

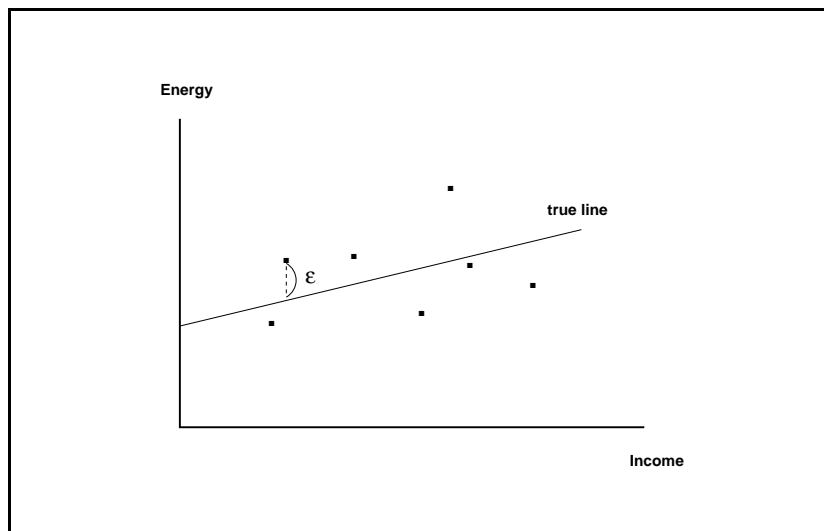
LECTURE / DISCUSSION

Statistical Properties of OLS

## Statistical Properties of OLS

$$y = \alpha + \beta x + \varepsilon$$

|                    |                    |  
dependent        included        omitted  
variable        explanatory        variables  
                  variables

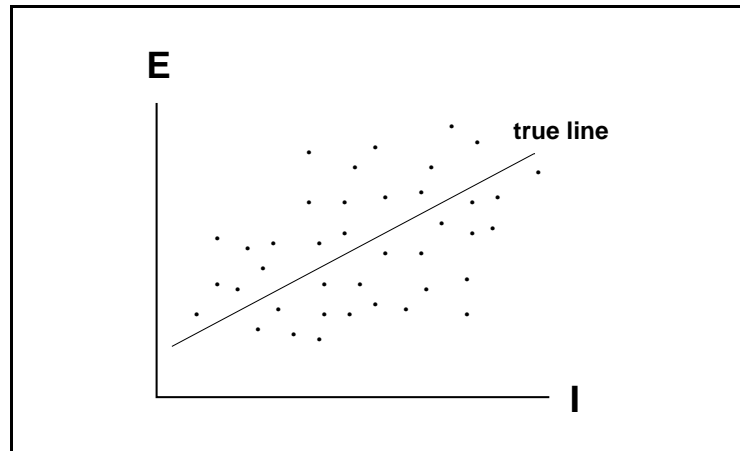


Elements of  $\varepsilon$  :

- size of dwelling
- number of members
- insulation level
- cracks around windows and doors
- attitudes toward conservation

☞ Some omitted variables **could** be included, and some cannot be measured completely and so **cannot** be included completely.

Entire population:



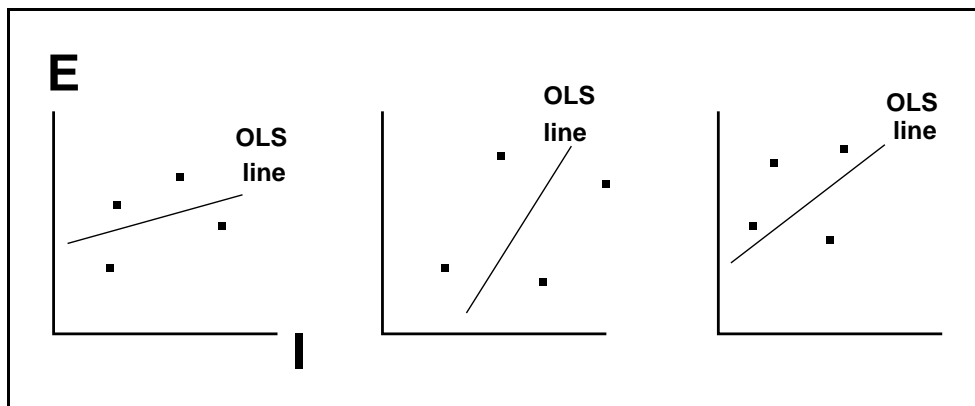
true  $\beta = 9$

Usually we sample from the population. Get a different **estimated** line for each sample.

Sample A

Sample B

Sample C

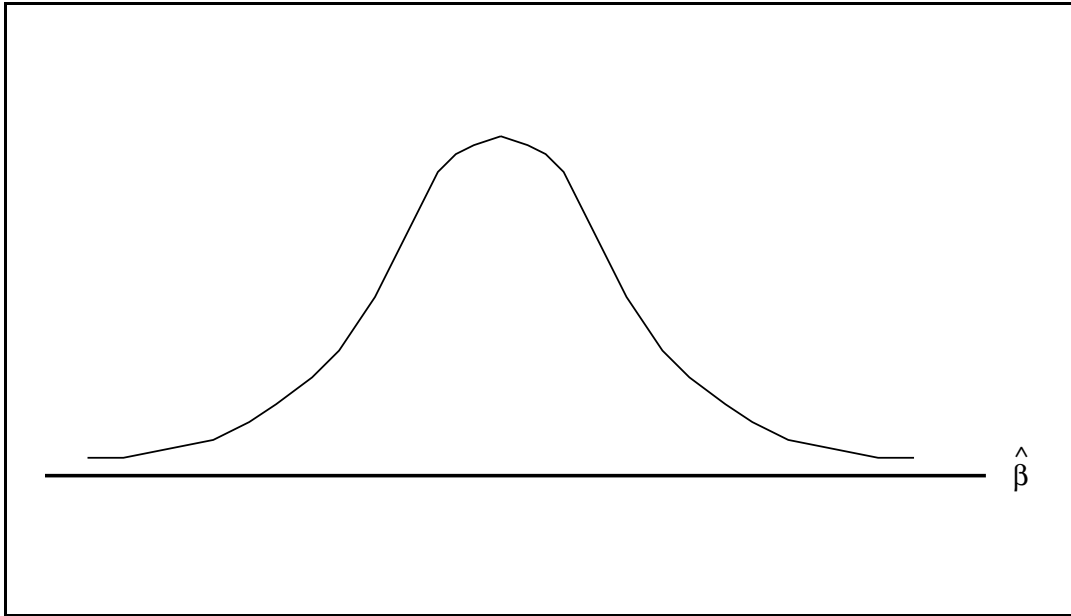


$\hat{\beta} = 5$

$\hat{\beta} = 12$

$\hat{\beta} = 9.5$

There is a distribution of  $\hat{\beta}$ 's :



