

# EE 325 Introductory Econometrics

By Wanwiphang Manachotphong<sup>1</sup>

Semester 2, 2019

<sup>1</sup>Email: wanwiphang@econ.tu.ac.th; Office: 525 Faculty of Economics, Tha Prachan; Office hours: Tuesdays, 11.00 – 12 noon (or by appointment)



# Contents

<b>1</b>	<b>Introduction and Motivation</b>	<b>3</b>
1	What is econometrics? . . . . .	3
1.1	Steps in Empirical Economics Analysis . . . . .	5
2	Types of economic data . . . . .	6
<b>2</b>	<b>Review of Some Statistical Concepts</b>	<b>7</b>
1	Random variables and distributions . . . . .	7
1.1	Types of Random Variables . . . . .	8
1.2	Examples of Some Distributions . . . . .	9
2	Joint Distributions, Conditional Distributions and Independence . . . . .	10
2.1	Joint Distributions . . . . .	10
2.2	Conditional Distribution and Marginal Distribution . . . . .	11
2.3	Independence . . . . .	14
3	Expectation, variance, covariance and correlation . . . . .	15
3.1	Expected value of a Random Variable – $E(X)$ or $\mu_X$ . . . . .	15
3.2	Properties of Expected Values . . . . .	16
3.3	Conditional and Marginal Expectations . . . . .	17
3.4	Variance and Standard Deviation of a Random Variable . . . . .	18
3.5	Properties of Variances . . . . .	19
3.6	Properties of Standard Deviations . . . . .	19
3.7	Covariance and Correlation of Two Variables . . . . .	21
4	Estimators and desirable properties of estimators . . . . .	23
4.1	Desirable properties of estimators . . . . .	24
<b>3</b>	<b>The Simple Regression Model</b>	<b>27</b>
1	Principle, assumptions and derivation of ordinary least squares (OLS) estimators . . . . .	29
1.1	Terminology for the Linear Regression . . . . .	29
1.2	Derivation of Ordinary Least Squares (OLS) Estimators . . . . .	30
2	Properties of OLS estimators . . . . .	33
2.1	Algebraic Properties . . . . .	33
2.2	Properties proving BLUE . . . . .	34
2.3	Assumptions on the simple linear regression (SLR) model . . . . .	34
2.4	Homoskedasticity VS. Heteroskedasticity . . . . .	35
2.5	Expectation of Estimators . . . . .	36
2.6	Variance of OLS Estimators . . . . .	37
2.7	Some Concepts to be emphasized (Population vs. Sample; PRF vs. SRF; error vs. residual) . . . . .	38
2.8	Goodness of Fit ( $R^2$ ) . . . . .	41
2.9	Incorporating Nonlinearities in Sample Regression . . . . .	42
3	Regression Through the Origin . . . . .	43

<b>4</b>	<b>Multiple Regression Analysis (Estimation)</b>	<b>45</b>
1	Motivation . . . . .	45
1.1	Assumption SLR 4 ( $E(u X) = 0$ ) in the Multiple Regression Context	46
2	The Model with k Independent Variables . . . . .	47
2.1	Assumptions on the multiple linear regression (MLR) model . . . .	48
2.2	Accounting Nonlinearity . . . . .	50
3	Estimation of parameters and properties of estimators . . . . .	51
3.1	Deriving OLS Estimators . . . . .	51
3.2	How could the multiple regression analysis enable ceteris paribus analysis? . . . . .	52
4	Expected Value of the OLS Estimators . . . . .	53
4.1	Issue #1: Including Irrelevant Variable (Overspecifying the Model)	53
4.2	Issue #2: Excluding Relevant Variable (Underspecifying the Model → omitted variable bias. This is a serious problem!) . . . . .	54
5	Variance of the OLS Estimators . . . . .	55
6	Estimator of the OLS Variance . . . . .	56
<b>5</b>	<b>Stata Lab 1 – Introduction</b>	<b>57</b>
1	What is STATA? . . . . .	57
1.1	STATA supports . . . . .	57
1.2	Data files and Do-files . . . . .	57
2	Tutorial 1: Exploring the Data and Running a Simple Regression . . . .	58
2.1	Examples from Wooldridge(2009) . . . . .	60
<b>6</b>	<b>Multiple Regression Analysis (Inference)</b>	<b>61</b>
1	Sampling Distribution of the OLS estimators ( $\hat{\beta}_{OLS}$ ) . . . . .	62
2	Testing Hypotheses about an individual regression coefficient "the t-test"	63
2.1	Testing Against Two-Sided Alternatives . . . . .	63
2.2	Testing Against One-Sided Alternatives . . . . .	65
3	Testing other hypotheses about $\beta_j$ . . . . .	66
4	Testing Hypotheses about a Single Linear Combination of the Parameter	67
5	Computing p-Values for t-Tests . . . . .	68
6	Confidence Intervals (CI) . . . . .	69
7	Testing Multiple Linear Restrictions: The F-test . . . . .	71
8	How the Hypothesis Testing is done in Practice . . . . .	74
<b>7</b>	<b>Multiple Regression Analysis : Further Issues</b>	<b>75</b>
1	Data scaling on OLS statistics . . . . .	75
2	More on functional forms . . . . .	76
3	Models with Interaction Terms . . . . .	78
4	More on the Goodness-of-Fit and Selection of Regressors . . . . .	79
<b>8</b>	<b>Multiple Regression Analysis with Qualitative Information:</b>	<b>81</b>
1	Outline . . . . .	81
2	Describing Qualitative Information . . . . .	81
3	Models with a single dummy independent variable . . . . .	82
4	It is not possible to include all of the dummy alternatives in the same model	83
5	Using dummy variables for multiple categories . . . . .	84

6	Interactions involving dummy variables . . . . .	87
7	Testing for Differences in Regression Functions across Groups . . . . .	89
	7.1 We can use the "Chow statistics" to test this type of hypothesis as well . . . . .	90
8	A Binary Dependent Variable (y variable): The Linear Probability Model . . . . .	92
<b>9</b>	<b>Heteroscedasticity Problem</b>	<b>95</b>
1	Nature and Consequences of heteroscedasticity for OLS . . . . .	95
	1.1 Nature of Heteroskedasticity . . . . .	95
	1.2 Consequences of Heteroskedasticity . . . . .	95
	1.3 How can the estimated value of $Var(\hat{\beta}_{OLS})$ be wrong? . . . . .	96
	1.4 Two types of remedies . . . . .	97
2	Testing for heteroskedasticity . . . . .	98
	2.1 Breusch-Pagan test (BP test) . . . . .	99
	2.2 The White Test . . . . .	102
3	Remedial measures . . . . .	103
	3.1 Weighted Least Squares (WLS) . . . . .	103
	3.2 Feasible GLS . . . . .	104
	3.3 What if the assumed $h_i$ function is wrong? . . . . .	105
<b>10</b>	<b>Lab 2 – Dummy, Heteroskedasticity, Specification Issues</b>	<b>107</b>
1	Does "beauty" help increase wage? . . . . .	107
2	Fixing Heteroskedasticity . . . . .	109
3	Labor Force Participation of Female . . . . .	110
<b>11</b>	<b>Serial Correlation and Heteroskedasticity in Time Series Regressions</b>	<b>113</b>
1	The Nature of Time Series Data . . . . .	113
2	Examples of Time Series Regression Models . . . . .	114
	2.1 Static Models . . . . .	114
	2.2 Finite Distributed Lag Models . . . . .	114
3	Properties of OLS under classical assumptions . . . . .	115
4	Properties of OLS with Serially Correlated Errors . . . . .	117
5	Unbiasedness and Consistency . . . . .	117
6	Efficiency and Inference . . . . .	117
7	Testing for Serial Correlation . . . . .	118
	7.1 A "t-test" for AR(1) serial correlation with strictly exogenous regressors . . . . .	118
	7.2 The Durbin-Watson Test (DW test) . . . . .	119
	7.3 Testing for AR(1) serial correlation "without" strictly exogenous regressors . . . . .	120
	7.4 Testing for AR(q) serial correlation "without" strictly exogenous regressors . . . . .	121
8	Correcting for serial correlation . . . . .	121
	8.1 Passive way . . . . .	121
	8.2 Active way – . . . . .	121
<b>12</b>	<b>Multicollinearity</b>	<b>123</b>
1	The Nature of Multicollinearity . . . . .	123
2	Consequences of Multicollinearity . . . . .	123

2.1	The OLS estimator will still be BLUE. . . . .	123
2.2	The variances and covariances will be very large. This makes precise estimation difficult. . . . .	123
3	Detection of multicollinearity . . . . .	124
4	Remedial Measures . . . . .	125



# Introduction and Motivation

## 1 What is econometrics?

Statistical method for estimating economic relationships, testing economic theories, and evaluating an implementing government and business policies (Wooldridge, 2009)

- Estimating economic relationships – How much discount would a hotel have to give in order to achieve full occupancy? (Price vs. Demand), How many cars would be sold if the car tax is reduced by half? (Tax vs. Consumption), etc.
  - Testing economic theories – For example, is it true that the demand curve is always downward sloping? etc.
  - Evaluating and implementing government policies – Does universal health care (e.g. 30 baht program) help decrease infant mortality rate? Which method is more effective in convincing rural students to come to school, free lunch or subsidy to the parents?
  - Evaluating and implementing business policies – Should firm pay the manager a fixed salary or a variable compensation in order to achieve the highest profit? Are part-time workers more productive than full-time workers? etc.
-

#### 4 1. Introduction and Motivation

- In econometrics, we believe that there are actually "true" answer(s) to the above questions.
  - Econometricians collect data (sample data) and use the sample data to answer the questions.
  - Econometric methods help justify that the answers that we get from analyzing the sample is comparable to the "true" answers.
-

### 1.1 Steps in Empirical Economics Analysis

- – Empirical -> data and numbers.
  1. Form a question, define the population of interest.
  2. In some cases, construct an economic model.

$$y = f(x_1, x_2, x_3, x_4)$$

where  $y$  denotes the number of days/week that a student (in rural Thailand) would go to school,  $x_1$  denotes the availability of school lunch,  $x_2$  denotes the provision of subsidies to parents,  $x_3$  denotes parents' income,  $x_4$  denotes parents' occupation.

3. If the analysis is based on a real economic model, one has to adapt it in such a way they can perform econometrics analysis.
4. In many cases, one bypasses 2. and 3., and construct an econometric model right away
5. Let the "Number of days a student goes to school" =  $y$

$$y = \beta_0 + \beta_1 \text{lunch} + \beta_2 \text{subsidies} + \beta_3 \text{parents\_inc.} + \beta_4 \text{parents\_occ.}$$


---

## 2 Types of economic data

- Data is usually a "subset" of the population of interest. For example, some students in the village of interest.
- Cross-sectional data – More than 1 individual, households, cities, villages. But 1 time period.
- Let the "Number of days a student goes to school" =  $y$

Student no.	$y$	lunch	subsidies	parents_inc.	parents_occ.
1	5	1	1	3,000	farmer
2	2	0	1	4,500	hair-dresser
3	3	0	0	6,000	farmer
4	4	0	1	3,500	driver

- Time-series data – 1 individual, households, cities, villages. More than 1 time period.

Student no.	Time	$y$	lunch	subsidies	parents_inc.	parents_occ.
1	1/02/10	5	1	1	3,000	farmer
1	2/02/10	5	0	1	3,000	farmer
1	3/02/10	3	0	0	3,000	farmer
1	4/02/10	2	0	0	3,000	farmer

- Panel or Longitudinal data – Several individual, households, cities, villages. More than 1 time period.

Student no.	Time	$y$	lunch	subsidies	parents_inc.	parents_occ.
1	1/02/10	5	1	1	3,000	farmer
1	2/02/10	5	0	1	3,000	farmer
2	1/02/10	2	0	0	4,500	hair-dresser
2	2/02/10	4	0	1	4,500	hair-dresser
...	...	...	...	...	...	...
4	2/02/10	3	0	1	3,500	driver

---

# Review of Some Statistical Concepts

- Suggested readings – Wooldridge, Appendix B and C **OR** Gujarati, Appendix A, pp.869-912

## 1 Random variables and distributions

- Suppose you work for Nok Air and is assigned to work on the reservation policy. You know that it is possible for a passenger to have bought that ticket and not show up. Therefore, it is possible to sell more than 100% of the tickets (let the flight overbooked). But by how much? What is the probability that each one passenger would not show up (on each day of the year)? By studying probability, random variables and their distribution could help you make this decision.
  - Here, each incident (a passenger's decision to show up) is a *random variable*. This random variable turn out to be two outcomes 1) show up, 2) not show up.
  - The summary of the probability that each outcome could happen is called the *probability distribution*.
-

1.1 *Types of Random Variables*

1. Discrete Random Variables – outcomes are only countable numbers, i.e. 0,1,2,3...
  - Bernoulli Random Variable is the one that takes only two values, coin flipping, passengers show up vs. not show up, etc.
  - For example:

$$\begin{aligned}
 X &= \begin{cases} 1 & \text{if the passenger shows up} \\ 0 & \text{otherwise} \end{cases} \\
 P(X = 1) &= \theta \\
 P(X = 0) &= 1 - \theta.
 \end{aligned}$$

where  $P(X = 1) = \theta$  is the probability that the passenger will show up. Since there are only two possible outcome, the probability that the passenger will not show up is  $1 - \theta$ .

