JEFFREY M. WOOLDRIDGE

Introductory Econometrics
A Modern Approach

SIXTH EDITION

Chapter 2

The Simple Regression Model

The Simple Regression Model

- **Definition of the simple linear regression model**

"Explains variable y in terms of variable x"

Intercept

Slope parameter

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

Dependent variable,
explained variable,
response variable,…

Independent variable,
explanatory variable,
regressor,…

Error term,
disturbance,
unobservables,…

The Simple Regression Model

- **Interpretation of the simple linear regression model**

"Studies how y varies with changes in x:"

holding other factors fixed (including other x, u)

$$\frac{\Delta y}{\Delta x} = \beta_1$$ as long as $$\frac{\Delta u}{\Delta x} = 0$$

By how much does the dependent variable change if the independent variable is increased by one unit?

Interpretation only correct if all other things remain equal when the independent variable is increased by one unit

- **The simple linear regression model is rarely applicable in practice but its discussion is useful for pedagogical reasons**

(methods suitable for teaching)

The Simple Regression Model

- **Example: Soybean yield and fertilizer**

$$yield = \beta_0 + \beta_1 fertilizer + u$$

Measures the effect of fertilizer on yield, holding all other factors fixed

Rainfall,
land quality,
presence of parasites, …

- **Example: A simple wage equation**

$$wage = \beta_0 + \beta_1 educ + u$$

Measures the change in hourly wage given another year of education, holding all other factors fixed

Labor force experience,
tenure with current employer,
work ethic, intelligence, …

The Simple Regression Model

- **When is there a causal interpretation?**

correlation ≠ causality
causality here → ceteris paribus

- **Conditional mean independence assumption**

$$E(u|x) = 0$$

The explanatory variable must not contain information about the mean of the unobserved factors

- **Example: wage equation**

$$wage = \beta_0 + \beta_1 educ + u$$

e.g. intelligence …

The conditional mean independence assumption is unlikely to hold because individuals with more education will also be more intelligent on average.

The Simple Regression Model

- **Population regression function (PFR)**
  - The conditional mean independence assumption implies that

