ECONOMETRICS

2017-2018

Ainara González de San Román

agod@faculty.ie.edu

UNIT II. LINEAR REGRESSION MODEL

(Chapter 2 - Wooldridge)

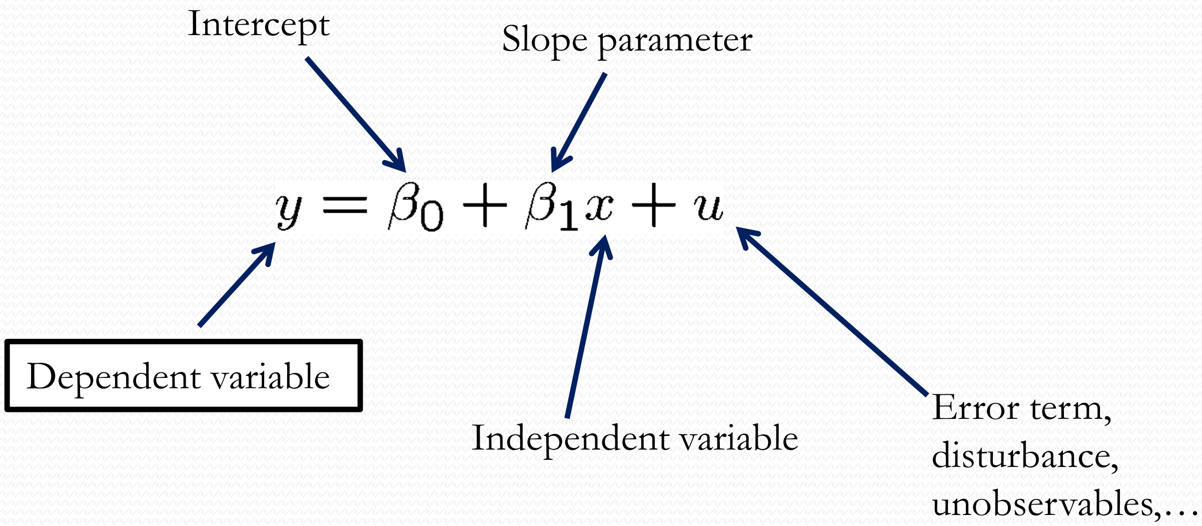
Outline

- The Simple Linear Regression Model (LRM)
- Estimation – Ordinary Least Squares (OLS)
- Properties of the Regression Coefficients
- Transformation of Variables

The Simple Linear 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

Independent variable

Error term,
disturbance,
unobservables,…

The Simple Linear Regression Model

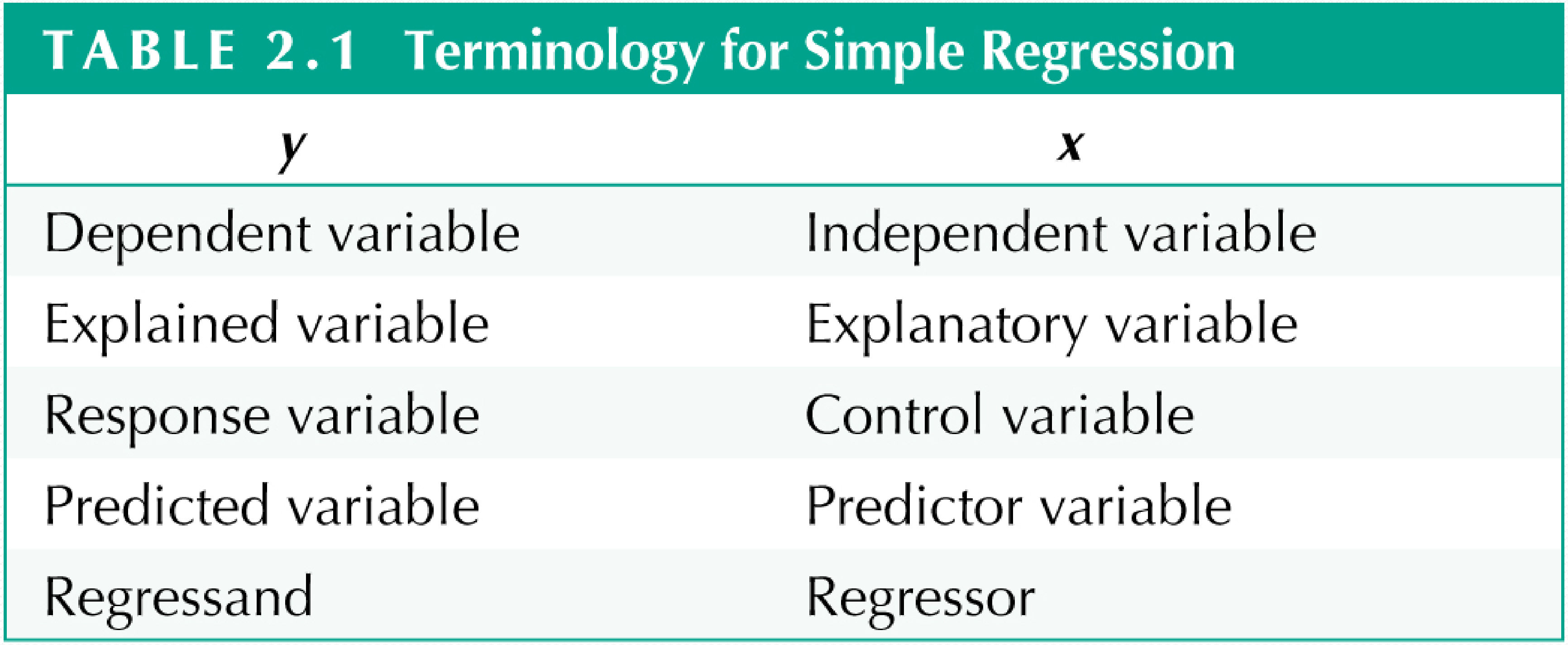


TABLE 2.1 Terminology for Simple Regression

y	*x*
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

The Simple Linear Regression Model

Interpretation of the simple linear regression model

"Studies how y varies with changes in x"

$$\frac{\partial y}{\partial x} = \beta_1 \qquad \text{as long as} \qquad \frac{\partial u}{\partial x} = 0$$

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

- **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, holding all other factors fixed

- **Limitation:** linearity implies that a one-unit change in x has the same effect on y regardless of the initial value of x This is unrealistic for many economic applications.

