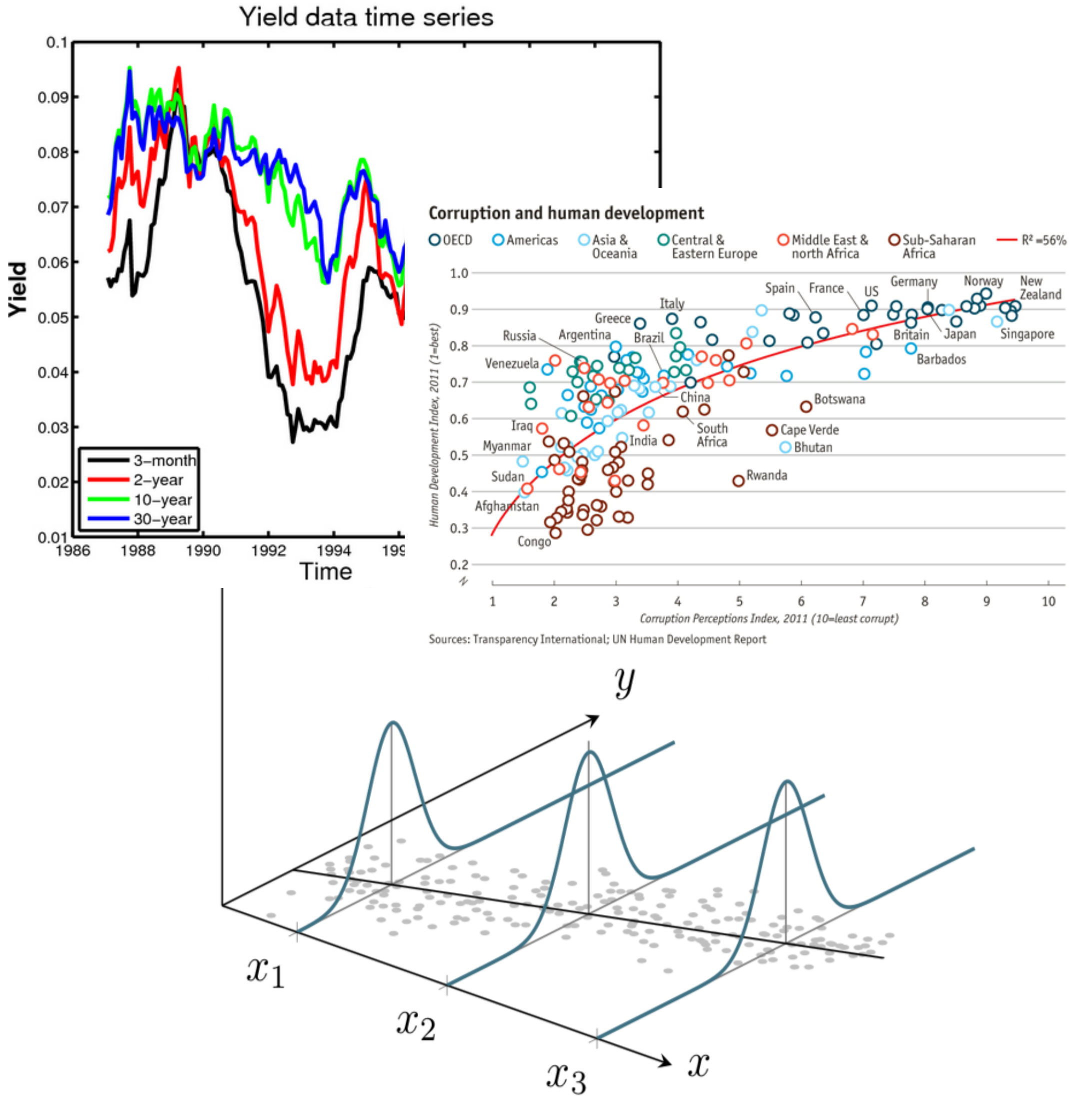
Chapter 2

The Simple Regression Model



The Simple Regression Model (1 of 39)

- **Definition of the simple 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 (2 of 39)

- **Interpretation of the simple linear regression model**
  - Explains how y varies with changes in x

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

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.

