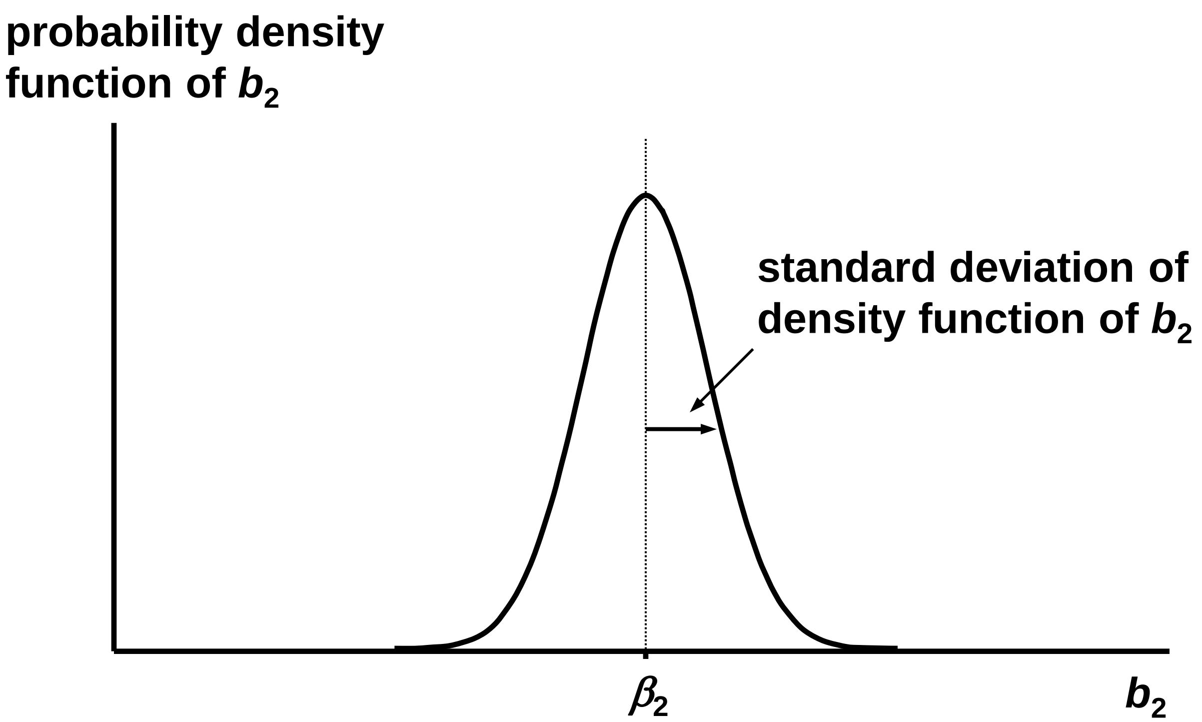
ECONOMETRICS NOTES

Unit 0 – LECTURE 4

PRECISION OF THE REGRESSION COEFFICIENTS

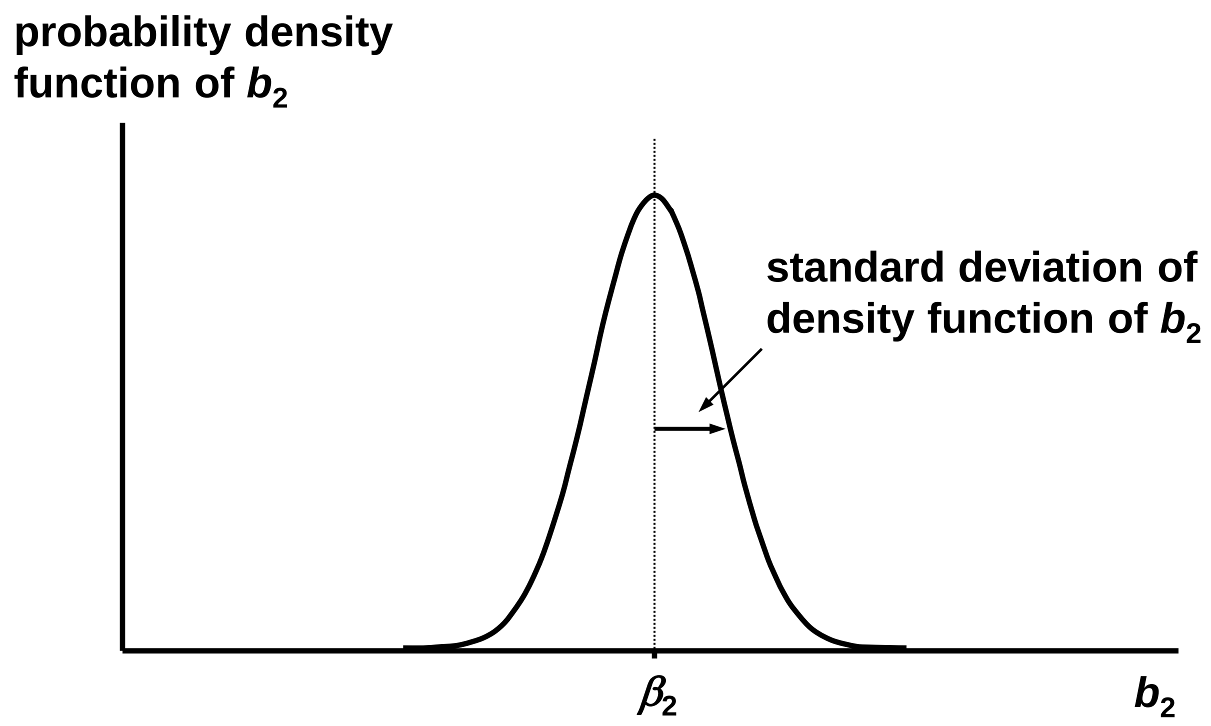
Simple regression model: $Y = \beta_1 + \beta_2 X + u$