$$E(y|x) = E(\beta_0 + \beta_1 x + u|x)$$

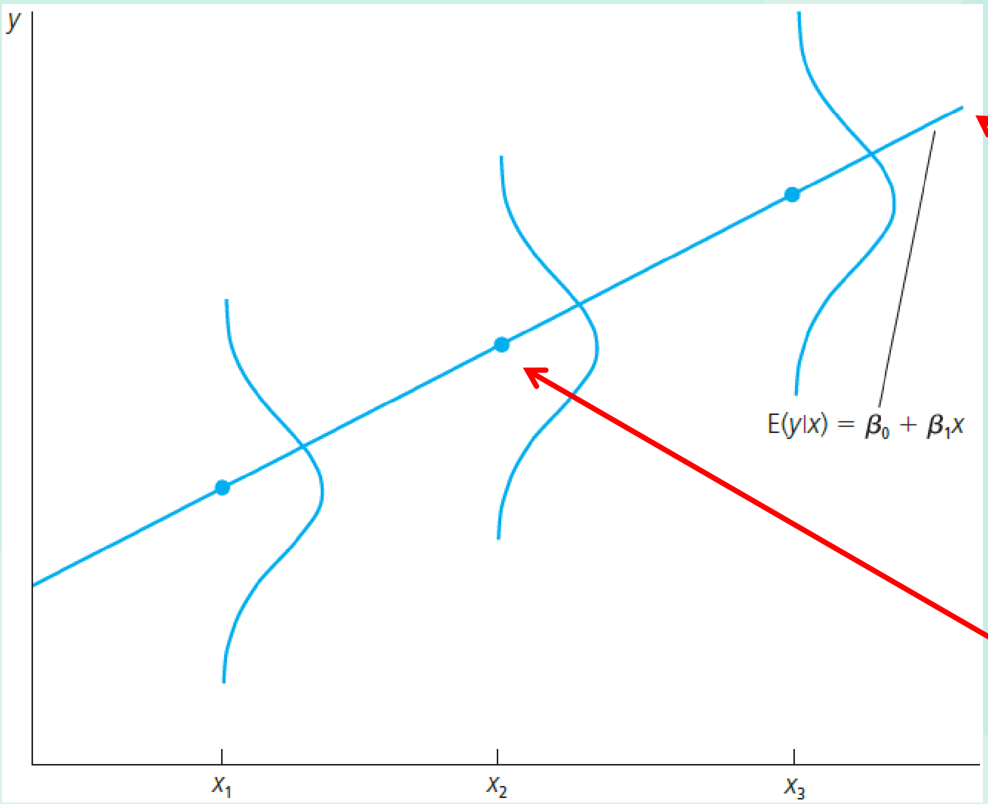
$$= \beta_0 + \beta_1 x + E(u|x)$$

$$= \beta_0 + \beta_1 x$$

  - This means that the average value of the dependent variable can be expressed as a linear function of the explanatory variable

linear regression
(not to be confused with linear correlation)

The Simple Regression Model



PRF

Population regression function

For individuals with $x = x_2$, the average value of y is $\beta_0 + \beta_1 x_2$

The Simple Regression Model

- **Deriving the ordinary least squares (OLS) estimates**
- **In order to estimate the regression model one needs data**
- **A random sample of n observations**

$(x_1, y_1) \leftarrow$ First observation

$(x_2, y_2) \leftarrow$ Second observation

$(x_3, y_3) \leftarrow$ Third observation

$\vdots$

$(x_n, y_n) \leftarrow$ n-th observation

$$\{(x_i, y_i) : i = 1, \ldots, n\}$$

Value of the <u>explanatory variable</u> of the i-th observation

Value of the <u>dependent</u> variable of the i-th observation

The Simple Regression Model

- **What does "as good as possible" mean?**
- **Regression residuals**

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- **Minimize sum of squared regression residuals**

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1$$

- **Ordinary Least Squares (OLS) estimates**

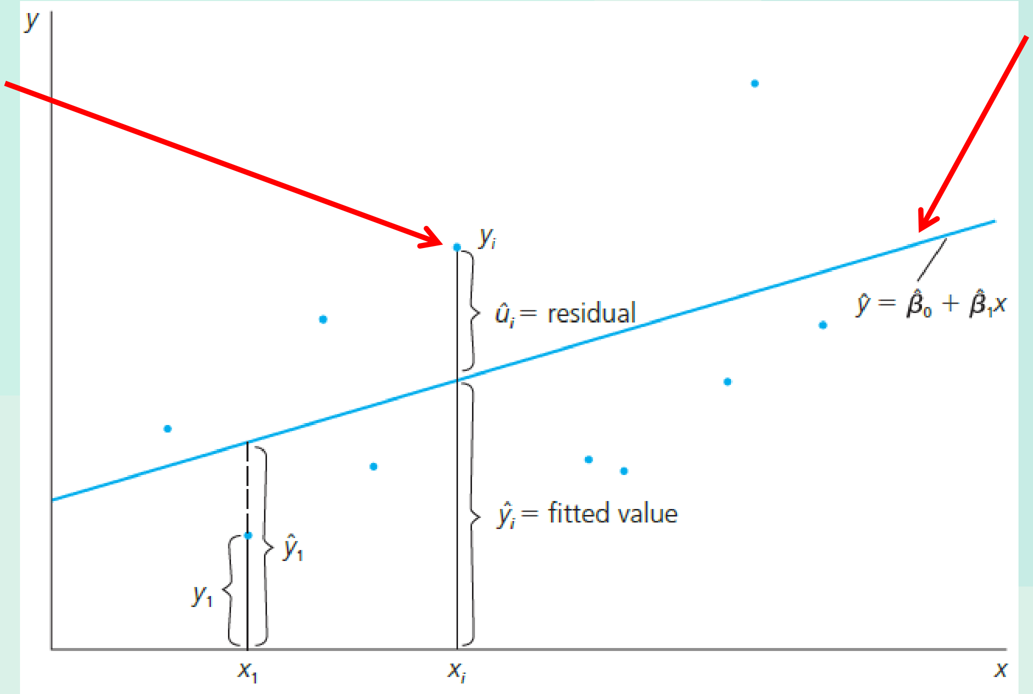
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The Simple Regression Model