The Simple Linear Regression Model

When is there a causal interpretation?

- **Conditional mean independence assumption**

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

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

- **Example**: wage equation

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

The Simple Linear Regression Model

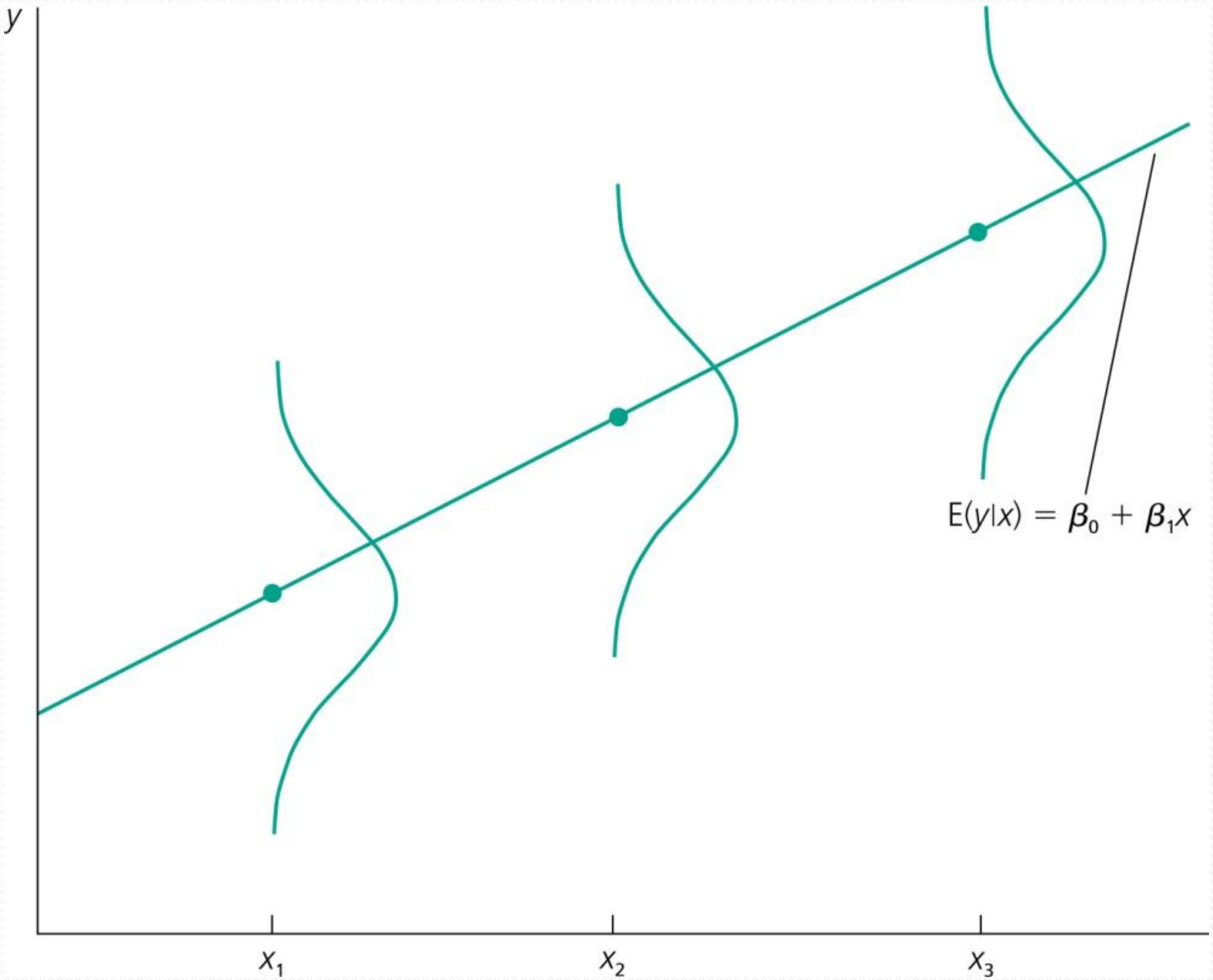
Population regression function (PFR)

- 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 Linear Regression Model



The Simple Linear 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 Linear Regression Model

Standard assumptions for the linear regression model

- **Assumption SLR.3 (Sample variation in 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)**