The Simple Regression Model (3 of 39)

- **Example: Soybean yield and fertilizer**

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

Rainfall,
land quality,
presence of parasites, …

Measures the effect of fertilizer on yield

- **Example: A simple wage equation**

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

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

Measures the change in hourly wage
given another year of education

The Simple Regression Model (4 of 39)

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

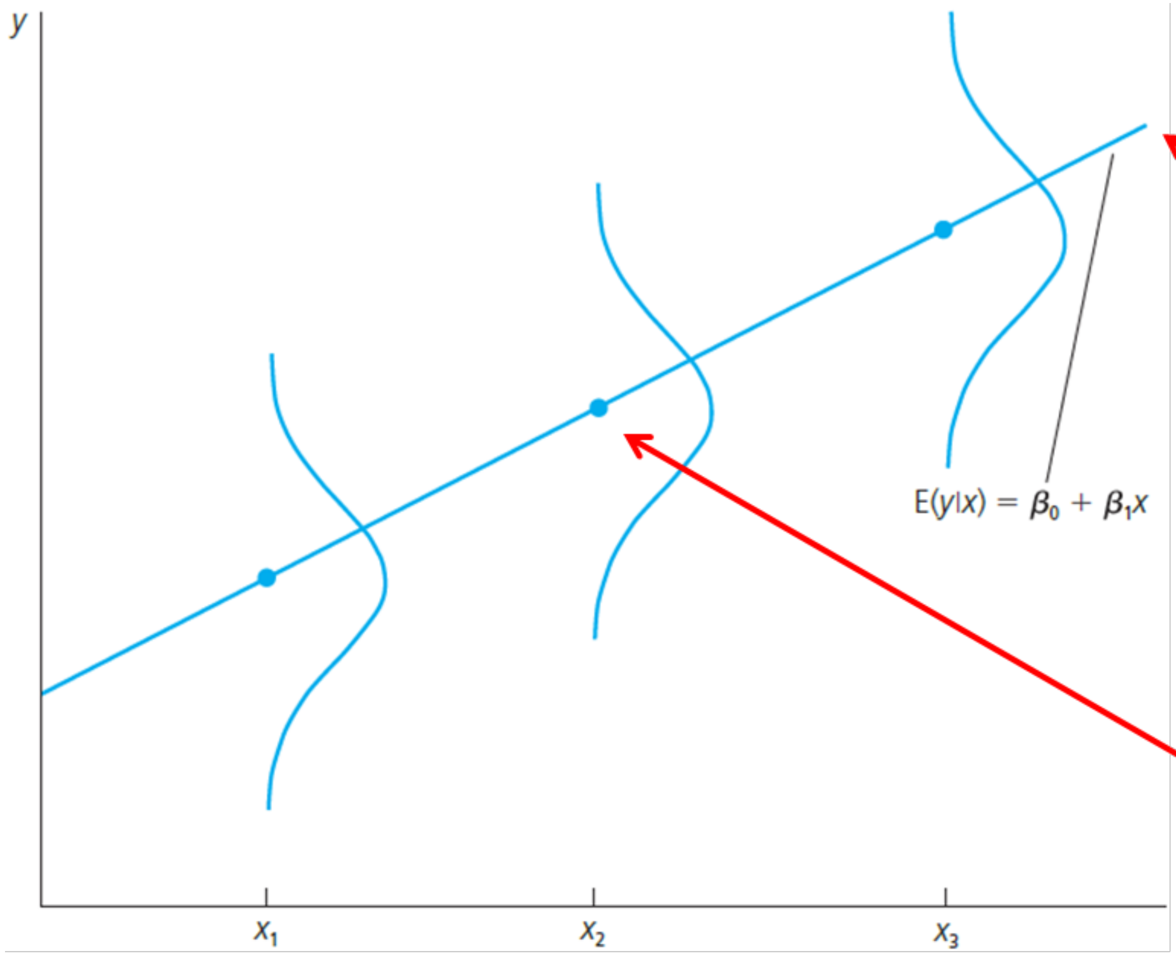
The Simple Regression Model (5 of 39)

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

$$\begin{aligned} E(y|x) &= E(\beta_0 + \beta_1 x + u|x) \\ &= \beta_0 + \beta_1 x + E(u|x) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

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

The Simple Regression Model (6 of 39)



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 (7 of 39)

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

$(x_1, y_1) \longleftarrow$ First observation

$(x_2, y_2) \longleftarrow$ Second observation

$(x_3, y_3) \longleftarrow$ Third observation

$\vdots$

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

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

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 (8 of 39)

- **Deriving the ordinary least squares (OLS) estimators**
- Defining regression residuals

