

Simple Regression Model (Part II)

Pedram Jahangiry



Basic Definitions

- ❑ **Population**: The entire group of subjects that interests us (Individuals, firms, cities, and ...)
- ❑ **Sample**: Part of this population that we actually observe
- ❑ **parameter** : A characteristic of the **population** whose value is **unknown**, but can be estimated.
- ❑ **Estimator**: A sample statistic (a rule) that will be used to estimate the value of the population parameter
- ❑ **Estimate**: The specific value of the estimator that is obtained in one particular sample.

- ❑ **Statistical inference**: involves using the sample to draw conclusions about the characteristics of the population. We use samples to draw inferences about a population because it is often impractical to scrutinize the entire population.

Biased sample

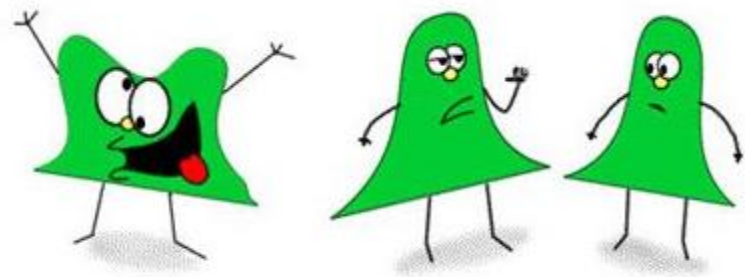


**Bad
Research**

- ❑ Any sample that differs systematically from the **population** that it is intended to represent is called a **biased sample**.
- ❑ **Selection bias** occurs when the selection of the sample systematically excludes or underrepresents certain groups
- ❑ **Self-selection bias** can occur when we examine data for a group of people who have chosen to be in that group
- ❑ **Survivors bias** occurs when we look only at the survivors
- ❑ **Nonresponse bias**: The systematic refusal of some groups to participate in an experiment or to respond to a poll
- ❑ **Random Selection** can address these biases! However, small sample bias cannot be addressed by random sampling

Sampling distribution

- ❑ It is said that the three most important factors in real estate are location, location, location. The three most important concepts in statistics are **sampling distributions, sampling distributions, sampling distributions**.
- ❑ We cannot say whether a particular sample estimate is above or below the population parameter because we **don't know** the value of the population parameter!
- ❑ The sampling distribution of a statistic, is the probability distribution or density curve that describes the all possible values of an estimator.



"KEEP YOUR EYE ON THAT GUY, TOM. HE'S NOT, YOU KNOW...NORMAL!"

Standard assumptions for the linear regression model

Assumption SLR.1

Linear in Parameters

In the population model, the dependent variable, y , is related to the independent variable, x , and the error (or disturbance), u , as

$$y = \beta_0 + \beta_1 x + u, \quad [2.47]$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

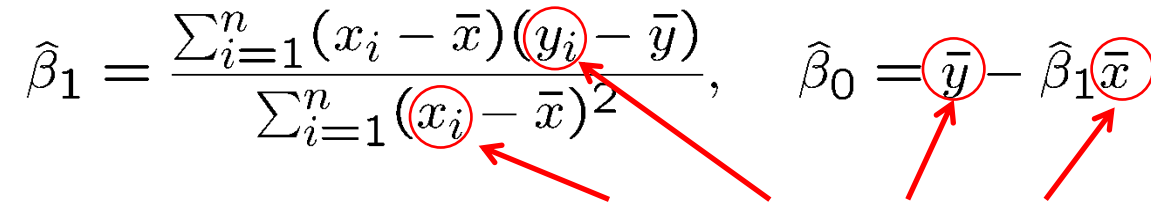
Assumption SLR.2

Random Sampling

We have a random sample of size n , $\{(x_i, y_i): i = 1, 2, \dots, n\}$, following the population model in equation (2.47).

Expected values and variances of the OLS estimators

The **estimated regression coefficients** are **random variables** themselves. Because they are calculated from a random sample

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


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.