2. Continuous Random Variables – outcomes can be any possible value.
  - For example:

$$\begin{aligned}
 X &= \text{PM2.5 level tomorrow} \\
 P(70 < X < 120) &= 1 - P(X \leq 70) - P(X \geq 120).
 \end{aligned}$$


---

## 1.2 Examples of Some Distributions

• Normal Distribution  $N(\mu, \sigma^2)$ 

-  $\mu$  = mean ... or the expected value of the random variable  $X$  when we draw  $X$  repeatedly for many many times (like 1,000 time).

-  $\sigma^2$  = variance ... or how far the random variable  $X$  is from its mean  $\mu$  on average.

$$- f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$- F(x) = \int_{-\infty}^x f(x)$$

## • Bernoulli Distribution

$$- f(x) = \begin{cases} \theta & \text{if } X = 1 \\ 1 - \theta & \text{if } X = 0 \\ 0 & \text{otherwise} \end{cases}$$

• -  $F(x) = ?$ 

-  $mean(x) = ?$

-  $variance(x) = ?$

---

## 2 Joint Distributions, Conditional Distributions and Independence

### 2.1 Joint Distributions

- It sometimes rains in August and I sometimes forget to bring my umbrella. Let  $X$  be a discrete random variable which takes the value of 1 when it rains, 0 otherwise. Let  $Y$  be a discrete random variable which takes the value of 1 when I bring my umbrella. What is the probability that it rains and I happen to bring my umbrella?
- Variables  $X$  and  $Y$  have a joint distribution.
- $f_{X,Y}(x,y) = P(X = x, Y = y)$  or  $P(X, Y)$  is the joint density function of  $(X, Y)$ .
- Suppose the probability that it rains ( $X = 1$ ) on a given day in August is 0.4, and the probability that I bring my umbrella ( $Y = 1$ ) is 0.7. What is the probability that I bring my umbrella on a rainy day?

Answer: It depends...

---

## 2.2 Conditional Distribution and Marginal Distribution

- Conditional probability  $P(Y|X)$ , Marginal probability  $P(Y)$
- Usually, in economics, we are interested in variables that are *not* independent from each other. Thus, the independence assumption does not usually hold.

For example

- the probability that I bring my umbrella may depend on what the weather forecast says.
  - the probability that a basketball player could score may depend on whether he could score in the previous attempt.
  - the probability that a student can pass a university entrance examination could depend on which high school he/she goes to, his/her parents' education level, his/her effort, etc.
  - Conditional distribution/probability is the distribution/probability of a random variable given the outcome of another (other) random variable(s).
    - What is the probability that I bring my umbrella given that it rains?
    - What is the probability that a basketball player could score given that he could not score in the previous attempt?
    - What is the probability that a student can pass university entrance exam given her high school, her parent education level and her effort?
-

- Marginal distribution (probability) is the distribution (probability) of  $Y$  regardless of the value of  $X$ .
  - Like, the probability that it rains
  - or the probability that I bring my umbrella.

$$P(Y = y).$$

$$P(X = x)$$

---

- The relationship between conditional, joint and marginal probability.

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

or in the continuous context, we can write

$$f_{X,Y}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

---

*2.3 Independence*

- If  $X$  and  $Y$  are independent, then knowing the outcome of  $X$  would not change the probabilities of the possible outcomes of  $Y$ .
- Like.. if "rain" and "my decision to bring an umbrella" are independent, then, knowing that it's going to rain would not change my probability to bring an umbrella.
- Thus, **if**  $X$  and  $Y$  are independent:

$$P(X = 1, Y = 1) = P(X = 1) \times P(Y = 1) = 0.4 \times 0.7 = 0.28$$

or  $P(X = x, Y = y) = P(X = x) \times P(Y = y)$

- If  $X$  and  $Y$  are **not** independent, then:

$$P(X = 1, Y = 1) \neq P(X = 1) \times P(Y = 1)$$

or  $P(X = x, Y = y) \neq P(X = x) \times P(Y = y)$ .

---

### 3 Expectation, variance, covariance and correlation

#### 3.1 Expected value of a Random Variable – $E(X)$ or $\mu_X$

- An expectation is a measure of central tendency.
- The expected value of a random variable  $X$ , denoted  $E(X)$  or  $\mu_X$ , is the average value of the random variable over many repeated draws.
  - Discrete case:  $E(X) = \sum_{i=1}^N x_i P(X = x_i)$
  - Continuous case:  $E(X) = \int_{-\infty}^{\infty} x f(x) dx$

Example: What grade would student A get from EE325?

$x_i$	$P(X = x_i)$	$x_i P(X = x_i)$
0( <i>F</i> )	0.05	0
1( <i>D</i> )	0.05	0.05
2( <i>C</i> )	0.25	0.50
3( <i>B</i> )	0.40	1.20
4( <i>A</i> )	0.25	1.00
Total ( $\sum_{i=1}^N$ )	1.00	2.75 ( $\approx B-$ )

---

3.2 *Properties of Expected Values*

1. For any constant  $c$ ,  $E(c) = c$
2. For any constant  $a$  and  $b$ ,  $E(aX + b) = aE(X) + b$
3. If  $\{a_1, a_2, \dots, a_n\}$  are constants and  $\{X_1, X_2, \dots, X_n\}$  are random variables, then

$$E(a_1X_1 + a_2X_2 + \dots + a_nX_n).$$

Or using summation notation

$$E \sum_{i=1}^n a_i X_i = \sum_{i=1}^n a_i E(X_i).$$

\*\* It is important to be reminded that  $E(X^2) \neq E(X)E(X)$ !

---

## 3.3 Conditional and Marginal Expectations

- Example: What grade would student A get from EE325 given the number of hours/week she spent on studying EE325.

	$P(X = x_i, H = h_i)$				
$x_i \backslash h_i$	0	3	6	9	12
0( <i>F</i> )	0.0	0.0	0.0	0.0	0.0
1( <i>D</i> )	0.1	0.1	0.0	0.0	0.0
2( <i>C</i> )	0.0	0.1	0.1	0.0	0.0
3( <i>B</i> )	0.0	0.05	0.15	0.1	0.05
4( <i>A</i> )	0.0	0.0	0.05	0.1	0.1

where  $x_i$  indicates grade,  $h_i$  indicates number of hours/week spent on studying EE325.

- What are the expectations of grade ( $x_i$ ) and hours ( $h_i$ )?
  - Are ( $x_i$ ) and ( $h_i$ ) independent? why or why not?
  - What is the conditional expectation of  $x_i$  given  $h_i = 9$ ?
-

## 3.4 Variance and Standard Deviation of a Random Variable

- The variance and standard deviation measure the "variability", the "dispersion" or the "spread" of a probability distribution

$$\begin{aligned}
 \text{Var}(X) \text{ or } \sigma^2 &= E[(X - \mu_x)^2] \\
 &= \\
 &= \\
 &= \\
 &= \\
 &= E(X^2) - \mu^2.
 \end{aligned}$$

- The standard deviation, denoted  $sd(X)$ , is the positive square root of the variance:

$$sd(X) \text{ or } \sigma = +\sqrt{\text{Var}(X)}.$$

Exercise:

- What would happen if variance = 0?
  - Depict probability distributions with different values of variance.
-

### 3.5 Properties of Variances

1.  $Var(X) = 0$  if, and only if, there is a constant  $c$ , such that  $P(X = c) = 1$ , in which case,  $E(X) = c$ .
2. For any constant  $a$  and  $b$ ,  $Var(aX + b) = a^2Var(X)$

### 3.6 Properties of Standard Deviations

1. For any constant  $c$ ,  $sd(c) = 0$
2. For any constants  $a$  and  $b$ ,

$$sd(aX + b) = |a|sd(X).$$

In particular, if  $a > 0$ , then  $sd(aX) = a \times sd(X)$ .

---

Exercise: What is  $Var(aX + b) = ?$ ,  $Var(aX + bY) = ?$

- Well, how do we compare and assess one random variable against another? One way to do this is through standardization.
- You may have heard of the  $Z$ -score

$$Z = \frac{X - \mu_X}{\sigma}.$$

- If  $X$  is normally distributed, then  $Z$  would be normally distributed. We can use the well-known  $Z$  statistic table for hypothesis-testing.
- In many cases, where the sample size gets very large,  $Z = \frac{X - \mu_X}{\sigma}$  is approximately normally distributed regardless of the original distribution of  $X$ .

Exercise:

- What is  $E(Z)$  and  $Var(Z)$ ?
-

### 3.7 Covariance and Correlation of Two Variables

- We talked about joint distribution and independence earlier. How do these two concepts relate to covariance and correlation?
- When the movement of one variable can give some information about the movement of another variable, these two variables are dependent.
- We can also say that they are "correlated" or their "covariance" is not zero.
- Covariance and correlation measure the amount of *linear* association between variables.
- \*\*It is worth noting that zero correlation does not imply independence.

$$\begin{aligned}
 \text{covariance or } Cov(X, Y) \text{ or } \sigma_{X,Y} &= E[(X - \mu_X)(Y - \mu_Y)] \\
 &= \\
 &= \\
 &= \\
 &= E(XY) - \mu_X\mu_Y.
 \end{aligned}$$

Exercise: What's the unit of  $Cov(X, Y)$ ?

---

- Correlation – makes the unit of dependence more standardized.

$$\text{Corr}(X, Y) \text{ or } \rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y}.$$

- $\rho_{X,Y} = 1$ ; perfect positive linear relationship
  - $\rho_{X,Y} = -1$ ; perfect negative linear relationship
  - $\rho_{X,Y} = 0$ ; no linear relationship.
-

#### 4 Estimators and desirable properties of estimators

- In statistics and econometrics, we hardly have the complete information of the population.
- Most data that we deal with are from a subset of the population or a "sample".
- We would like to learn about the "population" as we can using the "sample" that we have.
  - It is important to identify the population of interest
  - Once the population is identified, we can specify the model for the population relationship of interest.
- Examples of Estimators...

Population Parameter	Estimator(s)
population mean ( $\mu$ )	$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$
population variance ( $\sigma^2$ )	$S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$
$\alpha, \beta_1, \beta_2, \beta_N$	$\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_N$

- Miguel and Kremer(2004) "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities."

$$\text{school attendance rate} = \alpha + \beta_1 \times \text{deworming} + \beta_2 x_2 + \dots + \beta_N x_N$$

How do we know the true value of  $\alpha, \beta_1, \beta_2, \beta_N$ ? Without having the population and the correct model, it may be impossible to know the true value of  $\alpha, \beta_1, \beta_2, \beta_N$ . But the econometricians try their best to come up with the estimators, often denoted  $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_N$ . The better the estimator, the more it satisfies the desirable properties.

---

4.1 *Desirable properties of estimators*

- From now on, let's denote the population parameter of interest " $\theta$ " and its estimator " $W$ ". The desirable properties of estimators are: unbiasedness, efficiency and consistency.

**1. Unbiasedness** – the expected value of the estimator is equal to the value of the parameter it tries to estimate.

-  $E(W) = \theta$

-  $Bias(W) \equiv E(W) - \theta$

- Exercise: is  $\bar{X}$  a biased estimator? What about  $X_1$  ?

---

**2. Efficiency** – an estimator with a lower variance is said to be "more efficient" than another estimator with a higher variance

- If  $Var(W_1) \leq Var(W_2)$ , then  $W_1$  is a more efficient estimator of  $\theta$  than  $W_2$ .
  - Exercise: Which estimator is more efficient,  $\bar{X}$  or  $X_1$ ?
-

**3. Consistency** – when the sample size gets large, the estimator  $W$  can do better and better in estimating  $\theta$ .

- Large sample properties can also be called "asymptotic" properties.

- For  $W_n$ , which is an estimator of  $\theta$  based on a sample  $X_1, X_2, \dots, X_n$  of sample size  $n$ .

Then,  $W_n$  is a consistent estimator of  $\theta$ , if for every  $\varepsilon > 0$ ,

$$P(|W_n - \theta| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

if not, then we can say that  $W_n$  is inconsistent.

---

# The Simple Regression Model

- Some questions
  - How often ( $Y$ ) would a person like to go to starbucks if the price ( $X$ ) increases?
  - How many less cars ( $Y$ ) would be on the street in Bangkok if the government increases car tax ( $X$ )?
  - How much would my income ( $Y$ ) increase if I obtain a master's degree ( $X$ )?
  
- The concept
  - How would you measure the "change in family expenditure" ( $Y$ ) when the "family income" ( $X$ ) changes<sup>1</sup>.