- **Fit as good as possible a regression line through the data points:**

For example, the i-th data point (x_i, y_i)

Fitted regression line



The Simple Regression Model

- **CEO Salary and return on equity**

$$salary = \beta_0 + \beta_1 roe + u$$

Salary in thousands of dollars

Average return on equity of the CEO's firm

- **Fitted regression**

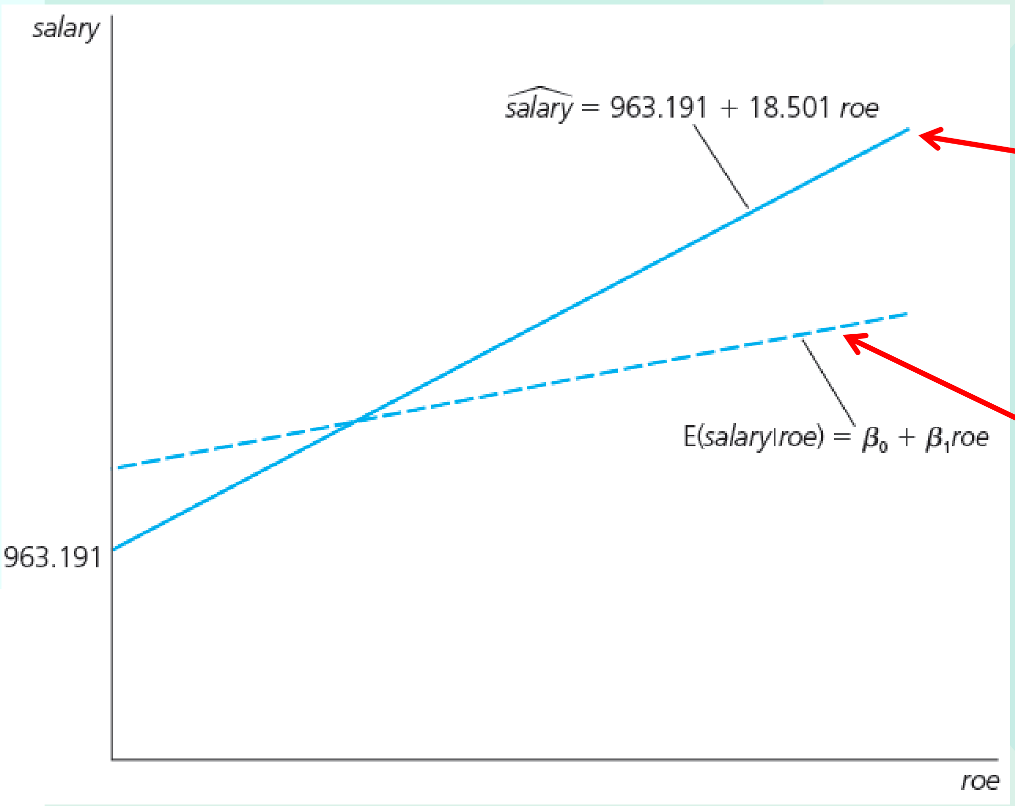
$$\widehat{salary} = 963.191 + 18.501\ roe$$

Intercept

If the return on equity increases by 1 percent, then salary is predicted to change by $18,501

- **Causal interpretation?**

The Simple Regression Model



SRF

Fitted regression line (depends on sample)

