The sampling distribution of the OLS estimator, its standard deviation, and how to estimate it

Econometrics for minor Finance, Lecture 4

Tunga Kantarcı, Fall 2025

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

Suppose our population regression function is

$$y = \beta_0 + \beta_1 x_1 + u$$

This is the true relationship we assume exists in the population, under the model assumptions we made such as

$$\mathrm{E}\left[u \mid x_1\right] = 0$$

and

$$\mathrm{Var}\left[u \mid x_1\right] = \sigma^2$$

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

The OLS estimator of

$$\beta_1$$

is

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

and it is a function of the sample data which is random. Hence the estimator is random. From one sample to another, its value varies. Therefore the estimator has a sampling distribution.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

In this lecture we will conduct a conceptual simulation exercise. We will repeatedly draw samples from the population to mimic repeated sampling and reveal the sampling distribution of the OLS estimator.

In reality, this distribution is unobservable. The purpose of the simulation is to demonstrate that such a sampling distribution always exists conceptually. We will use the simulated distribution to understand what is going on in econometrics in the rest of this course.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

The population model is

$$y = \beta_0 + \beta_1 x_1 + u$$

β_0 : Assume a value for the intercept.

β_1 : Assume a value for the slope.

x_1 : Draw a random sample of size N from a chosen PDF.

u : Draw a random sample of size N from a chosen PDF.

y : Generate observations using the above of same sample size: the data generating process.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

The generated y and x give us a paired sample. Using this sample, and the OLS estimator, we obtain the estimate

$$\hat{\beta}_1$$

We of course also obtain $\hat{\beta}_0$, but let's focus on the slope parameter.

By repeating this procedure across many samples, we obtain many such estimates. This gives rise to the sampling distribution of the OLS estimator.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

In the simulation, we shall keep the N observations of x_1 fixed while repeatedly generating new y. This simplifies the experiment: the variation in the sampling distribution can then be attributed to counterfactual scenarios other than the sampling variance of x_1. This mirrors what we do in statistical derivations. We condition on x_1, meaning we treat it as fixed. This greatly simplifies those derivations. In reality, x_1 is random unless it comes from an experimental design where the researcher chooses x_1 before y is realized. We justify treating x_1 as fixed by invoking the random sampling assumption.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

In the population model, we assume

$$\mathrm{E}\left[u \mid x_1\right] = 0$$

In the simulation, we enforce this assumption by generating u independently of x_1 so that the simulated data is in line with one of the assumptions of the data generating process.

We also enforce all the other assumptions we made.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

```
N_sim = 1000
N_obs = 9000
B_0 = 0.2
B_1 = 0.5
x = random('Uniform', -1, 1, [N_obs 1])
B_hat_0_sim = NaN(1, N_sim)
B_hat_1_sim = NaN(1, N_sim)
for i = 1:N_sim
    u = random('Normal', 0, 1, [N_obs 1])
    y = B_0 + B_1 * x + u
    B_hat_1 = sum((x-mean(x)).*(y-mean(y))) /
              sum((x-mean(x)).^2);
    B_hat_0 = mean(y) - B_hat_1 * mean(x);
    B_hat_1_sim(1,i) = B_hat_1
    B_hat_0_sim(1,i) = B_hat_0
end
```

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

```
B_hat_1_sim(1,:)
```

stores the simulated estimates from 1000 repeated samples. The first 6 estimates look like:

1×1000 double

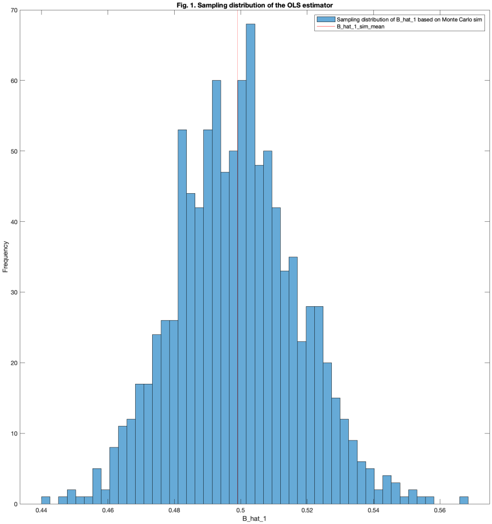
	1	2	3	4	5	6
1	0.5139	0.5155	0.5186	0.5220	0.4602	0.4782