Table 2.1: Weekly Family Income and Expenditure of all households on the Island

	$X = \text{Weekly Family Income (THB)}$					
	500	600	700	800	900	1000
$Y = \text{Weekly Family Expenditure (THB)}$	360	376	458	610	600	700
	313	475	422	468	531	679
	322	380	498	575	670	730
	310	382	560	542	630	591
	390	290	442	588	544	550
	315	425	440	466	565	620
	390	442	-	461	-	695
	400	-	-	-	-	635
Total	2800	2870	2820	3710	3540	5200
Conditional means of $Y, E(Y X)$	350	410	470	530	590	650

Table 2.2: Conditional Probabilities  $p(Y|X_i)$  for the weekly family income and expenditure

	$X = \text{Weekly Family Income (THB)}$					
	500	600	700	800	900	1000
$Y = \text{Weekly Family Expenditure (THB)}$	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	-	1/7	-	1/8
	1/8	-	-	-	-	1/8
Conditional means of $Y, E(Y X)$	350	410	470	530	590	650

<sup>1</sup>This example is obtained from Aj.Wasin's lecture notes.

- How do we calculate the conditional expected value of weekly consumption expenditure given the income level ( $E(Y|X)$ )?

- How do we calculate the expected value of weekly consumption  $E(Y)$ ?

Figure 2.1 Conditional Distribution of Expenditure for Various Levels of Income

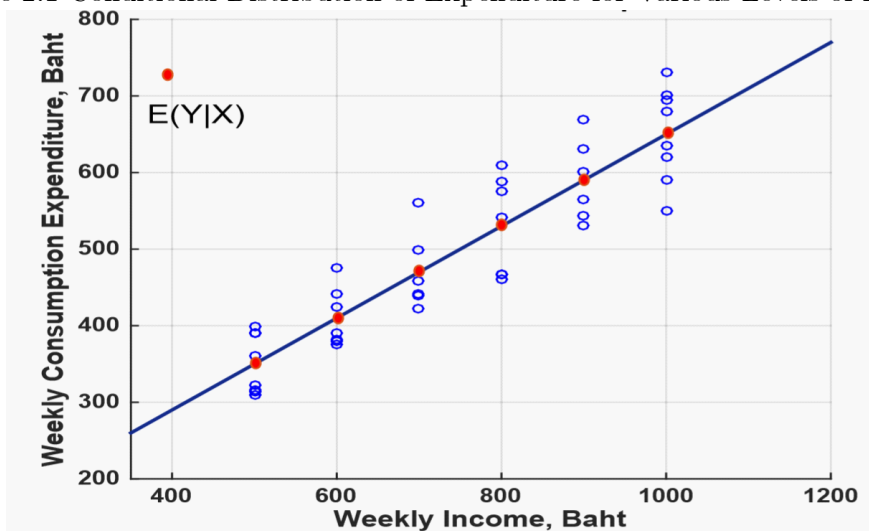
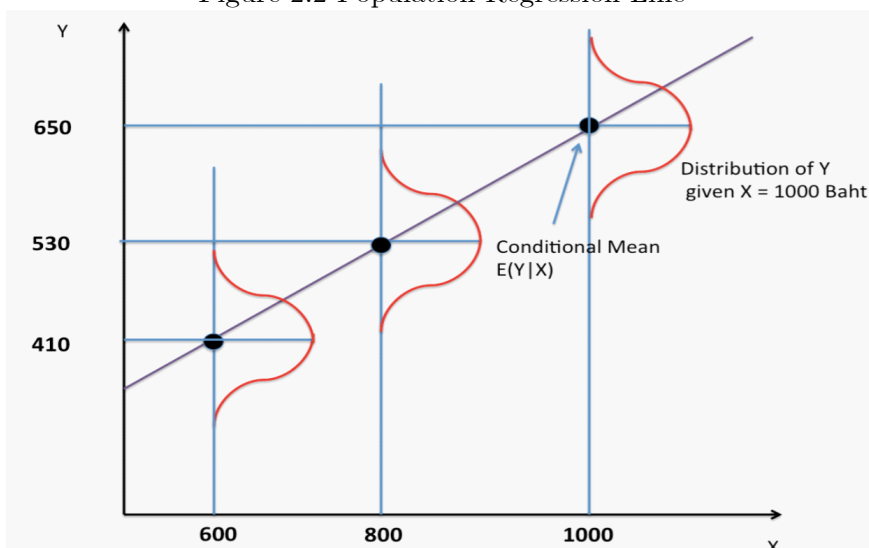


Figure 2.2 Population Regression Line



# 1 Principle, assumptions and derivation of ordinary least squares (OLS) estimators

## 1.1 Terminology for the Linear Regression

- The simple regression model can be defined as

**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

- $Y_i$  and  $X_i$  are variables
- $\beta_0$  and  $\beta_1$  are parameters

## 1.2 Derivation of Ordinary Least Squares (OLS) Estimators

- If we do not have the data point of the entire population, but rather on a subset of samples, what should we do to derive  $\beta_0$  and  $\beta_1$ ?

Example:

How often would a TU undergraduate student go to Fuji restaurant ( $Y$ ) if he/she receives an ( $X$ ) % discount in price?

TABLE 3.1. Number of visits to Fuji Restaurant per year and percentage discount

$i$	$Y_i =$ visit to Fuji restaurant (times/year)	$X_i =$ percentage discount in price
1	95	78
2	42	55
3	56	67
4	83	70
...	...	...
100	32	46

Source: data collected from a random survey of 100 TU undergraduate students

- We surveyed 100 TU undergraduate students (sample)
- We hope that we can use this data to explain the frequency of TU undergraduate students' (population) visit to Fuji restaurant given the price discount.

- The relation of the frequency of TU undergraduate students' (population) visit to Fuji restaurant ( $Y$ ) given the price discount ( $X$ ) can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- But since we don't have the data of the entire population and we don't know the true value of  $\beta_0$  and  $\beta_1$ , we need to find estimators of  $\beta_0$  and  $\beta_1$ .
- The estimators are often called  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . (like  $\bar{X}$  is an estimator of  $\mu$  and  $S^2$  is an estimator of  $\sigma^2$ ).

**How do we estimate  $\hat{\beta}_0$  and  $\hat{\beta}_1$ ?**

- The OLS suggests that we can find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that minimizes the sum of squared errors (deviation from the regression line).

- Mathematical Derivation of OLS (minimization the sum of squared errors)

$$\arg \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n (\hat{u}_i^2) = \arg \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

First Order Condition (F.O.C):

$$w.r.t. \hat{\beta}_0 \Rightarrow 0 = -2(\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)) \quad (3.1)$$

$$w.r.t. \hat{\beta}_1 \Rightarrow 0 = -2(\sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)) \quad (3.2)$$

Divide 3.1 by  $-2$ , we get

## 2 Properties of OLS estimators

### 2.1 Algebraic Properties

- Thus far, the OLS estimators are

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

this gives

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i \\ \hat{u}_i &= Y_i - \hat{Y}_i.\end{aligned}$$

This implies the following algebraic properties

1.  $\sum_{i=1}^n \hat{u}_i = 0$  – the calculation of OLS  $\hat{\beta}_0$  and  $\hat{\beta}_1$  is done such that it minimizes  $\sum_{i=1}^n \hat{u}_i^2$ . This is true when  $\sum_{i=1}^n \hat{u}_i = 0$
  2.  $\sum_{i=1}^n X_i \hat{u}_i = 0$  – we get this from the first order condition deriving  $\hat{\beta}_1$ . This implies that  $Cov(X_i, \hat{u}_i) = 0$ . (Why?)
  3. The point  $\bar{Y}$  and  $\bar{X}$  are always on the regression line – we know this from  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ .
-

2.2 *Properties proving BLUE*

- The OLS estimator is the Best Linear Unbiased Estimator (BLUE). How?
- The OLS'  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are unbiased estimators of  $\beta_0$  and  $\beta_1$  respectively

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

or  $E(\hat{\beta}_{OLS}) = \beta.$

- The OLS'  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are the most efficient among all the linear estimators

$$Var(\tilde{\beta}_{non-OLS} | X) - Var(\hat{\beta}_{OLS} | X) \geq 0.$$

2.3 *Assumptions on the simple linear regression (SLR) model*

- Some assumptions (or certain conditions) are required for the OLS to be BLUE. This set of assumptions are often called the "Gauss-Markov Assumptions for Simple Regression"

**Assumption SLR1.** Linear in Parameter –  $Y$  is linear in  $X$ .

**Assumption SLR2.** Random Sampling – We have a random sample size  $n$ ,  $\{(x_i, y_i) : i = 1, \dots, n\}$ , This, the model \_\_\_\_\_ becomes:

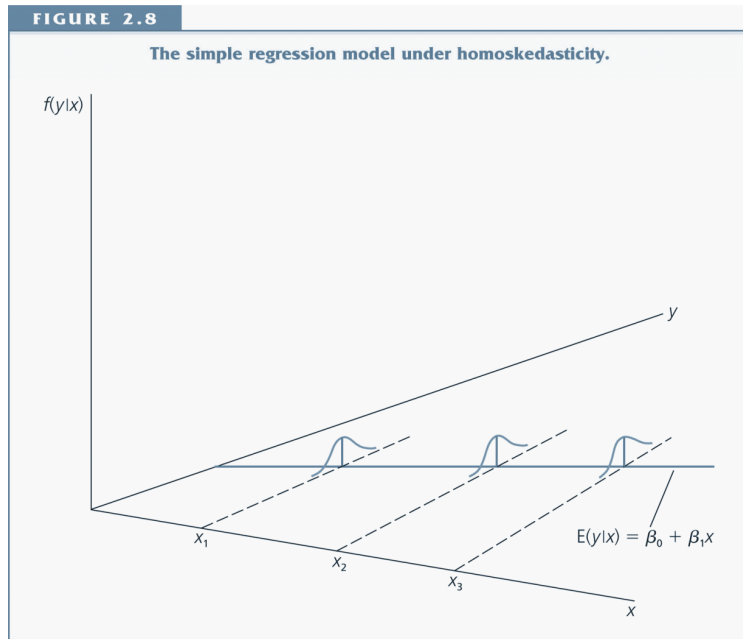
**Assumption SLR3.** Sample Variation in the Explanatory Variable – The sample outcomes are not all the same value

**Assumption SLR4.** Zero Conditional Mean –

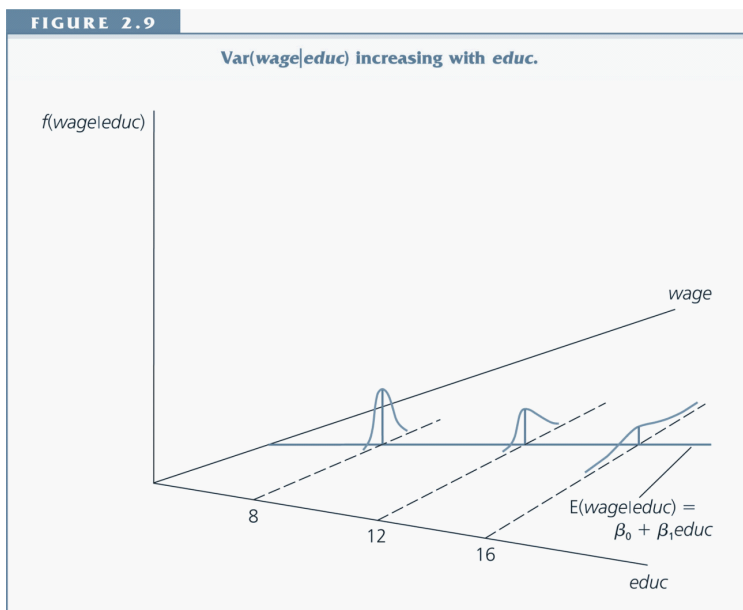
**Assumption SLR5.** Homoskedasticity –

- In other words, The OLS estimator of  $\beta$  in the linear model when  $u_i$  is *i.i.d.*( $0, \sigma^2$ ) is the best (minimum variance) estimator within the class of linear unbiased estimator.
- The conditional concept, which implies that  $X_i$  is predetermined (being conditional upon, or fixed) is very crucial for the OLS estimators to be unbiased.
- We must not forget that there is no reason that all these assumptions should be true!

2.4 Homoskedasticity VS. Heteroskedasticity



Homoskedasticity



Heteroskedasticity

2.5 *Expectation of Estimators*

- Proof for  $E(\hat{\beta}_1) = \beta_1$ : We need to use assumption SLR 1,2,3,4.

From

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (3.3)$$

For calculation tractability, let

## 2.6 Variance of OLS Estimators

- From eq.  $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n u_i k_i}{\sum_{i=1}^n k_i^2}$ , we can write

$$\begin{aligned}\hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n u_i k_i}{\sum_{i=1}^n k_i^2} \\ \text{Var}(\hat{\beta}_1) &= \text{Var}(\beta_1) + \text{Var}\left(\frac{\sum_{i=1}^n u_i k_i}{\sum_{i=1}^n k_i^2}\right)\end{aligned}$$

Here,  $\beta_1$  (the true  $\beta_1$ ) is a constant. And since we are conditioning on  $X_i$ , the values of  $k_i$  are also non-random. (This is not to be confused with the random sampling of  $X_i$ ,  $i = 1, \dots, n$ .) When you are conditioning on some variables, you take those variables as non-random (or as given). In this case, we can write

The proof that  $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n k_i^2}$  is the minimum under the class of "linear unbiased estimator" is complicated without relying on some matrix simplifications. This proof is beyond the scope of this course.