Unknown population regression line

PRF

The Simple Regression Model

- **Wage and education**

$$wage = \beta_0 + \beta_1 educ + u$$

Hourly wage in dollars

Years of education

- **Fitted regression** (fitted to sample data)

$$\widehat{wage} = -0.90 + 0.54\ educ$$

Intercept

In the sample, one more year of education was associated with an increase in hourly wage by $0.54

- **Causal interpretation?**

correlation ≠ causality
causality here → ceteris paribus

The Simple Regression Model

- **Voting outcomes and campaign expenditures (two parties)**

$$voteA = \beta_0 + \beta_1 shareA + u$$

Percentage of vote for candidate A

Percentage of campaign expenditures candidate A

- **Fitted regression**

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

Intercept

If candidate A's share of spending increases by one percentage point, he or she receives 0.464 percentage points more of the total vote

- **Causal interpretation?**

The Simple Regression Model

- **Properties of OLS on any sample of data**
- **Fitted values and residuals**

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad \widehat{u}_i = y_i - \widehat{y}_i$$

Fitted or predicted values

Deviations from regression line (= residuals)

- **Algebraic properties of OLS regression**

$$\sum_{i=1}^{n} \widehat{u}_i = 0 \qquad \sum_{i=1}^{n} x_i \widehat{u}_i = 0 \qquad \bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$$

Deviations from regression line sum up to zero

Covariance between deviations and regressors is zero

Sample averages of y and x lie on regression line

The Simple Regression Model

x_i

y_i

$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$

$\widehat{u}_i = y_i - \widehat{y}_i$

TABLE 2.1 Fitted Values and Residuals for the First 15 CEOs

obsno	roe	salary	salaryhat	uhat
1	14.1	1095	1224.058	–129.0581
2	10.9	1001	1164.854	–163.8542
3	23.5	1122	1397.969	–275.9692
4	5.9	578	1072.348	–494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	–188.2151
7	16.4	1078	1266.611	–188.6108
8	16.3	1094	1264.761	–170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	–616.7726
11	25.9	567	1442.372	–875.3721
12	26.8	933	1459.023	–526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	–438.7678
15	56.3	2011	2004.808	6.191895

For example, CEO number 12's salary was $526,023 lower than predicted using the the information on his firm's return on equity

The Simple Regression Model

- **Goodness-of-Fit** (think: ρ-squared and R-squared)

"How well does the explanatory variable explain the dependent variable?"

- **Measures of Variation** (SS = sum of squares)

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad SSR \equiv \sum_{i=1}^{n} \hat{u}_i{}^2$$

Total sum of squares, represents total variation in the dependent variable

(Total)

Explained sum of squares, represents variation explained by regression

(Explained)

Residual sum of squares, represents variation not explained by regression

(Unexplained)

The Simple Regression Model

- **Decomposition of total variation**

$$SST = SSE + SSR$$

Total variation Explained part Unexplained part

- **Goodness-of-fit measure (R-squared)**

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

R-squared measures the fraction of the total variation that is explained by the regression

The Simple Regression Model

- **CEO Salary and return on equity**

$$\widehat{salary} = 963.191 + 18.501\ roe$$

$$n = 209, \qquad R^2 = 0.0132$$

The regression explains only 1.3% of the total variation in salaries

- **Voting outcomes and campaign expenditures**

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

$$n = 173, \qquad R^2 = 0.856$$

The regression explains 85.6% of the total variation in election outcomes

- **Caution: A high R-squared does not necessarily mean that the regression has a causal interpretation!**

correlation ≠ causality

（相關不等於因果）

The Simple Regression Model

- **Incorporating nonlinearities: Semi-logarithmic form**
- **Regression of log wages on years of education**

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Natural logarithm of wage

- **This changes the interpretation of the regression coefficient:**

$$\beta_1 = \frac{\Delta \log(wage)}{\Delta educ} = \frac{1}{wage} \cdot \frac{\Delta wage}{\Delta educ} = \frac{\frac{\Delta wage}{wage}}{\Delta educ}$$