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

- Minimize the sum of the squared regression residuals

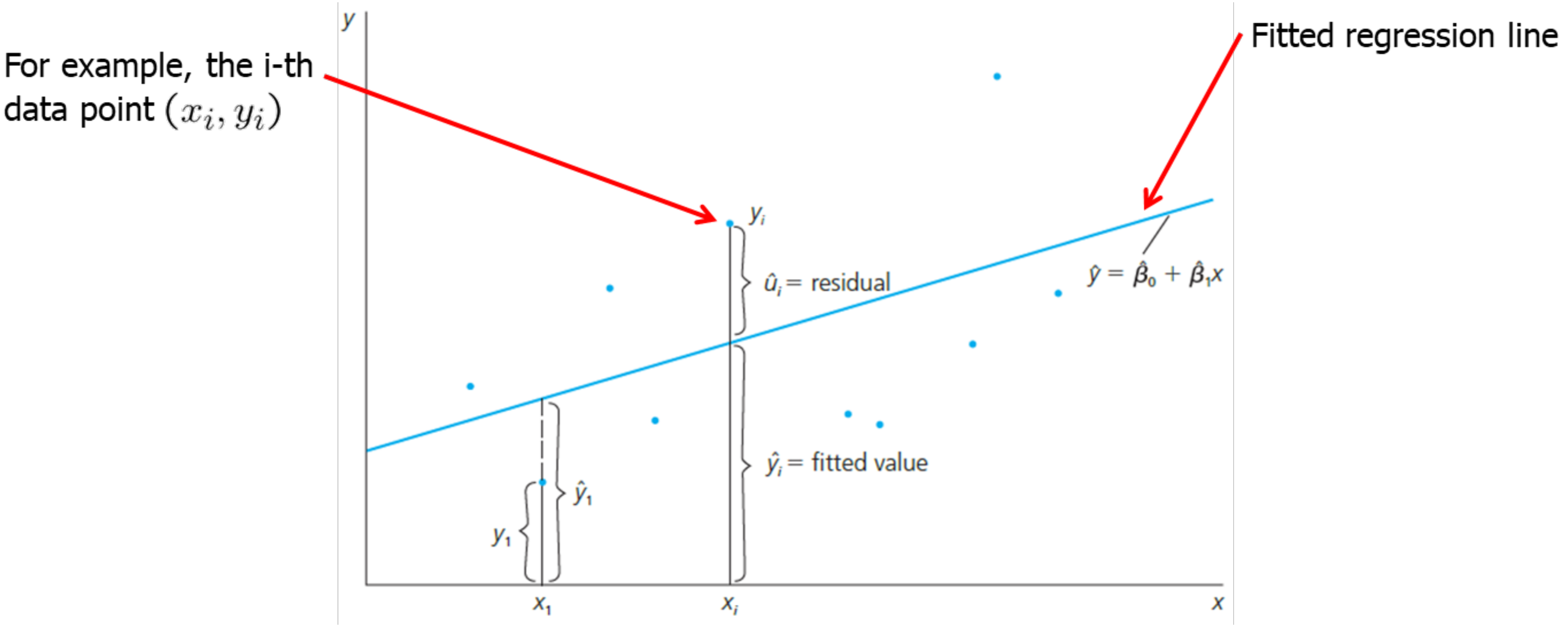
$$\min \sum_{i=1}^{n} \widehat{u}_i^2 \quad \rightarrow \quad \widehat{\beta}_0, \widehat{\beta}_1$$

- OLS estimators

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

The Simple Regression Model (9 of 39)

- **OLS fits as good as possible a regression line through the data points**



The Simple Regression Model (10 of 39)

- **Example of a simple regression**
- 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