---

2.7 *Some Concepts to be emphasized (Population vs. Sample; PRF vs. SRF; error vs. residual)*

- Population – is the "truth" and in econometrics, we believe that there is one set of "truth". Therefore, the Population Regression Function (PRF) is fixed. In almost all cases, we do not know "exactly" what PRF is.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (3.4)$$

where  $i = 1, 2, \dots, \Psi$ .  $\Psi$  represents the total number of the "population" we are interested in.

- Sample – is a subset of "truth", a subset of "population". We run, or construct, the Sample Regression Function (SRF) to estimate the PRF.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

where  $i = 1, 2, \dots, n, n < \Psi$ .

- The true "error term"  $u_i$  in eq.3.4 is **unobserved** because we never know what  $\beta_0$  and  $\beta_1$  is.
- We can, however, observe the "residual" or "predicted residual"  $\hat{u}_i$  from the following calculation

$$Y_i - \hat{Y}_i = \hat{u}_i.$$


---

**Examples:** two different sample regression models:

In practice, if the population is large, we can only obtain a subset of samples in order to obtain estimators of  $\beta_0$  and  $\beta_1$ . The estimated  $\hat{\beta}_0$  and  $\hat{\beta}_1$  will be unbiased if assumptions SLR1-5 are satisfied. However, the EXACT value of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  will be different according to the SAMPLE INFORMATION.

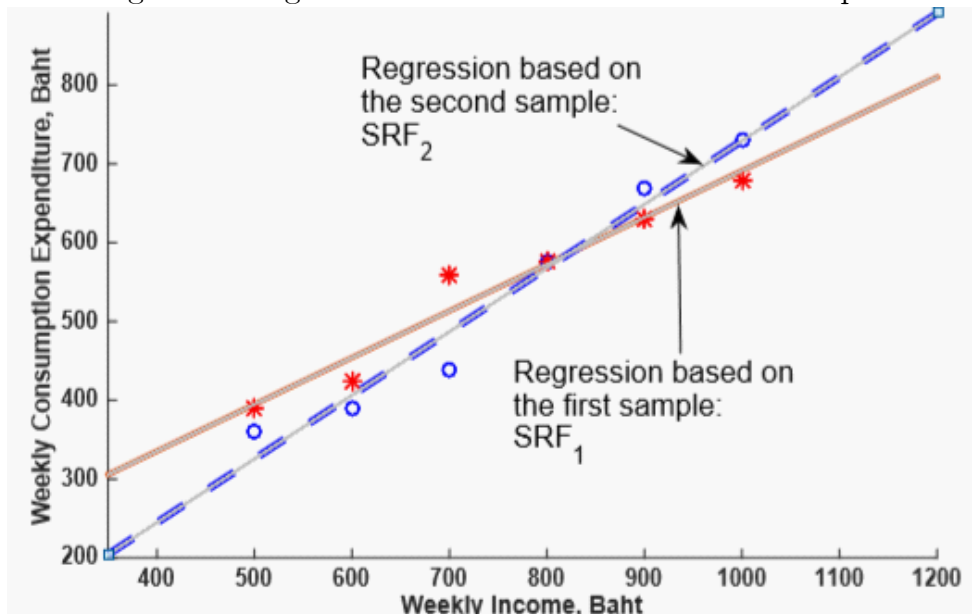
Table 2.3 Random sample no. 1 (collected on 3/Jan/2020)

$X$ (Weekly Family Income)	$Y$ (Weekly Family Expenditure)
500	360
600	390
700	440
800	575
900	670
1000	730

Table 2.4 Random sample no. 2 (collected on 14/Jan/2020)

$X$ (Weekly Family Income)	$Y$ (Weekly Family Expenditure)
500	350
600	410
700	470
800	530
900	590
1000	650

Figure 2.3 Regression lines based on two different samples





### 2.8 Goodness of Fit ( $R^2$ )

- We can never measure how well SRF can estimate PRF because PRF is not known to allow us to compare.
- But we can assess how well different SRFs fit with the "sample" data that we collect – this can also be called the "goodness of fit".

Residual Concepts:-

$$\begin{aligned} \text{Total Sum of Squares (SST)} &= \\ \text{Explained Sum of Squares (SSE)} &= \\ \text{Residual Sum of Squares (SSR)} &= \\ &SST = \end{aligned}$$

- The  $R^2$ , or the coefficient of determination, is defined as
-

2.9 *Incorporating Nonlinearities in Sample Regression*

- So far, we have only mentioned the "linear" relation.
- OLS can actually incorporate non-linearities as long as the non-linearities is in the "variables".
- For examples:

$$\log Y_i = \beta_0 + \beta_1 \log X_i + u_i \quad (\text{This is called a constant elasticity model. Why?})$$

or

$$\log Y_i = \beta_0 + \beta_1 X_i^2 + u_i$$

or

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + u_i.$$

etc. etc...

---

### 3 Regression Through the Origin

- If you know that the regression line goes through the origin ( $X = 0, Y = 0$ ) for sure, we can impose this restriction
- This restriction makes sense in some contexts. For example,  $X = \textit{income}$  and  $Y = \textit{income tax}$ .
- The SRF then becomes

$$Y_i = \tilde{\beta}_1 X_i$$

where  $\tilde{\beta}_1$  is an OLS estimator. It can be calculated through finding  $\tilde{\beta}_1$  that minimizes the sum of squared residual

---



# Multiple Regression Analysis (Estimation)

## 1 Motivation

The SLR4. assumption,  $E(u_i|X_i) = 0$ , is unrealistic. This implies that that  $u_i$  is uncorrelated with  $X_i$  because no matter what the value of  $X_i$  is, the expected value of  $u_i$  would still be 0! Thus, when this assumption does not hold, OLS estimates ( $\beta_0$  and  $\beta_1$  will be biased). (Why? – note:  $Cov(X_i, \hat{u}_i) = 0$  is always true by the OLS calculation. This does not mean that  $Cov(X_i, u_i) = 0$  in reality.)

In this case, the multiple regression analysis is introduced in order to achieve the condition  $E(u_i|X_i) = 0$ . It also enables us to explain the dependent variable better and to conduct the "ceteris paribus" or "holding all other things constant" analysis.

Example: If we want to find the relation between wage and education in the simple linear regression, would our  $\beta_0$  and  $\beta_1$  be biased? Probably!

- Consider a simple linear regression

1.1 Assumption SLR 4 ( $E(u|X) = 0$ ) in the Multiple Regression Context

Consider the *wage* equation.

- In the case of simple regression, the assumption SLR4 ( $E(u|X) = 0$ ) has to be satisfied in order to achieve an unbiased estimator of  $\beta_0$  and  $\beta_1$ .
  - In this two-variable regression of assumption SLR4 ( $E(u|X) = 0$ ) becomes  $E(u|X_1, X_2) = 0$ .
  - Therefore, the OLS estimator of  $\beta_0, \beta_1$  and  $\beta_2$  would be unbiased if  $E(u|educ, inc) = 0$ .
  - For example, "*innate ability*" is not included in the equation above. Thus, if "*innate ability*" can explain *wage*, it would be in  $u$ .
  - If it is true that  $E(\text{innate ability}|educ, inc) = 0$ , or the expected value of *innate ability* is the same and (equal to zero) for all education and income levels, then the OLS estimators  $\hat{\beta}_0, \hat{\beta}_1$  and  $\hat{\beta}_2$  would be unbiased. (Is this likely?)
-

## 2 The Model with k Independent Variables

- The "population" version of the multiple linear regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u,$$

where

$\beta_0$  is the intercept.

$\beta_1$  is the parameter associated with  $X_1$ .

$\beta_2$  is the parameter associated with  $X_2$ , and so on.

$u$  is the error term

- $Y, X_1, X_2, \dots, X_k$  are variables
- $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  are parameters

**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

2.1 *Assumptions on the multiple linear regression (MLR) model*

- Some assumptions (or certain conditions) are required for the OLS to be BLUE. This set of assumptions are often called the "Gauss-Markov Assumptions for Simple Regression"

**Assumption MLR1.** Linear in Parameter –  $Y$  is linear in  $X$ .

**Assumption MLR2.** Random Sampling –

**Assumption MLR3.** No Perfect Collinearity

**Assumption MLR4.** Zero Conditional Mean –

**Assumption MLR5.** Homoskedasticity –

- In other words, The OLS estimator of  $\beta$  in the linear model when  $u_i$  is *i.i.d.*( $0, \sigma^2$ ) is the best (minimum variance) estimator within the class of linear unbiased estimator.
- The conditional concept, which implies that  $X_i$  is predetermined (being conditional upon, or fixed) is very crucial for the OLS estimators to be unbiased.
- We must not forget that there is no reason that all these assumptions should be true!

## 2.2 Accounting Nonlinearity

- The regression model requires linearity in parameters.
- Similar to the Simple Regression Model, the Multiple Regression Model can also take into account the nonlinear relationships between variables.

Example: In the CEO Salary example, we could write the relations between CEO salary (*salary*), firm sales (*sales*) and CEO age (*age*) as follows:

$$\log(\textit{salary}) = \beta_0 + \beta_1 \log(\textit{sales}) + \beta_2 \textit{age} + \beta_3 \textit{age}^2 + u$$

- - This model has  $k = 3$  because there are 3 regressors.  $Y = \log(\textit{salary})$ ,  $X_1 = \log(\textit{sales})$ ,  $X_2 = \textit{age}$ ,  $X_3 = \textit{age}^2$ .
  - $\beta_1$  measures the change in  $\log(\textit{salary})$  with respect to  $\log(\textit{sales})$ , holding other factors fixed.  $\beta_1$  is the sales elasticity of CEO salary.
  - How do we measure the change in  $\log(\textit{salary})$  with respect to age, holding other factors fixed?
  - How do we measure the change in *salary* with respect to age, holding other factors fixed?
- In any case, the OLS estimates of  $\beta$  would be unbiased if

$$E(u|X_1, X_2, \dots, X_k) = 0.$$

This is the Multiple Regression version of assumption SLR 4—all factors in the unobserved error term should be uncorrelated with the explanatory variables.

---

### 3 Estimation of parameters and properties of estimators

#### 3.1 Deriving OLS Estimators

- We begin with a multiple regression with 2 regressors. Regressions with more regressors can be analyzed in the exact same fashion. Let the population regression model be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where the estimated OLS equation (sample version) of the above regression can be written as

- As before, the OLS estimators are the ones that minimize the sum of residual squared given the observations  $i = 1, 2, \dots, n$  in the sample.

$$\arg \min_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^n (\hat{u}_i)^2 = \arg \min_{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

First Order Condition (F.O.C):

$$w.r.t. \hat{\beta}_0 \Rightarrow 0 = -2(\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})) \quad (4.1)$$

$$w.r.t. \hat{\beta}_1 \Rightarrow 0 = \quad (4.2)$$

$$w.r.t. \hat{\beta}_2 \Rightarrow 0 = \quad (4.3)$$

- Solving equations 4.1, 4.2 and 4.3 simultaneously, we can derive the solution for  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  (the calculation is beyond the scope of this course).
-

3.2 How could the multiple regression analysis enable *ceteris paribus* analysis?

- Consider a multiple regression function of *wage*

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + u \quad (4.4)$$

Here,

- $\beta_0$  is the intercept.  
 $\beta_1$  measures the change in *wage* with respect to *educ*, holding other factors fixed.  
 $\beta_2$  measures the change in *wage* with respect to *inc*, holding other factors fixed.
- What if the function of *wage* is, instead written as

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + \beta_3 educ^2 + u$$

Then,

$\beta_0$  is the intercept.

The change in *wage* with respect to *educ* (holding other factors fixed) is measured by:

The change in *wage* with respect to *inc* (holding other factors fixed) is measured by:

---

#### 4 Expected Value of the OLS Estimators

- Under assumptions MLR 1 to 4 (see Wooldridge),  $\hat{\beta}_{OLS}$  are unbiased.
- 2 issues should be considered regarding the biasedness of  $\hat{\beta}_{OLS}$

##### 4.1 Issue #1: Including Irrelevant Variable (Overspecifying the Model)

- Suppose we specify the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (4.5)$$

**and this model satisfies the multiple regression assumptions 1 to 4**

---

4.2 *Issue #2: Excluding Relevant Variable (Underspecifying the Model → omitted variable bias. This is a serious problem!)*

- Suppose we the **TRUE** model is actually

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where none of the  $\beta$  is zero **and this model satisfies the multiple regression assumptions 1 to 4.**

- But we omit variable  $X_2$  and estimate the following equation using OLS

## 5 Variance of the OLS Estimators

- The  $\hat{\beta}_{OLS}$  would be the most efficient among the linear unbiased estimators if assumption 5 is satisfied
- Multiple Linear Regression (MLR) assumption 5: Homoskedasticity

The error term  $u$  has the same variance given any values of the explanatory variables.

$$Var(u|X_1, X_2, \dots, X_k) = \sigma^2$$

- Example:

- If the MLR assumption 5 is true, then

## 6 Estimator of the OLS Variance

- Since we don't know what  $\sigma^2$  is (population concept), we need to find an estimator of it.