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

```
histogram(B_hat_1_sim(1,:))
```

plots the histogram of these estimates, visualizing the sampling distribution of the OLS estimator. This illustrates that the estimator is a random variable whose values differ across samples. The shape is approximately normal, a point we will return to later.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

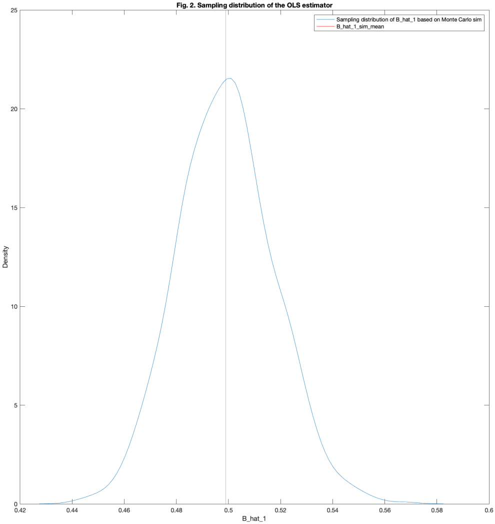


Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

```
kdensity(B_hat_1_sim(1,:))
```

produces the kernel density estimate, a smoothed version of the histogram. We adopt it because it is easier to compare across scenarios. For example, we can overlay sampling distributions from different sample sizes to observe how the distribution changes, something that is cumbersome with histograms.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population



Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

```
B_hat_1_sim(1,1000)
```

returns a simulated estimate from the sampling distribution as 0.5237. It uses the 1000th generated sample.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

The sampling distribution reminds us that there is uncertainty around an OLS estimate obtained from a sample.

The standard deviation of the sampling distribution of the OLS estimator provides a summary measure of this uncertainty.

Sampling distribution of the OLS estimator: Simulation: Repeated sampling from the population

```
std(B_hat_1_sim(1,:))
```

computes the standard deviation of the simulated OLS estimates from $N_{\text{sim}} = 1000$ repeated samples. Formally,

$$\text{SD}\left[\hat{\beta}_{1,\text{sim}}\right] = \sqrt{\frac{1}{N_{\text{sim}}-1}\sum_{n_{sim}=1}^{N_{\text{sim}}}\left(\hat{\beta}_{1,n_{sim}} - \overline{\hat{\beta}_1}\right)^2}$$

The result is 0.0186.

Sampling distribution of the OLS estimator: Reality: One sample from the population

In reality, we do not observe the sampling distribution of the OLS estimator, because repeated sampling from the population is not feasible. In reality, we only have one sample at hand, and hence one OLS estimate.

Sampling distribution of the OLS estimator: Reality: One sample from the population

```
. regress y x_1

      Source |       SS           df       MS      Number of obs   =     9,000
-------------+----------------------------------   F(1, 8998)      =    796.00
       Model |  810.94006         1   810.94006   Prob > F        =    0.0000
    Residual |  9166.89214     8,998  1.01876996   R-squared       =    0.0813
-------------+----------------------------------   Adj R-squared   =    0.0812
       Total |   9977.8322     8,999  1.10877122   Root MSE        =    1.0093

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         x_1 |   .5236875   .0185616    28.21   0.000     .4873025    .5600725
       _cons |   .1918419   .0106413    18.03   0.000     .1709825    .2127012
------------------------------------------------------------------------------
```

Sampling distribution of the OLS estimator: Reality: One sample from the population

This is standard Stata regression output. It shows the OLS estimate of the slope parameter based on a single sample, which is the typical situation in practice. Here, the sample happens to be the 1000th draw in the simulation.

