots & x_{ki} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{2n} & \dots & x_{jn} & \dots & x_{kn} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_i' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'.
\end{aligned}
$$

Statistical properties in large samples, consistency, proof

$$
\begin{aligned}
\boldsymbol{X}'\varepsilon &= \begin{bmatrix} 1 & 1 & \dots & 1 & \dots & 1 \\ x_{21} & x_{22} & \dots & x_{2i} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \dots & x_{ji} & \dots & x_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{ki} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i \varepsilon_i .
\end{aligned}
$$

Statistical properties in large samples, consistency, proof

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, proof

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

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

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

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

Statistical properties in large samples, consistency, proof

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, proof

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, proof

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

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

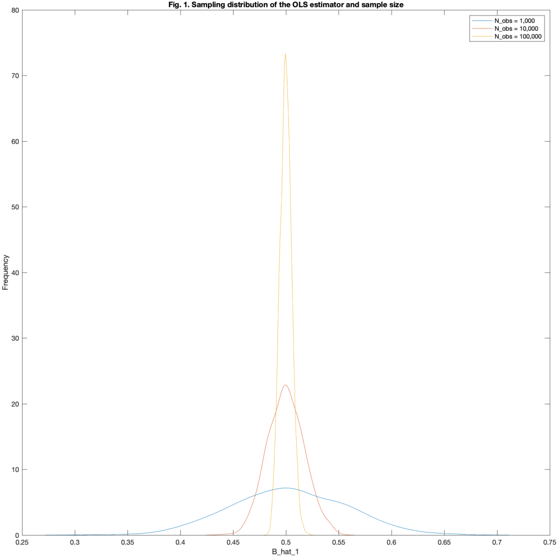
Statistical properties in large samples, consistency, proof

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

Hence,

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

Statistical properties in large samples, consistency



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. We do not discuss the former here. The latter is as follows. To prove 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

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

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

Assuming that $\mathrm{E}\left[\boldsymbol{x}_i \varepsilon_i\right] = 0$ (A3); assuming that $\boldsymbol{x}_i$ and ε_i are both i.i.d. (A5), and not assuming that ε_i is normal (A6) but applying the central limit theorem (Greene, Theorem D.19A),

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

This is the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. That is, if we let n go to infinity, $\hat{\boldsymbol{\beta}}$ has an exact normal distribution.

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.