- Thus, STATA's calculation of the std.err. of  $\hat{\beta}_j$  is

$$\widehat{std.deviation}.\hat{\beta}_j = std.err.\hat{\beta}_j = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^N (X_{ij} - \bar{X}_j)^2 (1 - R_j^2)}}.$$

---

Comments:

# Stata Lab 1 – Introduction

## 1 What is STATA?

- A statistical software package used mostly in economics, sociology, political science and epidemiology.
- Stata can be used to manage database, run regressions, generate graphics, do simulations, etc.
- The user should have their own dataset. The Stata data file is usually saved in the .dta format.
- Data of any other formats (like excel) can be imported and/or converted into .dta format.

### 1.1 STATA supports

- Stata's own website: <http://www.stata.com/support/faqs/>
- Stata program's help function: For example, suppose you would like to know more about the "regress" command, then... Open the stata program > in the "command" box > type "help regress" without the " "> press enter.
- Stata's official manual (can be found in the library and embeded in the program)
- Other Stata's user's manual: My favorite one is "An Introduction to Modern Econometrics Using Stata" by Christopher F. Baum.
- or... simply type your question(s) into a search engine.

### 1.2 Data files and Do-files

- The Stata's data file keeps all the data points. For example, each  $Y_i$  and the corresponding  $X_i, \forall i = 1, 2, 3, \dots, n$ .
- The do-file records all the commands that you use to analyze the data.
- Once the data is cleaned, it is best not to keep on re-saving the original data file. If several steps have to be done before analyzing the data (like running a regression), do it on the do-file.

## 2 Tutorial 1: Exploring the Data and Running a Simple Regression

- **Download Wooldridge datasets**

1. Download Wooldridge's data – go to thomsonedu's website (or go to your BE Moodle: EE325):

"[http://www.thomsonedu.com/aise/economics/wooldridge\\_2e\\_datasets/](http://www.thomsonedu.com/aise/economics/wooldridge_2e_datasets/)".

- **To open the STATA program**

1. Double click on the STATA icon.
2. Click on the "Do-file" icon on the top panel of the Stata program. "Save As" your Do-file (and name it "EE325") on your computer.

- **Open the data file using Do-file**

1. file -> open -> then, direct the program to the file "CEOSAL2.DTA".
2. On the command window, you will see a command to open this file. If you would like to open the file from your do-file in the future, you can use this command.

- **To explore and understand the data**

1. type: browse
2. type: describe
3. type: summarize
4. type: sum
5. type: codebook
6. type: describe salary
7. type: tabulate college
8. type: tab college
9. to use the "if" command to find conditional mean (average) type: sum if grad == 1
10. type: sum salary if age <= 40
11. type: correlate salary sales profits
12. type: correlate salary sales profits, covariance
13. type: plot salary profits
14. type: twoway scatter salary profits

- **To run a simple (OLS) regression (one explanatory variable)**

1. type: regress salary profits

- **To create the fitted value ( $\hat{Y}_i$ ) and the residual ( $\hat{u}_i$ )**

1. type: predict y\_hat, xb

2. type: predict u\_hat, residual

- **To see how well we do at finding a Sample Linear Function**

1. type: twoway scatter salary profits || line y\_hat profits

2. type: twoway scatter u\_hat profits

3. To check if the OLS estimation makes  $X_i$  uncorrelated with  $\hat{u}_i$  (by the OLS calculation, they should not correlate), type: correlate sales u\_hat

- **To execute mathematical operations**

1. type: generate log\_salary = log(salary)

2. type: gen log\_profit = log(profit)

3. type: gen profit\_2 = 3+5\*profit

4. type: regress salary profits profit\_2

5. type: gen profit\_sq = profit^2

6. type: regress salary profits profit\_sq

- **To perform a multiple regression analysis**

1. type: regress salary profits sales

2. type: regress salary profits sales ceoten

- **To exit your Stata**

1. Save your do-file

2. file -> exit -> don't save (never ever modify your master dataset!)

- **To find out what all the above commands mean** – (type in the command box) help summarize, help predict, help twoway, etc etc.

## 2.1 Examples from Wooldridge(2009)

C2.2 The data set in CEOSAL2.dta contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollar, and *ceoten* is prior number of years as company CEO.

1. Find the average salary and the average tenure in the sample
2. How many CEOs are in their first year as CEO (that is, *ceoten* = 0)? What is the longest tenure as a CEO?
3. Estimate the simple regression model

$$\log(\textit{salary}) = \beta_0 + \beta_1 \textit{ceoten} + u.$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

C2.1 Use the data in WAGE2.dta to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (*IQ*).

1. Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ score are standardized so that the average in the population is 100 with a standard deviation equal to 15)
2. Estimate a simple regression model where a one percentage point increase in IQ changes *wage* by a constant dollar amount. Use this model to find the predicted increase in *wage* for an increase in IQ of 15 percentage points. Does IQ explain most of the variation in *wage*?
3. Now, estimate a model where each one percentage point increase in IQ has the same percentage effect on *wage*. If IQ increases by 15 percentage points, what is the approximate percentage increase in predicted *wage*?

# Multiple Regression Analysis (Inference)

## Objectives

1. Students know how to test hypotheses about the parameter ( $\beta$ )
2. Students can test for the validity of the proposed population model.

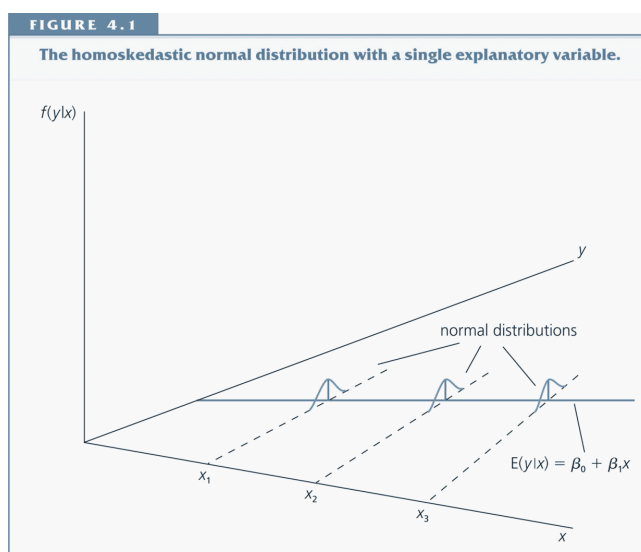
## Econometrics Analysis - The Steps

1 Sampling Distribution of the OLS estimators ( $\hat{\beta}_{OLS}$ )

To be able to test hypotheses about the parameter ( $\hat{\beta}$ ), we need an assumption about the distribution of  $u$ .

**Assumption 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 variable  $\sigma^2$ .



- By normality of the error term ( $u$ ), we have normality of  $\hat{\beta}$

Or, using the standardization  $\frac{Z-\mu}{\sigma} \sim \text{normal}(0,1)$ , we have

## 2 Testing Hypotheses about an individual regression coefficient "the t-test"

- But we do not have  $sd.(\hat{\beta}_j)$ , we can only calculate the estimator of it :

- for degree of freedom  $> 30$ ,  $t \approx z$ . We can use the  $Z$  - score table to find the critical value(s).

### 2.1 Testing Against Two-Sided Alternatives

Consider the wage equation :

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

Suppose we want to test whether experience has a partial effect on wage:

$$\begin{aligned} H_0 & : \beta_2 = 0 \quad (\text{experience has no partial effect}) \\ H_a & : \beta_2 \neq 0. \quad (\text{experience has a partial effect}) \end{aligned}$$

\*\* Note that

Suppose  $n = 34$ , degree of freedom =  $34 - 3 - 1 = 30$ .

From the table t-table, the 5% critical value for a 2-tailed test with 30 d.f. is       
    

- If we change the significance level to 10%, then the 2-tailed critical value becomes
- If we cannot reject  $H_0 : \beta_2 = 0$  (at a given significance level – e.g. 5%), we say that  $X_2$  is statistically significant at the 5% level.
- In other words  $X_2$  has a partial effect on the expected value of  $Y$ .

```
. regress log_wage educ exper tenure
```

Source	SS	df	MS			
Model	46.8741776	3	15.6247259	Number of obs =	526	
Residual	101.455574	522	.194359337	F( 3, 522) =	80.39	
Total	148.329751	525	.28253286	Prob > F =	0.0000	
				R-squared =	0.3160	
				Adj R-squared =	0.3121	
				Root MSE =	.44086	

log_wage	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	.0796756	.4890435

## 2.2 Testing Against One-Sided Alternatives

Suppose the rule for rejecting  $H_0$  becomes

$$H_a : \beta_j > 0.$$

We could have the  $H_0$  as

$$\begin{aligned} H_0 & : \beta_j \leq 0 & \text{or} \\ H_0 & : \beta_j = 0 \end{aligned}$$

depending on the context. For  $H_0 : \beta_j = 0$ , it implies that we are ruling out population values of  $\beta_j$  less than zero. For  $H_0 : \beta_j \leq 0$ , it implies that we are not ruling out population values of  $\beta_j$  less than zero. To give an example of a test against the one-sided alternative hypothesis, consider the wage equation again :

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

If we want to test whether experience has a positive partial effect on wage

$$\begin{aligned} H_0 & : \beta_2 = 0 \\ H_a & : \beta_2 > 0. \end{aligned}$$

We implicitly rule out the possibility that  $\beta_2 < 0$  here.

3 Testing other hypotheses about  $\beta_j$ 

- Most common  $H_0$  is  $H_0 : \beta_j = 0$ .
- However, we can test other types of hypotheses. For example,  $H_0 : \beta_j = a_j$  where  $a_j$  is a constant number.

If we want to test the  $H_0$  using a 5% significance level, then we reject  $H_0$  if

$$t_{wife\_income} > \text{-----} \quad \text{or}$$
$$t_{wife\_income} < \text{-----}$$

#### 4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

$$\log(\text{wage}) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 \text{exp } er + u$$

where  $jc$  = number of years attending a two-year college

$univ$  = number of years at a four-year college

$\text{exp } er$  = months in the workforce.

We want to test whether  $\beta_1 = \beta_2$ .

another possible hypothesis test (one-tailed alternative)

## 5 Computing p-Values for t-Tests

- What is the significance level given the computed t-statistics?

- p-value :  $P(|T| > |t|)$

Example 1:  $H_0 : \beta_j \geq 0, H_a : \beta_j < 0, \text{ d.f.} = 140.$

suppose the calculated  $t_{\hat{\beta}_j} = -2.75$

- From the z-table, the value -2.75 corresponds to area = \_\_\_\_\_  
—.
- Thus, p-value = \_\_\_\_\_.
- Would we reject  $H_0$  if we use the significance level = 5%?

Example 2:  $H_0 : \beta_j = a_j, H_a : \beta_j \neq a_j, \text{ d.f.} = 18.$

suppose the calculated  $t_{\hat{\beta}_j} = -2.18$

- From the t-table, the value -2.18 corresponds to area = \_\_\_\_\_  
—.
- Thus, p-value = \_\_\_\_\_.
- Would we reject  $H_0$  if we use the significance level = 5%?

## 6 Confidence Intervals (CI)

- Confidence Intervals for the POPULATION PARAMETER ( $\beta_j$ )
- A 95% CI of  $\beta_j$  is given by

Example 1: **95% CI**

Example 2: **99% CI**

## 7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$\begin{aligned}y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \\H_1 &: H_0 \text{ is not true}\end{aligned}$$

We can use the F-test to test this type of "multiple hypotheses".