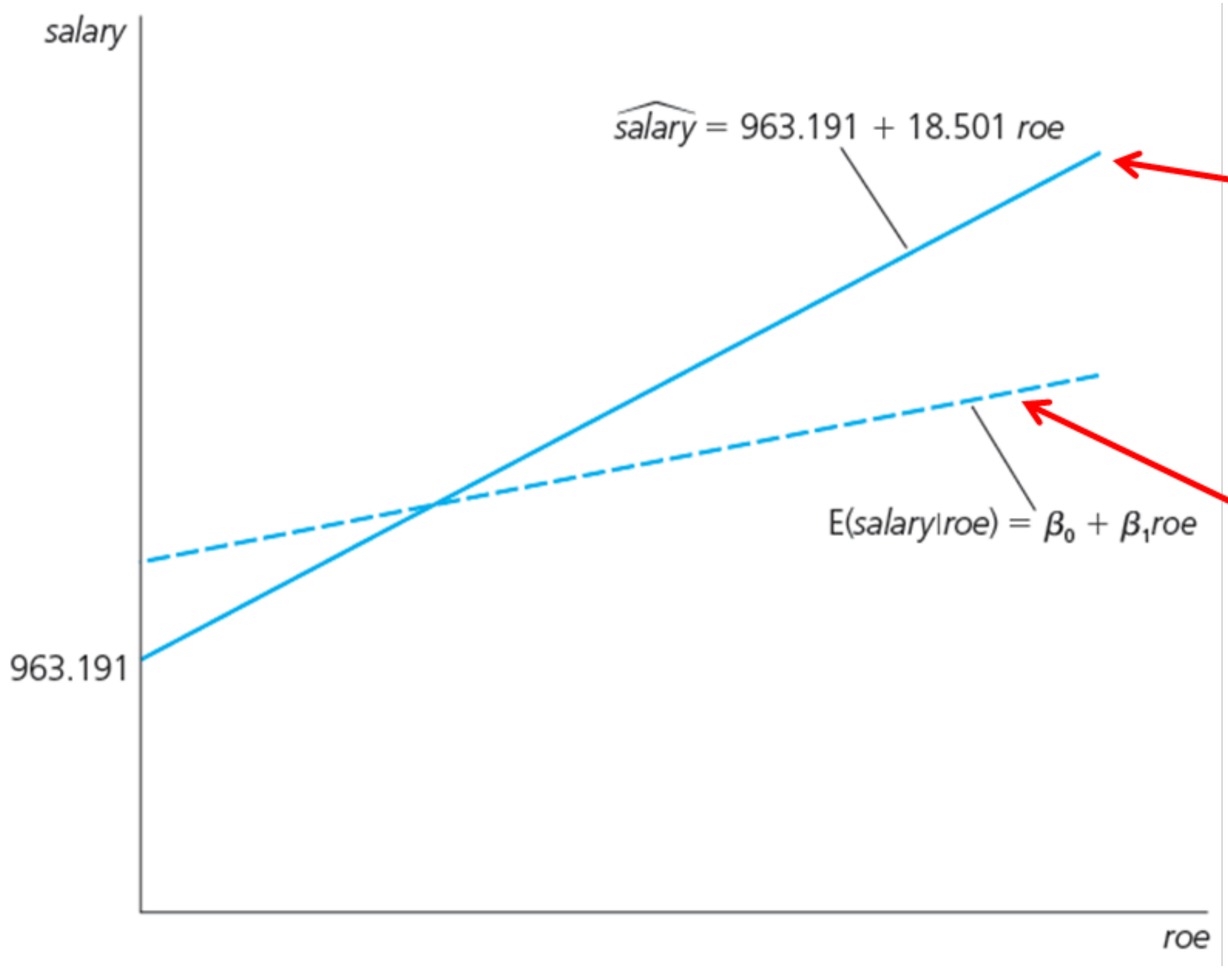
$$\widehat{salary} = 963.191 + 18.501\ roe$$

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 (11 of 39)



Fitted regression line (depends on sample)

Unknown population regression line

The Simple Regression Model (12 of 39)

- **Example of a simple regression**
- Wage and education

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

Hourly wage in dollars

Years of education

- Fitted regression

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

Intercept

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

- Causal interpretation?

The Simple Regression Model (13 of 39)

- **Example of a simple regression**
- 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 (14 of 39)

- **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 (15 of 39)

obsno	roe	salary	salaryhat	uhat
1	14.1	1095	1224.058	-129.058
2	10.9	1001	1164.854	-163.854
3	23.5	1122	1397.960	-275.969
4	5.9	578	1072.348	-494.348
5	13.8	1368	1218.508	149.493
6	20.0	1145	1333.215	-188.215
7	16.4	1078	1266.611	188.611
8	16.3	1094	1264.761	-170.761
9	10.5	1237	1157.454	79.546
10	26.3	833	1449.773	-616.773
11	25.9	567	1442.372	-875.372
12	26.8	933	1459.023	-526.023
13	14.8	1339	1237.009	101.991
14	22.3	937	1375.768	-438.768
15	56.3	2011	2004.808	6.192

- This table presents fitted values and residuals for 15 CEOs.
- For example, the 12th CEO's predicted salary is \$526,023 higher than their actual salary.
- By contrast the 5th CEO's predicted salary is \$149,493 lower than their actual salary.

The Simple Regression Model (16 of 39)

- **Goodness of fit**
  - How well does an explanatory variable explain the dependent variable?

- Measures of variation:

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

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

Explained sum of squares, represents variation explained by regression

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^{\,2}$$

Residual sum of squares, represents variation not explained by regression

The Simple Regression Model (17 of 39)

- **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 (18 of 39)

- **CEO Salary and return on equity**

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

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

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

- **Voting outcomes and campaign expenditures**

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

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

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!