We have seen that the regression coefficients b_1 and b_2 are random variables. They provide point estimates of β_1 and β_2, respectively. In the last sequence we demonstrated that these point estimates are unbiased.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$



In this sequence we will see that we can also obtain estimates of the standard deviations of the distributions. These will give some idea of their likely reliability and will provide a basis for tests of hypotheses.

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\}$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2}$$

Expressions (which will not be derived) for the variances of their distributions are shown above. Prove the expression for the variance of b_2.

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\}$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2}$$

We will focus on the implications of the expression for the variance of b_2. Looking at the numerator, we see that the variance of b_2 is proportional to σ_u^2. This is as we would expect, why??

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\}$$

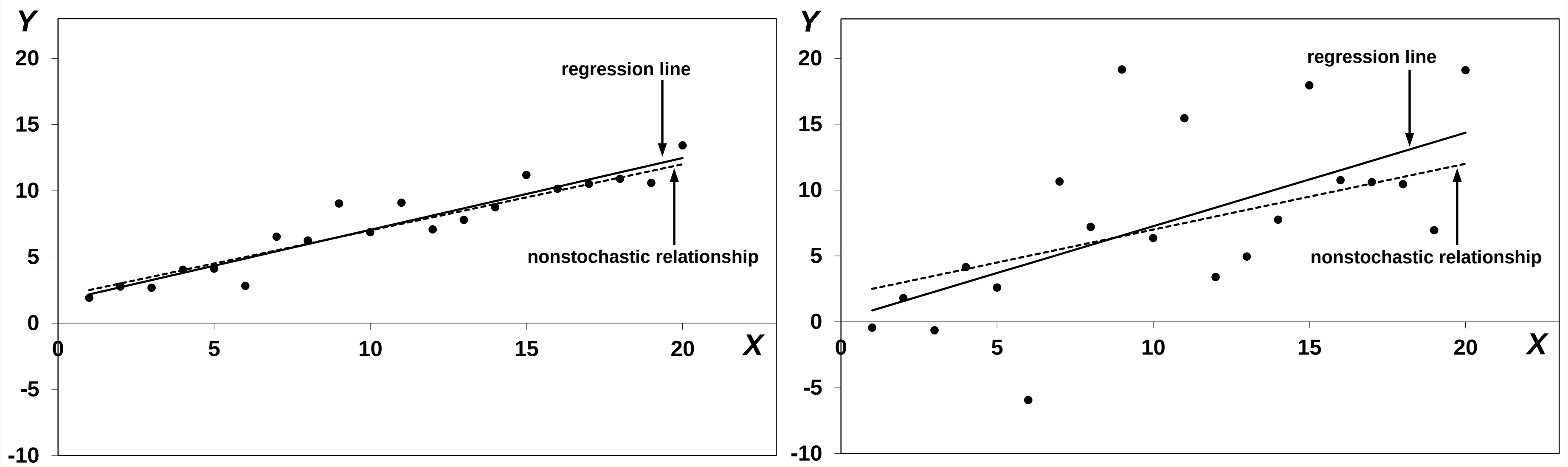
$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2}$$

The more noise there is in the model, the less precise will be our estimates.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$