Percentage change of wage

… if years of education are increased by one year

The Simple Regression Model

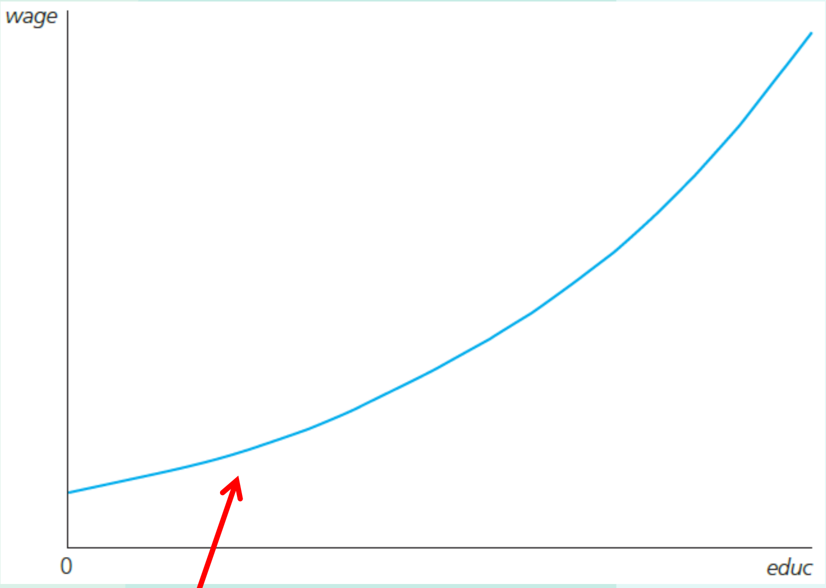
- **Fitted regression**

$$\widehat{\log}(wage) = 0.584 + 0.083\ educ$$

The wage increases by 8.3% for every additional year of education (= return to another year of education)

For example:

$$\frac{\frac{\Delta wage}{wage}}{\Delta educ} = \frac{\frac{+0.83\$}{10\$}}{+1\ year} = 0.083 = +8.3\%$$



Growth rate of wage is 8.3% per year of education

The Simple Regression Model

- **Incorporating nonlinearities: Log-logarithmic form**

- **CEO salary and firm sales**

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u$$

Natural logarithm of CEO salary

Natural logarithm of his/her firm‵s sales

- **This changes the interpretation of the regression coefficient:**

$$\beta_1 = \frac{\Delta \log(salary)}{\Delta \log(sales)} = \frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}}$$

Percentage change of salary

… if sales increase by 1%

Logarithmic changes are always percentage changes

The Simple Regression Model

- **CEO salary and firm sales: fitted regression**

$$\widehat{\log}(salary) = 4.822 + 0.257 \log(sales)$$

+ 1% sales; + 0.257% salary

- **For example:**

$$\frac{\frac{\Delta salary}{salary}}{\frac{\Delta sales}{sales}} = \frac{\frac{+2{,}570\$}{1{,}000{,}000\$}}{\frac{+10{,}000{,}000\$}{1{,}000{,}000{,}000\$}} = \frac{+0.257\%\ \text{salary}}{+1\%\ \text{sales}} = 0.257$$

- **The log-log form postulates a constant elasticity model, whereas the semi-log form assumes a semi-elasticity model**

The Simple Regression Model

- **Expected values and variances of the OLS estimators**
  (biasedness and efficiency)
- **The estimated regression coefficients are random variables** (why?) **because they are calculated from a random sample**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Data is random and depends on particular sample that has been drawn

- **The question is what the estimators will estimate on average and how large their variability in repeated samples is**

$$E(\hat{\beta}_0) = ?,\ E(\hat{\beta}_1) = ? \qquad Var(\hat{\beta}_0) = ?,\ Var(\hat{\beta}_1) = ?$$

The Simple Regression Model

- **Standard assumptions for the linear regression model**
- **Assumption SLR.1 (Linear in parameters)**