The Simple Regression Model (19 of 39)

- **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

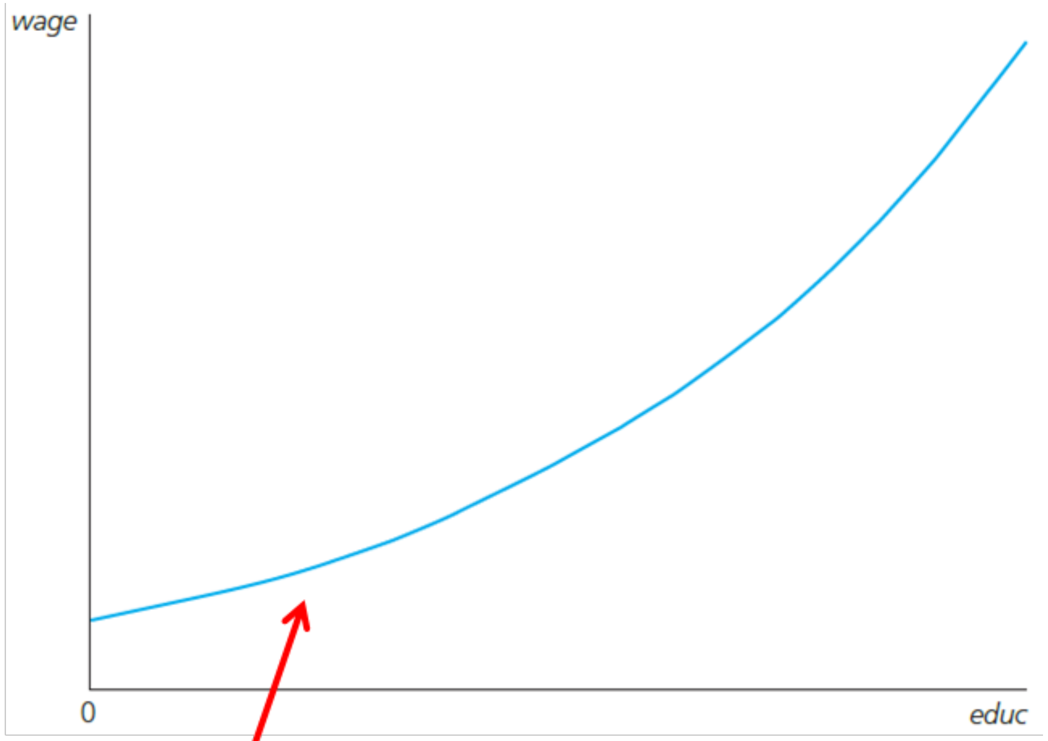
... if years of education are increased by one year

The Simple Regression Model (20 of 39)

- **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)



Growth rate of wage is 8.3% per year of education

The Simple Regression Model (21 of 39)

- **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 in salary if sales increase by 1%

Logarithmic changes are always percentage changes

The Simple Regression Model (22 of 39)

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

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

$$+1\%\ sales \rightarrow +.257\%\ salary$$

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

The Simple Regression Model (23 of 39)

- **Expected values and variances of the OLS estimators**
- The estimated regression coefficients are random variables because they are calculated from a random sample

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

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 will their variability be in repeated samples

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

The Simple Regression Model (24 of 39)

- **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)

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

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 (25 of 39)