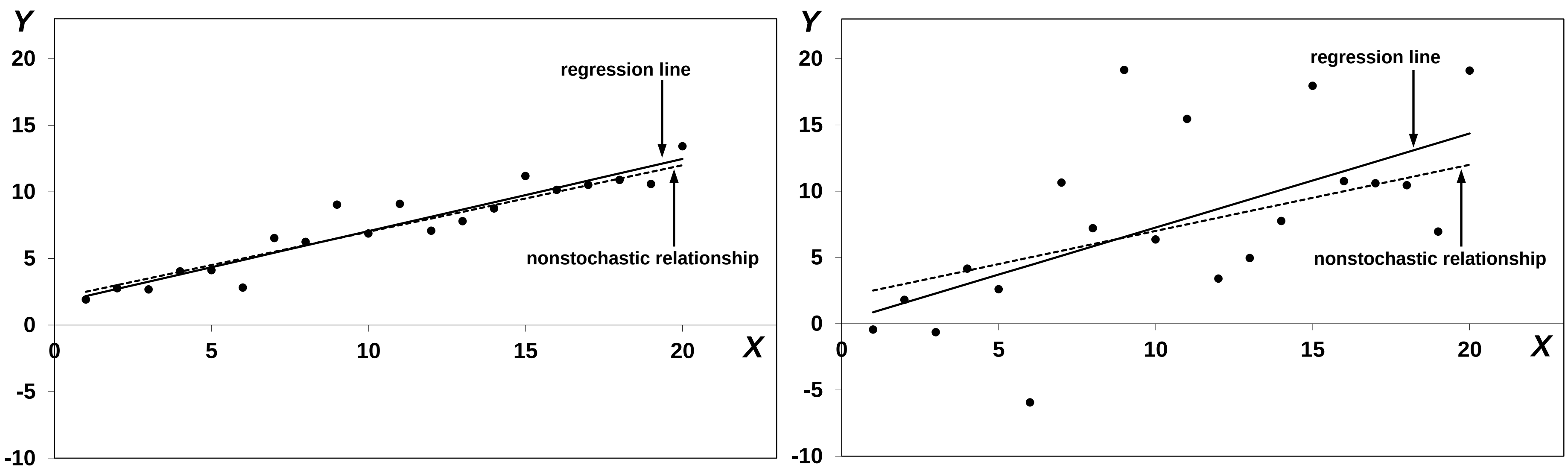


This is illustrated by the diagrams above. The nonstochastic component of the relationship, $Y = 3.0 + 0.8X$, represented by the dotted line, is the same in both diagrams.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$

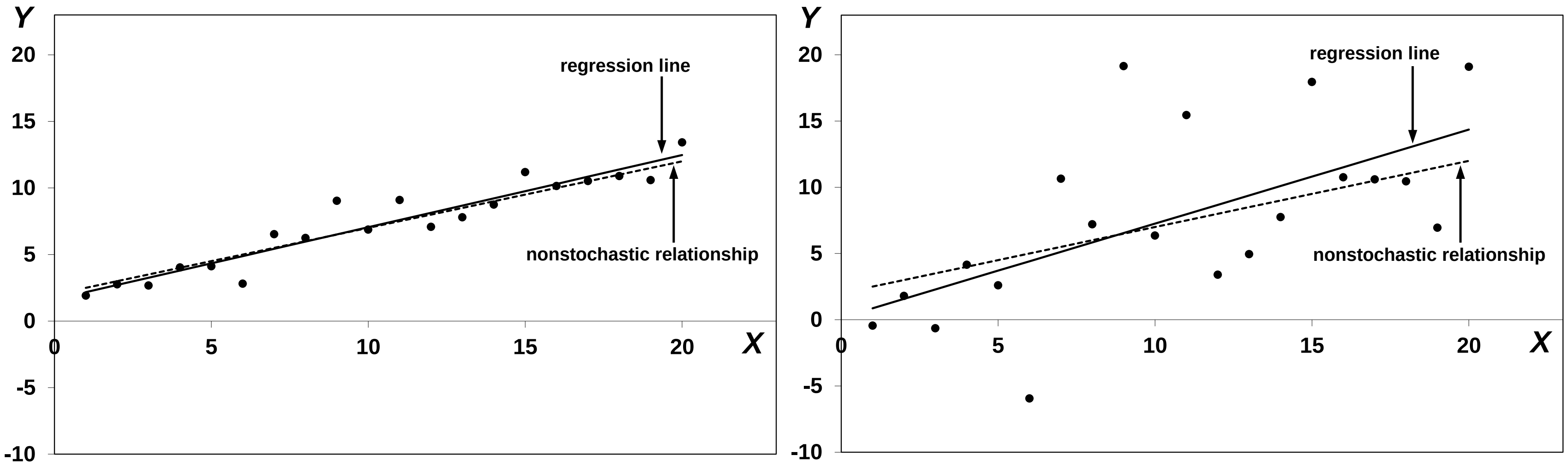


The values of *X* are the same, and the same random numbers have been used to generate the values of the disturbance term in the 20 observations.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$



However, in the right-hand diagram the random numbers have been multiplied by a factor of 5. As a consequence, the regression line, the solid line, is a much poorer approximation to the nonstochastic relationship.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\}$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2}$$

Looking at the denominator of the expression for the variance of b_2, the larger is the sum of the squared deviations of X, the smaller is the variance of b_2.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\}$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)}$$

$$\mathrm{MSD}(X) = \frac{1}{n} \sum (X_i - \bar{X})^2$$

However, the size of the sum of the squared deviations depends on two factors: the number of observations, and the size of the deviations of X_i about its sample mean. To discriminate between them, it is convenient to define the mean square deviation of X, MSD(X).

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\}$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma_u^2}{n \, \text{MSD}(X)}$$

$$\text{MSD}(X) = \frac{1}{n} \sum (X_i - \bar{X})^2$$

From the expression as rewritten, it can be seen that the variance of b_2 is inversely proportional to *n*, the number of observations in the sample, controlling for MSD(*X*). The more information you have, the more accurate your estimates are likely to be.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\}$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X)}$$