1. Our full model is called the "unrestricted" model (ur). Suppose it can be expressed as:

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

2. The model which takes out  $x$  (which we think its associated  $\beta = 0$ ) is called the restricted model ( $r$ ).

3. Some useful facts

4. Other ways to calculate the F-statistics:

**Example:** Suppose we are interested in understanding the determinant of a baseball player's salary.

- salary* = season salary
- years* = years in major leagues
- gamesyr* = games per year in the league
- bavg* = career batting average
- hrunsyr* = homeruns per year
- rbisyr* = runs batted in per year

- the unrestricted model (ur) is defined by

```
. regress log_salary years gamesyr bavg hrunsyr rbisyr
```

Source	SS	df	MS	Number of obs = 353		
Model	308.989208	5	61.7978416	F( 5, 347)	=	117.06
Residual	183.186327	347	.527914487	Prob > F	=	0.0000
				R-squared	=	0.6278
				Adj R-squared	=	0.6224
Total	492.175535	352	1.39822595	Root MSE	=	.72658

log_salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years	.0688626	.0121145	5.68	0.000	.0450355	.0926898
gamesyr	.0125521	.0026468	4.74	0.000	.0073464	.0177578
bavg	.0009786	.0011035	0.89	0.376	-.0011918	.003149
hrunsyr	.0144295	.016057	0.90	0.369	-.0171518	.0460107
rbisyr	.0107657	.007175	1.50	0.134	-.0033462	.0248776
_cons	11.19242	.2888229	38.75	0.000	10.62435	11.76048

- the restricted model (r) is defined by

```
. regress log_salary years gamesyr
```

Source	SS	df	MS	Number of obs = 353		
Model	293.864058	2	146.932029	F( 2, 350)	=	259.32
Residual	198.311477	350	.566604221	Prob > F	=	0.0000
				R-squared	=	0.5971
				Adj R-squared	=	0.5948
Total	492.175535	352	1.39822595	Root MSE	=	.75273

log_salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years	.071318	.012505	5.70	0.000	.0467236	.0959124
gamesyr	.0201745	.0013429	15.02	0.000	.0175334	.0228156
_cons	11.2238	.108312	103.62	0.000	11.01078	11.43683

Now, our  $H_0$  and  $H_a$  becomes

## 8 How the Hypothesis Testing is done in Practice

1. Check the values of  $t$  – *statistic* reported by the statistical software (i.e. STATA, SPSS, SAS)

⇒ These  $t$  – *statistics* are to test  $H_0 : \beta_i = 0$

⇒ If the d.f.  $> 30$ , then when  $t > 1.96$ , we can reject  $H_0$

⇒ **When  $t > 1.96$** , we can say that  $\beta_i$  is **statistically significant** at 5% level.  
(value of  $\beta_i \neq 0$ )

⇒ **When  $t < 1.96$**  we can say that  $\beta_i$  is **not statistically significant** at 5% level.

⇒ If  $t < 1.96$  we can drop  $x_i$  from the model

⇒ After we drop  $x_i$ , we estimate the new regression function and obtain a new set of  $\hat{\beta}$ .

2. We can also perform other hypothesis testings of interest.

e.g.  $H_0 : \beta_i = \beta_j$

or  $H_0 : \beta_i = 5$  etc.

or perform an F-test for testing multiple linear restrictions

3. Usually, in economics, the estimation results are reported using this form

Dependent Variable: $\log(\text{salary})$			
Independent Variables	(1)	(2)	(3)
$\log(\text{sales})$	.224 (.027)	.158 (.040)	.188 (.040)
$\log(\text{mktval})$	—	.112 (.050)	.100 (.049)
$\text{profmarg}$	—	-.0023 (.0022)	-.0022 (.0021)
$\text{ceoten}$	—	—	.0171 (.0055)
$\text{comten}$	—	—	-.0092 (.0033)
$\text{intercept}$	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)
Observations	177	177	177
R-squared	.281	.304	.353

# Multiple Regression Analysis : Further Issues

## 1 Data scaling on OLS statistics

When we change the unit of measurement of a variable, the value of estimators would change accordingly. For example

$$\widehat{bweght} = \widehat{\beta}_0 + \widehat{\beta}_1 cigs + \widehat{\beta}_2 fa\ min\ c,$$

where

*bwght* = child birth weight, in grams.

*cigs* = number of cigarettes smoked by the mother while pregnant, per day.

*fa min c* = annual family income, in thousands of dollars.

## 2 More on functional forms

- Logarithmic Functional Form

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u$$

- Models with Quadratics

**Example :** Effects of Pollution on Housing Prices

$$\log(\text{price}) = \beta_0 + \beta_1 \log(\text{nox}) + \beta_2 \log(\text{dist}) + \beta_3 \text{rooms} + \beta_4 \text{room}^2 + \beta_5 \text{stratio} + u$$

where

- price* = housing price
- nox* = level of pollution
- dist* = distance from downtown
- rooms* = number of rooms
- stratio* = average student per teacher ratio

The estimation result is given by

**regress lprice lnox dist rooms rooms\_sq stratio**

Source	SS	df	MS			
Model	51.4933152	5	10.298663	Number of obs =	506	
Residual	33.0889098	500	.06617782	F( 5, 500) =	155.62	
Total	84.582225	505	.167489554	Prob > F =	0.0000	
				R-squared =	0.6088	
				Adj R-squared =	0.6049	
				Root MSE =	.25725	

lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnox	-.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	-.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	-.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	.0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	-.0486667	.0058131	-8.37	0.000	-.0600879	-.0372455
_cons	13.59154	.5650901	24.05	0.000	12.4813	14.70178

Consider the effect of "room"

What would be the % change in price when the number of room increases from 5 to 6?

### 3 Models with Interaction Terms

Consider

$$price = \beta_0 + \beta_1 sqr\ ft + \beta_2 bdrms + \beta_3 sqr\ ft \times bdrms + \beta_4 bthrms + u$$

where

*price* = housing price

*sqr ft* = house size (square feet)

*bdrms* = number of bedrooms

*bthrms* = number of bathrooms

## 4 More on the Goodness-of-Fit and Selection of Regressors

- Adding more regressors ALWAYS improve fit

Using adjusted R-squared to choose between non-nested models (one model is not a subset of another).

Consider Model 1

$$\begin{aligned} \widehat{salary} &= 830.63 + 0.0163sales + 19.63roe \\ &\quad (223.90) \quad (0.0089) \quad (11.08) \\ n &= 209, \quad R^2 = 0.029, \quad \bar{R}^2 = 0.020 \end{aligned}$$

Consider Model 2

$$\begin{aligned} \log(\widehat{salary}) &= 4.36 + 0.2751 \log(sales) + 0.0179roe \\ &\quad (0.29) \quad (0.033) \quad (0.004) \\ n &= 209, \quad R^2 = 0.282, \quad \bar{R}^2 = 0.275 \end{aligned}$$



# Multiple Regression Analysis with Qualitative Information:

## 1 Outline

- Describing qualitative information
- Using a single dummy independent variable
- Using dummy variables for multiple categories
- Interactions involving dummy variables
- A binary dependent variable (Y variable): The linear probability model

## 2 Describing Qualitative Information

- "Female" and "Married" are qualitative variable.
- We arbitrarily assign a dummy variable to describe them.

$$\begin{aligned}
 \textit{female} &= \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases} \\
 \textit{married} &= \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases}
 \end{aligned}$$

**TABLE 7.1**

**A Partial Listing of the Data in WAGE1.RAW**

<i>person</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

## 3 Models with a single dummy independent variable

Consider

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u.$$

where

$$female = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

In this case, the  $\delta_0$  notation is used to highlight the interpretation of the parameters multiplying dummy variables. In other cases, we can use any notation that is the most convenient.

4 It is not possible to include all of the dummy alternatives in the same model

- If we include all alternatives of a dummy variable in the same model, we will face the "perfect collinearity" problem.

For example:

$$1 = female + male$$

$$female = male + 1$$

or

$$1 = winter + spring + summer + fall$$

$$winter = 1 - spring - summer - fall$$

- At least one alternative has to be dropped. We treat the dropped alternative as the "BASE GROUP" or "BASELINE" or "BENCHMARK GROUP".

```
. regress lwage female male married educ exper
note: male omitted because of collinearity
```

Source	SS	df	MS	Number of obs = 526		
Model	54.3265253	4	13.5816313	F( 4, 521) = 75.27		
Residual	94.0032262	521	.180428457	Prob > F = 0.0000		
Total	148.329751	525	.28253286	R-squared = 0.3663		
				Adj R-squared = 0.3614		
				Root MSE = .42477		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.3251146	.0377061	-8.62	0.000	-.3991892	-.25104
male	0	(omitted)				
married	.1380145	.0411197	3.36	0.001	.0572338	.2187953
educ	.0872644	.0071554	12.20	0.000	.0732075	.1013213
exper	.0076213	.0015314	4.98	0.000	.0046129	.0106297
_cons	.4690918	.1040575	4.51	0.000	.264668	.6735156

## 5 Using dummy variables for multiple categories

**Case 1** We can use many dummy variables in the same model

Consider a model which includes 2 dummy variables– *female* and *married*.

$$\log(\text{wage}) = \beta_0 + \delta_0 \text{female} + \delta_1 \text{married} + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{tenure} + \beta_5 \text{tenure}^2 + u.$$

```
regress lwage female married educ exper expersq tenure tenursq
```

Source	SS	df	MS			
Model	65.6482326	7	9.37831895	Number of obs =	526	
Residual	82.6815188	518	.159616832	F( 7, 518) =	58.76	
Total	148.329751	525	.28253286	Prob > F =	0.0000	
				R-squared =	0.4426	
				Adj R-squared =	0.4351	
				Root MSE =	.39952	

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.2901838	.0361121	-8.04	0.000	-.3611279	-.2192396
married	.0529219	.0407561	1.30	0.195	-.0271456	.1329894
educ	.0791547	.0068003	11.64	0.000	.0657952	.0925143
exper	.0269535	.0053258	5.06	0.000	.0164907	.0374163
expersq	-.0005399	.0001122	-4.81	0.000	-.0007603	-.0003196
tenure	.0312962	.0068482	4.57	0.000	.0178426	.0447499
tenursq	-.0005744	.0002347	-2.45	0.015	-.0010355	-.0001134
_cons	.4177837	.0988662	4.23	0.000	.2235557	.6120116

Comments:

Consider a model which includes dummy variables for each gender/marital status combination— *marrmale*, *marrfem* and *singfem*.

$$\log(wage) = \beta_0 + \delta_0marrmale + \delta_1marrfem + \delta_3singfem + \beta_1educ + \beta_2exper + \beta_3exper^2 + \beta_4tenure + \beta_5tenure^2 + u. \tag{8.1}$$

`regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq`

Source	SS	df	MS	Number of obs = 526		
Model	68.3617623	8	8.54522029	F( 8, 517) =	55.25	
Residual	79.9679891	517	.154676961	Prob > F =	0.0000	
Total	148.329751	525	.28253286	R-squared =	0.4609	
				Adj R-squared =	0.4525	
				Root MSE =	.39329	

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Comments:

**Case 2** We can use dummy variables to represent multiple categories of a variable. Consider the relationship between law school rankings and starting salaries.

$$\log(\text{salary}) = \beta_0 + \delta_0 \text{top10} + \delta_1 r11\_25 + \delta_3 r26\_40 + \delta_4 r41\_60 + \beta_1 \text{LSAT} + \beta_2 \text{GPA} + \beta_3 \log(\text{libvol}) + \beta_4 \log(\text{cost}) + u.$$

where  $\text{top10}$ ,  $r11\_25$ ,  $r26\_40$ ,  $r41\_60$  would be equal to 1 when the variable  $\text{rank}$  falls into the appropriate range.

\*\* Rank below 60 would be the base case.

```
. regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost
```

Source	SS	df	MS	Number of obs =	136
Model	9.16538532	8	1.14567316	F( 8, 127) =	120.15
Residual	1.2109665	127	.009535169	Prob > F =	0.0000
Total	10.3763518	135	.076861865	R-squared =	0.8833
				Adj R-squared =	0.8759
				Root MSE =	.09765

lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
top10	.5393428	.053542	10.07	0.000	.4333927 .6452928
r11_25	.4716199	.0390921	12.06	0.000	.3942637 .548976
r26_40	.2790977	.0346972	8.04	0.000	.2104383 .3477571
r41_60	.182382	.0283098	6.44	0.000	.126362 .238402
LSAT	.0060482	.0034919	1.73	0.086	-.0008616 .012958
GPA	.1305893	.0818678	1.60	0.113	-.0314122 .2925908
llibvol	.0725522	.0289213	2.51	0.013	.0153221 .1297824
lcost	.0249169	.0283224	0.88	0.381	-.031128 .0809619
_cons	8.363103	.4457314	18.76	0.000	7.481081 9.245125

Comments:

## 6 Interactions involving dummy variables

### Case 1 Interactions among dummies

\*\* We can use interactions among dummies to account for the effect of each combination of dummies as well:

A different way to estimate eq.(8.1) is

$$\log(wage) = \beta_0 + \delta_0 female + \delta_1 married + \delta_3 female \cdot married + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u.$$

where  $female \cdot married = female \times married$ .

```
. gen female_married = female*married
. regress lwage female married female_married educ exper expersq tenure tenursq
```

Source	SS	df	MS	Number of obs = 526		
Model	68.3617623	8	8.54522029	F( 8, 517) = 55.25		
Residual	79.9679891	517	.154676961	Prob > F = 0.0000		
				R-squared = 0.4609		
				Adj R-squared = 0.4525		
Total	148.329751	525	.28253286	Root MSE = .39329		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
married	.2126757	.0553572	3.84	0.000	.103923	.3214284
female_married	-.3005931	.071767	-4.19	0.000	-.4415838	-.1596024
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Comments:

**Case 2** Interaction between a dummy and a continuous variable

$$\log(\text{wage}) = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + \delta_1 \text{female} \cdot \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{tenure} + \beta_5 \text{tenure}^2 + u.$$

```
. gen female_educ = female*educ
. regress lwage female married female_educ educ exper expersq tenure tenursq
```

Source	SS	df	MS	Number of obs = 526		
Model	65.677852	8	8.2097315	F( 8, 517) = 51.35		
Residual	82.6518994	517	.159868277	Prob > F = 0.0000		
Total	148.329751	525	.28253286	R-squared = 0.4428		
				Adj R-squared = 0.4342		
				Root MSE = .39984		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.2197774	.1675154	-1.31	0.190	-.5488721	.1093172
married	.0529779	.0407884	1.30	0.195	-.0271535	.1331092
female_educ	-.0056186	.0130532	-0.43	0.667	-.0312625	.0200254
educ	.0813472	.0085008	9.57	0.000	.0646469	.0980476
exper	.0268542	.005335	5.03	0.000	.0163733	.0373351
expersq	-.0005375	.0001124	-4.78	0.000	-.0007583	-.0003167
tenure	.0314803	.0068669	4.58	0.000	.0179898	.0449709
tenursq	-.0005792	.0002351	-2.46	0.014	-.0010412	-.0001173
_cons	.389629	.1186101	3.28	0.001	.1566119	.6226461

Comments:

### 7 Testing for Differences in Regression Functions across Groups

- Is it reasonable to believe that the population regression function that explains the dependent variable is the same across subsamples of populations?
- For example, is it reasonable to believe that the function that explains "GPA of college athlete" is the same for male and female students?