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

In the population, the relationship between y and x is linear

- **Assumption SLR.2 (Random sampling)**

$$\{(x_i, y_i) : i = 1, \ldots, n\}$$

The data is a random sample drawn from the population

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

Each data point therefore follows the population equation

The Simple Regression Model

- **Discussion of random sampling: Wage and education**
  - The population consists, for example, of all workers of country A
  - In the population, a linear relationship between wages (or log wages) and years of education holds
  - Draw completely randomly a worker from the population
  - The wage and the years of education of the worker drawn are random because one does not know beforehand which worker is drawn
  - Throw back worker into population and repeat random draw n times
  - The wages and years of education of the sampled workers are used to estimate the linear relationship between wages and education

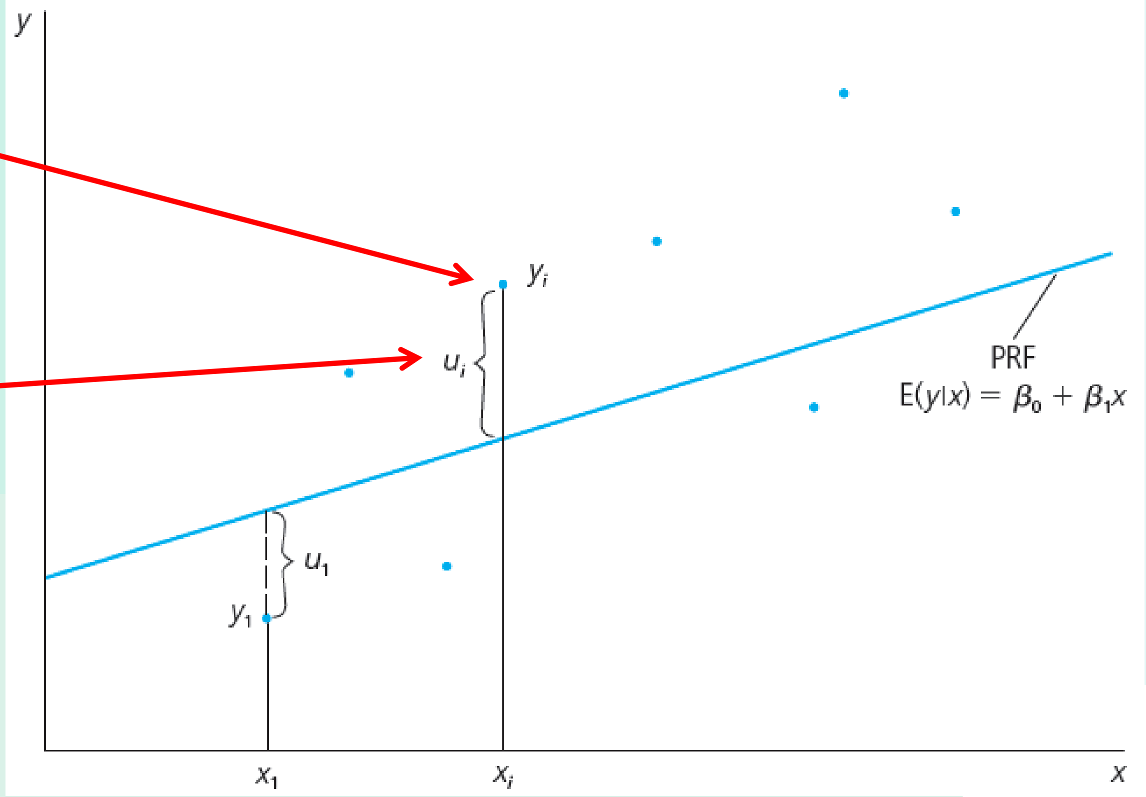
The Simple Regression Model

The values drawn for the i-th worker

(x_i, y_i)

The implied deviation from the population relationship for the i-th worker:

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



The Simple Regression Model

- **Assumptions for the linear regression model (cont.)**
- **Assumption SLR.3 (Sample variation in the explanatory variable)**