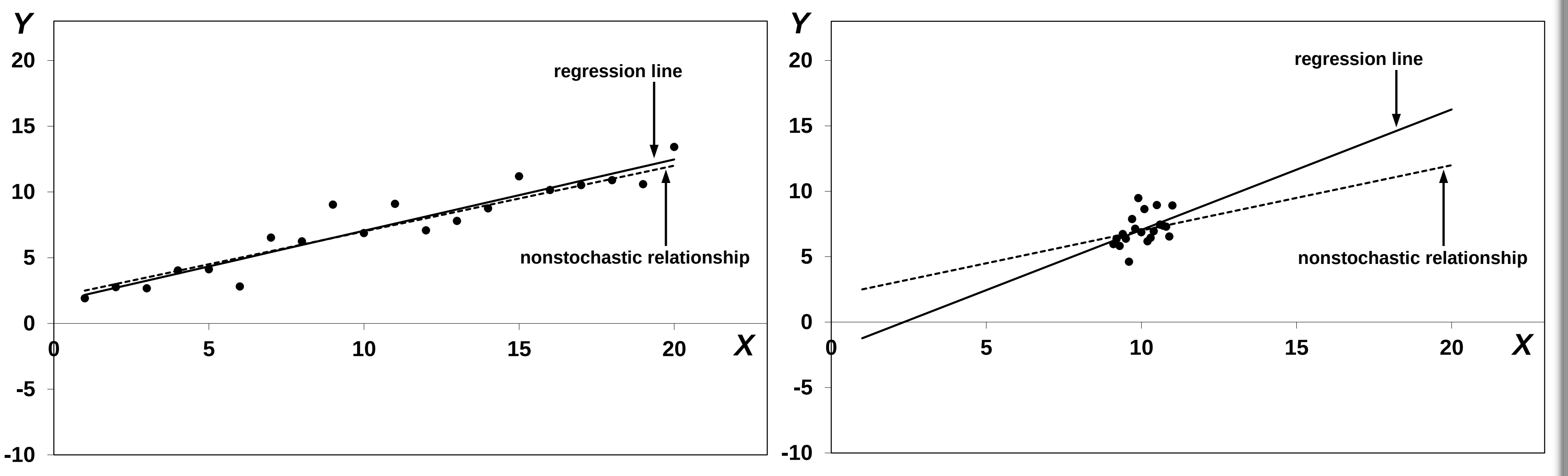
$$\mathrm{MSD}(X) = \frac{1}{n} \sum (X_i - \overline{X})^2$$

A third implication of the expression is that the variance is inversely proportional to the mean square deviation of *X*. What is the reason for this?

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$

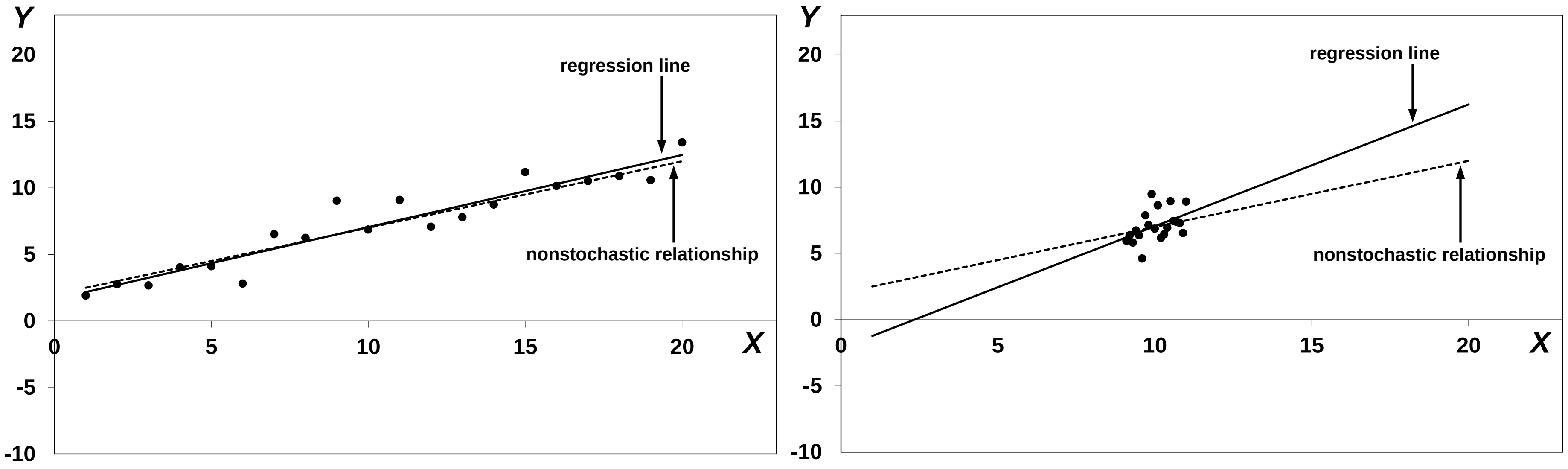


In the diagrams above, the nonstochastic component of the relationship is the same and the same random numbers have been used for the 20 values of the disturbance term.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$