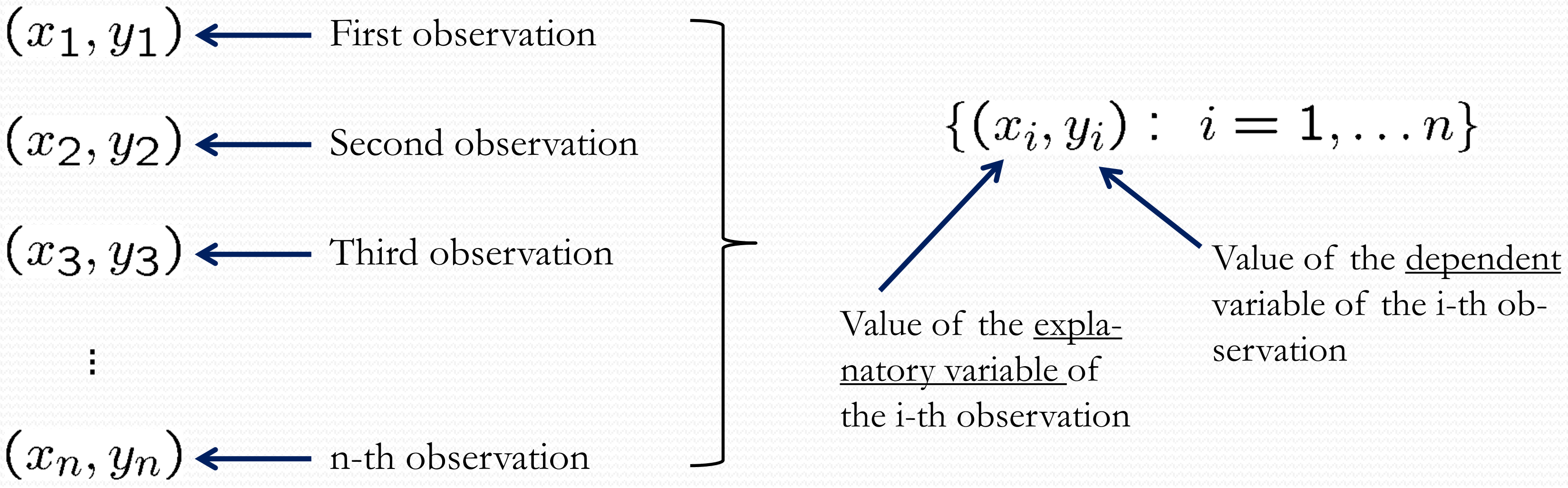
$$E(u_i|x_i)=0$$

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

The Simple Linear Regression Model

HOW TO ESTIMATE THE`PARAMETERS OF THE MODEL

- In order to estimate the regression model we need data: A random sample from the population



$(x_1, y_1) \longleftarrow$ 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

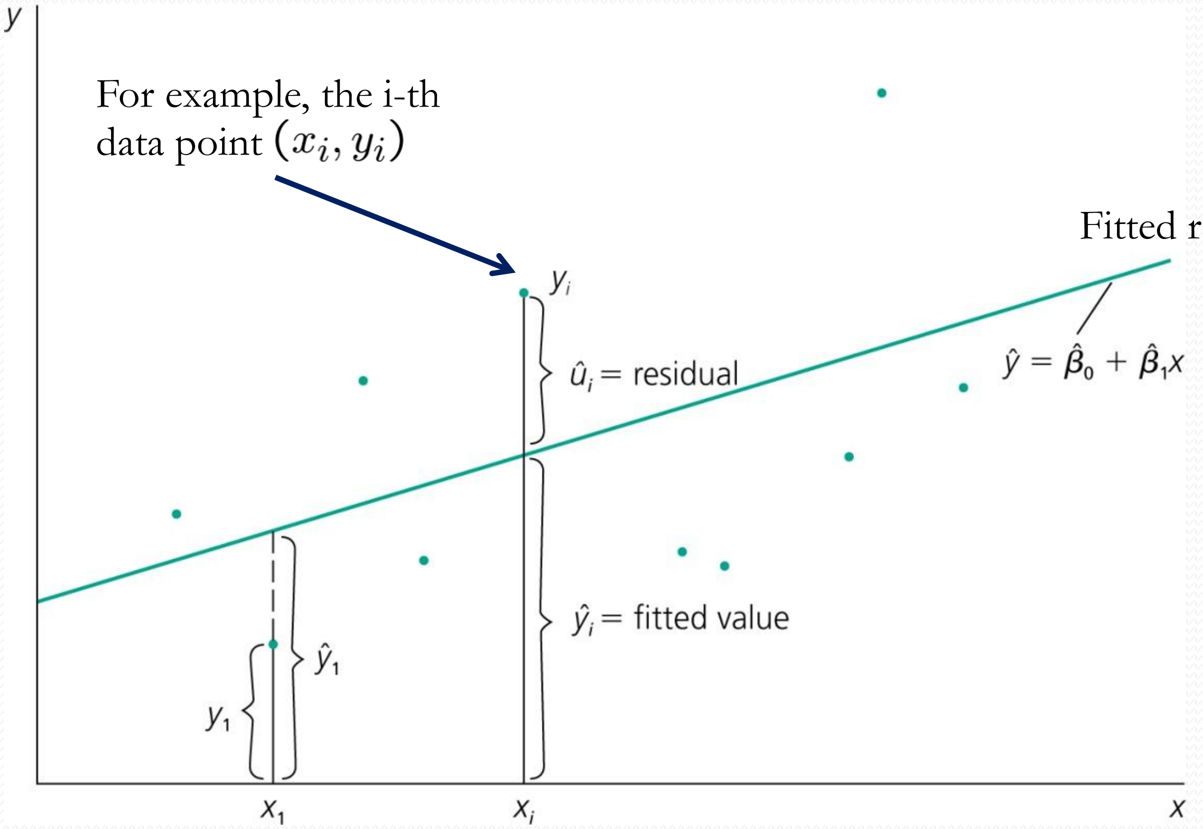
$$\{(x_i, y_i) : \ i = 1, \ldots n\}$$

Value of the explanatory variable of the i-th observation

Value of the dependent variable of the i-th observation

The Simple Linear Regression Model

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


y
y_i
$\hat{u}_i$ = residual
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y}_i$ = fitted value
$\hat{y}_1$
y_1
x_1
x_i
x

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

Fitted regression line

The Simple Linear Regression Model

HOW TO ESTIMATE THE PARAMETERS OF THE MODEL

Two unknowns and two equations:

$$E(y-\beta_0-\beta_1 x)=0$$

$$E[x(y-\beta_0-\beta_1 x)]=0$$

Given the data, we choose the estimates that solve the sample counterpart of the system of equations above.

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

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

The Simple Linear Regression Model

HOW TO ESTIMATE THE PARAMETERS OF THE MODEL

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

These estimates are called the **Ordinary Least Squares (OLS)** estimates of β_0 and β_1. In this session we will learn why they receive this name.

Estimation – Ordinary Least Squares

- **Fitted value:** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- **Regression residuals :** difference between the actual and the fitted value.

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

- **Minimize sum of squared regression residuals**

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

- **First Order Conditions lead to Ordinary Least Squares (OLS) estimates**

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

CHECK!!!