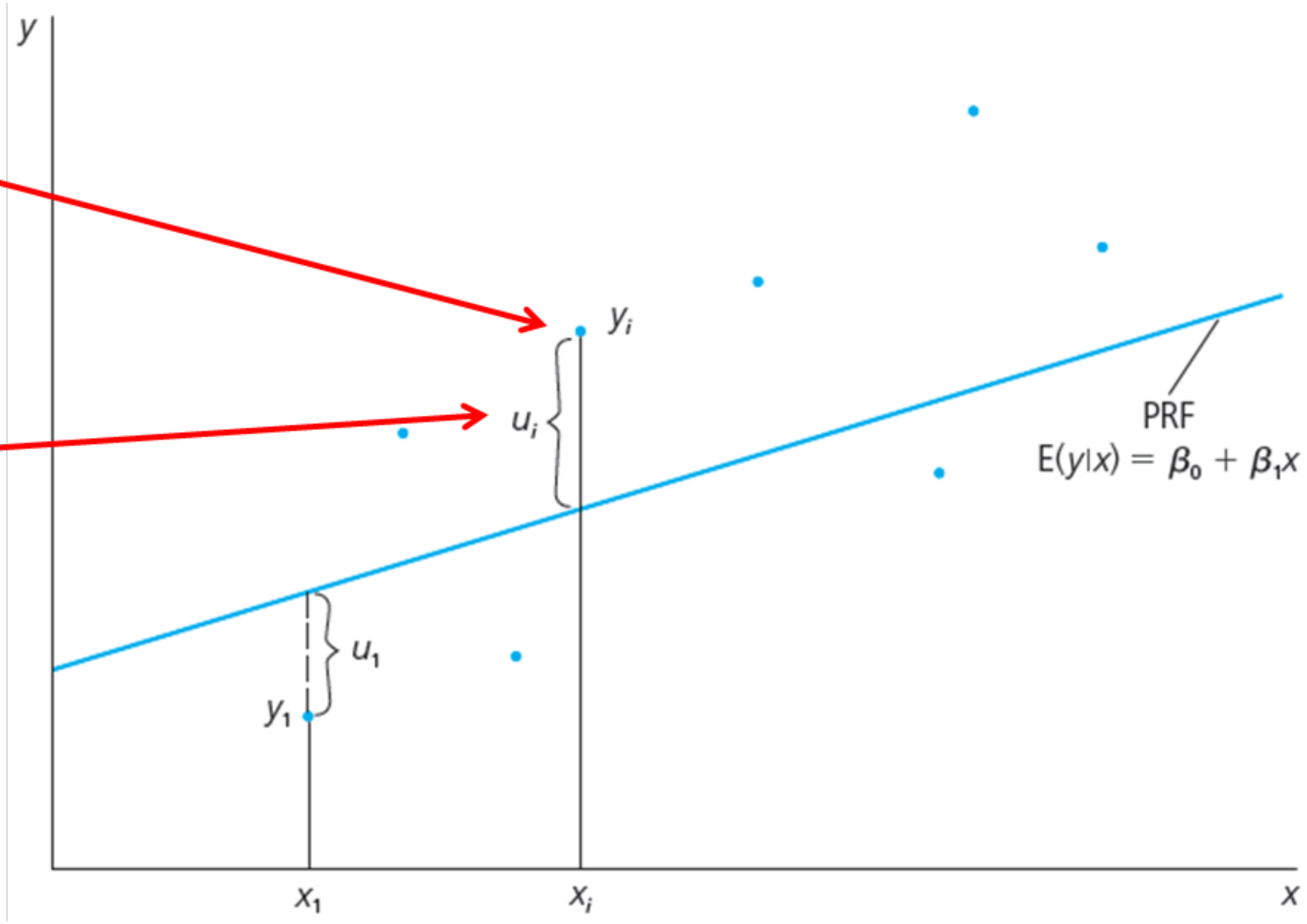
- **Discussion of random sampling: Wage and education**
  - The population consists, for example, of all workers of country A
  - In the population, there is a linear relationship between wages (or log wages) and years of education.
  - Draw completely randomly a worker from the population
  - The wage and the years of education of the worker drawn are random because one does not know beforehand which worker is drawn.
  - Throw that worker back into the population and repeat the 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 (26 of 39)

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 (27 of 39)

- **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 (28 of 39)

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

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

- 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 values that characterize 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 (29 of 39)