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

Standard assumptions for the linear regression model (cont'd)

Assumption SLR.3 Sample Variation in the Explanatory Variable

The sample outcomes on x , namely, $\{x_i, i = 1, \dots, n\}$, are not all the same value.

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

← 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

The error u has an expected value of zero given any value of the explanatory variable. In other words,

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

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

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

Unbiasedness of OLS

THEOREM 2.1

UNBIASEDNESS OF OLS:

Using Assumptions SLR.1 through SLR.4,

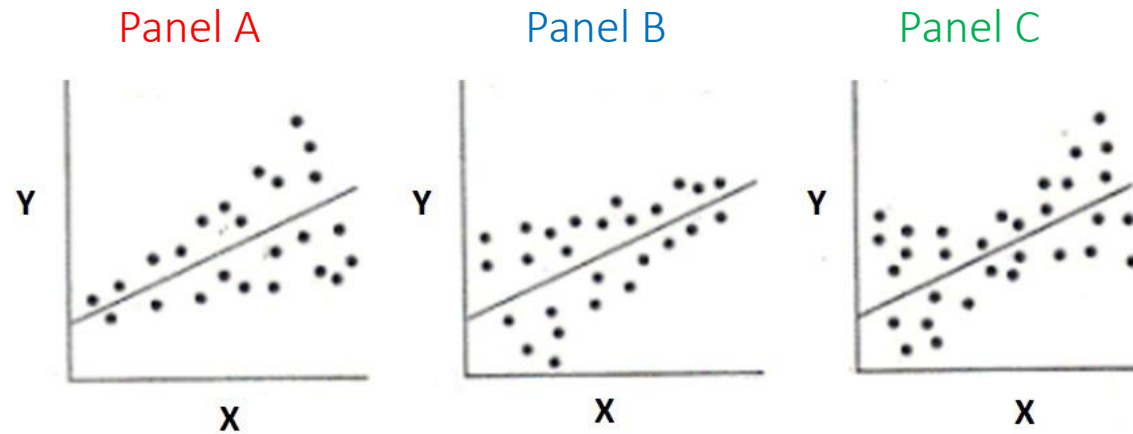
$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1,$$

for any values of β_0 and β_1 . In other words, $\hat{\beta}_0$ is unbiased for β_0 , and $\hat{\beta}_1$ is unbiased for β_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. We can **never** know for sure whether this is the case

Heteroskedasticity (Unequal variances)



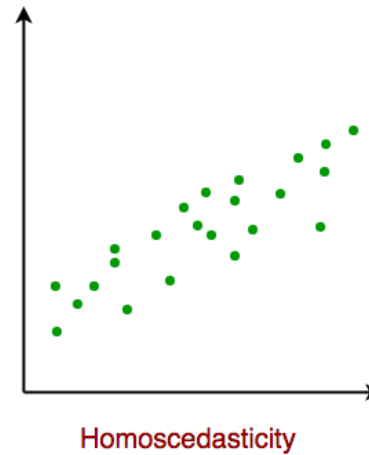
Examples:

Panel A: Income vs consumption

Panel B: Learning by doing (practice vs mistakes) or better data collection methods (GDP pre/post war)

Panel C: Outliers at both ends!

Homoskedasticity (Equal variances)