Estimation – Ordinary Least Squares

- The name "Ordinary Least Squares" comes from the fact that these estimates minimize the sum of squared residuals.
- Once we have determined the OLS intercept and slope estimates, we form the **OLS regression line:**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Intercept → $\hat{\beta}_0$ **:** is the predicted value of y when $x = 0$

Slope → $\hat{\beta}_1$ **:** it tells us the amount by which $\hat{y}$ changes when x increases by 1 unit.

$$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x} \quad \textit{or alternatively} \quad \Delta \hat{y} = \hat{\beta}_1 \, \Delta x$$

Given any change in x we can compute the predicted change in y

Estimation – Ordinary Least Squares

- We next examine several examples of simple regression obtained by using real data.
- Since these examples involve many observations, the calculations were done using an econometrics software package.
- At this point, you should be careful not to read too much into these regressions; they are not necessarily uncovering a causal relationship.
- We could interpret much better the estimates once we establish the statistical properties of them.

Estimation – Ordinary Least Squares

EXAMPLE 1: CEO SALARY AND RETURN ON EQUITY

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

Salary in thousands of dollars

Average Return on equity for the CEO‘s firm for the previous 3 years (%)

- **OLS regression – 209 observations (CEOs) in 2013**

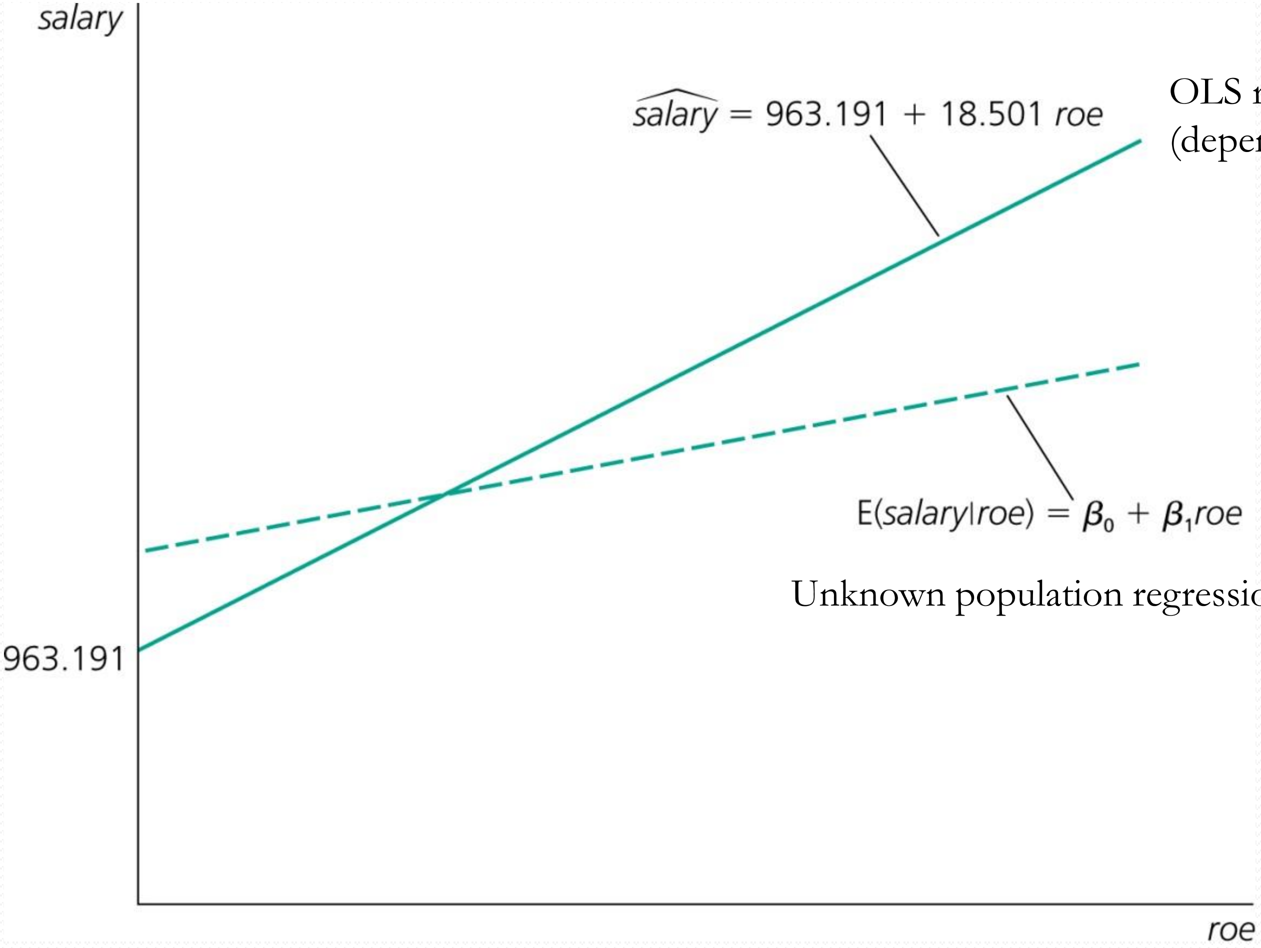
$$\widehat{salary} = 963.191 + 18.501\ roe$$

Intercept

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

- **Causal interpretation?**

Estimation – Ordinary Least Squares



OLS regression line
(depends on sample)

Unknown population regression line

Estimation – Ordinary Least Squares

EXAMPLE 2: WAGE AND EDUCATION

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

Hourly wage in dollars

Years of schooling

- **OLS regression - 526 individuals in 1976**

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

Intercept

In the sample, one more year of education was associated with an increase in hourly wage by 54 cents

Estimation – Ordinary Least Squares

EXAMPLE 3: Voting outcomes and campaign expenditures (two parties)

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

Percentage of votes for candidate A

Percentage of campaign expenditures candidate A

- **OLS regression – 173 two-party races for a country social election in 2013**

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

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

Estimation – Ordinary Least Squares

Properties of OLS on any sample of data

1. **Fitted values and residuals:** we assume that the intercept and slope estimates have been obtained for a given sample of data. Given $\hat{\beta}_0$ and $\hat{\beta}_1$ we can obtain the fitted value for each observation:

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

The OLS residual associated to each observation is the difference between the actual value and the fitted value of the dependent variable:

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

- If $\hat{u}_i$ is positive, the line underpredicts y_i; if negative, the line overpredictsy_i.