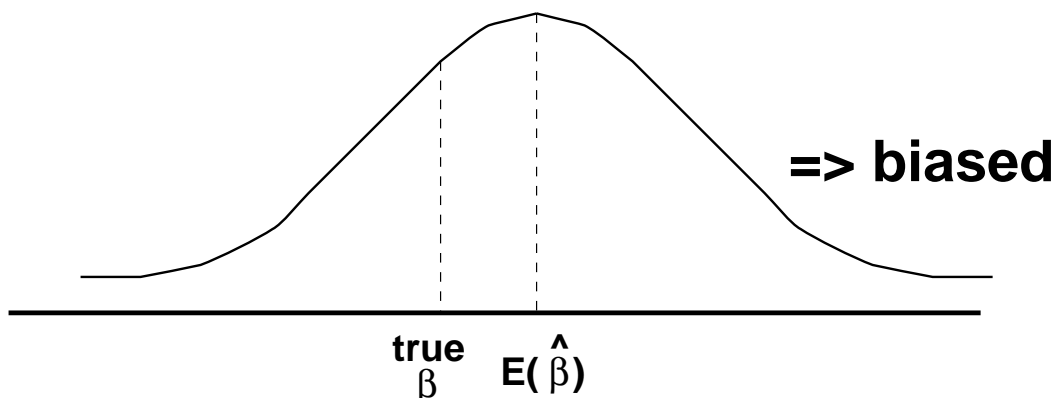
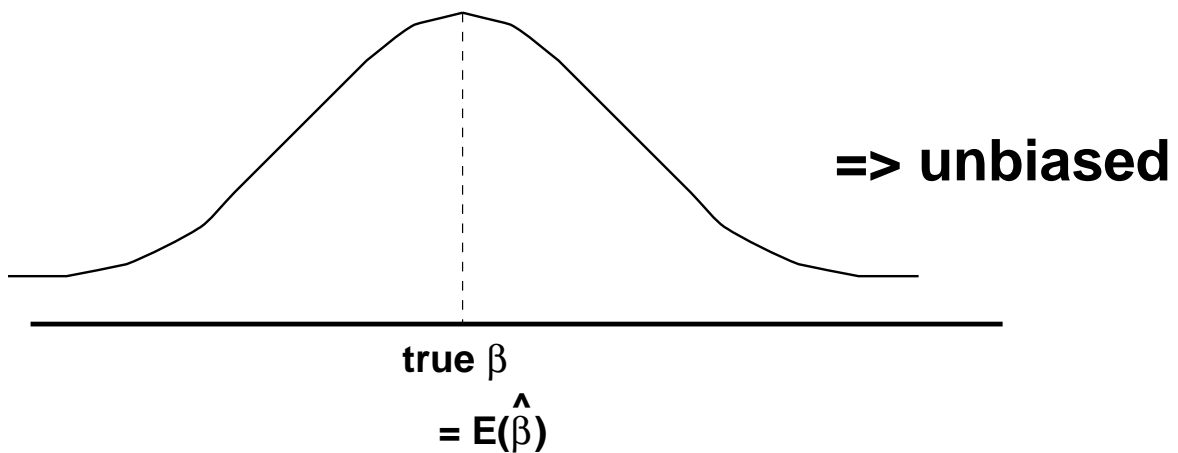
Issues About Distribution of OLS  $\hat{\beta}$  .

- Mean
- Variance
- Shape

## Unbiasedness

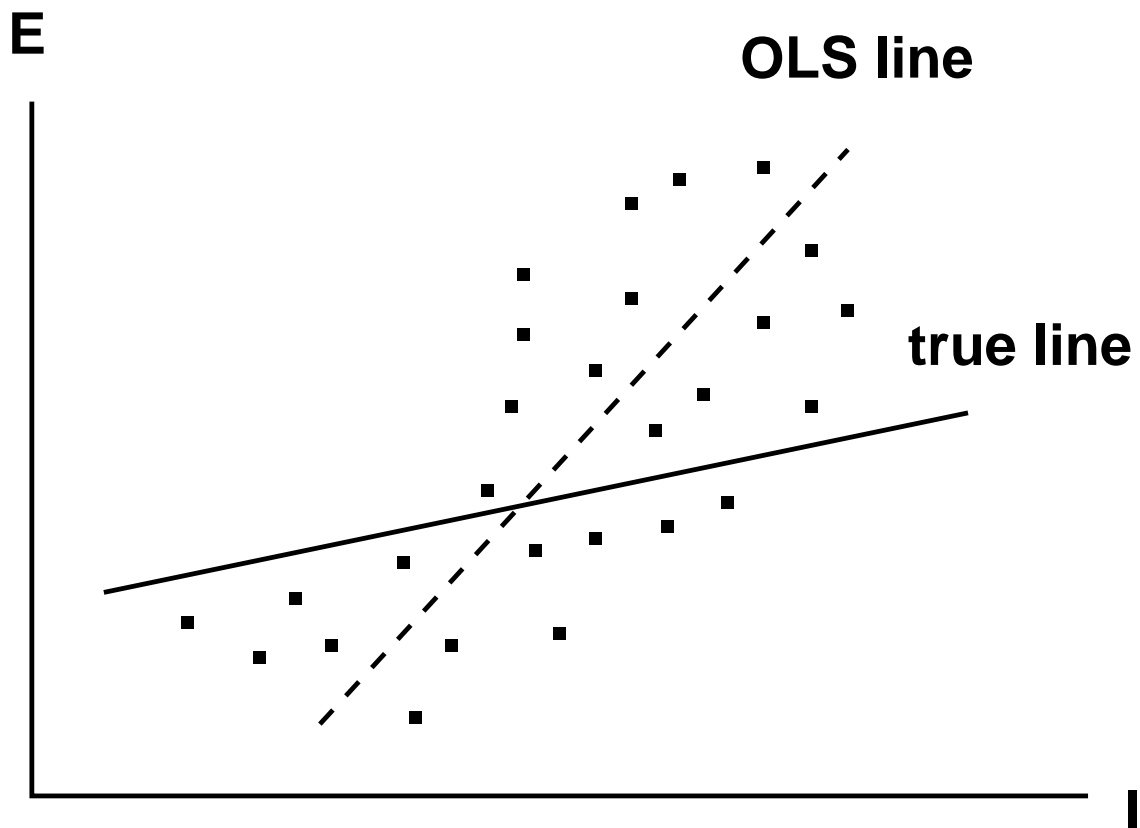
An estimator is unbiased if the mean of its frequency distribution is the true value.

That is, if  $E(\hat{\beta}) = \beta$ , then  $\hat{\beta}$  is unbiased.



OLS  $\hat{\beta}$  is biased if omitted variables are correlated with included variables in population.

Example:



Higher income households tend to have more members.

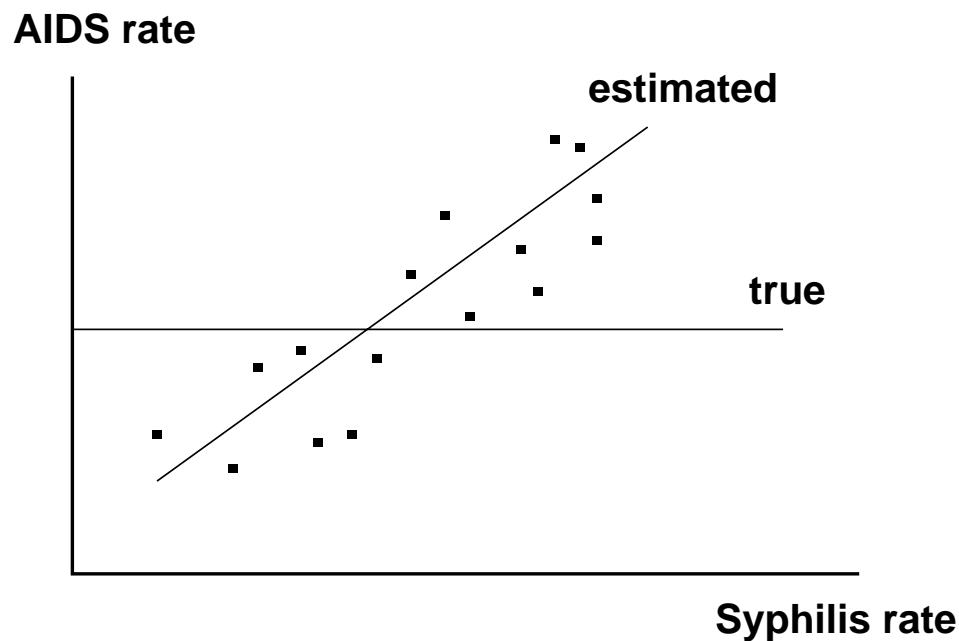
True line: Effect of extra income holding other factors, like number of members, constant.

Problem: OLS  $\hat{\beta}$  for income picks up the effects of household size.

Solutions:

1. Add the correlated omitted variables to the regression.
2. If option 1 is not possible, use instrumental variables estimation.

