



2024 Capacity Development on
IMPACT EVALUATION

APRIL 24, 2024 | 13:00–17:00




Introduction to Statistical Concepts and Econometric Methods

ALELLIE B. SOBREVÍÑAS, Ph.D.

Associate Professor/Assistant Dean for Research and Advanced Studies
School of Economics, De La Salle University Manila

OBJECTIVE

To discuss the necessary **statistical concepts** and **econometric methods** to aid the participants in **understanding impact evaluation** (i.e., experimental and quasi-experimental methods) and **in preparation** for the more in-depth discussion of these methodologies and corresponding analyses



COVERAGE

- 1. Basic Econometrics and Regression**
- 2. Interpretation of Regression Results**
- 3. Hypothesis Testing**

Basic Econometrics and Regression

What is Econometrics?

- the application of **statistical** and **mathematical** theories to economics for the purpose of testing hypotheses and forecasting future trends
- based upon the development of statistical methods for estimating **economic relationships, testing economic theories, and evaluating and implementing government and business policy** (Wooldridge, 2013)

Simple Linear Regression Model

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

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

- **Explaining y in terms of x**

Examples:

y = palay yield; x = amount of fertilizer

y = hourly wage; x = years of education

Simple linear regression model: $y = \beta_0 + \beta_1 x + \mu$

- If the other factors in μ are held fixed, so that the change in μ is zero, $\Delta\mu = 0$, then x has a **linear** effect on y :

$$\Delta y = \beta_1 \Delta x \quad \text{if } \Delta\mu = 0$$

- The change in y is simply β_1 multiplied by the change in x .
- β_1 is the **slope parameter** in the relationship between y and x , holding the other factors in μ fixed; it is of primary interest in applied economics
- β_0 is the **intercept parameter**; the *constant term*
- treats all factors affecting y other than x as being unobserved
- μ is the **error term** or **disturbance**: represents factors other than x that affect y ; standing for “unobserved”

Example: Palay yield and fertilizer

Suppose that palay yield is determined by the model

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

- The agricultural researcher is interested in the **effect of fertilizer on yield**, holding other factors fixed. This effect is given by β_1 .
 - The coefficient β_1 measures the effect of fertilizer on yield, holding other factors fixed: $\Delta yield = \beta_1 fertilizer$.
- The error term μ contains factors such as **land quality, rainfall**, and so on.

Example: A simple wage equation

A model relating a person's wage to observed education and other unobserved factors is

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

- If *wage* is measured in pesos per day and *educ* is years of education, then β_1 measures the change in daily wage given another year of education, holding all other factors fixed.
- Some of those other factors include **labor force experience, innate ability, tenure with current employer, work ethic**, and numerous other things.

Deriving the Ordinary Least Squares (OLS) Estimates

- We need a sample from the population
- Let $\{(x_i, y_i): i = 1, \dots, n\}$ denote a random sample of size n from the population

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

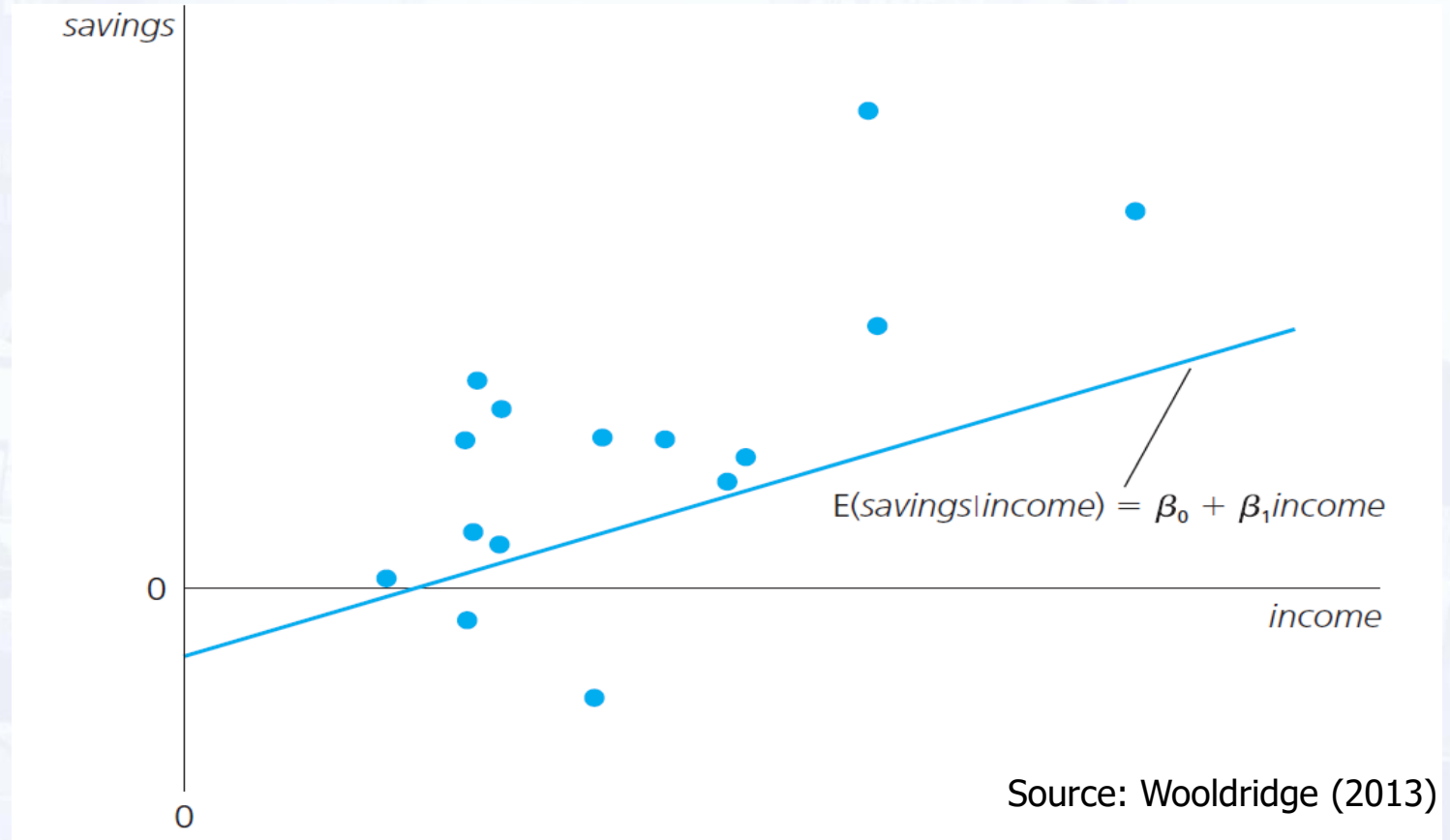
For each i ,

u_i is the error term for observation i because it contains all factors affecting y_i other than x_i