Estimation – Ordinary Least Squares

EXAMPLE 2 (Continued) – the 15 first observations

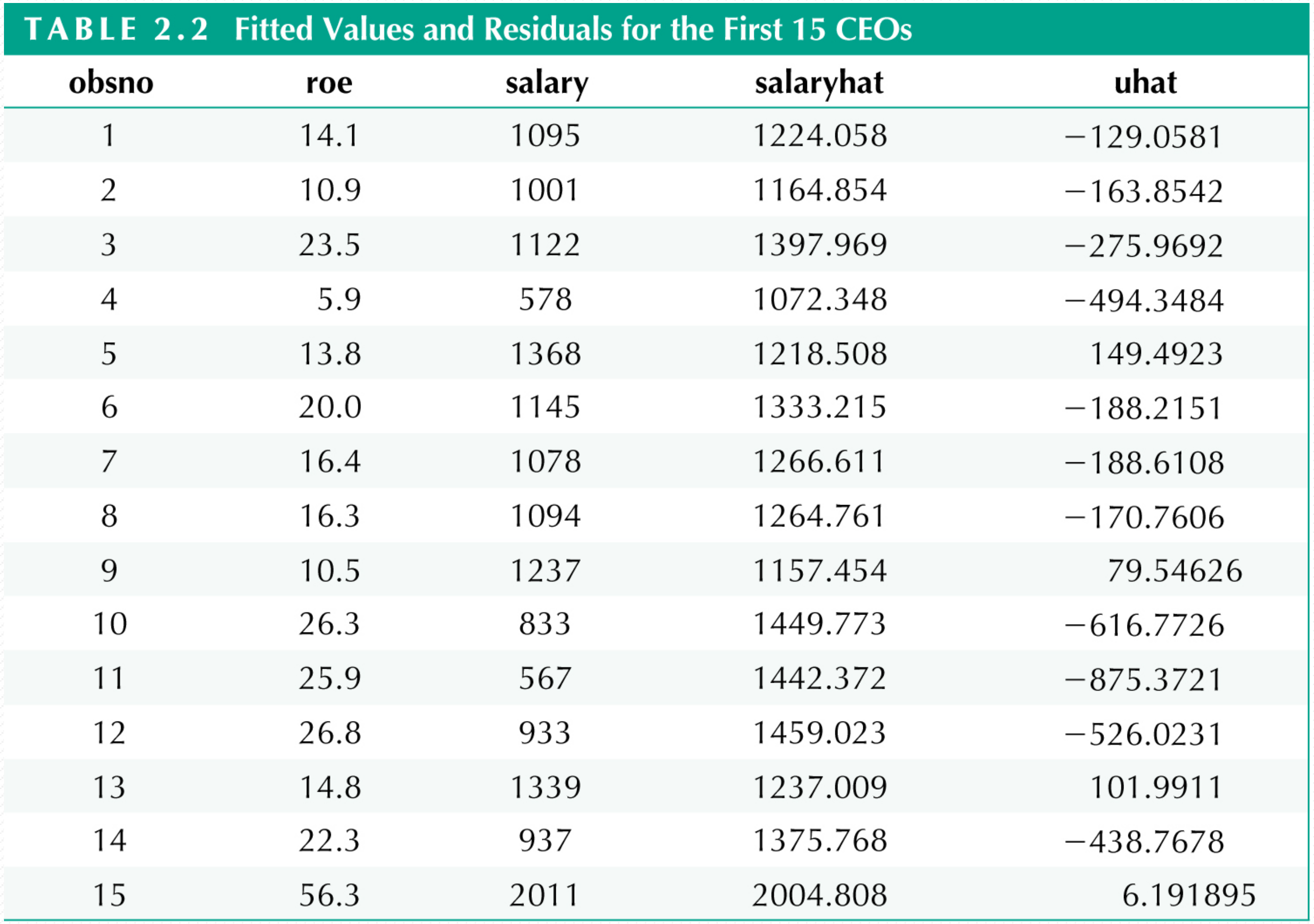


TABLE 2.2 Fitted Values and Residuals for the First 15 CEOs

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

Estimation – Ordinary Least Squares

Properties of OLS on any sample of data

2. **Algebraic properties of OLS regression – 3 important properties!**

(1) The sum, and therefore the sample average of the OLS residuals, is zero.

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

(2) The sample covariance between the regressors and the OLS residuals is zero

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

Estimation – Ordinary Least Squares

Properties of OLS on any sample of data

2. **Algebraic properties of OLS regression – 3 important properties!**

(3) The point $(\bar{x}, \bar{y})$ is always on the OLS regression line

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Remarks: these properties need no proof. Property (1) and (2) follow from the OLS first order conditions. Property (3) comes from the OLS estimation of the intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Estimation – Ordinary Least Squares

GOODNESS-OF-FIT

How well does the explanatory variable explain the dependent variable?