## Example



$$(\text{AIDS rate}) = \alpha + \beta(\text{syphilis rate}) + \varepsilon$$

Actually: frequency of unprotected sex causes transmission of syphilis and HIV.

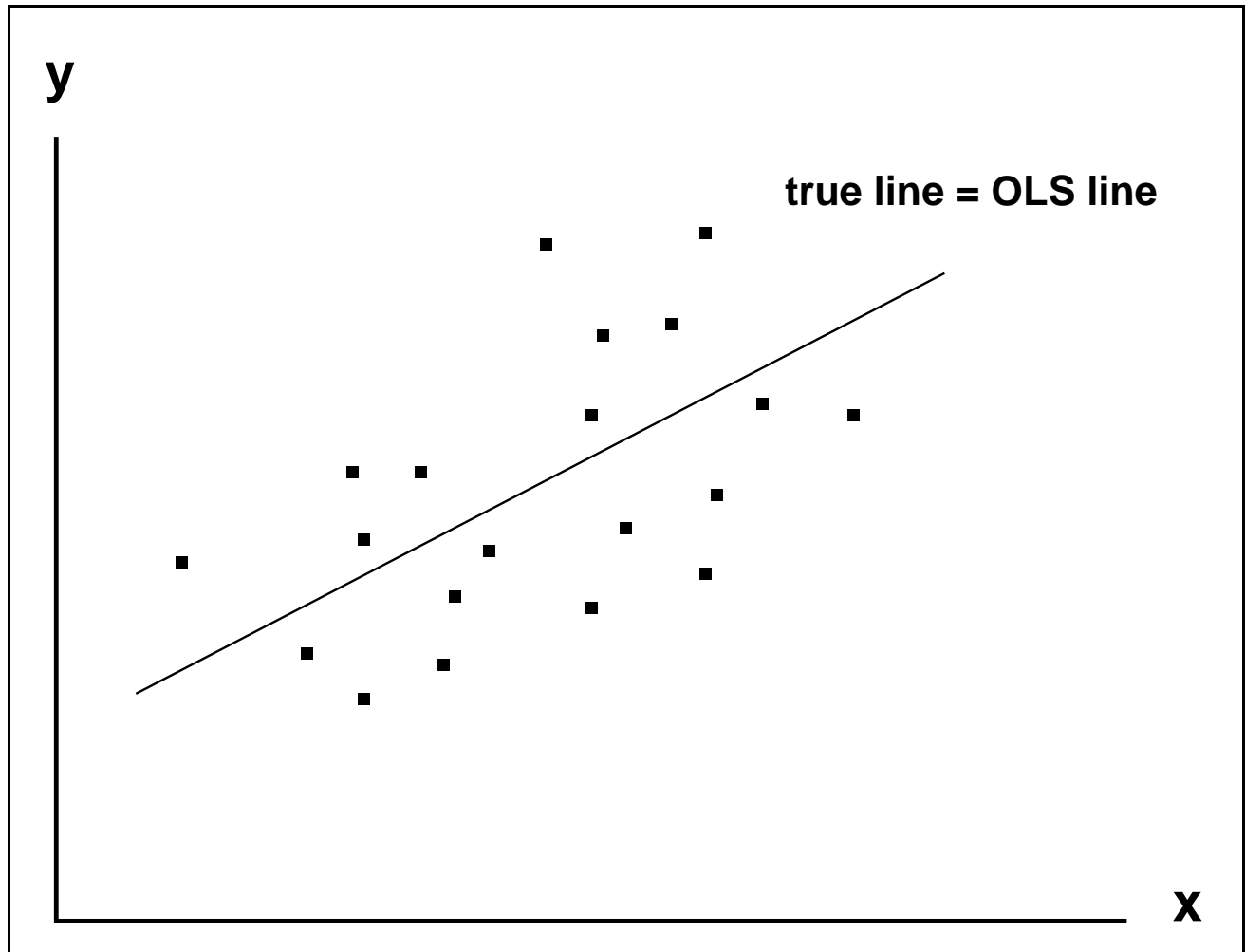
By omitting frequency of unprotected sex, OLS regression makes it **look** like syphilis causes AIDS.

## Another Way of Seeing the Problem

OLS finds the line for which the residuals are uncorrelated with the explanatory variables.

If residuals are in reality correlated with the explanatory variables, then OLS will give the wrong line.

OLS  $\hat{\beta}$  is **unbiased** if omitted variables are uncorrelated with included variables in population.



## Proof of Unbiasedness

Assume:  $Y_n = \alpha + \beta X_n + \varepsilon_n$

$\text{Corr}(\varepsilon, x) = 0$  in population

Proof: Because  $\alpha$  is included, the equation can be rewritten as deviations:

$$y_n = \beta x_n + \varepsilon_n$$

$$\begin{aligned}\hat{\beta} &= \frac{\sum y_n x_n}{\sum x_n^2} = \frac{\sum (\beta x_n + \varepsilon_n) x_n}{\sum x_n^2} \\ &= \frac{\beta \sum x_n^2 + \sum \varepsilon_n x_n}{\sum x_n^2} \\ &= \beta + \frac{\sum \varepsilon_n x_n}{\sum x_n^2}\end{aligned}$$

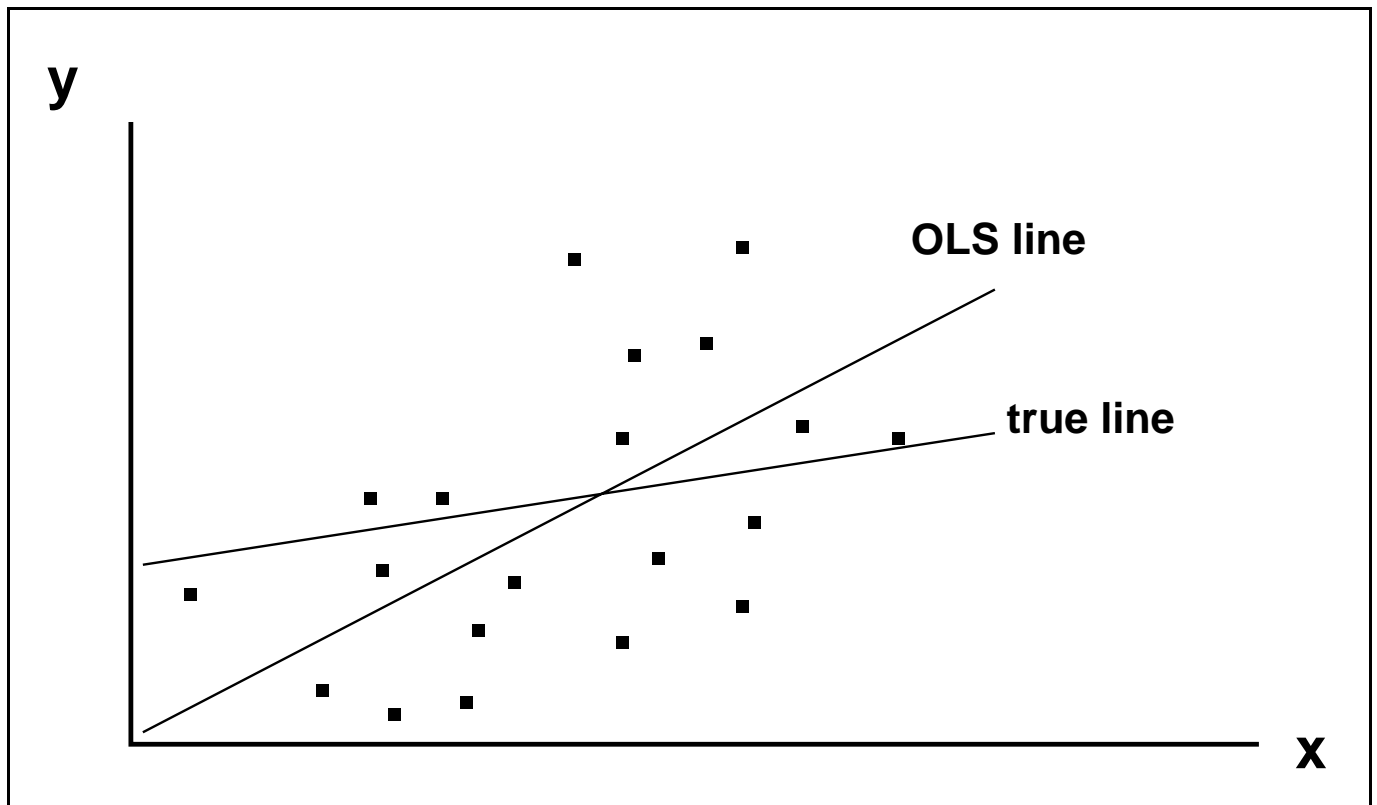
So:

$$\begin{aligned}E(\hat{\beta}) &= \beta + E\left[\frac{\sum \varepsilon_n x_n}{\sum x_n^2}\right] \\ &= \beta + \frac{\text{Cov}(\varepsilon_n, x_n)}{\text{Var}(x_n)} \\ &= \beta\end{aligned}$$