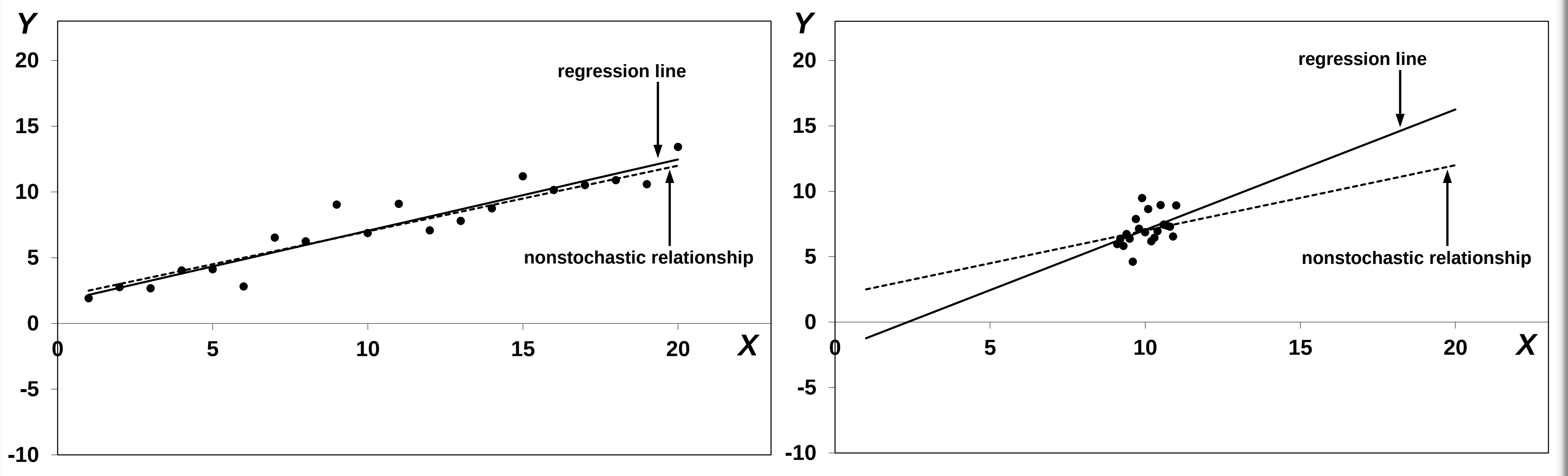


However, MSD(*X*) is much smaller in the right-hand diagram because the values of *X* are much closer together.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$

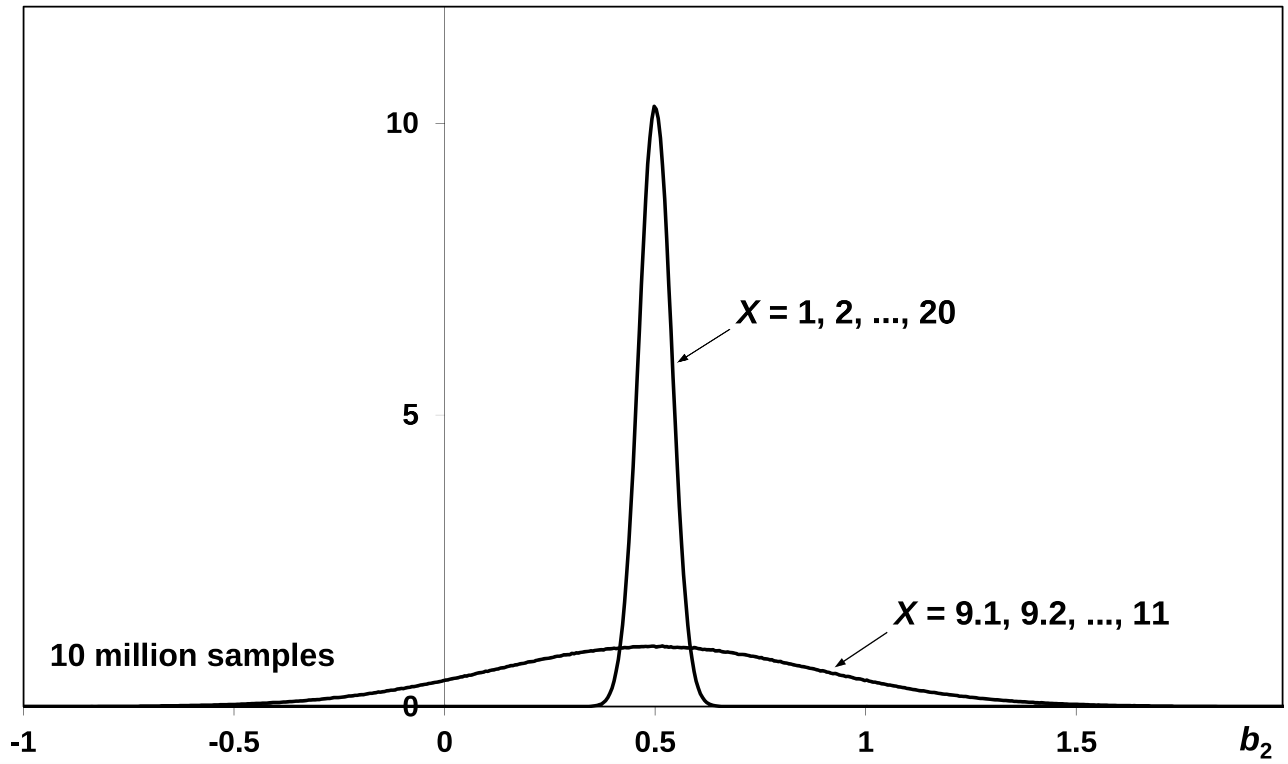


Hence, in that diagram, the position of the regression line is more sensitive to the values of the disturbance term, and as a consequence the regression line is likely to be relatively inaccurate.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$