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$$

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable lead to different values of the dependent variable)

- **Assumption SLR.4 (Zero conditional mean)**

$$E(u_i|x_i) = 0$$

The value of the explanatory variable must contain no information about the mean of the unobserved factors

The Simple Regression Model

- **Theorem 2.1 (Unbiasedness of OLS)**

$$SLR.1 - SLR.4 \quad \Rightarrow \quad E(\widehat{\beta}_0) = \beta_0,\ E(\widehat{\beta}_1) = \beta_1$$

- **Interpretation of unbiasedness**
  - The estimated coefficients may be smaller or larger, depending on the sample that is the result of a random draw
  - However, on average, they will be equal to the (true) values that charac-terize the true relationship between y and x in the population
  - "On average" means if sampling was repeated, i.e. if drawing the random sample and doing the estimation was repeated many times
  - In a given sample, estimates may differ considerably from true values

The Simple Regression Model

- **Variances of the OLS estimators**
  - Depending on the sample, the estimates will be nearer or farther away from the true population values
  - How far can we expect our estimates to be away from the true population values on average (= sampling variability)?
  - Sampling variability is measured by the estimator's variances

$$Var(\widehat{\beta}_0),\ Var(\widehat{\beta}_1)$$

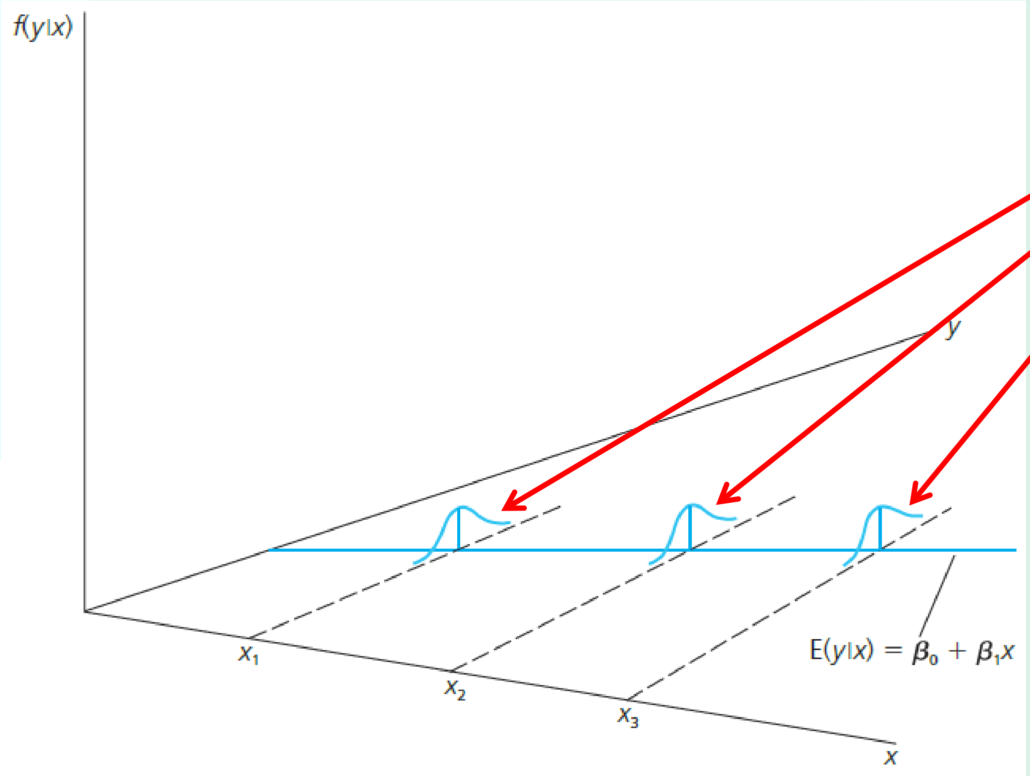
- **Assumption SLR.5 (Homoskedasticity)**

$$Var(u_i|x_i) = \sigma^2$$

The value of the explanatory variable must contain no information about the variability of the unobserved factors