Example: Savings and Income

Scatterplot of savings and income for 15 families and the population regression

$$E(\text{savings}|\text{income}) = \beta_0 + \beta_1 \text{income}$$



Fitted value and residuals

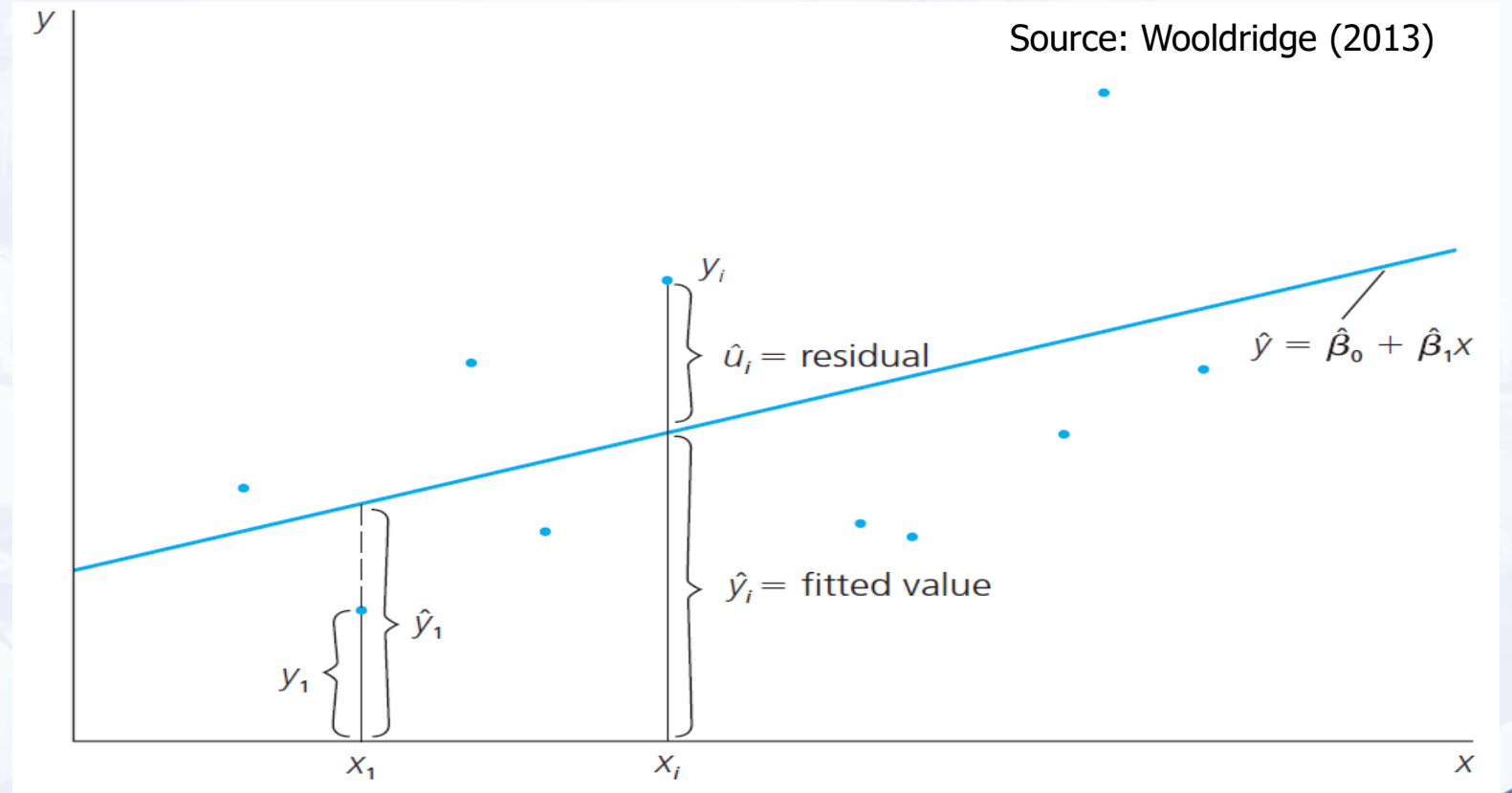
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- This is the value we **predict** for y when $x = x_i$ for the given intercept and slope.
- For any β_0 and β_1 , we define a **fitted value** for y when $x = x_i$.
- The **residual** for observation i is the difference between the actual y_i and its fitted value:

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

Graph of fitted values and residuals

- The name “**ordinary least squares**” comes from the fact that the estimates minimize the sum of squared residuals.



OLS Regression Line

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

\hat{y} = predicted values; estimates

$\hat{\beta}_0$ = predicted value of y when $x = 0$

- When using this equation to compute **predicted values** of y for various values of x , we must account for the intercept in the calculations.
- also called the **sample regression function (SRF)** because it is the estimated version of the population regression function

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

- In most cases, the **slope estimate**, which we can write as follows is of primary interest:

$$\hat{\beta}_1 = \Delta \hat{y} / \Delta x$$

- the amount by which \hat{y} changes when x increases by one unit

Equivalently,
$$\Delta \hat{y} = \hat{\beta}_1 \Delta x$$

- given any change in x (whether positive or negative), we can compute the predicted change in y

Example: Wage and Education

For the population of people in the workforce in 1976, let $y = \textit{wage}$, where *wage* is measured in dollars per hour. Thus, for a particular person, if $\textit{wage} = 6.75$, the hourly *wage* is \$6.75.

Let $x = \textit{educ}$ denote years of schooling; for example, $\textit{educ} = 12$ corresponds to a complete high school education. Since the average wage in the sample is \$5.90, the Consumer Price Index indicates that this amount is equivalent to \$19.06 in 2003 dollars.


```
use http://fmwww.bc.edu/ec-p/data/wooldridge/wage1
```

```
sum wage educ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	526	5.896103	3.693086	.53	24.98
educ	526	12.56274	2.769022	0	18

$n = 526$ individuals

Average wage in the sample= \$5.90

Average years of schooling= 13

reg wage educ

Source	SS	df	MS	Number of obs = 526		
Model	1179.73204	1	1179.73204	F(1, 524) = 103.36		
Residual	5980.68225	524	11.4135158	Prob > F = 0.0000		
Total	7160.41429	525	13.6388844	R-squared = 0.1648		
				Adj R-squared = 0.1632		
				Root MSE = 3.3784		

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

OLS regression line (or sample regression function)

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

- only **18 people** in the sample of 526 have less than eight years of education
- the regression line does poorly at very low levels of education

tab educ

educ	Freq.	Percent	Cum.
0	2	0.38	0.38
2	1	0.19	0.57
3	1	0.19	0.76
4	3	0.57	1.33
5	1	0.19	1.52
6	6	1.14	2.66
7	4	0.76	3.42
8	22	4.18	7.60
9	17	3.23	10.84
10	30	5.70	16.54
11	29	5.51	22.05
12	198	37.64	59.70
13	39	7.41	67.11
14	53	10.08	77.19
15	21	3.99	81.18
16	68	12.93	94.11
17	12	2.28	96.39
18	19	3.61	100.00
Total	526	100.00	

OLS regression line (or sample regression function)

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

- For a person with **8 years of education**, the predicted *wage* is $-0.90 + 0.54(8) = 3.42$, or \$3.42 per hour (in 1976 dollars).

$$\widehat{wage} = -0.90 + 0.54 (8) = 3.42 \text{ (in 1976 dollars)}$$

- The slope estimate implies that one more year of education increases hourly wage by **54 ¢ an hour**.

Goodness-of-Fit

***R*-squared** of the regression

- sometimes called the **coefficient of determination**
- the ratio of the explained variation compared to the total variation
- the *fraction of the sample variation in y that is explained by x*
- always between zero and one
- when interpreting, we usually multiply it by 100 to change it into a percent:
 $100 \times R^2$ is the ***percentage of the sample variation in y that is explained by x*** .

reg wage educ

Source	SS	df	MS
Model	1179.73204	1	1179.73204
Residual	5980.68225	524	11.4135158
Total	7160.41429	525	13.6388844

Number of obs = 526

F(1, 524) = 103.36

Prob > F = 0.0000

R-squared = 0.1648

Adj R-squared = 0.1632

Root MSE = 3.3784

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.5413593	.053248	10.17	0.000	.4367534	.6459651
_cons	-.9048516	.6849678	-1.32	0.187	-2.250472	.4407687

Linearity in Simple Regression

- A one-unit change in x has the same effect on y , **regardless of the initial value of x**
- May be unrealistic for many economic applications
 - Example:** Wage-education model \rightarrow increasing returns
 - the next year of education has **larger** effects on wages than did the previous year

Incorporating Nonlinearities in Simple Regression

- Probably, a better characterization of how wage changes with education is that each year of education increases wage by a constant *percentage*

Example: An increase in education from 5 years to 6 years increases wage by, say, 8% (*ceteris paribus*), and an increase in education from 11 to 12 years also increases wage by 8%.

A model that gives (approximately) a constant percentage effect is

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

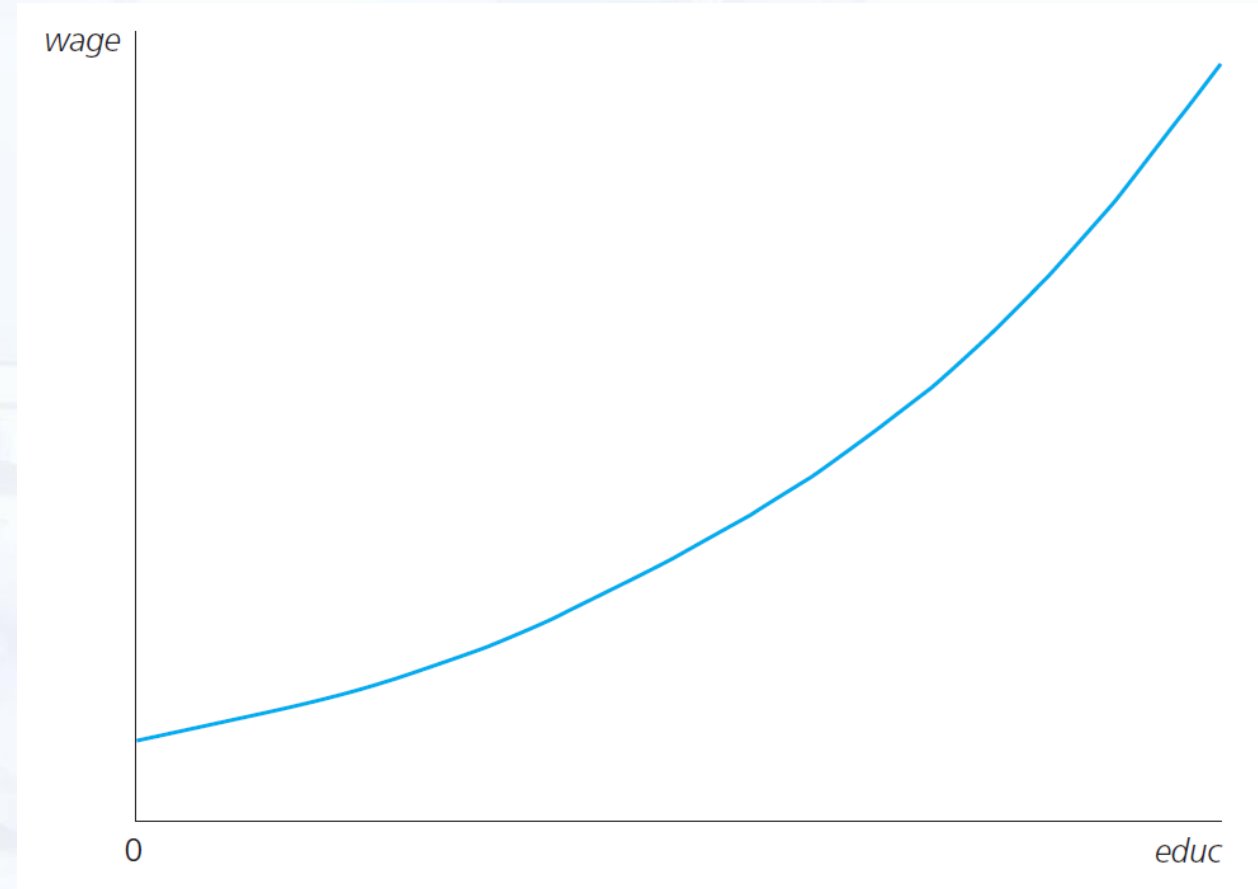
where $\log(\cdot)$ refers to the ***natural logarithm***.

- If $\Delta u = 0$, then $\% \Delta wage \approx (100 \cdot \beta_1) \Delta educ$
=> Increasing returns to education
- By exponentiating, we can write

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

Graph of

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



Source: Wooldridge (2013)