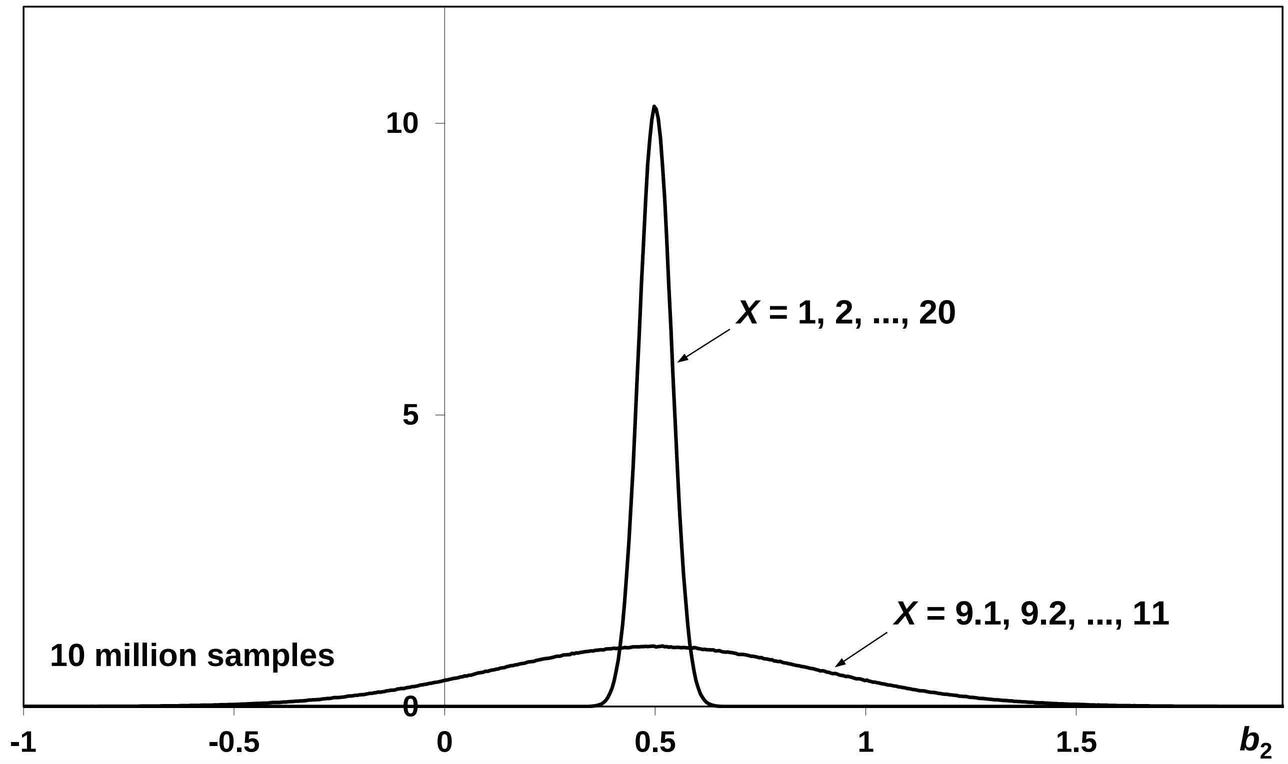


The figure shows the distributions of the estimates of b_2 for X = 1, 2, ..., 20 and X = 9.1, 9.2, ..., 11 in a simulation with 10 million samples.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$Y = 2.0 + 0.5X$$



It confirms that the distribution of the estimates obtained with the high dispersion of *X* has a much smaller variance than that with the low dispersion of *X*.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)}$$

Of course, as can be seen from the variance expressions, it is really the ratio of the MSD(*X*) to the variance of *u* which is important, rather than the absolute size of either.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)}$$

We cannot calculate the variances exactly because we do not know the variance of the disturbance term. However, we can derive an estimator of σ_u^2 from the residuals.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X)}$$

Clearly the scatter of the residuals around the regression line will reflect the unseen scatter of *u* about the line $Y_i = \beta_1 + b_2 X_i$, although in general the residual and the value of the disturbance term in any given observation are not equal to one another.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)}$$

$$\mathrm{MSD}(e) = \frac{1}{n} \sum (e_i - \bar{e})^2 = \frac{1}{n} \sum e_i^2$$

One measure of the scatter of the residuals is their mean square error, MSD(*e*), defined as shown. (Remember that the mean of the OLS residuals is equal to zero). Intuitively this should provide a guide to the variance of *u*.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X)}$$

$$\mathrm{MSD}(e) = \frac{1}{n} \sum (e_i - \bar{e})^2 = \frac{1}{n} \sum e_i^2$$

Before going any further, ask yourself the following question. Which line is likely to be closer to the points representing the sample of observations on X and Y, the true line $Y = \beta_1 + \beta_2 X$ or the regression line $\hat{Y} = b_1 + b_2 X$?