## Now consider the intercept.

Suppose true  $\alpha$  is not zero.

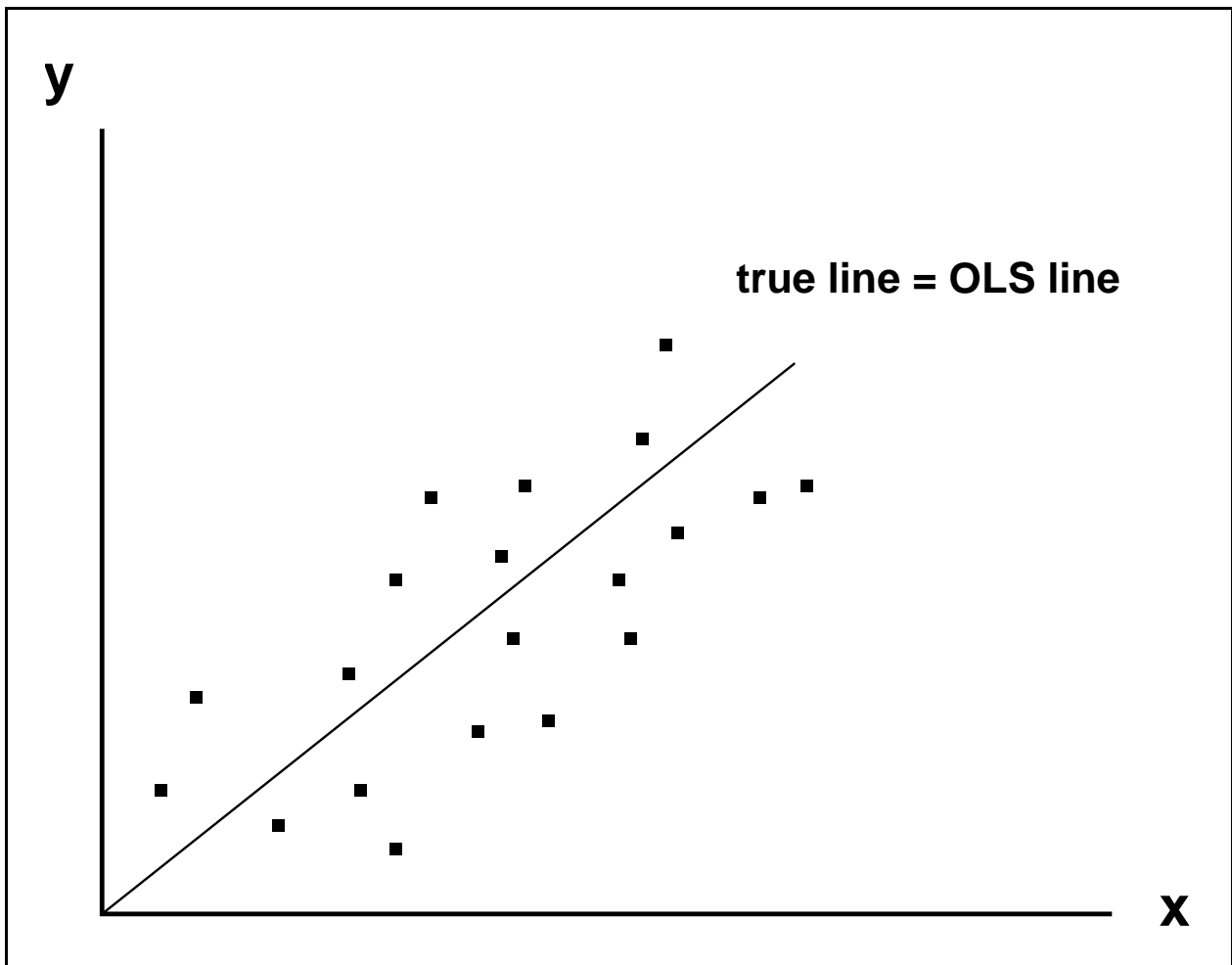
Omitting intercept in the regression makes the OLS  $\hat{\beta}$  biased.



Solution: Include an intercept.

Exception: Suppose true  $\alpha$  is zero.

OLS  $\hat{\beta}$  is unbiased with or without intercept.



## Summary

OLS  $\hat{\beta}$  is **unbiased** if

1. An intercept is included, or true intercept is zero

**and**

2. Omitted variables are uncorrelated with included variables.

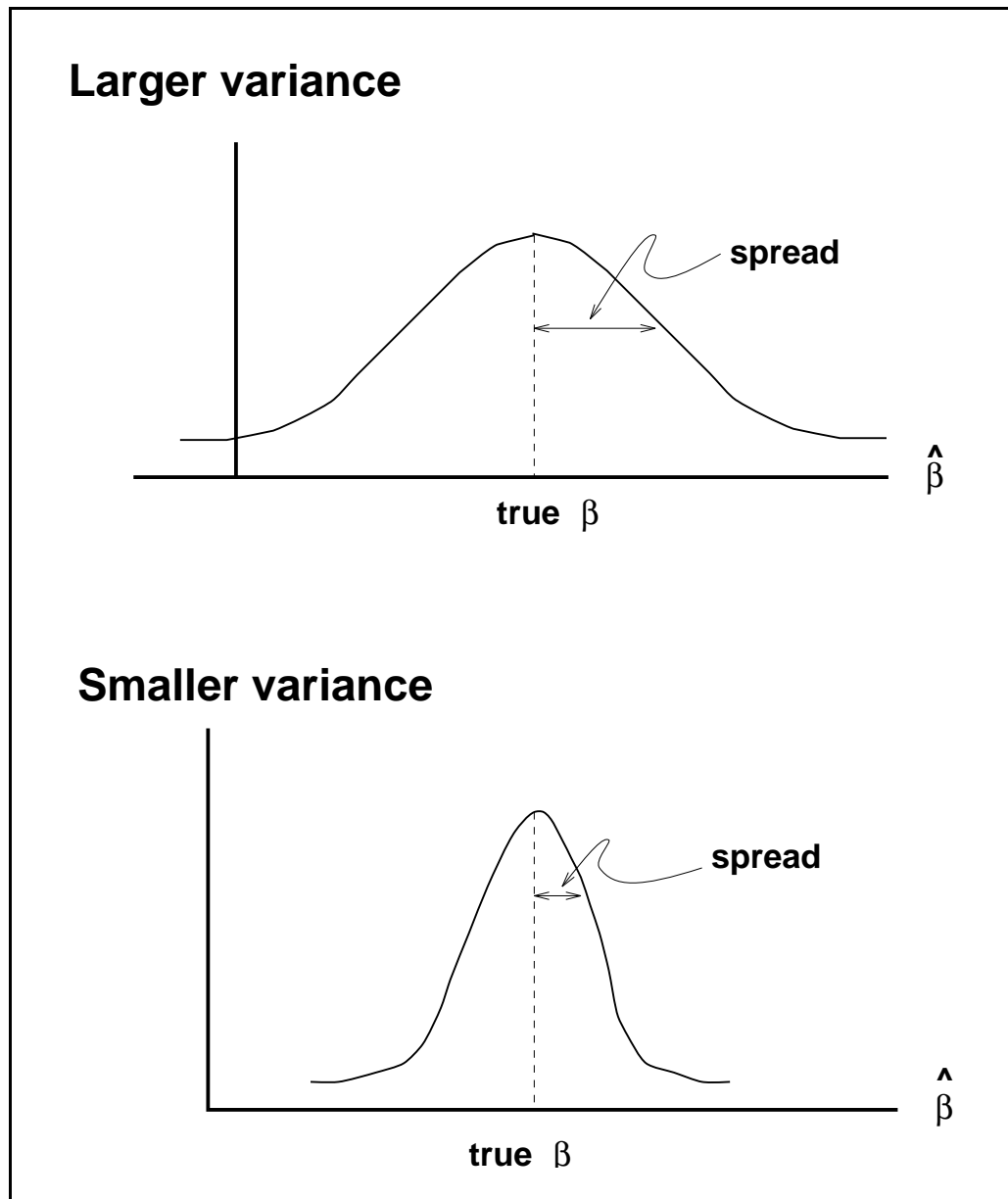
Formally:

$$E(\varepsilon|x) = 0$$

which implies

1.  $E(\varepsilon) = 0$
2.  $\text{Corr}(\varepsilon, x) = 0$

## Variance of Distribution of OLS $\hat{\beta}$



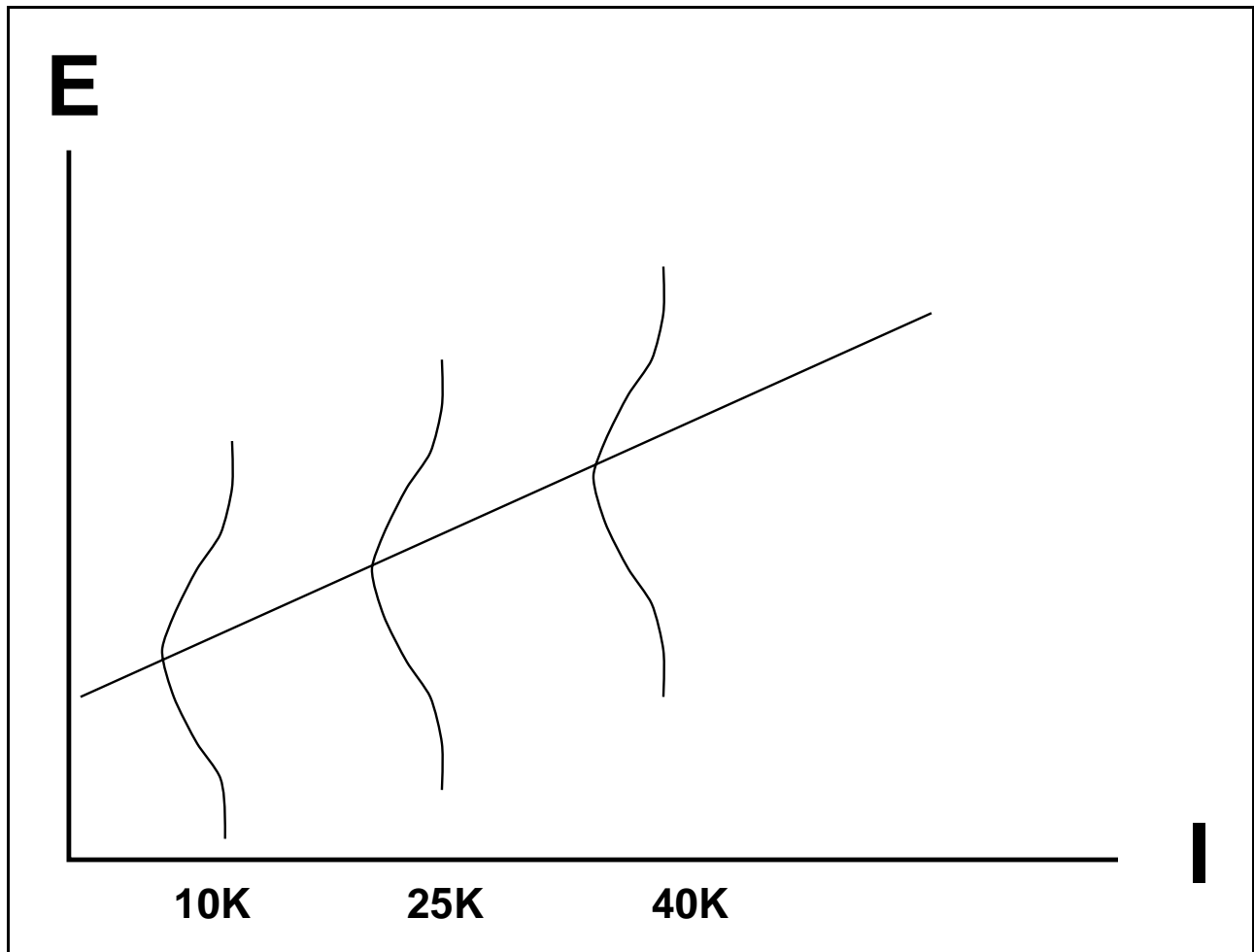
Want as small variance as possible.

## Variance of $\hat{\beta}$ is Lower

- When fewer variables are omitted, and more variables are included.
- When sample size is larger
- When variance of included variables is larger

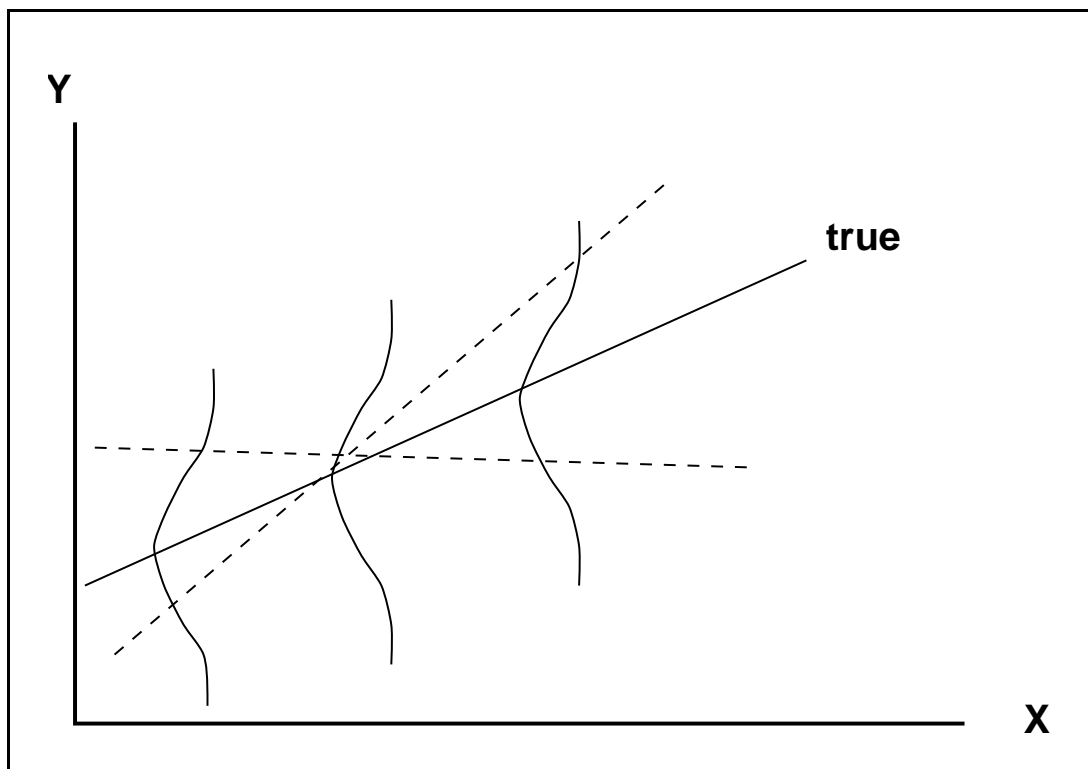
Think of population as follows:

For any value of  $X$ , there are different  $Y$ 's because of  $\epsilon$ 's.