```
reg lwage educ
```

Source	SS	df	MS
Model	27.5606296	1	27.5606296
Residual	120.769132	524	.230475443
Total	148.329762	525	.28253288

```
Number of obs =      526
F( 1, 524) = 119.58
Prob > F      = 0.0000
R-squared     = 0.1858
Adj R-squared = 0.1843
Root MSE     = .48008
```

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0827444	.0075667	10.94	0.000	.0678796	.0976092
_cons	.5837726	.0973358	6.00	0.000	.3925562	.774989

- \widehat{wage} increases by **8.3%** for every additional year of education (*return to another year of education*)

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

$$n = 526$$

$$R^2 = 0.186$$

Assumptions of Simple Linear Regression

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 $y = \beta_0 + \beta_1 x + \mu$ where β_0 and β_1 are the population intercept and slope parameters, respectively.

- This equation is linear in *parameters* β_0 and β_1
- This equation is not as restrictive as it initially seems; by choosing y and x appropriately, we can obtain interesting nonlinear relationships

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 Assumption SLR.1.

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.

SLR.4: Zero Conditional Mean


The error u , has an expected value of zero given any value of the explanatory variable. In other words, $E(u|x) = 0$.

SLR.5: Homoskedasticity (“constant variance” assumption)

The error u , has the same variance given any value of the explanatory variable. In other words, $\text{Var}(u|x) = \sigma^2$.

NOTE: Only **SLR.1** through **SLR.4** are needed to show $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased**. **SLR.5** is added to obtain the usual OLS variance formulas.

Multiple Regression Analysis

- More amenable to ceteris paribus analysis because it allows us to ***explicitly control*** for many other factors that simultaneously affect the dependent variable
 - Important for testing economic theories and for evaluating policy effects when we must rely on nonexperimental data
 - **More variations** in y can be explained
 - Can be used to **build better models** for predicting the dependent variable
 - Can incorporate fairly **general functional form** relationships
 - Most widely used for **empirical analysis** in economics and other social sciences
 - The method of **ordinary least squares** is popularly used for estimating the parameters of the multiple regression model
- 

The Model with Two Independent Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

where

β_0 is the intercept.

β_1 measures the change in y with respect to x_1 , holding other factors fixed.

β_2 measures the change in y with respect to x_2 , holding other factors fixed.

Example: Wage equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

where *wage* is dollars per hour

educ is years of schooling

exper is years of labor experience

- *wage* is determined by two explanatory or independent variables (*educ* and *exper*), and by other unobserved factors, which are contained in *u*
- **Primary interest:** effect of *educ* on *wage*, holding fixed all other factors affecting *wage*; β_1
- Effectively takes *exper* out of the error term and puts it **explicitly** in the equation
- β_2 measures the **ceteris paribus effect** of *exper* on *wage*

Key Assumption in Models with Two Independent Variables

$$E(u|x_1, x_2) = 0$$

- For any any values of x_1 and x_2 in the population, the average of the **unobserved factors** is equal to zero.
- The expected value of u is the same for all combinations of x_1 and x_2 ; that this common value is zero is no assumption at all as long as the intercept β_0 is included in the model.

Example: Wage equation

$$E(u|educ, exper) = 0$$

- Other factors affecting *wage* are not related on average to *educ* and *exper*
- If we think that **innate ability** is part of u , then we will need average ability levels to be the same across all combinations of education and experience in the working population (may or may not be true)

Obtaining and Interpreting the OLS Estimates

The estimated **OLS equation** is written as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad \text{where} \quad \begin{aligned} \hat{\beta}_0 &= \text{estimate of } \beta_0 \\ \hat{\beta}_1 &= \text{estimate of } \beta_1 \\ \hat{\beta}_2 &= \text{estimate of } \beta_2 \end{aligned}$$

- To obtain $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$: The method of ordinary least squares chooses the estimates to **minimize the sum of squared residuals**.
- $\hat{\beta}_0$ = predicted value of y when $x_1 = 0$ and $x_2 = 0$
- The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have partial or ceteris paribus interpretations.

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

CHAPTER BREAK

The Model with k Independent Variables

The general **multiple linear regression model** (also called the *multiple regression model*) can be written in the population as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

where β_0 is the intercept


β_1 is the parameter associated with x_1

β_3 is the parameter associated with x_2 , and so on

- Since there are k independent variables and an intercept, this equation contains $k + 1$ (unknown) population parameters.
- The terminology for multiple regression is similar to that for simple regression.

Key Assumption for General Multiple Regression Model

$$E(u|x_1, x_2, \dots, x_k) = 0$$

- All factors in the unobserved error term should be **uncorrelated** with the explanatory variables.
 - It means that we have **correctly accounted for the functional relationships** between the explained and explanatory variables.
 - Any problem that causes u to be correlated with any of the independent variables causes this assumption to fail.
 - Implies that OLS is **unbiased**.
- 

Obtaining and Interpreting the OLS Estimates

The estimated OLS equation is written as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \quad \text{where} \quad \begin{aligned} \hat{\beta}_0 &= \text{the estimate of } \beta_0 \\ \hat{\beta}_1 &= \text{the estimate of } \beta_1 \\ \hat{\beta}_2 &= \text{the estimate of } \beta_2 \end{aligned}$$

- Written in terms of **changes**: $\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \cdots + \hat{\beta}_k \Delta x_k$
- $\hat{\beta}_1$ = measures the change in \hat{y} due to a one-unit increase in x_1 , holding all other independent variables fixed, i.e., $\Delta \hat{y} = \hat{\beta}_1 \Delta x_1$, holding x_2, x_3, \dots, x_k fixed.
- We have ***controlled for*** the variables x_2, x_3, \dots, x_k when estimating the effect of x_1 on y .

Example: Wage Equation

wage = hourly wage

lwage = natural
logarithm of wage

educ = years of
education

exper = years of
labor market
experience

tenure = years with
the current
employer

```
sum wage lwage educ exper tenure
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	526	5.896103	3.693086	.53	24.98
lwage	526	1.623268	.5315382	-.6348783	3.218076
educ	526	12.56274	2.769022	0	18
exper	526	17.01711	13.57216	1	51
tenure	526	5.104563	7.224462	0	44

```
reg lwage educ exper tenure
```

Source	SS	df	MS	Number of obs = 526		
Model	46.8741805	3	15.6247268	F(3, 522) = 80.39		
Residual	101.455581	522	.194359351	Prob > F = 0.0000		
Total	148.329762	525	.28253288	R-squared = 0.3160		
				Adj R-squared = 0.3121		
				Root MSE = .44086		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796755	.4890435

The estimated wage equation

$$\log(wage) = .284 + .092\widehat{educ} + .0041exper + .022tenure$$

- The coefficients have a **percentage interpretation**; but have ceteris paribus interpretation.
- Example of interpretation:
 - The **coefficient .092** means that, holding *exper* and *tenure* fixed, another year of education is predicted to increase $\log(wage)$ by .092, which translates into an approximate 9.2% [$100(.092)$] increase in *wage*.
 - If we take two people with the same levels of experience and job tenure, the coefficient on *educ* is the **proportionate difference in predicted wage** when their education levels differ by one year. (This measure of the return to education at least keeps two important productivity factors fixed.)