- **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

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

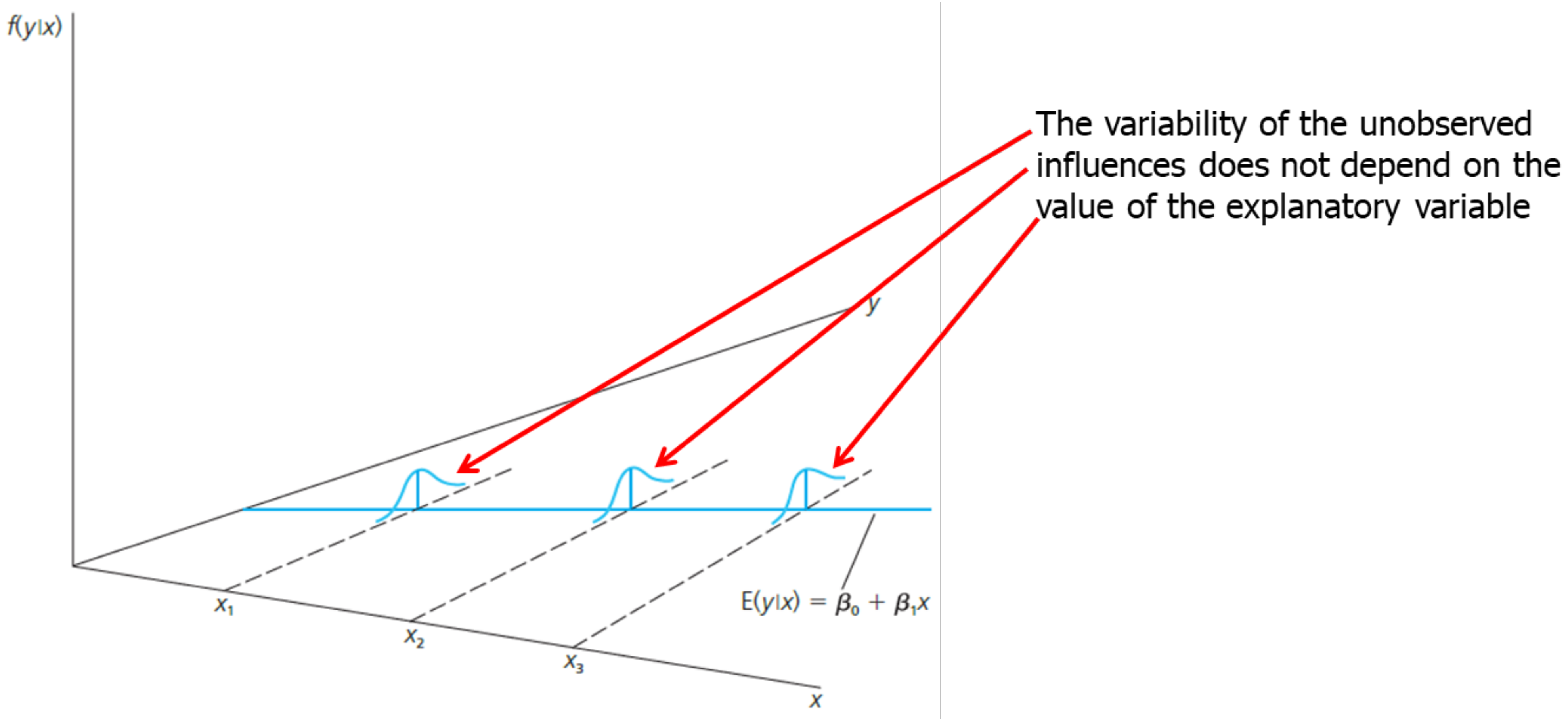
- Assumption SLR.5 (Homoskedasticity)

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

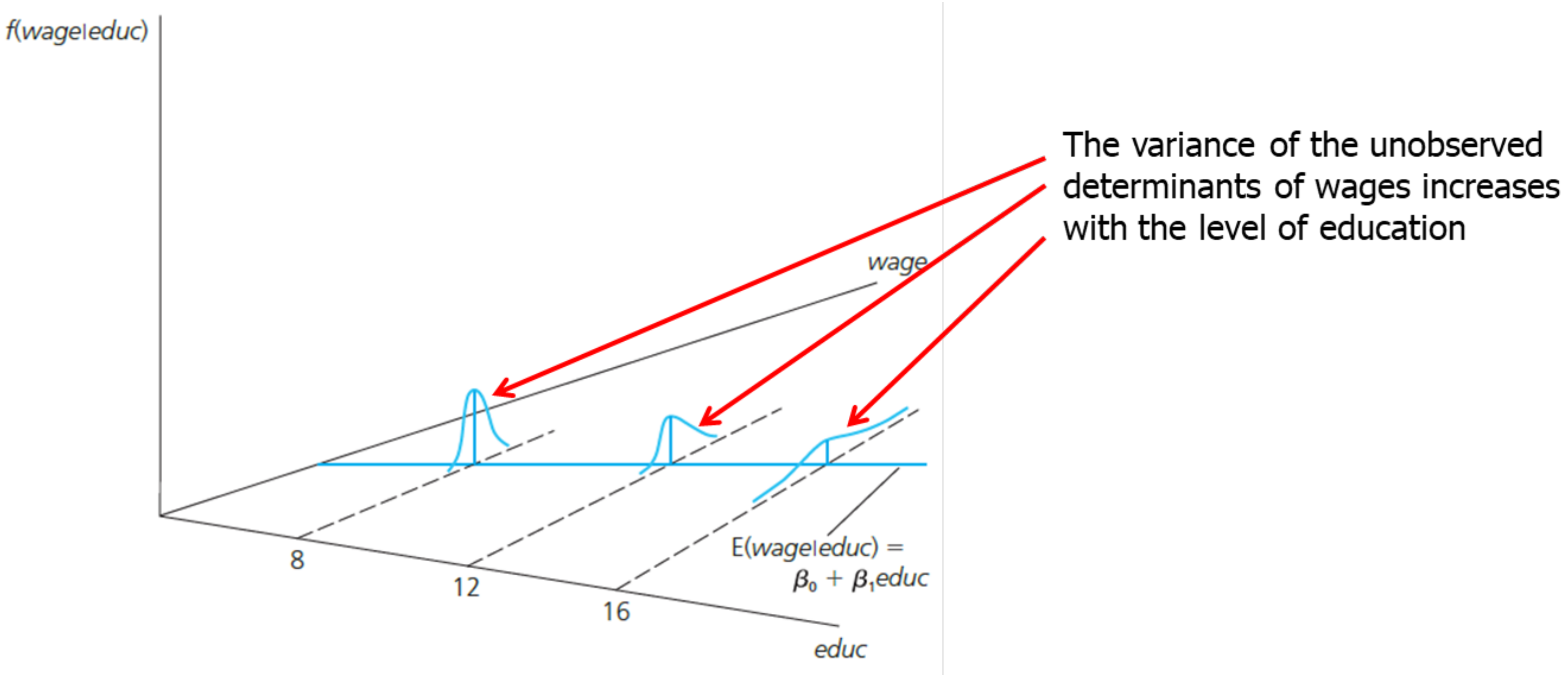
The value of the explanatory variable must contain no information about the variability of the unobserved factors