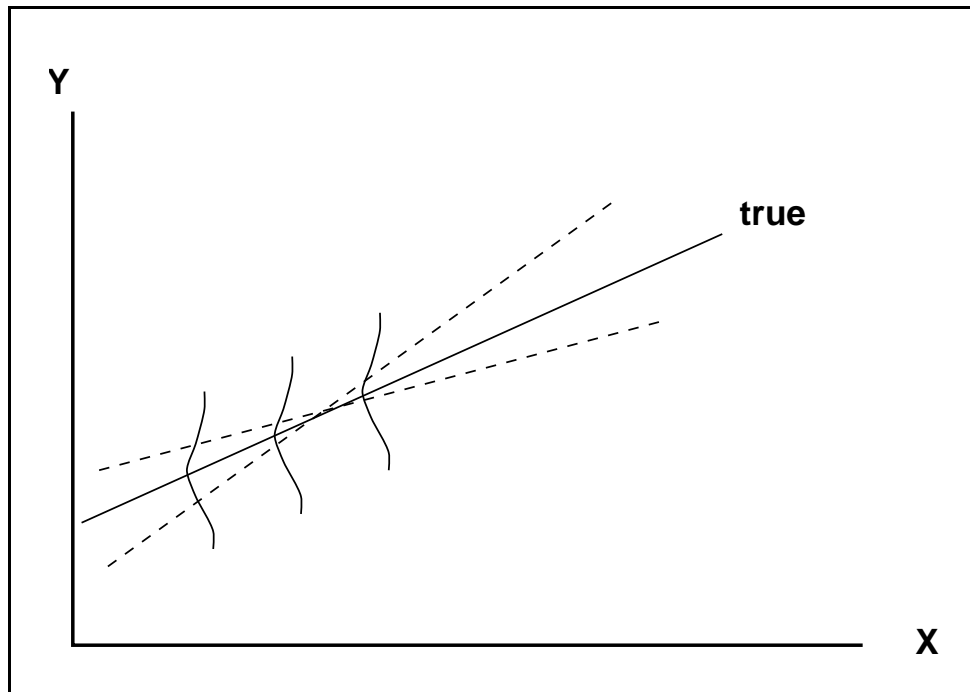
## Variance of $\hat{\beta}$ is Lower When Fewer Variables are Omitted

Large influence from omitted variables.



Large change in  $\hat{\beta}$  from one sample to next.

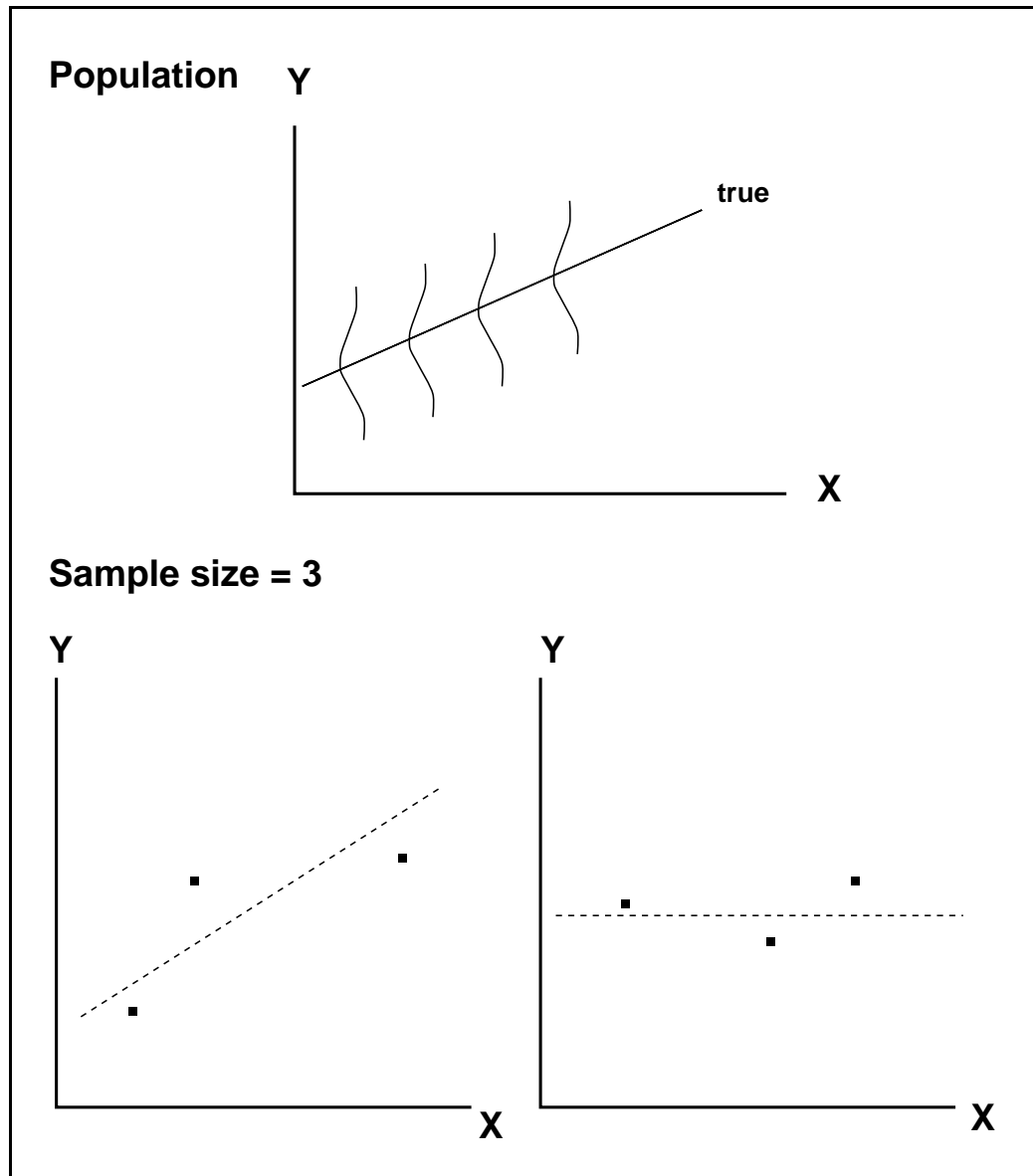
Small influence from omitted variables.



Small change in  $\hat{\beta}$  from one sample to next.