Assumptions of Multiple Linear Regression

MLR.1: Linear in Parameters


The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \dots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR.2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, 2, \dots, n\}$, following the population model in Assumption MLR.1.



MLR.3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

- If an independent variable is an **exact linear combination** of the other independent variables, then we say the model suffers from **perfect collinearity**, and it cannot be estimated by OLS

Examples: MLR.3 will fail

1. when one variable is a **constant multiple** of another
2. when a researcher inadvertently puts the **same variable measured in different units** into a regression equation (e.g., income measured in pesos and income measured in thousand pesos)
3. when one independent variable can be expressed as an **exact linear function** of two or more of the other independent variables (e.g., $x_3 = x_1 + x_2$)

MLR.4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words, $E(u|x_1, x_2, \dots, x_k) = 0$

Examples: MLR.4 will fail

1. if the functional relationship between the explained and explanatory variables is **misspecified**
 - when we use the level of a variable when the log of the variable is what actually shows up in the population model, or vice versa
Example: if the true model has $\log(wage)$ as the dependent variable but we use $wage$ as the dependent variable in our regression analysis, then the estimators will be biased
2. **Omitting** an important factor that is correlated with any of the x_1, x_2, \dots, x_k (omitted variables are less likely to be a problem in multiple regression analysis than in simple regression analysis)

- **When MLR.4 holds, we often say that we have **exogenous explanatory variables****
- If x_j is correlated with u for any reason, then x_j is said to be an **endogenous explanatory variable**.
- The term “**endogenous explanatory variable**” has evolved to cover any case in which an **explanatory variable may be correlated with the error term**.

MLR.5: Homoskedasticity (“constant variance” assumption)

The error u has the same variance given any values of the explanatory variables. In other words, $\text{Var}(u|x_1, \dots, x_k) = \sigma^2$

Example: $wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$

$$\text{Var}(u|educ, exper, tenure) = \sigma^2$$

NOTE: Under **MLR.1** through **MLR.4** , the OLS estimators are **unbiased**.
Under **MLR.1** through **MLR.5**, the OLS estimators are the **best linear unbiased** .


Unbiasedness of OLS

Under Assumptions **MLR.1** through **MLR.4**, $E(\hat{\beta}_j) = \beta_j, j = 0, 1, \dots, k$

for any values of the population parameter β_j . In other words, the OLS estimators are unbiased estimators of the population parameters.

- When we say that OLS is **unbiased** under Assumptions **MLR.1** through **MLR.4**, we mean that the **procedure** by which the OLS estimates are obtained is **unbiased** when we view the procedure as being applied across all possible random samples.
- We hope that we have obtained a sample that gives us an estimate **close to the population value**.

MLR.6: Normality

- The population error u is independent of the explanatory variables x_1, x_2, \dots, x_k and is normally distributed with zero mean and variance σ^2 : $u \sim \text{Normal}(0, \sigma^2)$
 - It is added to obtain the exact sampling distributions of t statistics and F statistics, so that we can carry out exact **hypotheses tests**
 - It can be dropped if we have a reasonably **large sample size**
 - It implies a stronger **efficiency** property of OLS: the OLS estimators have **smallest variance** among *a//* unbiased estimators
 - It is much stronger than any of our previous assumptions
- 

- Since u is independent of the x_j under MLR.6, $E(u | x_1, \dots, x_k) = 0$, $E(u) = 0$ and $\text{Var}(u | x_1, \dots, x_k) = \text{Var}(u) = \sigma^2$

- For cross-sectional regression applications, Assumptions MLR.1 through MLR.6 are called the **classical linear model (CLM) assumptions**

$y | \mathbf{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \sigma^2)$, where \mathbf{x} is shorthand for (x_1, \dots, x_k)

- Normality of the error term translates into **normal sampling distributions** of the OLS estimators.
- Under Assumption MLR.6 (and the random sampling Assumption MLR.2), the errors are **independent, identically distributed** $\text{Normal}(0, \sigma^2)$ random variables.
- The normality of the OLS estimators is still *approximately* true in **large samples** even without normality of the errors.

Testing Hypotheses about a Single Population Parameter: The t Test

Population model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

t Distribution for the Standardized Estimators

Under the CLM Assumptions MLR.1 through MLR.6,

$$(\hat{\beta}_j - \beta_j) / \text{se}(\hat{\beta}_j) \sim t_{n-k-1} = t_{df}$$

where $k + 1$ is the number of unknown parameters in the population model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ (k slope parameters and the intercept β_0) and $n-k-1$ is the degrees of freedom (df)

- It allows us to **test hypotheses** involving the β_j .
- In most applications, our primary interest lies in testing the **null hypothesis**.

$$H_0: \beta_j = 0 \quad \text{where } j \text{ corresponds to any of the } k \text{ independent variables}$$

Example: Wage Equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

$H_0: \beta_2 = 0 \Rightarrow$ Once education and tenure have been accounted for, the number of years of labor market experience has no effect on hourly wage

The statistic we use to test H_0 (against any alternative) is called “the” ***t* statistic** or “the” ***t* ratio** of $\hat{\beta}_j$ and is defined as

$$t_{\hat{\beta}_j} \equiv \hat{\beta}_j / \text{se}(\hat{\beta}_j)$$

Note: We are testing hypotheses about the ***population parameters***. We are *not* testing hypotheses about the estimates from a particular sample.

Example: Wage Equation

wage = hourly wage

lwage = natural logarithm of wage

educ = years of education

exper = years of labor market experience

tenure = years with the current employer


```
reg lwage educ exper tenure
```

Source	SS	df	MS
Model	46.8741805	3	15.6247268
Residual	101.455581	522	.194359351
Total	148.329762	525	.28253288

Number of obs = 526
F(3, 522) = 80.39
Prob > F = 0.0000
R-squared = 0.3160
Adj R-squared = 0.3121
Root MSE = .44086

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796755	.4890435

Test of Hypotheses

1. Specification of the **null hypothesis**, H_0
 2. Specification of the **alternative hypothesis**, H_1
 3. Specification of the **test statistic** and its **distribution** under the null hypothesis
 4. Selection of the **significance level α** in order to determine the rejection region
 5. Calculation of the **test statistic** from the data sample
 6. **Conclusions**, which are based on the test statistic and the **rejection region**
- 

Testing Against One-sided Alternatives

1. **Null hypothesis** $H_0: \beta_j = 0$
2. **Alternative hypothesis: One-sided alternative** $H_1: \beta_j > 0$
3. t -distribution with $n-k-1$ degrees of freedom
4. **Significance Level** (the probability of rejecting H_0 when it is in fact true)
Suppose that we choose **5% significance level**: We are willing to mistakenly reject H_0 when it is true 5% of the time