- **Measures of Variation**

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

Total sum of squares, represents total variation in dependent variable

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

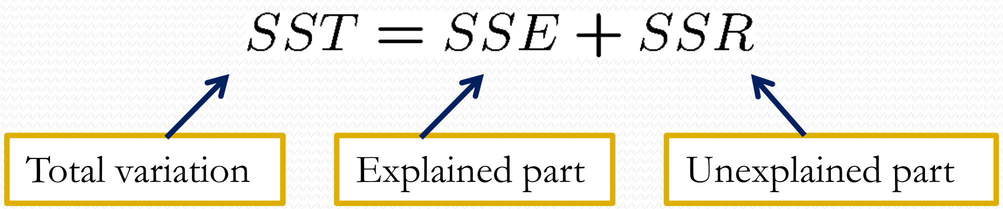
Explained sum of squares, represents variation explained by regression

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

Residual sum of squares, represents variation not explained by regression

Estimation – Ordinary Least Squares

- **Decomposition of total variation**



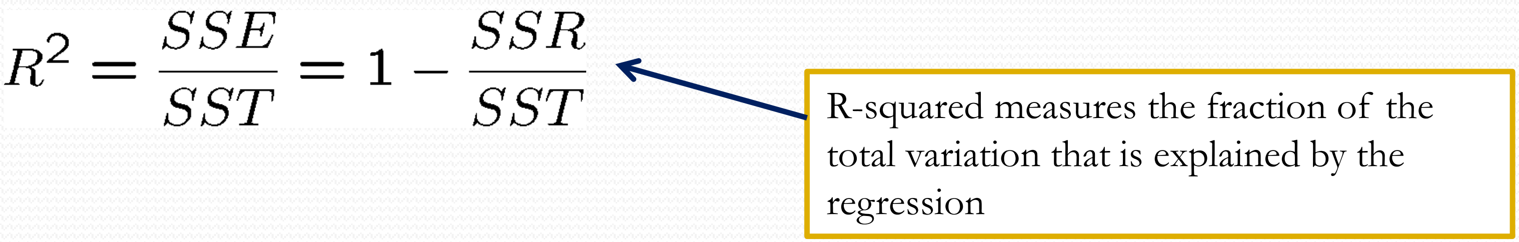
$$SST = SSE + SSR$$

Total variation

Explained part

Unexplained part

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



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

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

Estimation – Ordinary Least Squares

- **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!

Estimation – Ordinary Least Squares

Example: The following table contains the *ACT* scores and the *GPA* for eight college students. Grade point average is based on a four-point scale and has been rounded to one digit after decimal.

Student	*GPA*	*ACT*
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

Estimation – Ordinary Least Squares

Example:

- Estimate the relationship between *GPA* and *ACT* using OLS; that is, obtain the intercept and the slope estimates in the equation.

$$\widehat{GPA} = 0.5681 + 0.1022ACT$$

- Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain.

- How much higher is the *GPA* predicted to be if the *ACT* score is increased by five points? . If *ACT* is 5 points higher, $\widehat{GPA}$ increases by 0.1022(5) = 0.511.

Estimation – Ordinary Least Squares

Example:

- Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum up to zero.

i	GPA	$\widehat{GPA}$	$\hat{u}$
1	2.8	2.7143	0.0857
2	3.4	3.0209	0.3791
3	3.0	3.2253	–0.2253
4	3.5	3.3275	0.1725
5	3.6	3.5319	0.0681
6	3.0	3.1231	–0.1231
7	2.7	3.1231	–0.4231
8	3.7	3.6341	0.0659

Estimation – Ordinary Least Squares

Example:

- How much of the variation in *GPA* for these eight students is explained by *ACT*? Explain.

$R^2 = 1 - \text{SSR}/\text{SST} \quad 1 - (0.4347/1.0288) = 0.577.$

Therefore, about 57.7% of the variation in *GPA* is explained by *ACT* in this small sample of students.

Properties of the Regression Coefficients

- Recall the properties of any estimator $\hat{\theta}$ we learnt in Unit I
- Now the question is...

...What the estimators will estimate on average and how large their variability in repeated samples is going to be???

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

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

Properties of the Regression Coefficients

Theorem 1.1: 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 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.

Properties of the Regression Coefficients

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

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

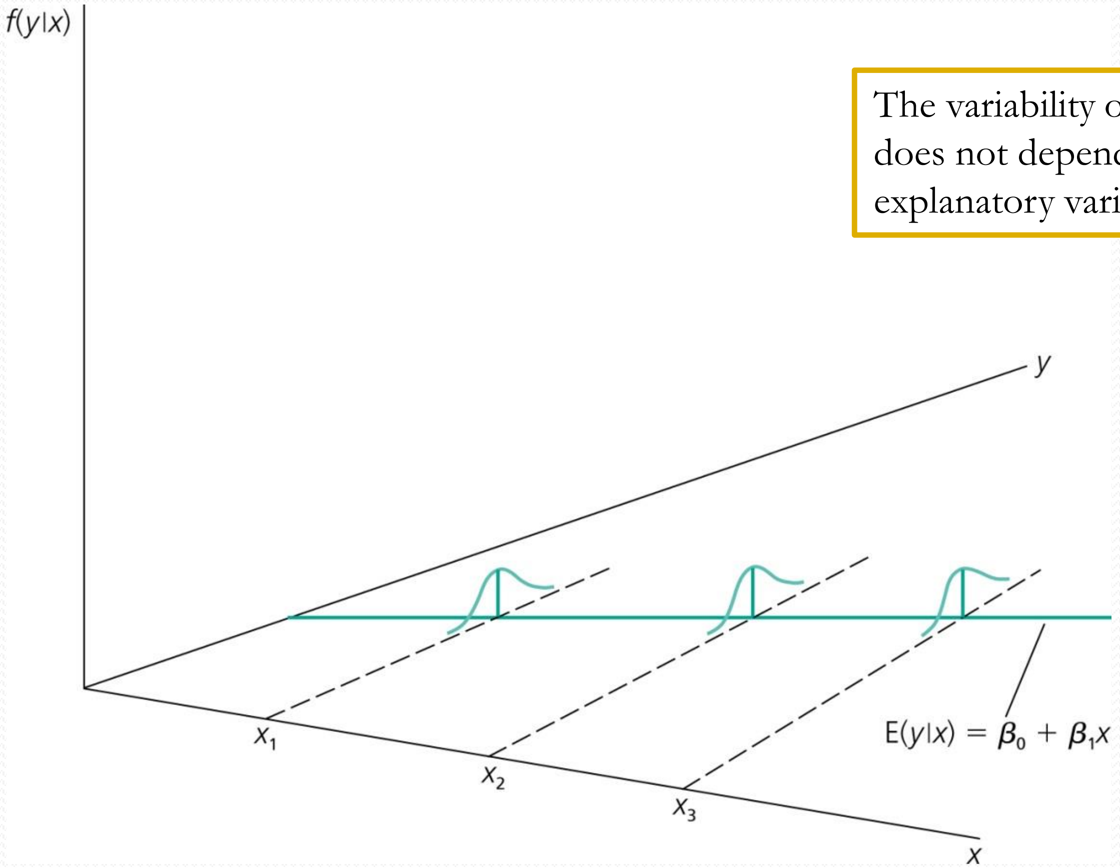
- **Assumption SLR.5 (Homoskedasticity)**