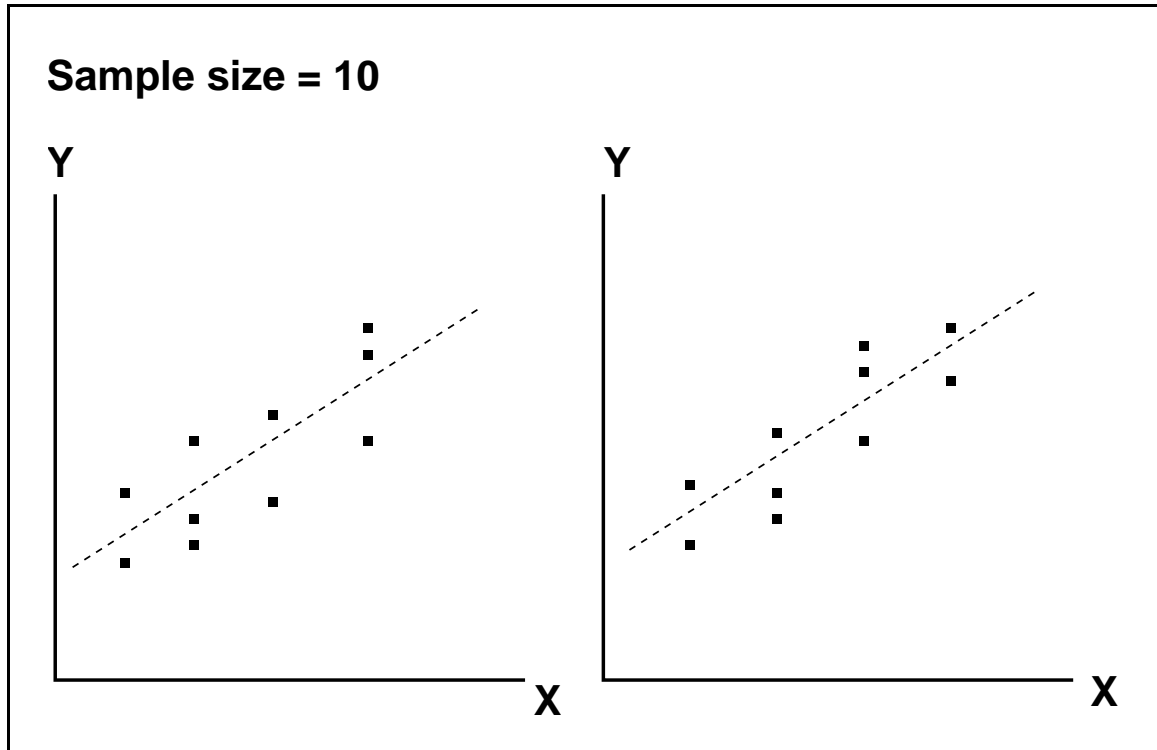
Implication: Include as many causal variables as possible.

Variance of  $\hat{\beta}$  is lower with larger samples.



Large change in  $\hat{\beta}$  from one sample to next.

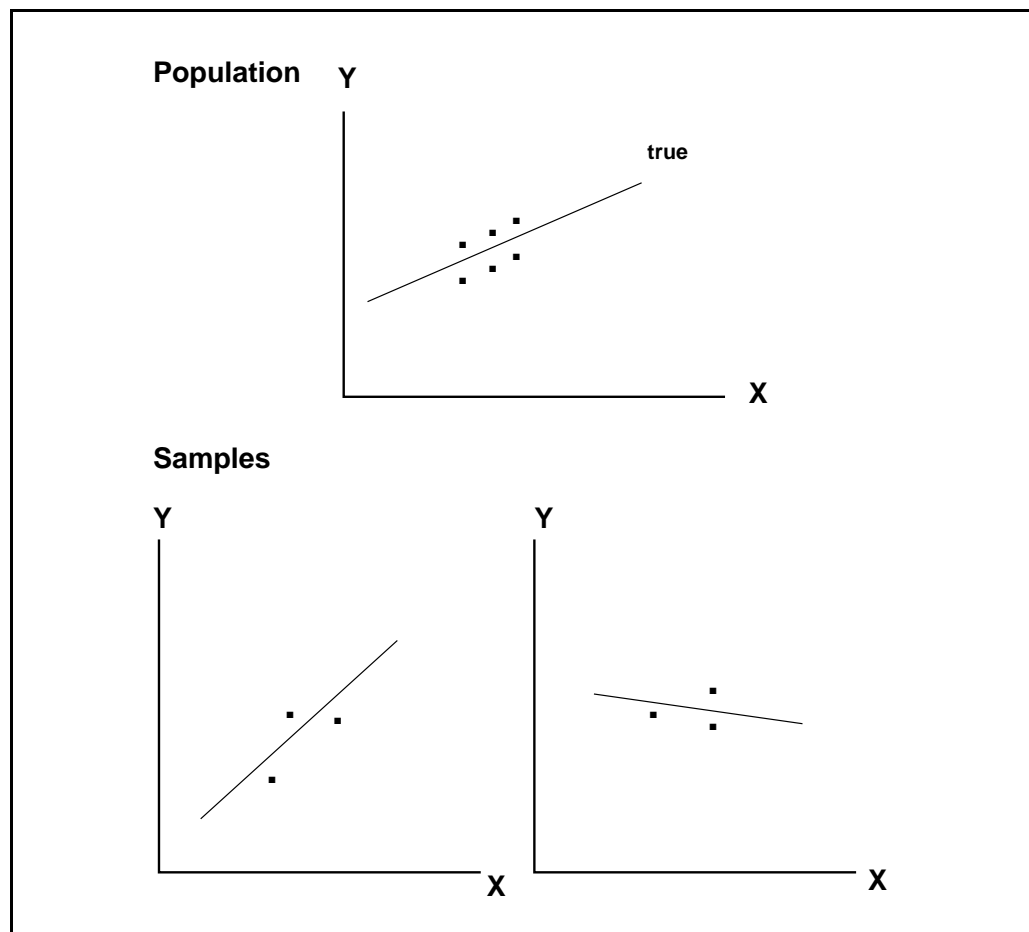
Smaller change in  $\hat{\beta}$  from one sample to next.



Implication: Use as large samples as possible.

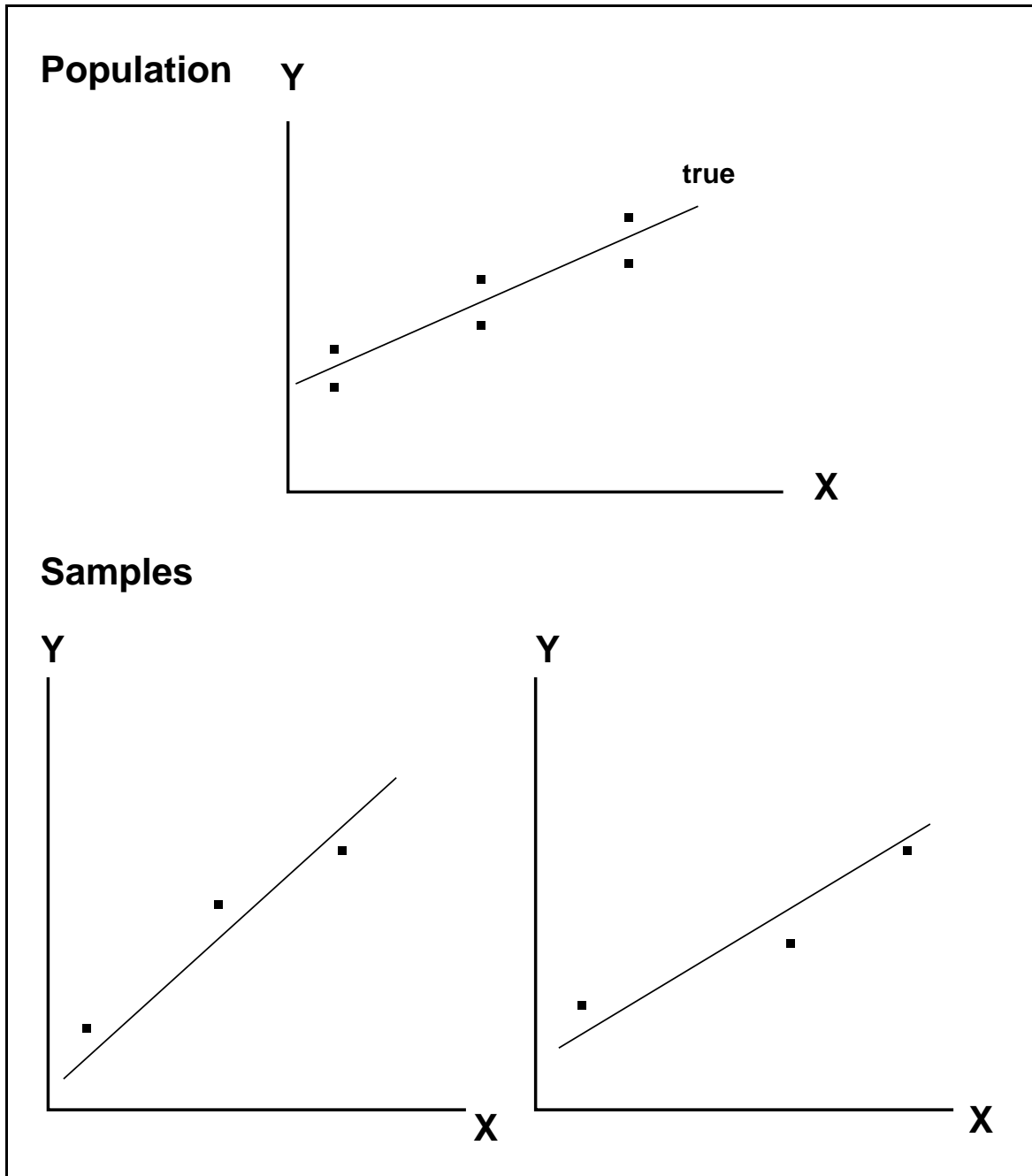
## Variance of $\hat{\beta}$ is Lower When the Variance in the Included Variables is Higher