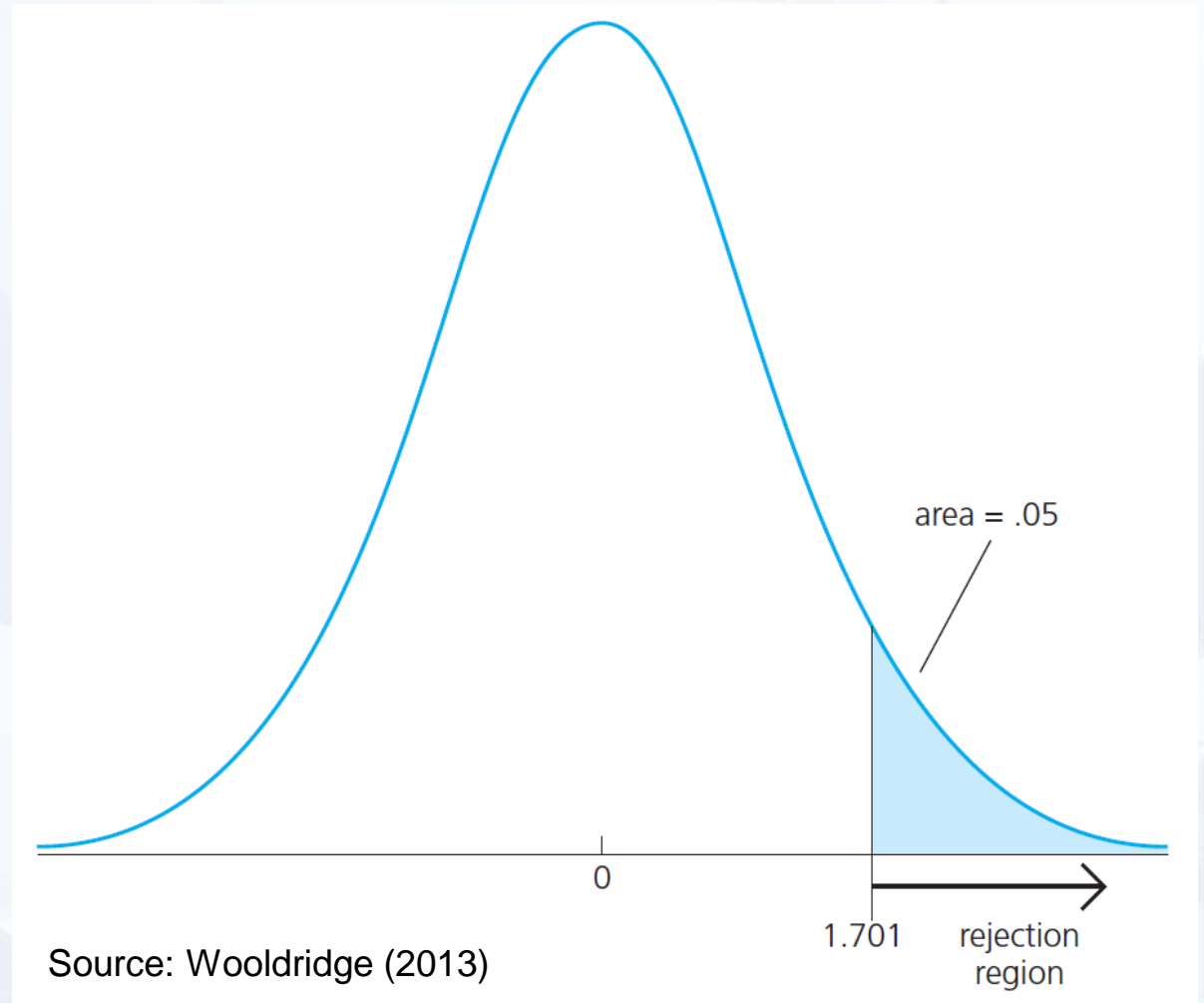
5. Calculation of the t -statistic $t_{\hat{\beta}_j}$

6. Conclusion based on the **rejection rule**: H_0 is rejected in favor of H_1 at the 5% significance level if $t_{\hat{\beta}_j} > c$ where c is the **critical value**

To obtain c , we only need the significance level and the degrees of freedom.

Example: For a 5% level test and with $n - k - 1 = 28$ degrees of freedom, the critical value is $c = 1.701$. If $t_{\hat{\beta}_j} \leq 1.701$, then we fail to reject H_0 at the 5% level

**5% rejection rule for the
alternative $H_1: \beta_j > 0$ with 28 df**



Example: Wage Equation

wage = hourly wage

lwage = natural
logarithm of wage

educ = years of
education

exper = years of
labor market
experience

tenure = years with
the current
employer

```
reg lwage educ exper tenure
```

Source	SS	df	MS
Model	46.8741805	3	15.6247268
Residual	101.455581	522	.194359351
Total	148.329762	525	.28253288

Number of obs = 526
F(3, 522) = 80.39
Prob > F = 0.0000
R-squared = 0.3160
Adj R-squared = 0.3121
Root MSE = .44086

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796755	.4890435

$n = 526$

$R^2 = 0.316$

$$\log(\widehat{wage}) = 0.284 + 0.092educ + 0.0041exper + 0.022tenure$$

(0.104) (0.007) (0.0017) (0.003)

$$\log(\widehat{wage}) = 0.284 + 0.092educ + 0.0041exper + 0.022tenure$$

(0.104) (0.007) (0.0017) (0.003)

Null hypothesis: $H_0: \beta_{exper} = 0$ **Alternative hypothesis:** $H_0: \beta_{exper} > 0$

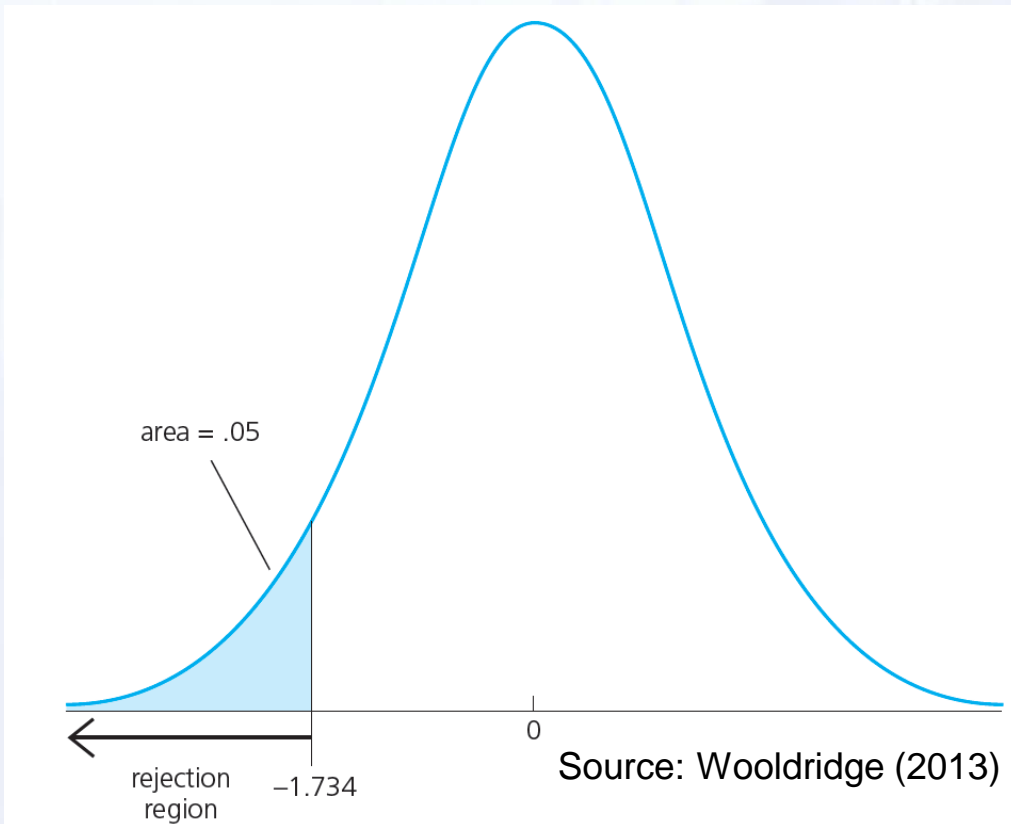
Degrees of freedom: 522

- We can use the **standard normal** critical values. The 5% critical value is 1.645, and the 1% critical value is 2.326

t-statistic for $\hat{\beta}_{exper}$ is $t_{exper} = 0.0041/0.0017 \approx 2.41$

- $\hat{\beta}_{exper}$, or *exper*, is statistically significant even at 1% level
- $\hat{\beta}_{exper}$ is statistically greater than 0 at 1% significance level

5% rejection rule for the alternative $H_1: \beta_j < 0$ with 18 *df*



$$H_0: \beta_j = 0$$

$$H_1: \beta_j < 0$$

Rejection rule: $t_{\hat{\beta}_j} < -c$

Example: If the significance level is 5% and the **degrees of freedom** is 18, then $c = 1.734$

$H_0: \beta_j = 0$ is rejected in favor of $H_1: \beta_j < 0$ at the 5% level if $\beta_j < -1.734$.