The Simple Regression Model

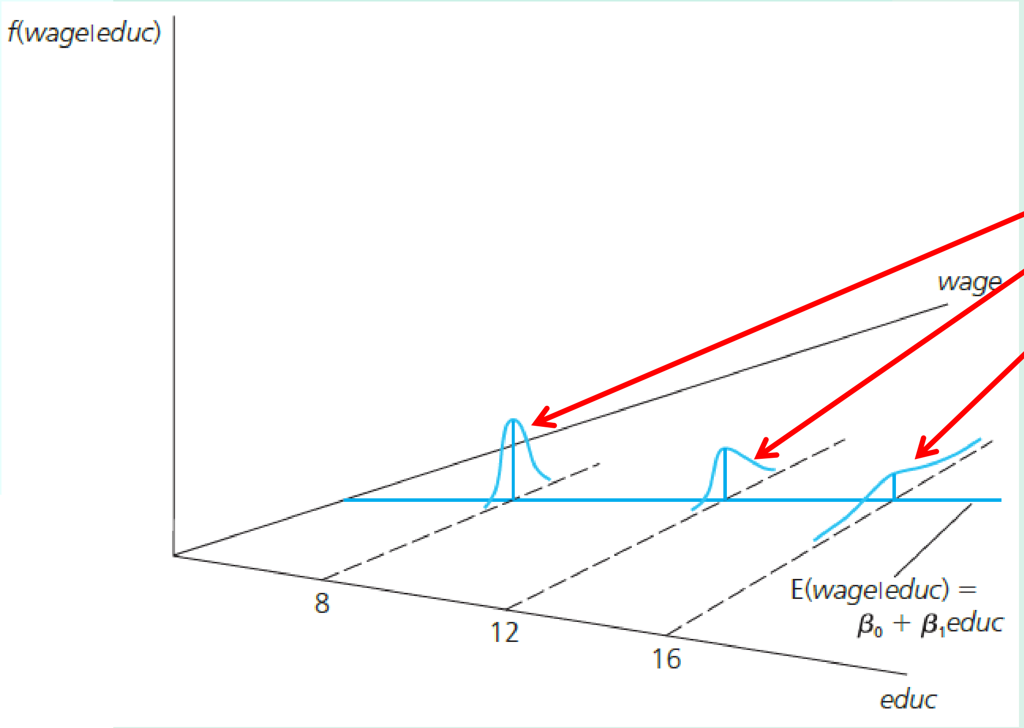
- **Graphical illustration of homoskedasticity**



The variability of the unobserved influences does not depend on the value of the explanatory variable

The Simple Regression Model

- **An example for heteroskedasticity: Wage and education**



The variance of the unobserved determinants of wages increases with the level of education

The Simple Regression Model

- **Theorem 2.2 (Variances of the OLS estimators)**

  Under assumptions SLR.1 – SLR.5:

$$Var(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$Var(\widehat{\beta}_0) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{SST_x}$$

- **Conclusion:**
  - The sampling variability of the estimated regression coefficients will be the higher, the larger the variability of the unobserved factors, and the lower, the higher the variation in the explanatory variable

The Simple Regression Model

- **Estimating the error variance**

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

The variance of u does not depend on x, i.e. equal to the unconditional variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

One could estimate the variance of the errors by calculating the variance of the residuals in the sample; unfortunately this estimate would be biased

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

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients from the number of observations

The Simple Regression Model

- **Theorem 2.3 (Unbiasedness of the error variance)**

$$SLR.1 - SLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

- **Calculation of standard errors (s.e.) for regression coefficients**

$$se(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2 / SST_x}$$

Plug in $\hat{\sigma}^2$ for the unknown σ^2

$$se(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2 / SST_x}$$

The estimated standard deviations of the regression coefficients are called "standard errors." They measure how precisely the regression coefficients are estimated.