Consider

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

where

*cumgpa* = cumulative GPA

*sat* = SAT score

*hsperc* = high school rank percentile

*tothrs* = total hours of college courses

- If we want to test whether "male" students and "female" students have the same values of  $\beta_0, \beta_1, \beta_2, \beta_3$  we can estimate the following model

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u,$$

and the null hypothesis would be

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0 \tag{8.2}$$

$$H_a : otherwise \text{ (at least one } \delta_j = 0)$$

We can use the F-test to test this type of null hypothesis:

1. The restricted model (r)

`. regress cumgpa sat hsperc tothrs`

Source	SS	df	MS	Number of obs =	732
Model	168.533658	3	56.1778861	F( 3, 728) =	74.72
Residual	547.364897	728	.751874858	Prob > F =	0.0000
				R-squared =	0.2354
				Adj R-squared =	0.2323
Total	715.898555	731	.979341389	Root MSE =	.86711

cumgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sat	.0009028	.0002079	4.34	0.000	.0004947 .0013109
hsperc	-.0063791	.0015678	-4.07	0.000	-.0094572 -.0033011
tothrs	.0119779	.0009314	12.86	0.000	.0101494 .0138064
_cons	.9291105	.2285515	4.07	0.000	.4804118 1.377809

## 2. The unrestricted model (ur)

```
. gen female_sat = female*sat
. gen female_hsperc = female*hsperc
. gen female_tothrs = female*tothrs
. regress cumgpa female sat female_sat hsperc female_hsperc tothrs female_tothrs
```

Source	SS	df	MS	Number of obs = 732		
Model	181.589407	7	25.9413439	F( 7, 724) = 35.15		
Residual	534.309148	724	.73799606	Prob > F = 0.0000		
Total	715.898555	731	.979341389	R-squared = 0.2537		
				Adj R-squared = 0.2464		
				Root MSE = .85907		

cumgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-1.113638	.528539	-2.11	0.035	-2.15129	-.0759859
sat	.0006113	.000235	2.60	0.009	.0001499	.0010727
female_sat	.0011167	.0005	2.23	0.026	.0001351	.0020984
hsperc	-.0059675	.0017765	-3.36	0.001	-.0094551	-.0024798
female_hsperc	.0000508	.0041025	0.01	0.990	-.0080035	.008105
tothrs	.0103004	.0010928	9.43	0.000	.0081549	.0124459
female_tothrs	.0055599	.0020696	2.69	0.007	.0014968	.009623
_cons	1.213984	.2648281	4.58	0.000	.6940617	1.733907

Comments:

### 7.1 We can use the "Chow statistics" to test this type of hypothesis as well

- When there are many variables in the model, adding an interaction for every explanatory variable would make the regression analysis messy.
- In which case, we can use the "Chow test" or "Chow statistic" to test the hypothesis expressed in (8.2).
- Chow statistic is a type of F-statistic.

$$F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k + 1)]}{k + 1},$$

where  $n$  is the total number of observations.

$SSR_p = SSR$  from the pooled model (include observations from both subsamples)

$SSR_1 = SSR$  from subsample 1

$SSR_2 = SSR$  from subsample 2

`regress cumgpa sat hsperc tothrs if female == 0`

Source	SS	df	MS	Number of obs =	552
Model	89.6937042	3	29.8979014	F( 3, 548) =	41.94
Residual	390.619421	548	.712809162	Prob > F =	0.0000
Total	480.313125	551	.871711661	R-squared =	0.1867
				Adj R-squared =	0.1823
				Root MSE =	.84428

cumgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sat	.0006113	.000231	2.65	0.008	.0001576 .001065
hsperc	-.0059675	.0017459	-3.42	0.001	-.0093969 -.002538
tothrs	.0103004	.001074	9.59	0.000	.0081907 .0124101
_cons	1.213984	.2602697	4.66	0.000	.7027359 1.725233

`regress cumgpa sat hsperc tothrs if female == 1`

Source	SS	df	MS	Number of obs =	180
Model	83.4816253	3	27.8272084	F( 3, 176) =	34.08
Residual	143.689727	176	.816418902	Prob > F =	0.0000
Total	227.171352	179	1.2691137	R-squared =	0.3675
				Adj R-squared =	0.3567
				Root MSE =	.90356

cumgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sat	.0017281	.0004642	3.72	0.000	.0008119 .0026442
hsperc	-.0059167	.0038895	-1.52	0.130	-.0135927 .0017594
tothrs	.0158603	.0018485	8.58	0.000	.0122122 .0195085
_cons	.1003465	.4810947	0.21	0.835	-.8491105 1.049803

Comments:

## 8 A Binary Dependent Variable (y variable): The Linear Probability Model

- So far, our Y variables are continuous.
- What if we are interested in explaining a qualitative Y variable (that is, Y is a dummy variable)?

Consider

$$\begin{aligned}y &= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u \\E(y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k,\end{aligned}$$

where  $\mathbf{x}$  denotes all of the explanatory variables  $(x_1, \dots, x_k)$ .

```
. regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6
```

Source	SS	df	MS	Number of obs = 753		
Model	48.8080578	7	6.97257969	F( 7, 745) = 38.22		
Residual	135.919698	745	.182442547	Prob > F = 0.0000		
Total	184.727756	752	.245648611	R-squared = 0.2642		
				Adj R-squared = 0.2573		
				Root MSE = .42713		

inlf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nwifeinc	-.0034052	.0014485	-2.35	0.019	-.0062488	-.0005616
educ	.0379953	.007376	5.15	0.000	.023515	.0524756
exper	.0394924	.0056727	6.96	0.000	.0283561	.0506287
expersq	-.0005963	.0001848	-3.23	0.001	-.0009591	-.0002335
age	-.0160908	.0024847	-6.48	0.000	-.0209686	-.011213
kidslt6	-.2618105	.0335058	-7.81	0.000	-.3275875	-.1960335
kidsge6	.0130122	.013196	0.99	0.324	-.0128935	.0389179
_cons	.5855192	.154178	3.80	0.000	.2828442	.8881943

where

*inlf* = 1 if the woman reports working for a wage outside the home at some point during the year, zero otherwise.

*nwifeinc* = husband's earnings (in thousands of dollars)

*educ* = years of education

*exper* = past years of labor market experience

*age* = age

*kidslt6* = number of children less than 6 years old

*kidsage6* = number of kinds between 6 - 18 years old



# Heteroscedasticity Problem

## 1 Nature and Consequences of heteroscedasticity for OLS

- Heteroskedasticity (broadly) -
  
- Heteroskedasticity (in econometrics) -

### *1.1 Nature of Heteroskedasticity*

### *1.2 Consequences of Heteroskedasticity*

1. Does not affect the biasedness of the OLS estimators

2. Does not affect the value of  $R^2$  and  $adj.R^2$

3. Make the estimated value of  $Var(\hat{\beta}_{OLS})$  wrong

4. Affect the correctness of our inference

*1.3 How can the estimated value of  $Var(\hat{\beta}_{OLS})$  be wrong?*

Suppose

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

Given that assumption 1 to 4 are true, but assumption 5 (homoskedasticity) is violated. Thus,

$$Var(u_i|x_i) =$$

And from the OLS estimation steps, we can write

$$\hat{\beta}_1 = \beta_1 +$$

*1.4 Two types of remedies*

1. Passive

## 2. Active

## 2 Testing for heteroskedasticity

- The main point -

Suppose

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

Assume that assumption 1 to 4 are true. Our hypotheses to test for heteroskedasticity would be

We know that  $Var(u|\mathbf{x}) = E(u^2|\mathbf{x}) - [E(u|\mathbf{x})]^2$ . But \_\_\_\_\_  
according to assumption 4. Thus,  $H_0$  and  $H_a$  can be written as

## 2.1 Breusch-Pagan test (BP test)

**To perform the Breusch-Pagan Test in STATA**

STATA commands (in case  $k = 4$ ):

```
regress y x1 x2 x3 x4
predict u_hat, residual
generate u_hat_sq = u_hat^2
regress u_hat_sq x1 x2 x3 x4
```

\*\* Then, check the F-statistic on the top right-hand corner of the result table.

Example: Finding the determinants of GPA.

```
. regress termgpa attend priGPA final frosh soph
```

Source	SS	df	MS			
Model	226.077541	5	45.2155081	Number of obs =	680	
Residual	142.319996	674	.211157264	F( 5, 674) =	214.13	
Total	368.397537	679	.542558964	Prob > F =	0.0000	
				R-squared =	0.6137	
				Adj R-squared =	0.6108	
				Root MSE =	.45952	

termgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
attend	.046594	.0036082	12.91	0.000	.0395093	.0536787
priGPA	.5329307	.0403281	13.21	0.000	.4537468	.6121146
final	.0503197	.0040339	12.47	0.000	.0423992	.0582403
frosh	.0974307	.0560211	1.74	0.082	-.0125662	.2074276
soph	.0689273	.0467006	1.48	0.140	-.0227689	.1606236
_cons	-1.361077	.1316861	-10.34	0.000	-1.619642	-1.102513

```

. predict u_hat, residual
. generate u_hat_sq = u_hat^2
. regress u_hat_sq attend priGPA final frosh soph
regress u_hat_sq attend priGPA final frosh soph

```

Source	SS	df	MS			
Model	8.22606613	5	1.64521323	Number of obs =	680	
Residual	76.2624962	674	.113149104	F( 5, 674) =	14.54	
Total	84.4885623	679	.124430872	Prob > F =	0.0000	
				R-squared =	0.0974	
				Adj R-squared =	0.0907	
				Root MSE =	.33638	

u_hat_sq	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
attend	-.0088079	.0026413	-3.33	0.001	-.0139941	-.0036218
priGPA	-.1454432	.029521	-4.93	0.000	-.2034074	-.0874791
final	.0061879	.0029529	2.10	0.036	.0003899	.0119859
frosh	-.1077493	.0410085	-2.63	0.009	-.1882692	-.0272294
soph	-.0975658	.0341858	-2.85	0.004	-.1646892	-.0304423
_cons	.7368933	.0963968	7.64	0.000	.5476191	.9261674

Alternatively, you can use the following set of STATA commands:

```

regress y x1 x2 x3 x4
estat hettest x1 x2 x3 x4

```

```
. regress termgpa attend priGPA final frosh soph
```

Source	SS	df	MS	Number of obs = 680		
Model	226.077541	5	45.2155081	F( 5, 674)	=	214.13
Residual	142.319996	674	.211157264	Prob > F	=	0.0000
				R-squared	=	0.6137
				Adj R-squared	=	0.6108
Total	368.397537	679	.542558964	Root MSE	=	.45952

termgpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
attend	.046594	.0036082	12.91	0.000	.0395093	.0536787
priGPA	.5329307	.0403281	13.21	0.000	.4537468	.6121146
final	.0503197	.0040339	12.47	0.000	.0423992	.0582403
frosh	.0974307	.0560211	1.74	0.082	-.0125662	.2074276
soph	.0689273	.0467006	1.48	0.140	-.0227689	.1606236
_cons	-1.361077	.1316861	-10.34	0.000	-1.619642	-1.102513

```
. estat hettest attend priGPA final frosh soph
```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: attend priGPA final frosh soph

```
chi2(5) = 93.90
Prob > chi2 = 0.0000
```

If the null hypothesis is rejected (we have the heteroskedasticity problem), we can use the "robust" option in STATA. This option gives us the correct standard error, or "heteroskedasticity-robust standard error". We can now use the t-statistics in this case.

```
. regress termgpa attend priGPA final frosh soph, robust
```

Linear regression

termgpa	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
attend	.046594	.0044101	10.57	0.000	.0379348	.0552532
priGPA	.5329307	.0426426	12.50	0.000	.4492023	.616659
final	.0503197	.0041066	12.25	0.000	.0422564	.058383
frosh	.0974307	.0633543	1.54	0.125	-.0269648	.2218262
soph	.0689273	.0520495	1.32	0.186	-.0332713	.1711259
_cons	-1.361077	.1448208	-9.40	0.000	-1.645431	-1.076723

2.2 *The White Test*

Similar to the Breush-Pagan test, but is stricter because it does not allow  $\hat{u}^2$  to be correlated with  $x^2$  or interactions among different  $x_s$ .