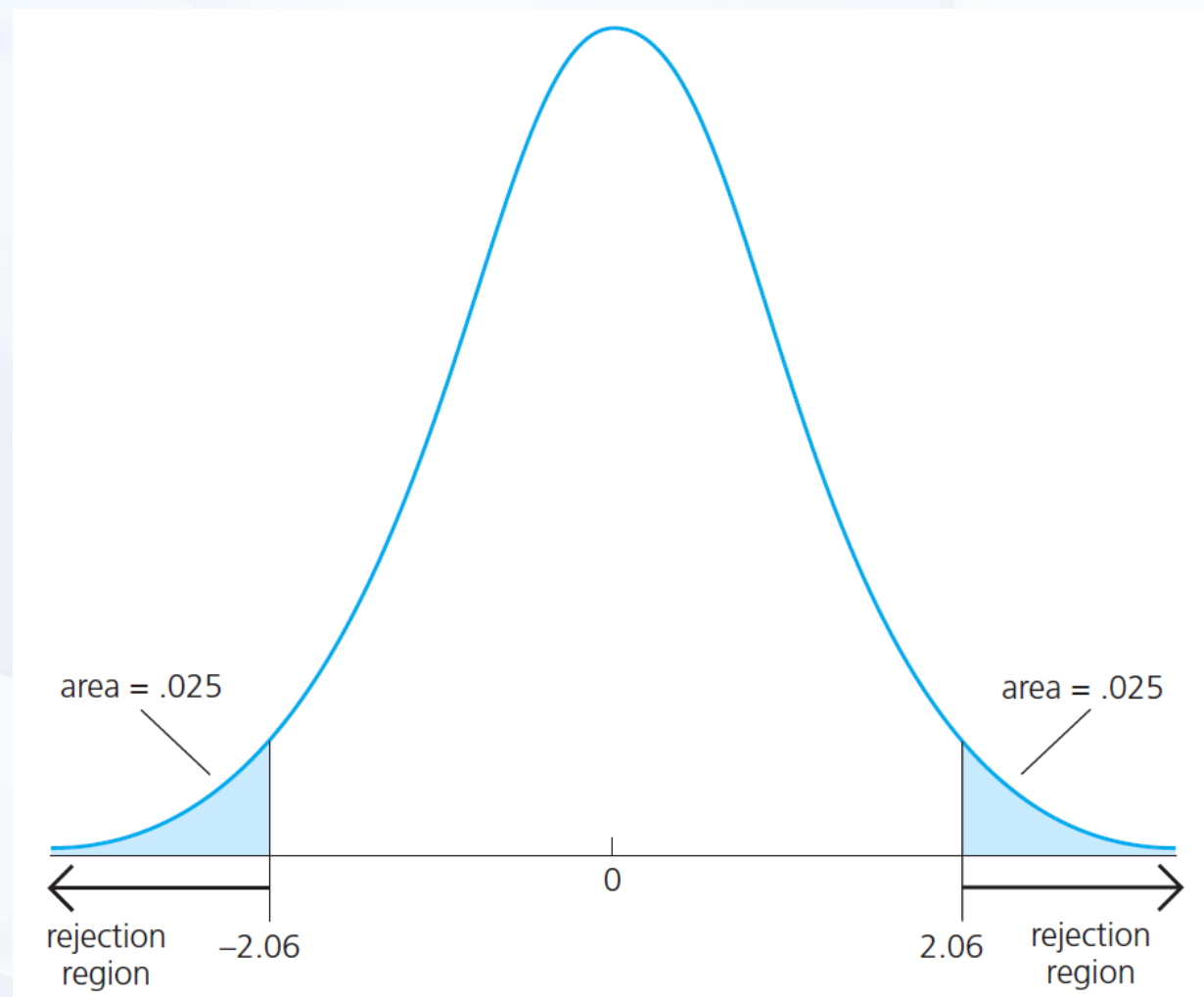
Two-Sided Alternatives

Null hypothesis $H_0: \beta_j = 0$ **Alternative hypothesis** $H_1: \beta_j \neq 0$

- We are interested in the absolute value of the t -statistic
- Rejection rule: $\left| t_{\hat{\beta}_j} \right| > c$
 - c is chosen to make the area in each tail of the t distribution equal 2.5%
 - c is the 97.5th percentile in the t distribution with $n - k - 1$ degrees of freedom

Example: When $n - k - 1 = 25$, the 5% critical value for a two-sided test is $c = 2.060$

**5% rejection rule for the
alternative $H_1: \beta_j \neq 0$ with 25 *df***



Source: Wooldridge (2013)

The ***p***-value

- The ***p***-value is defined as the probability, under the null hypothesis, of obtaining a result, which is equal to, or more extreme, than what was actually observed.
- the ***smallest*** significance level at which the null hypothesis would be rejected
- **small *p*-values** are evidence *against* the null; **large *p*-values** provide little evidence against H_0
- Having the *p*-value allows us to easier determine the outcome of the test, as we do not need to directly compare the critical values.
 - If $p \leq \alpha$, we **reject** H_0 .
 - If $p > \alpha$, we **fail to reject** H_0 .

Example: Wage Equation

wage = hourly wage

lwage = natural logarithm of wage

educ = years of education

exper = years of labor market experience

tenure = years with the current employer

```
reg lwage educ exper tenure
```

Source	SS	df	MS
Model	46.8741805	3	15.6247268
Residual	101.455581	522	.194359351
Total	148.329762	525	.28253288

Number of obs = 526
F(3, 522) = 80.39
Prob > F = 0.0000
R-squared = 0.3160
Adj R-squared = 0.3121
Root MSE = .44086

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796755	.4890435

Confidence Intervals (CI)

- **Interval estimates**; they provide a range of likely values for population parameter and not just a point estimate

- 95% CI for the unknown β_j : $\hat{\beta}_j \pm c \cdot se(\hat{\beta}_j)$

where the constant c is the 97.5th percentile in a t_{n-k-1} distribution

- lower bound: $\hat{\beta}_j - c \cdot se(\hat{\beta}_j)$
- upper bound: $\hat{\beta}_j + c \cdot se(\hat{\beta}_j)$

- If random samples were obtained over and over again, with $\underline{\beta}_j$ and $\overline{\beta}_j$ computed each time, then the (unknown) population value β_j would lie in the interval $(\underline{\beta}_j, \overline{\beta}_j)$ for 95% of the samples

The F statistic for Overall Significance of Regression

- In the model with k independent variables, we can write the null hypothesis as
 $H_0: x_1, x_2, \dots, x_k$ do not help to explain y $H_0: E(y|x_1, x_2, \dots, x_k) = E(y)$

- In terms of the parameters, the null is that all slope parameters are zero:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad H_1: \text{At least one of the } \beta_j \text{ is different from } 0$$

- the F statistic for testing H_0 can be written as (special form)

$$\frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

where R^2 is just the usual R -squared from the regression of y on x_1, x_2, \dots, x_k .

- Under H_0 (and assuming CLM assumptions hold), F is distributed as an **F random variable** with $(q, n-k-1)$ degrees of freedom

$$F \sim F_{q, n-k-1}$$

- The distribution of $F_{q, n-k-1}$ is readily tabulated and available in **statistical tables**; q is the number of variables excluded in the restricted model.
- Reject H_0 in favor of H_1 when F is **sufficiently large**. The critical value c depends on the chosen significance level, q and $n-k-1$.
- Once c has been obtained, we reject H_0 in favor of H_1 at the chosen significance level if $F > c$.