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)}$$

$$\mathrm{MSD}(e) = \frac{1}{n} \sum (e_i - \overline{e})^2 = \frac{1}{n} \sum e_i^2$$

The answer is the regression line, because by definition it is drawn in such a way as to minimize the sum of the squares of the distances between it and the observations.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum \left(X_i - \overline{X} \right)^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum \left(X_i - \overline{X} \right)^2} = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X)}$$

$$\mathrm{MSD}(e) = \frac{1}{n} \sum \left(e_i - \overline{e} \right)^2 = \frac{1}{n} \sum e_i^2$$

Hence the spread of the residuals will tend to be smaller than the spread of the values of *u*, and MSD(*e*) will tend to underestimate σ_u^2.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X)}$$

$$\mathrm{MSD}(e) = \frac{1}{n} \sum (e_i - \overline{e})^2 = \frac{1}{n} \sum e_i^2$$

$$E(\mathrm{MSD}(e)) = \frac{n-2}{n} \sigma_u^2$$

Indeed, it can be shown that the expected value of MSD(*e*), when there is just one explanatory variable, is given by the expression above.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)}$$

$$s_u^2 = \frac{n}{n-2}\mathrm{MSD}(e) = \frac{n}{n-2}\frac{1}{n}\sum e_i^2 = \frac{1}{n-2}\sum e_i^2$$

However, it follows that we can obtain an unbiased estimator of σ_u^2 by multiplying MSD(*e*) by *n* / (*n* – 2). We will denote this s_u^2.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X)}$$

$$s_u^2 = \frac{n}{n-2} \mathrm{MSD}(e) = \frac{n}{n-2} \frac{1}{n} \sum e_i^2 = \frac{1}{n-2} \sum e_i^2$$

$$\text{s.e.}(b_1) = s_u \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}} \qquad \text{s.e.}(b_2) = \sqrt{\frac{s_u^2}{\sum (X_i - \overline{X})^2}}$$

We can then obtain estimates of the standard deviations of the distributions of b_1 and b_2 by substituting s_u^2 for σ_u^2 in the variance expressions and taking the square roots.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n\,\mathrm{MSD}(X)}$$

$$s_u^2 = \frac{n}{n-2}\mathrm{MSD}(e) = \frac{n}{n-2}\frac{1}{n}\sum e_i^2 = \frac{1}{n-2}\sum e_i^2$$

$$\text{s.e.}(b_1) = s_u \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}} \qquad \text{s.e.}(b_2) = \sqrt{\frac{s_u^2}{\sum (X_i - \overline{X})^2}}$$

These are described as the standard errors of b_1 and b_2, ‘estimates of the standard deviations’ being a bit of a mouthful.

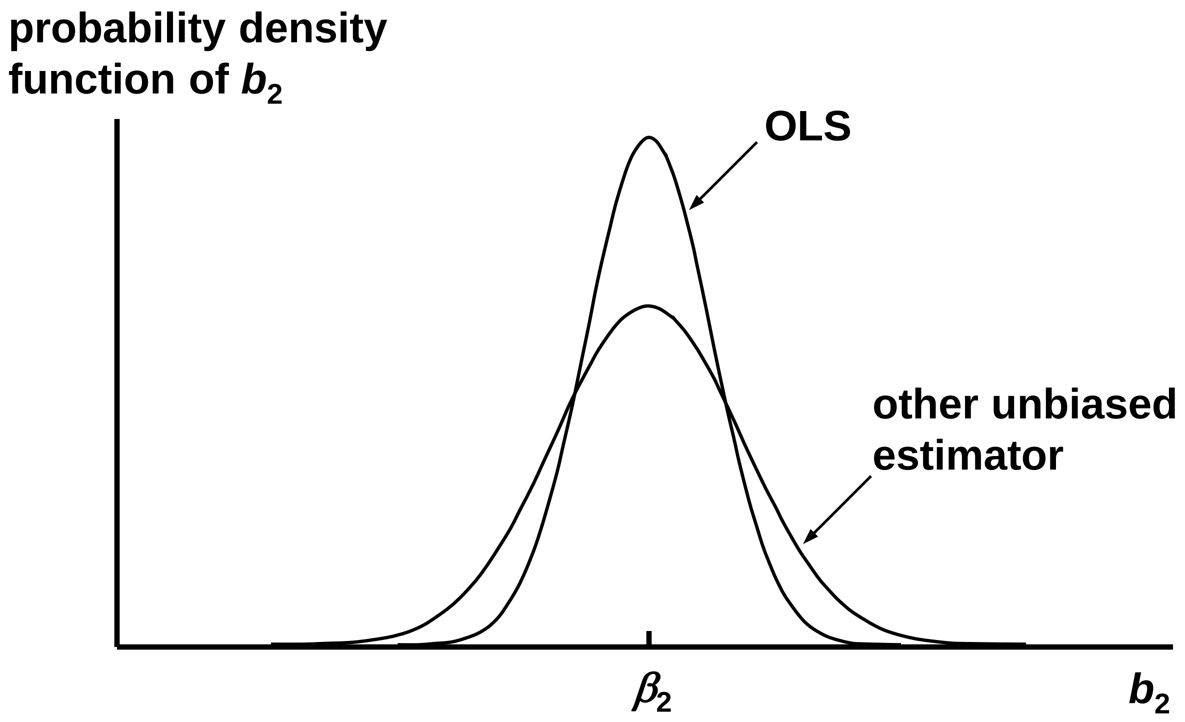
PRECISION OF THE REGRESSION COEFFICIENTS

```
. reg EARNINGS S

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  112.15
       Model |  19321.5589     1  19321.5589           Prob > F      =  0.0000
    Residual |  92688.6722   538  172.283777           R-squared     =  0.1725
-------------+------------------------------           Adj R-squared =  0.1710
       Total |  112010.231   539  207.811189           Root MSE      =  13.126

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   2.455321   .2318512    10.59   0.000     1.999876    2.910765
       _cons |  -13.93347   3.219851    -4.33   0.000    -20.25849   -7.608444
------------------------------------------------------------------------------
```