Assumption SLR.5

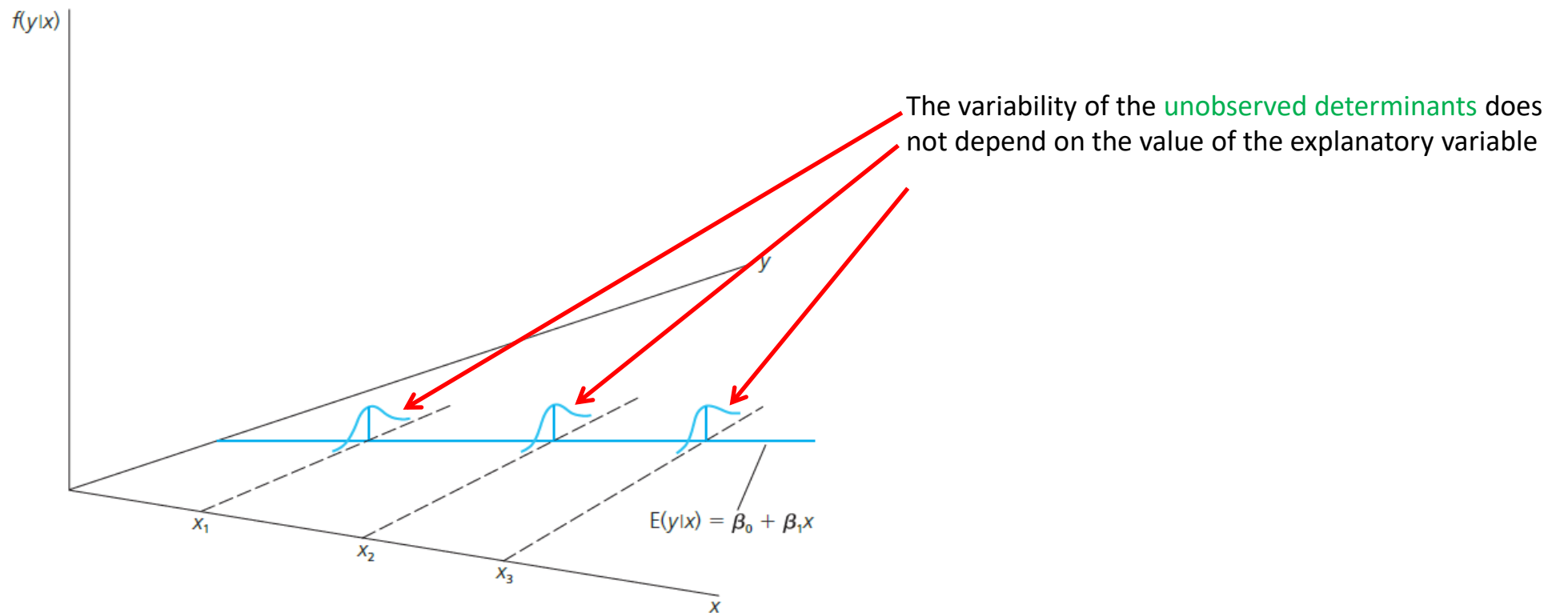
Homoskedasticity

The error u has the same variance given any value of the explanatory variable. In other words,