Example: Wage Equation

wage = hourly wage

lwage = natural logarithm of wage

educ = years of education

exper = years of labor market experience

tenure = years with the current employer

```
reg lwage educ exper tenure
```

Source	SS	df	MS
Model	46.8741805	3	15.6247268
Residual	101.455581	522	.194359351
Total	148.329762	525	.28253288

Number of obs = 526

F(3, 522) = 80.39

Prob > F = 0.0000

R-squared = 0.3160

Adj R-squared = 0.3121

Root MSE = .44086

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796755	.4890435

Regression Diagnostics

1. Zero Conditional Mean assumption

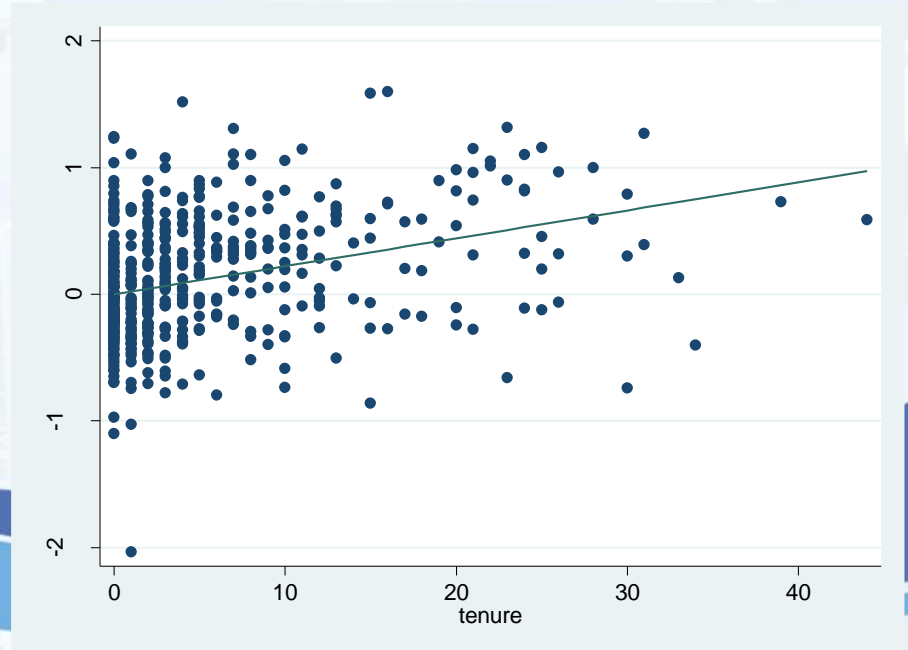
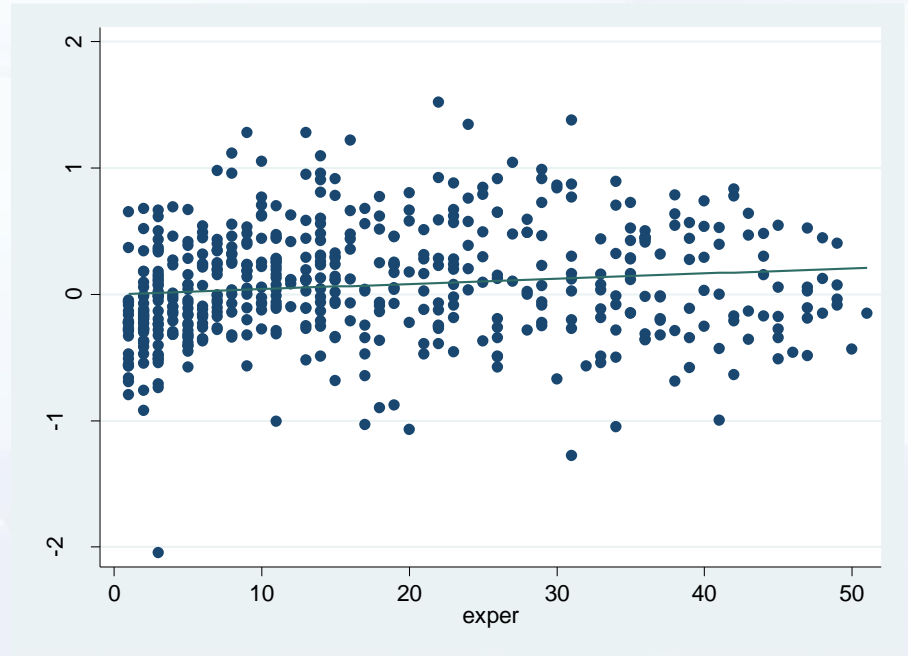
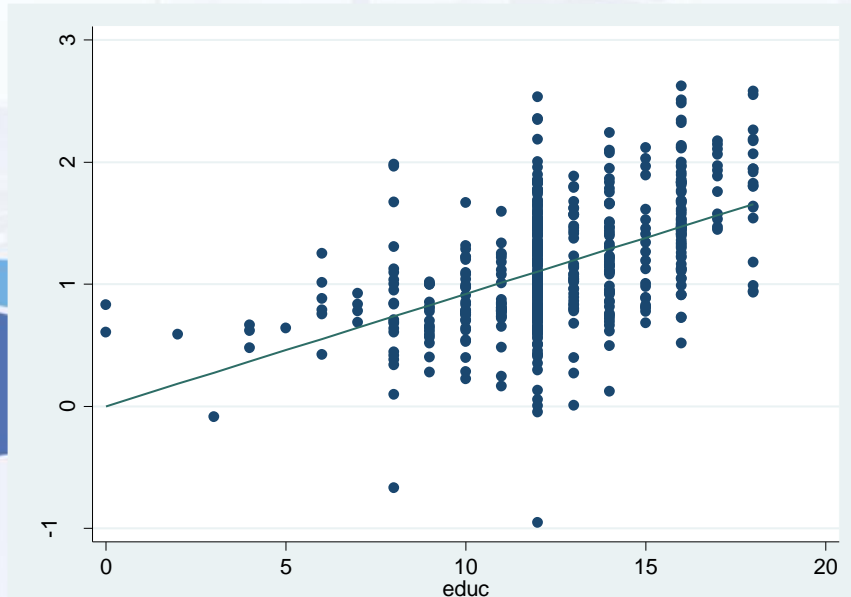
This assumption will be violated is

- a. The relationship between the dependent and independent variables is **nonlinear**;
- b. Some **outliers** have a strong effect on the estimated regression coefficients; or
- c. Some **influential factors** have been omitted that are in fact correlated with the included independent variables.

a. Linearity

Component-plus-residual plots (partial residual plots)

- Allow the determination of the **functional form** of the relationship
- The product of the **residual** and the **linear part of the independent variable** are plotted against the other independent variables



b. Outliers/Influential cases

- Observations that **heavily influence** the results of the regression model
- Mostly, these are observations that have **unusual combinations** of the regression variables included in the model

Example: A person with a huge income living in a very small home

A person with a very low level of education but earning a very high wage/salary

- Formal way to discover influential cases: **DFBETAs**
 1. Fit a **regression model** and fit it again with one observation deleted.
 2. **Compare** the two results. If there is a big difference in the estimated coefficients, the observation that was excluded in the second computation has a big influence on the coefficient estimates.
 3. **Repeat** the above technique for each observation to determine its influence on the estimated regression coefficients. Compute this for each of the k regression coefficients separately.

The equation for computing the **influence of the i th case** on the estimation of the k th regression coefficient is

$$\text{DFBETA}_{Ik} = \frac{b_k - b_{k(i)}}{s_{e(i)} / \sqrt{RSS_k}}$$

where b_k is the estimated coefficient of variable k

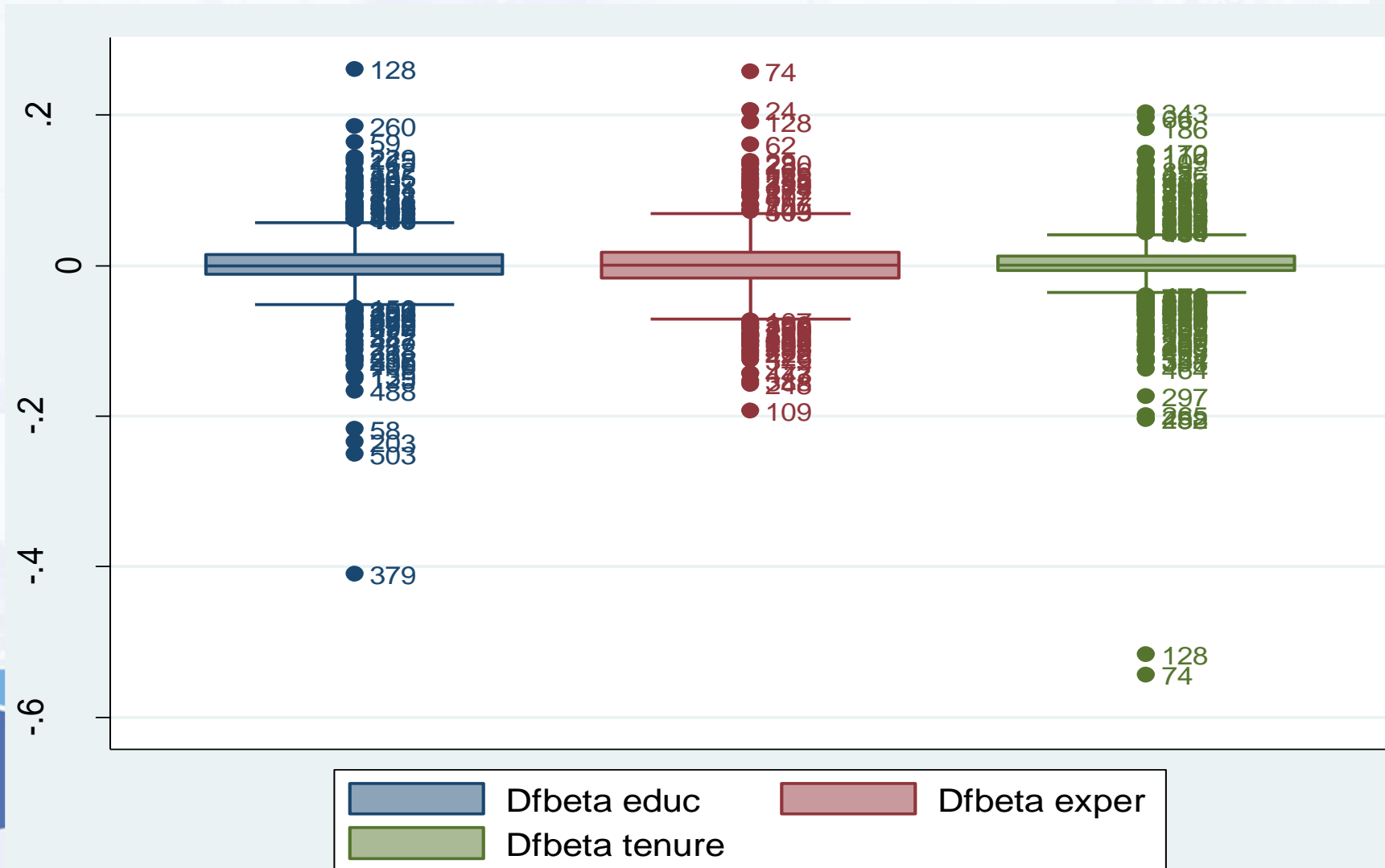
$b_{k(i)}$ is the corresponding coefficient without observation i