The standard errors of the coefficients always appear as part of the output of a regression. Here is the regression of hourly earnings on years of schooling discussed in a previous slideshow. The standard errors appear in a column to the right of the coefficients.

PRECISION OF THE REGRESSION COEFFICIENTS

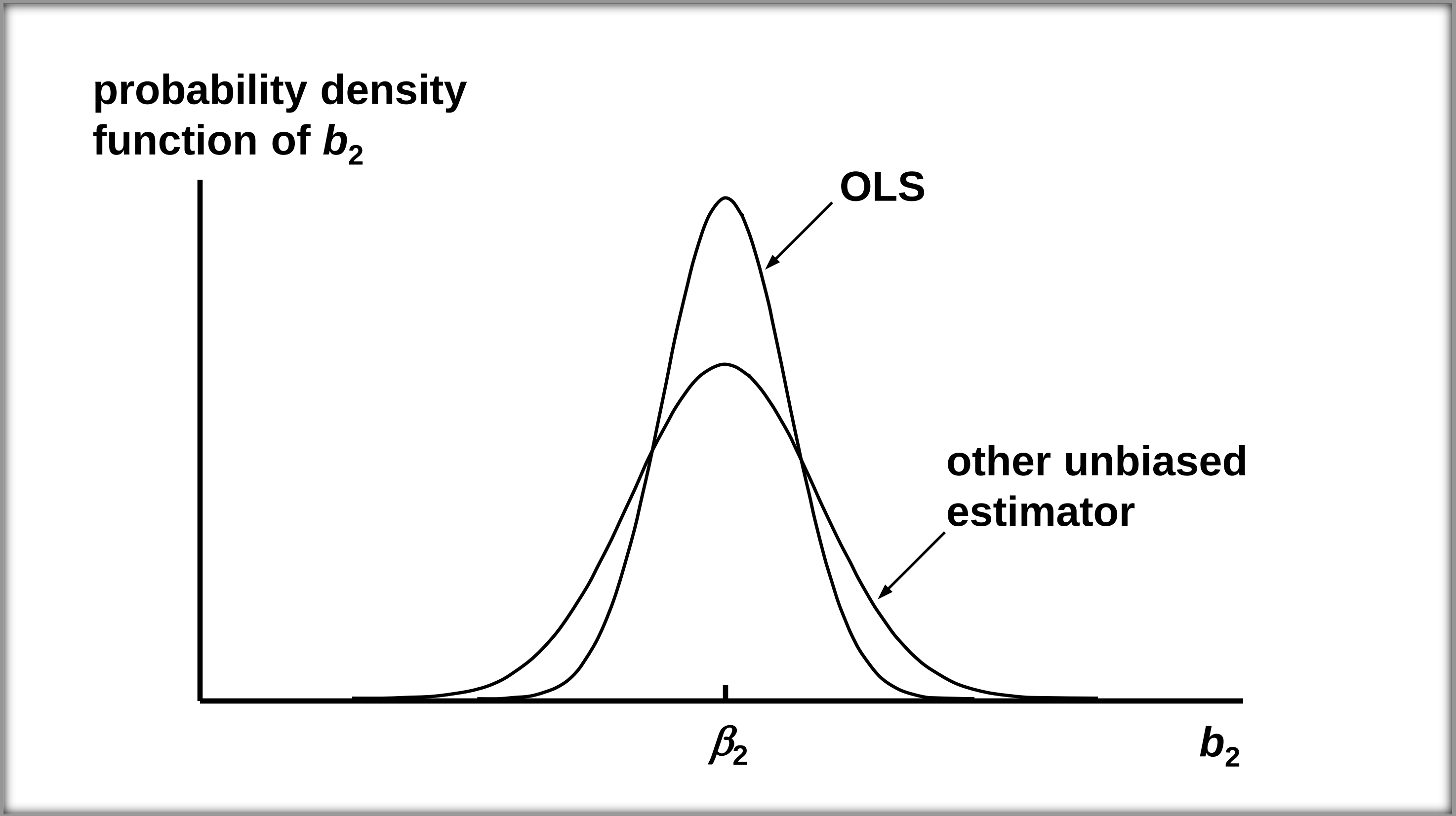
Simple regression model: $Y = \beta_1 + \beta_2 X + u$



The Gauss–Markov theorem states that, provided that the regression model assumptions are valid, the OLS estimators are BLUE: best (most efficient) linear (functions of the values of *Y*) unbiased estimators of the parameters.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$



The proof of the theorem is not difficult .

Home work

Prove the Gauss–Markov theorem

END OF LECTURE