Small spread in X:



Large change in  $\hat{\beta}$  from one sample to next.

Large spread in X:



Small change in  $\hat{\beta}$  from one sample to next.

Implication: Obtain as much variance as possible in explanatory variables.

## Proof of Implications for Variance

Assume  $y_n = \beta x_n + \varepsilon_n$  (intercept allows deviations)

- $x_n$  non-stochastic
- $E(\varepsilon_n | x_n) = 0$
- $\varepsilon_n$  independent over  $n$
- $V(\varepsilon_n) = \sigma^2$  (homoscedasticity)

$$\text{Recall } \hat{\beta} = \beta + \frac{\sum \varepsilon_n x_n}{\sum x_n^2}.$$

$$\begin{aligned} V(\hat{\beta}) &= V\left(\frac{\sum \varepsilon_n x_n}{\sum x_n^2}\right) \\ &= \left(\frac{1}{\sum x_n^2}\right)^2 V(\sum \varepsilon_n x_n) \\ &= \left(\frac{1}{\sum x_n^2}\right)^2 \sum x_n^2 V(\varepsilon_n) \\ &= \left(\frac{1}{\sum x_n^2}\right)^2 \sum x_n^2 \sigma^2 \\ &= \sigma^2 \frac{\sum x_n^2}{(\sum x_n^2)^2} \\ &= \sigma^2 / \sum x_n^2 \end{aligned}$$

$$V(\hat{\beta}) = \sigma^2 / \sum x_n^2$$

- decreases when  $\sigma^2$  decreases.
- decreases when sample size increases, since sum in denominator gets larger.
- decreases when the variance of  $x$  increases, since the denominator is proportional to the variance.

## Summary

To get lower variance in  $\hat{\beta}$

1. Include as many explanatory variables as possible.
2. Increase sample size.
3. Obtain as large a variance in explanatory variables as possible.

## How to Measure Variance?

### Measure of variance of $\hat{\beta}$

$$V(\hat{\beta}) = \sigma^2 / \sum x_n^2$$

where  $\sigma^2$  is variance in errors.

### Estimate of $\sigma^2$

$$s^2 = \frac{1}{N - K} \sum r_n^2$$

$s$  is called the "standard error of regression."

### Estimate of $V(\hat{\beta})$

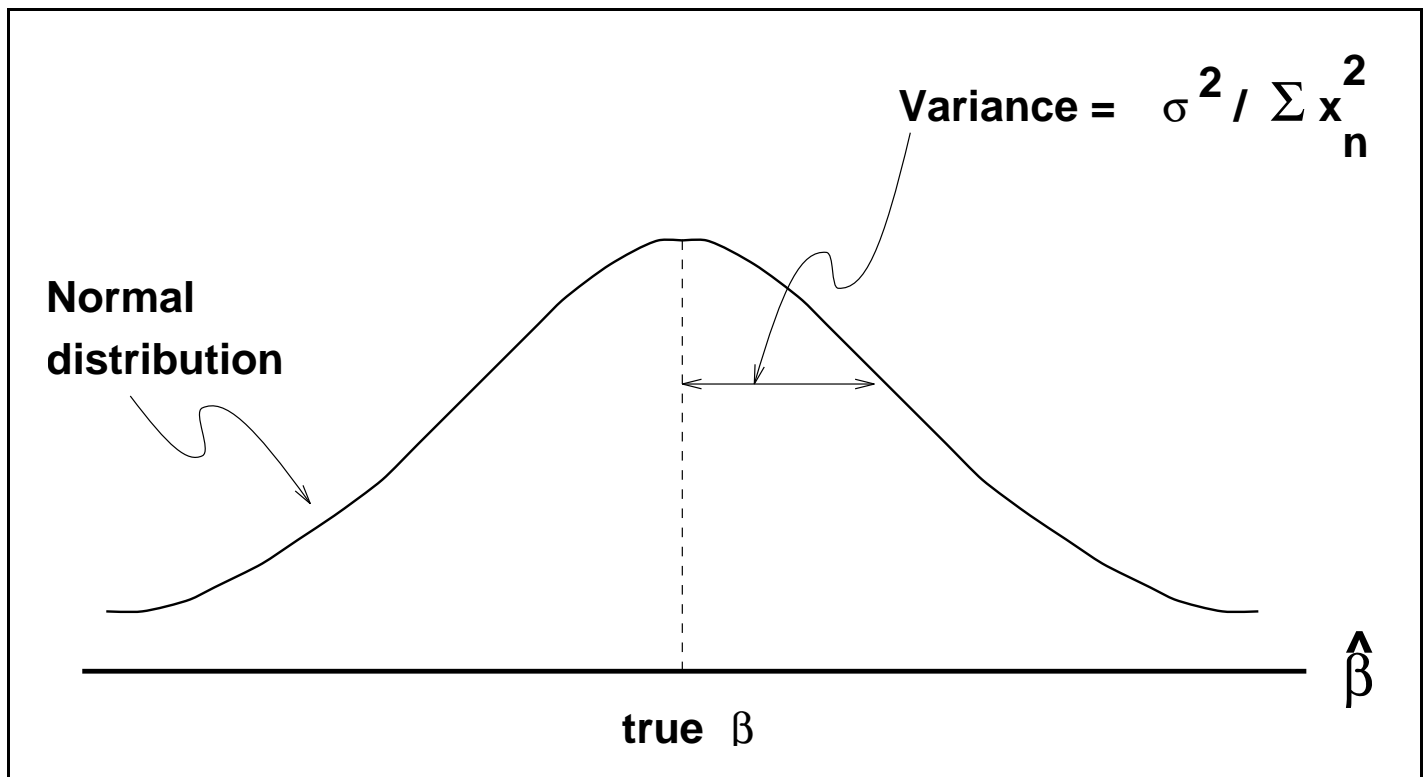
$$V(\hat{\beta}) = s^2 / \sum x_n^2$$

$\sqrt{V(\hat{\beta})}$  is called the "standard error of  $\hat{\beta}$ ."

## Shape of Distribution of $\hat{\beta}$

**Normal distribution if**

1.  $\varepsilon_n$ s are distributed normally
- or
2. Sample size is large.



$$\hat{\beta} \sim N(\beta, \sigma^2 / \sum x_n^2)$$

# Summary

1. OLS  $\hat{\beta}$  is unbiased if
  - intercept is included
  - omitted variables are uncorrelated with included variables
  
2. Variance of  $\hat{\beta}$  is smaller when
  - more variables are included
  - sample size is larger
  - explanatory variables have larger variance
  
3.  $\hat{\beta}$  has a normal distribution if
  - errors have normal distribution
  - or**
  - sample size is large
  
4. Standard error of  $\hat{\beta}$  is estimate of the standard deviation of the distribution of  $\hat{\beta}$ .