Sampling distribution of the OLS estimator: Reality: One sample from the population

Without access to the sampling distribution, we cannot compute its standard deviation to quantify the uncertainty of this estimate.

But then how can we still learn about the precision of an OLS estimate from a single sample?

We need an estimator of this uncerainity.

Estimating the standard deviation of the OLS estimator: SD estimator: Derivation

Our population model is

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The OLS slope estimator is

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

Deviation form of $y_i - \bar{y}$ from the model is

$$y_i - \bar{y} = (\beta_0 + \beta_1 x_i + u_i) - (\beta_0 + \beta_1 \bar{x} + \bar{u}) = \beta_1(x_i - \bar{x}) + (u_i - \bar{u})$$

Estimating the standard deviation of the OLS estimator: SD estimator: Derivation

Substitute $y_i - \bar{y}$ in the slope estimator:

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

$$= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Difference from the true parameter:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Estimating the standard deviation of the OLS estimator: SD estimator: Derivation

Take the variance conditional on X:

$$\text{Var}\left[\hat{\beta}_1 \;\middle|\; x\right] = \text{Var}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \;\middle|\; x\right]$$

Expand the numerator variance:

$$\begin{aligned}\text{Var}\left[\sum_{i=1}^{n}(x_i - \bar{x})u_i \;\middle|\; x\right] &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \text{Var}\left[u_i \mid x\right] \\ &\quad + \sum_{i \neq j}(x_i - \bar{x})(x_j - \bar{x}) \text{Cov}\left[u_i, u_j \mid x\right]\end{aligned}$$

Estimating the standard deviation of the OLS estimator: SD estimator: Derivation: Where assumptions enter

Assume that errors are uncorrelated:

$$\operatorname{Cov}[u_i, u_j \mid x] = 0$$

for all $i \neq j$.

Assume that errors are homoskedastic:

$$\operatorname{Var}[u_i \mid x] = \sigma^2$$

for all i.

Estimating the standard deviation of the OLS estimator: SD estimator: Derivation

With these assumptions:

$$\operatorname{Var}\left[\sum_{i=1}^{n}(x_i - \bar{x})\, u_i \,\middle|\, x\right] = \sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Estimating the standard deviation of the OLS estimator: SD estimator: Derivation

For the denominator, as we condition on x, the fraction acts as constant, and is out of the variance operator as a square using var. property 2 in the first lecture slides:

$$\text{Var}\left[\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\ \middle|\ x\right] = \left[\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^2$$

Estimating the standard deviation of the OLS estimator: SD estimator: Derivation

The variance becomes

$$\text{Var}\left[\hat{\beta}_1 \,\middle|\, x\right] = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

Estimating the standard deviation of the OLS estimator: SD estimator

The standard deviation is the square root of the variance:

$$\text{SD}\left[\hat{\beta}_1 \,\middle|\, x\right] = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

This is the standard deviation in the population. The error

$$u_i$$

is the population error. The variance of it

$$\text{Var}\left[u_i \mid x\right] = \sigma^2$$

is that in the population. We do not observe it.

Estimating the standard deviation of the OLS estimator: SD estimator

The key is the homoskedadticity assumption. Without it, the variance is

$$\text{Var}\left[u_i \mid x\right] = \sigma_i^2$$

That is, it varies across units i. If this assumption does not hold, we can not use the standard deviation estimator we are about the derive! Later in this course we will derive another estimator of the standard deviation of the OLS estimator that does not need this assumption.

Estimating the standard deviation of the OLS estimator: SD estimator

We cannot use

$$\text{SD}\left[\hat{\beta}_1 \,\middle|\, x\right] = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

because the variance of the error

$$\text{Var}\left[u_i \mid x\right] = \sigma^2$$

is unobserved.

Estimating the standard deviation of the OLS estimator: SD estimator

An unbiased estimator of σ is

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

where $\hat{u}_i$ is the residual for i. This is called the regression standard error estimator.