$s_{e(i)}$ is the standard deviation of the residuals without observation i

Note: The ratio in the denominator standardizes the difference so that the influence on the standard coefficients are comparable.

dfbeta

```
graph box _dfbeta*, marker (1,mlab(number)) marker(2, mlab(number)) marker(3, mlab(number))
```




Potential solutions for influential observations

1. If an influential case can be attributed unquestionably to a **measurement error**, you should either **correct the error** or **delete the observation** from the file.
2. If the influential observation results from **extreme values** of the dependent variable, it is reasonable to use **median regression**.
3. When observations with a high value for the independent variable influence the regression extraordinarily, you should ask if another **factor that is typically related to high (or low) value** for that variable influences the dependent variable.
 - **Example:** With right-skewed distribution of HH income, you may want to change the model to use the logarithm of HH income instead of HH income itself.

Note: If you decide to drop the highly influential observations from your dataset, run the regression again. Report both results (i.e., the one with and the one without the highly influential cases.)

c. Omitted variables (Omitted factors)

- They influence the dependent variable and are at the same time correlated with one or more of the independent variables of the regression model.
 - Strictly speaker, the following are also omitted factors
 - **Nonlinear relationships**- you may have overlooked the fact that an independent variable does not have the same influence on the dependent variable throughout the range of the dependent variable
 - **Influential cases** – you may have neglected to model your theory adequately or overlooked a mechanism that would explain the outliers.
 - The problem of **excluding a relevant variable** or **underspecifying the model**; Generally causes the OLS estimators to be **biased**
 - To know which variables have been omitted
 - Begin by graphing the residuals against all variables that are not included in the model – but this is possible only for those variables that are included in the data file.
- 

Multicollinearity

- If there is perfect linear relationship between two variables of the regression models, one of them will be excluded when calculating the model.
- Even if the two variables are not perfect linear combinations of each other, some problems can arise:
 - the **standard errors** of the estimated coefficients might increase
 - there might be an unexpected change in the size of the estimated coefficients or their signs
- If the model fits the data well (based on R^2) but nevertheless has a **few significant estimated coefficients**, then multicollinearity may be a problem.
- **Solving the multicollinearity problem:** dropping other independent variables from the model in an effort to reduce multicollinearity. Unfortunately, dropping a variable that belongs in the population model can lead to **bias**.

estat vif

Variable	VIF	1/VIF
exper	1.48	0.676765
tenure	1.35	0.741127
educ	1.11	0.898658
Mean VIF	1.31	

variance inflation factor (VIF) = the factor by which $\text{Var}(\hat{\beta}_j)$ is higher because x_j is not uncorrelated with the other explanatory variables

- If we had a choice we would like VIF_j to be smaller (other things equal)
- setting a **cutoff value for VIF** above which we conclude multicollinearity is a “problem” is arbitrary and not especially helpful. Sometimes the value **10** is chosen: if VIF_j is above 10 then we conclude that multicollinearity is a “problem” for estimating β_j

BUT: looking at the size of VIF_j is of limited use.

2. Homoskedasticity (“constant variance” assumption)

- Requires that the variance of the errors be the **same** for all values of the independent variables.
- Its violation is heteroskedasticity leading to the following:
 - **Inefficient estimate**- there is an increasing probability that a particular estimated regression coefficient **deviates from the true value** for the population
 - **Incorrect standard errors** of the coefficients and has an impact on any statistical inference


```
reg lwage educ exper tenure
```

```
. estat hettest
```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of lwage

chi2(1) = 7.62

Prob > chi2 = 0.0058

Possible solutions for heteroskedasticity

1. **Transforming** the dependent variable
2. If transforming the dependent variable does not remove the heteroskedasticity in the regression model, you may estimate the **robust standard errors** => the standard errors are computed so that homoskedasticity of the error terms need not be assumed.

```
. reg lwage educ exper tenure, vce(robust)
```

Linear regression

Number of obs = 526
F(3, 522) = 67.76
Prob > F = 0.0000
R-squared = 0.3160
Root MSE = .44086

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.092029	.0079212	11.62	0.000	.0764676	.1075903
exper	.0041211	.0017459	2.36	0.019	.0006913	.0075509
tenure	.0220672	.003782	5.83	0.000	.0146374	.0294971
_cons	.2843595	.1117069	2.55	0.011	.0649092	.5038098

Model Extensions and Other Issues

Categorical independent variable

```
reg lwage educ exper tenure female
```

Source	SS	df	MS	Number of obs = 526		
Model	58.1853046	4	14.5463261	F(4, 521) = 84.07		
Residual	90.1444572	521	.173021991	Prob > F = 0.0000		
Total	148.329762	525	.28253288	R-squared = 0.3923		
				Adj R-squared = 0.3876		
				Root MSE = .41596		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0874623	.0069389	12.60	0.000	.0738307	.101094
exper	.0046294	.0016271	2.85	0.005	.0014328	.007826
tenure	.017367	.0029762	5.84	0.000	.0115201	.0232138
female	-.3011459	.0372456	-8.09	0.000	-.3743158	-.2279759
_cons	.5013479	.1019024	4.92	0.000	.3011579	.701538


Including Irrelevant Variables in a Regression Model (overspecifying a model)

- One (or more) of the independent variables is included in the model even though it has **no partial effect** on y in the population.
- Suppose we specify the model as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ and this model satisfies Assumptions MLR.1 through MLR.4. However, x_3 has no effect on y after x_1 and x_2 have been controlled for, which means that $\beta_3 = 0$. Because we do not know that $\beta_3 = 0$, we are inclined to estimate the equation including x_3 .

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

What is the effect of including x_3 in when its coefficient in the population model is zero?
No effect in terms of unbiasedness of $\hat{\beta}_1$ and $\hat{\beta}_2$, but can have undesirable effects on the **variances** of the OLS estimators.

Concluding Remarks

- Econometrics as a powerful tool for analyzing economic data and in making informed policy decisions
 - can provide empirical evidence to support economic theories and to inform policymaking
 - Goals of any econometric analysis
 - to estimate the parameters in the model
 - to test hypotheses about these parameters (the values and signs of the parameters determine the validity of an economic theory)
 - Need to understand clearly the assumptions in estimating the models
 - Constantly evolving and adapting to new challenges in economic research.
- 

Main References

Kohler, U. and Kreuter, F. (2012). *Data Analysis Using Stata*, Third Edition, Stata Press, Texas.

Wooldridge, J.M. (2013) *Introductory Econometrics—A Modern Approach*. Cengage Learning, Boston, MA.

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Massachusetts



Introduction to Statistical Concepts and Econometric Methods

ALELLIE B. SOBREVINAS, Ph.D.

Associate Professor/Assistant Dean for Research and Advanced Studies
School of Economics, De La Salle University Manila

alellie.sobrevinas@dlsu.edu.ph