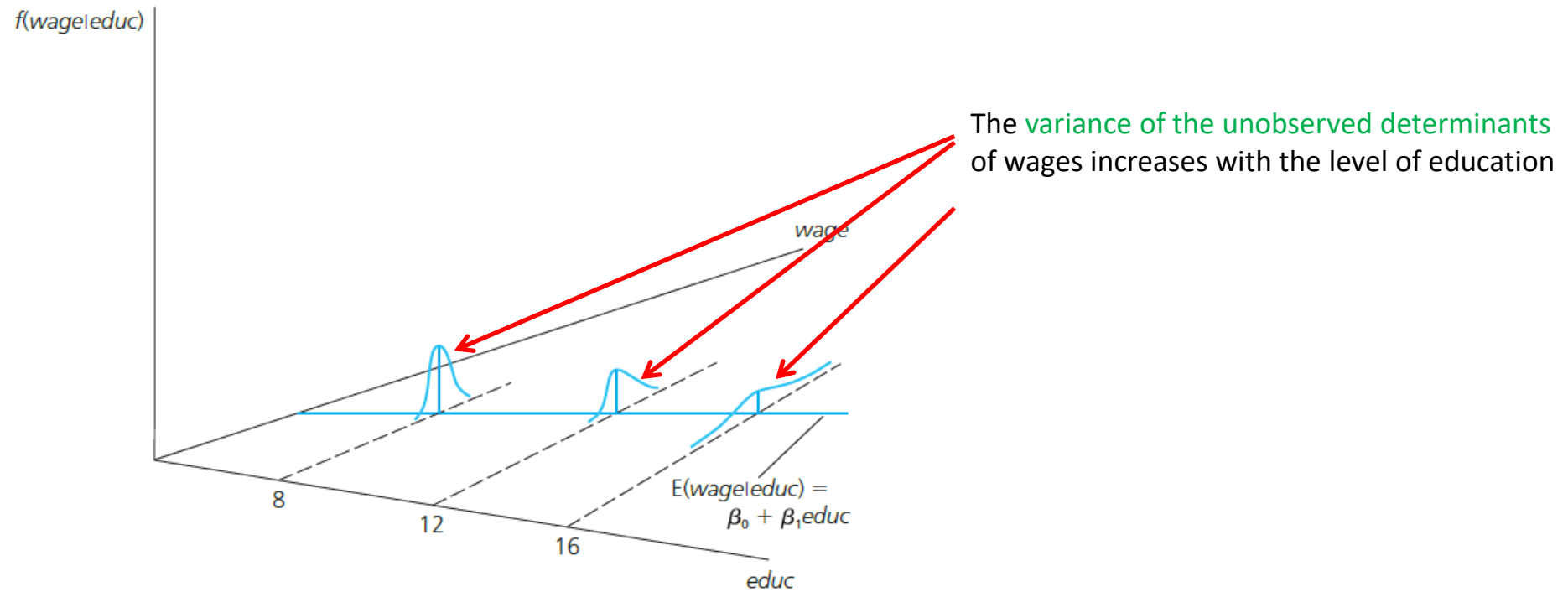
$$\text{Var}(u|x) = \sigma^2.$$

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

Graphical illustration of homoskedasticity



An example for heteroskedasticity: Wage and education



Incorporating nonlinearities

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	$\log(y)$	$\log(x)$	$\% \Delta y = \beta_1 \% \Delta x$

For successful empirical work, it is much more important to become proficient at **interpreting coefficients** than to become **efficient at computing formulas**

Incorporating nonlinearities: Log-Level model

Regression of **log wages** on years of education

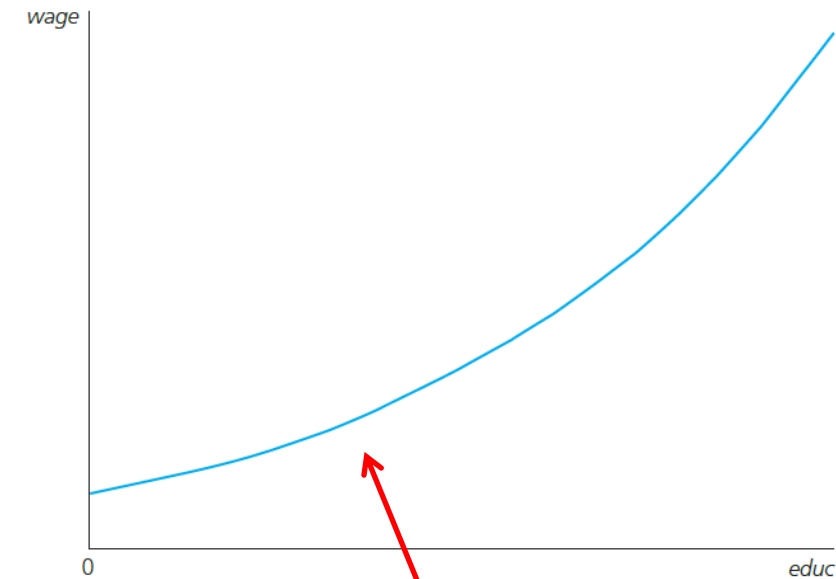
$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Fitted regression:

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

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

- $100 * \beta_1$ is called **semi-elasticity**
- Log-Level model is called **semi-elasticity model**



Growth rate of wage is 8.3%
per year of education

Incorporating nonlinearities: Log-Log model

Regression of **log salary** on **log sales**

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

Natural logarithm of CEO salary

Natural logarithm of firm's sales

Interpretation of the regression coefficient:

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

Percentage change of salary
... if sales increase **by 1%**

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

- Logarithmic changes are always percentage changes
- β_1 is the **elasticity**
- Log-Log model is called **constant elasticity model**

+ 1% sales; + 0.257% salary