'Regression standard error' is the conventional shorthand for regression standard error estimator. Sometimes also called the 'root mean squared error'.

Estimating the standard deviation of the OLS estimator: SD estimator

The estimator of the standard deviation of the OLS estimator is not called the standard deviation estimator but the standard error estimator. The word ‘error’ emphasizes that it is an estimate of uncertainty, not the true population spread. It is given by

$$\mathrm{SEE}\left[\hat{\beta}_1 \,\middle|\, x\right] = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

if errors are homoskedastic.

‘Standard error’ is the conventional shorthand for standard error estimator.

Estimating the standard deviation of the OLS estimator: SD estimator

The OLS estimator is a random variable. Its value changes across random samples, so it has a sampling distribution. The standard deviation of this distribution measures the uncertainty in a given OLS estimate. In practice, we do not observe the sampling distribution and therefore cannot compute its true standard deviation. Instead, we use the standard error estimator, which estimates this standard deviation using the sample data at hand.

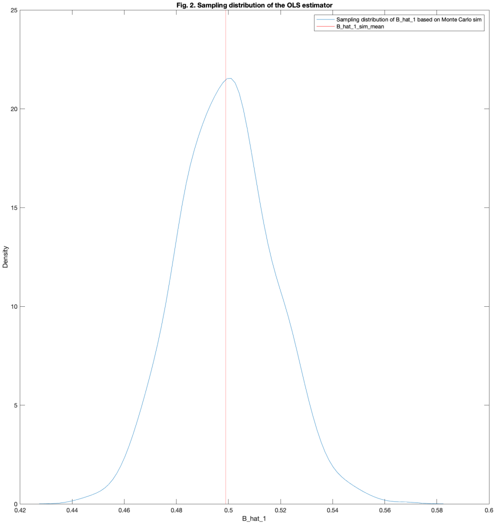
Estimating the standard deviation of the OLS estimator: SD estimator

Recall the sampling distribution of the OLS estimator

$$\hat{\beta}_1$$

that we created using simulation.

Estimating the standard deviation of the OLS estimator: SD estimator



Estimating the standard deviation of the OLS estimator: SD estimator

```
B_hat_1_sim(1,:)
```

was the vector containing all simulated OLS coefficient estimates from 1000 repeated samples that we used to create the sampling distribution.

Then

```
std(B_hat_1_sim(1,:))
```

took the standard deviation of these estimates. It gave 0.0186.

Estimating the standard deviation of the OLS estimator: SD estimator

Now, among all the repeated samples used to simulate OLS estimates, pick one of these samples to mimic reality, and use it to compute

$$\text{SEE}\left[\hat{\beta}_1 \,\middle|\, x\right] = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

which gives an estimate of the same standard deviation. It is 0.0185!

Estimating the standard deviation of the OLS estimator: SD estimator

See this estimate in the Stata regression output. This is what we have in practice.

Sampling distribution of the OLS estimator: Sample regression function

```
. regress y x_1

      Source |       SS           df       MS      Number of obs   =     9,000
-------------+----------------------------------   F(1, 8998)      =    796.00
       Model |  810.94006         1  810.94006   Prob > F        =    0.0000
    Residual |  9166.89214     8,998  1.01876996   R-squared       =    0.0813
-------------+----------------------------------   Adj R-squared   =    0.0812
       Total |  9977.8322     8,999  1.10877122   Root MSE        =    1.0093

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         x_1 |   .5236875   .0185616    28.21   0.000     .4873025    .5600725
       _cons |   .1918419   .0106413    18.03   0.000     .1709825    .2127012
------------------------------------------------------------------------------
```

SD estimator: Determinants

$$\text{SEE}\left[\hat{\beta}_1 \mid x\right] = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

The expression shows that the SEE of the OLS estimator is

i. higher if the estimated variance of the regression error $\hat{\sigma}^2$ is higher,

ii. lower if the sample size n is larger,

iiii. lower if the sample variation in the predictor $x_i - \bar{x}$ is larger.