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

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

Properties of the Regression Coefficients

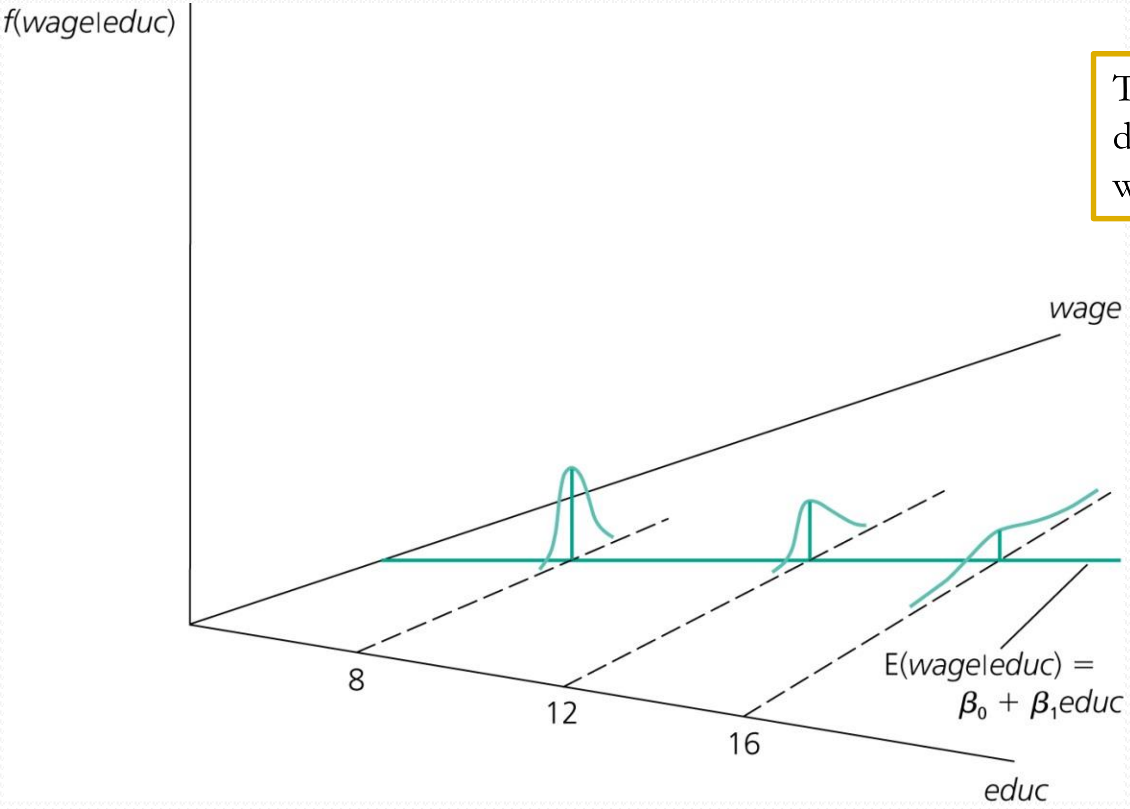
Graphical illustration of homoskedasticity



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

Properties of the Regression Coefficients

An example for <u>heteroskedasticity</u>: Wage and education



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

Properties of the Regression Coefficients

Theorem 2.2: Variances of 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 higher the larger the variability of the unobserved factors, and lower, the higher the variation in the explanatory variable.

Properties of the Regression Coefficients

- **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. is equal to the unconditional variance

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

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 subtracting the number of estimated regression coefficients from the number of observations

Properties of the Regression Coefficients

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**

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

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

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

Transformation of Variables

- **Linear Relationships:** So far we have been focused on linear relationships between the dependent and independent variables.
  - Fortunately, it is rather easy to incorporate many nonlinearities into simple regression analyisis by appropiately defining the dependent and independet variables.
  - We will cover two possibilities that often appear in applied work
    1. **Semi-Logarithmic form**: the dependent variable is transformed into logs
    2. **Log-Logarithmic form**: both the dependent and explanatory variables are transformed into logs.

Transformation of Variables

Incorporating nonlinearities: 1. Semi-logarithmic form

- The dependent variable appears in **logarithmic** form. *Why is this done?*
- Recall the **wage-education example.**
    - We obtained a slope estimate of 0.54, which means that each additional year of education is predicted to increase hourly wage by 54 cents.
    - Because of the linear nature of the that relationship, 54 cents is the increase for either the first year of education or the twentieth year. This may not be reasonable!
    - Probably a better characterization of how wage changes with education is that each year of education increases wage by a *constant percentage*. The **semi-log model** gives us that constant percentage effect.