The Simple Regression Model (30 of 39)

- **Graphical illustration of homoskedasticity**



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



The Simple Regression Model (32 of 39)

- **Theorem 2.2 (Variances of the OLS estimators)**
- Under assumptions SLR.1 – SLR.5:

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

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

- 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 (33 of 39)

- **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 for regression coefficients

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

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

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

The Simple Regression Model

- **Regression on a binary explanatory variable**
- Suppose that x is either equal to 0 or 1

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

$$E(y|x = 0) = \beta_0 \quad E(y|x = 1) = \beta_0 + \beta_1$$

- This regression allows the mean value of y to differ depending on the state of x

$$\beta_1 = E(y|x = 1) - E(y|x = 0)$$

- Note that the statistical properties of OLS are no different when x is binary

The Simple Regression Model (36 of 39)

- **Counterfactual outcomes, causality and policy analysis**
- In policy analysis, define a treatment effect as:

$$\tau_i = y_i(1) - y_i(0)$$

- Note that we will never actually observe this since we either observe $y_i(1)$ or $y_i(0)$ for a given i, but never both.

- Let the average treatment effect be defined as:

$$\tau_{ate} = E[y_i(1)] - E[y_i(0)]$$

The Simple Regression Model (37 of 39)

- **Counterfactual outcomes, causality and policy analysis (contd.)**
- Let x_i be a binary policy variable.

$$y_i = (1 - x_i)y_i(0) + x_i y_i(1)$$

- This can be written as:

$$y_i = \alpha_0 + \tau x_i + u_i(0)$$

Assume that $y_i(0) = \alpha_0 + u_i(0)$ and a constant treatment effect such that $y_i = y_i(0) + \tau x_i$

- Therefore, regressing y on x will give us an estimate of the (constant) treatment effect.
- As long as we have random assignment, OLS will yield an unbiased estimator for the treatment effect τ.

The Simple Regression Model (38 of 39)

- **Random assignment**
  - Subjects are randomly assigned into treatment and control groups such that there are no systematic differences between the two groups other than the treatment.
  - In practice, randomized control trials (RCTs) are expensive to implement and may raise ethical issues.
  - Though RCTs are often not feasible in economics, it is useful to think about the kind of experiment you would run if random assignment was a possibility. This helps in identifying the potential impediments to random assignment (that we could conceivable control for in a multivariate regression).

The Simple Regression Model (39 of 39)

- **Example: The effects of a job training program on earnings**
- Real earnings are regressed on a binary variable indicating participation in a job training program.

$$\widehat{re78} = 4.55 + 1.79 train$$
$$n = 445, R^2 = 0.018$$

$re78$ is real earnings in 1978 measured in thousands of dollars. $train$ is a binary variable equal to 1 if the individual participated in the job training program

- Those who participated in the training program have earnings $1,790 higher than those who did not participate.
- This represents a 39.3% increase over the $4,550 average earnings from those who did not participate.