Transformation of Variables

Incorporating nonlinearities: 1. 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:**

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

Percentage change of wage

… if years of education are increased by one year

Transformation of Variables

Incorporating nonlinearities: 1. Semi-logarithmic form

- **Fitted regression**

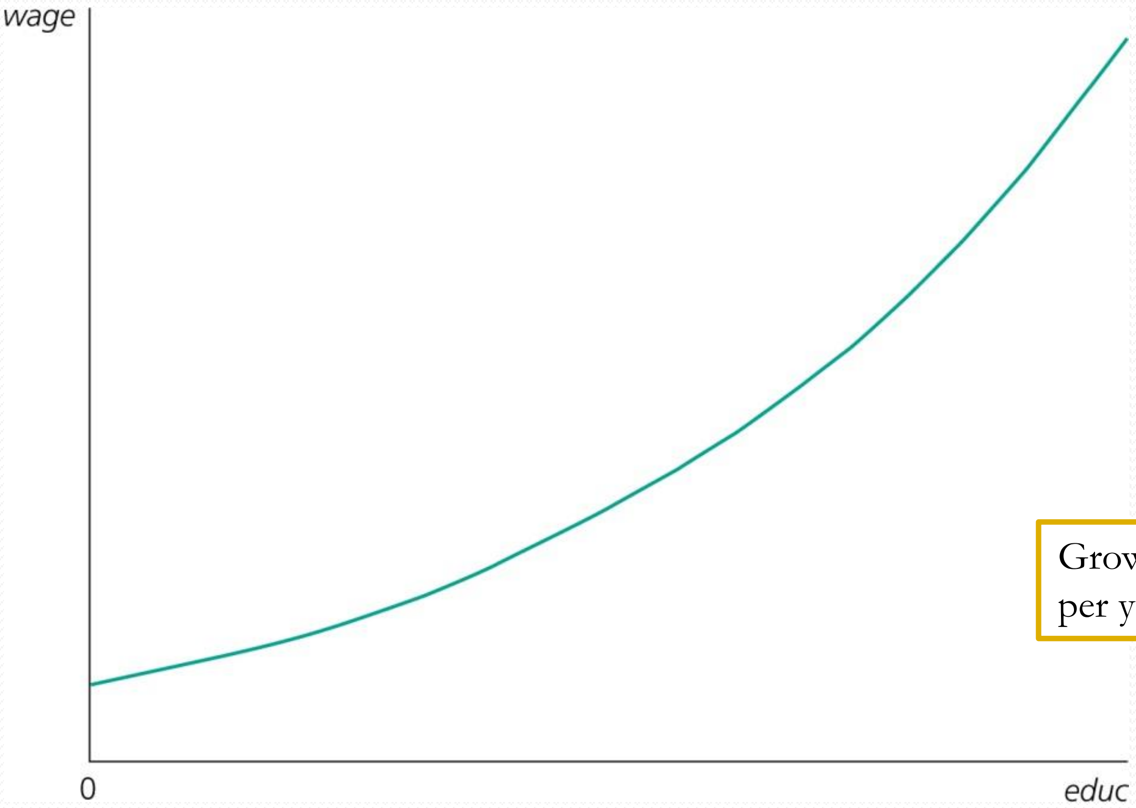
$$\widehat{\log}(wage) = 0.584 + 0.083\ educ$$

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

Remark: rememeber that the main reason for using the log of wage is to impose a **constant percentage effect** of education on wage. Once the above equation is obtained, the natural log of wage is rarely mentioned.

Transformation of Variables

Incorporating nonlinearities: 1. Semi-logarithmic form



Growth rate of wage is 8.3 % per year of education

Transformation of Variables

Incorporating nonlinearities: 2. 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:**

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

Percentage change of salary

… if sales increase by 1%

Logarithmic changes are always percentage changes

Transformation of Variables

Incorporating nonlinearities: 2. Log-logarithmic form

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

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

+ 1 % sales → + 0.257 % salary

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

Transformation of Variables

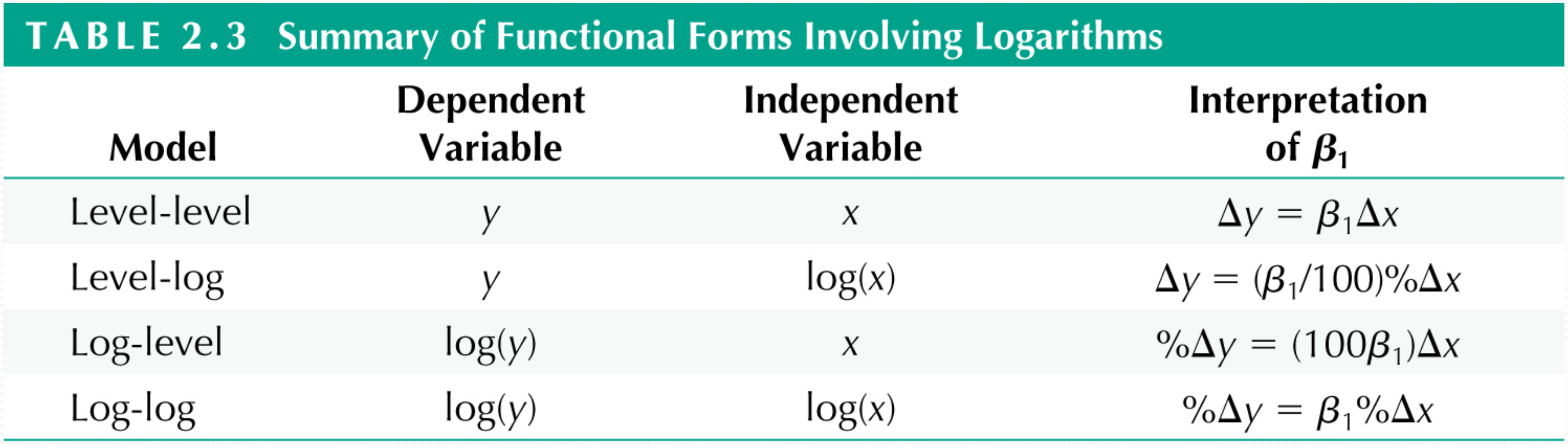


TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Summary

- We have introduced the simple linear regression model and cover its basic properties.
- Given a random sample, the method of ordinary least squares is used to estimate the slope and intercept parameters in the population model.
- We have demonstrated the algebra of the OLS regression line, including computation of fitted values and residuals, and the obtaining of predicted changes in the dependent variable for a given change in the independent variable.
- We discuss the use of the natural log to allow for constant elasticity and constant semi-elasticity models.
- We learnt that the OLS estimators are unbiased.
- We get simple formulas for the sampling variances of the OLS estimators when we add the assumption of homoscedasticity.