Suppose

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

The White Test (special case) (save degree of freedom)

1. Get  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$  by OLS.
2. Calculate  $\hat{u}_i^2 = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k)]^2$
3. Calculate  $\hat{y}_i = (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k)$
4. Calculate  $\hat{y}_i^2$
5. Estimate  $\hat{u}_i^2 = \gamma_0 + \gamma_1 \hat{y}_i + \gamma_2 \hat{y}_i^2 + \text{error}$  (keep  $R^2$  of this regression)
6.  $LM = nR^2$
7. If  $p - \text{value} > \text{significance level}$ , cannot reject  $H_0$ .

### 3 Remedial measures

As mentioned before, there are 2 types of remedies – passive and active.

- The passive remedies just re-calculate the *std.err.* or  $\hat{\beta}$  using the heteroskedasticity-robust standard error formula(s).
- The active remedies include the "weighted least squares (WLS) estimators", "generalized least squares (GLS) estimators", or "feasible GLS estimator".

#### 3.1 Weighted Least Squares (WLS)

We assume that the heteroskedasticity may take the pattern

From

$$\begin{aligned} \text{Var}(u_i|\mathbf{x}) &= E(u_i^2|\mathbf{x}) - [E(u_i|\mathbf{x})]^2 \\ &= \end{aligned}$$

We get

To make the error term become homoskedastic, we weight every term by  $\sqrt{h_i}$ .

How do we find  $h_i$  or  $\sqrt{h_i}$ , the heteroskedasticity function?

1. If the heteroskedasticity is "known" to be caused by a multiplicative constant, we can adjust using that constant.
2. If the heteroskedasticity pattern is not known, we can estimate it. This procedure is called "Feasible Generalized Least Squares" (also called Feasible GLS or FGLS)

### 3.2 Feasible GLS

Since  $Var(\hat{\beta}_j)$  would not be unbiased, we can make valid inferences about  $\beta_j$ , e.g. can use t-test, F-test, etc.

Feasible GLS in practice

1. Get  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$  by OLS. (regress  $y$   $x_1$   $x_2$  ...  $x_k$ )
2. Calculate  $\hat{u}_i = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_k x_k)]$  (predict `u_hat`, residual)
3. Create  $\log(\hat{u}_i^2)$  (generate `log_u_sq = log(u_hat^2)`)
4. Estimate  $\log(\hat{u}_i^2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k + error$  (regress `log_u_sq`  $x_1$   $x_2$  ...  $x_k$ )
5. Obtain the fitted value of  $\widehat{\log(\hat{u}_i^2)}$ , called  $\hat{g}$ . (predict `g_hat`, `xb`)
6. Create  $\hat{h} = \exp(\hat{g})$  (generate `h_hat = exp(g_hat)`)
7. Divide  $y$  and each  $x_{ij}$  by  $\sqrt{\hat{h}}$
8. Estimate  $\frac{y}{\sqrt{\hat{h}}} = \frac{\lambda_0}{\sqrt{\hat{h}}} + \lambda_1 \frac{x_1}{\sqrt{\hat{h}}} + \lambda_2 \frac{x_2}{\sqrt{\hat{h}}} + \dots + \lambda_k \frac{x_k}{\sqrt{\hat{h}}} + \frac{error}{\sqrt{\hat{h}}}$

Steps 7 & 8 on STATA would be: `regress y x1 x2 ... xk [aweight =  $\frac{1}{\sqrt{\hat{h}}}$ ]`

### 3.3 What if the assumed $h_i$ function is wrong?



## Lab 2 – Dummy, Heteroskedasticity, Specification Issues

### 1 Does "beauty" help increase wage?

1. Download the datafile "beauty.xlsx" from your EE325 Moodle page.
2. Open the STATA software program. Click on the "Data Editor" icon.
3. Copy the entire dataset from the excel file and paste it onto the STATA's Data Editor page.  
Choose "Treat first row as variable names".

4. Save the new STATA dataset.

Choose File -> Save As -> (then name the new dataset "beauty\_stata")

5. Open a new do-file and save it.

Choose "New Do-file Editor" icon

On the Do-file's top panel, choose File -> Save As -> (then name the new do-file "second\_stata\_lab")

- **To explore and understand the data**

6. type: sum
7. type: codebook
8. type: tab look
9. type: tab look female

- **We want to test whether the fraction of above-average-looking women and men are the same**

$$abvavg = \beta_1 female$$

$$H_0 : \beta_1 = 0.5$$

$$H_0 : \beta_1 \neq 0.5$$

10. type: gen abvavg = 0
11. type: replace abvavg = 1 if look > 3
12. type: regress abvavg female, nocon

13. Can we reject  $H_0$ ?

- We want to test whether "good look" has a positive impact on wage

14. type: gen belavg = 0

15. type: replace belavg = 1 if look < 3

16. type: gen log\_wage = log(wage)

17. regress log\_wage belavg abvavg

- Seems like we may have the omitted variable bias. Let's take into account other variables.

18. type: regress log\_wage abvavg belavg educ

19. type: regress log\_wage abvavg belavg educ exper expersq

20. type: regress log\_wage abvavg belavg educ exper expersq bigcity

21. type: regress log\_wage abvavg belavg educ exper expersq bigcity black

22. type: regress log\_wage abvavg belavg educ exper expersq bigcity black union

23. type: regress log\_wage abvavg belavg educ exper expersq bigcity black union female

- Check if we have the heteroskedasticity problem. (Let's use the White Test (special case))

24. type: regress log\_wage abvavg belavg educ exper expersq bigcity black union female

25. type: predict u\_hat, resid

26. type: predict y\_hat, xb

27. type: gen u\_hat\_sq = u\_hat^2

28. type: gen y\_hat\_sq = y\_hat^2

29. type: regress u\_hat\_sq y\_hat y\_hat\_sq

30. Calculate  $LM = nR^2$

31. Do we reject  $H_0$  : homoskedasticity at 5% level of confidence?

32. Now, try using the ready-made test by STATA
33. type: regress log\_wage abvavg belavg educ exper expersq bigcity black union female
34. type: estat hettest
35. Do we reject  $H_0$  : homoskedasticity at 5% level of confidence?
- **Should we believe that the value of  $\beta$  are the same for female and male?** (Chow Test)
  - Chow statistic is a type of F-statistic  $F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k + 1)]}{k + 1}$
36. To get  $SSR_p$ : regress log\_wage abvavg belavg educ exper expersq bigcity black union
37. To get  $SSR_1$ : regress log\_wage abvavg belavg educ exper expersq bigcity black union if female == 0
38. To get  $SSR_2$ : regress log\_wage abvavg belavg educ exper expersq bigcity black union if female == 1
39. What is the value of the F-statistic? Can we reject  $H_0$  (can use the same model)?

## 2 Fixing Heteroskedasticity

1. Download the "GPA1.DTA" dataset from your EE425 Moodle page and open it in STATA.  
Choose File -> Open -> (then direct to the location of the file)
2. type: des
3. type: regress colGPA hsGPA ACT skipped PC
  - **Check if we have the heteroskedasticity problem. (Let's use the White Test (special case))**
4. type: predict u\_hat, resid
5. type: predict y\_hat, xb
6. type: gen u\_hat\_sq = u\_hat^2

7. type: gen y\_hat\_sq = y\_hat^2
8. type: regress u\_hat\_sq y\_hat y\_hat\_sq
9. Calculate  $LM = nR^2$
10. Do we reject  $H_0$  : homoskedasticity at 5% level of confidence?
  - **Now we think the fitted value in #8. is a reasonable candidate for  $\hat{h}_i$**  So,
11. type: predict h\_hat, xb
12. To check if  $\hat{h}_i$  are all positive type: sum h\_hat
13. type: gen sqrt\_h = h\_hat^0.5
14. type: gen wcolGPA = colGPA/sqrt\_h
15. type: gen whsGPA = hsGPA/sqrt\_h
16. type: gen wACT = ACT/sqrt\_h
17. type: gen wskipped = skipped/sqrt\_h
18. type: gen wPC = PC/sqrt\_h
19. type: gen w = 1/sqrt\_h
20. type: regress colGPA hsGPA ACT skipped PC
21. type: regress wcolGPA w whsGPA wACT wskipped wPC, nocon
22. type: regress wcolGPA w whsGPA wACT wskipped wPC, nocon robust

### 3 Labor Force Participation of Female

1. Download the "MORA.DTA" dataset from your EE325 Moodle page and open it in STATA.  
Choose File -> Open -> (then direct to the location of the file)
2. type: des
3. type: regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6
4. type: estat hettest
5. type: regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6, robust
6. type: predict y\_hat, xb
7. type: twoway scatter inlf educ || line y\_hat educ
8. type: sort educ

9. type: twoway scatter inlf educ || line y\_hat educ

- **We need to "hold other things constant". Suppose  $nwifeinc = 30$ ,  $exper = 10$ ,  $age = 35$ ,  $kidslt6 = 0$ ,  $kidsage6 = 0$ .**

10. type: gen y\_new = 0.585 + 30\*(-0.0034) + educ\*(0.0379) + 10\*(0.0395) + 100\*(-0.0006) + 35\*(-0.0161)

11. type: twoway scatter inlf educ || line y\_new educ



# Serial Correlation and Heteroskedasticity in Time Series Regressions

## 1 The Nature of Time Series Data

**TABLE 10.1**

Partial Listing of Data on U.S. Inflation and Unemployment Rates, 1948–2003

Year	Inflation	Unemployment
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
⋮	⋮	⋮
1998	1.6	4.5
1999	2.2	4.2
2000	3.4	4.0
2001	2.8	4.7
2002	1.6	5.8
2003	2.3	6.0

## 2 Examples of Time Series Regression Models

There are many time series regression models. Different models would be suitable for different types of relationship we want to estimate. Some examples of time series models are Static Model, AR (Autoregressive), ADL (Autoregressive Distributed Lag), FDL (Finite Distributed Lag), ARMA (Autoregressive Moving Average), ARCH (Autoregressive Conditional Heteroskedasticity) etc.

In this class we will talk about 2 examples 1) Static Models and 2) FDL.

### *2.1 Static Models*

Studies a contemporaneous (occurring in the same period of time) relationship of variables.

For example:

### *2.2 Finite Distributed Lag Models*

For example,

In general,

$$y_t = \alpha_0 + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + u_t$$

$$\delta_0 = \frac{dy_t}{dx_t}$$

$$\delta_1 = \frac{dy_t}{dx_{t-1}}$$

### 3 Properties of OLS under classical assumptions

**Assumption TS1.** Linear in Parameter –  $Y$  is linear in  $X$ .

**Assumption TS2.** No Perfect Collinearity

**Assumption TS3.** Zero Conditional Mean

**\*\*\* Under Assumptions TS1 to TS3,  $\hat{\beta}_{OLS}$  would be unbiased \*\*\***

**Assumption TS4.** Homoskedasticity

**Assumption TS5.** No Serial Correlation

**\*\*\* Under Assumptions TS1 to TS5,  $\hat{\beta}_{OLS}$  would be BLUE (best linear unbiased estimators)\*\*\***

The variance of  $\hat{\beta}_{OLS}$  (under ass.TS1-TS5)

#### 4 Properties of OLS with Serially Correlated Errors

#### 5 Unbiasedness and Consistency

#### 6 Efficiency and Inference

With serial correlation,  $\hat{\beta}_{OLS}$  would not be BLUE ( $var(\hat{\beta}_{OLS})$  would not be minimized).  
Consider

$$u_t = \rho u_{t-1} + e_t ; t = 1, 2, \dots, n \quad \text{and} \quad |\rho| < 1$$

where  $u_t$  is from a regression model

$$y_t = \beta_0 + \beta_1 x_t + u_t.$$

..

## 7 Testing for Serial Correlation

Given the model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + u_t$$

### 7.1 A "t-test" for AR(1) serial correlation with strictly exogenous regressors

The most common type of serial correlation or autocorrelation is the AR(1) type:

To perform the test:

1. Estimate  $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + u_t$
2. Obtain  $\hat{u}_t, \hat{u}_{t-1}; \forall t = 1, 2, \dots, n$
3. Estimate  $\hat{u}_t = \rho \hat{u}_{t-1} + error$

4. Perform the  $t$  – test for

### 7.2 The Durbin-Watson Test (*DW test*)

This implies

$$\begin{aligned}\hat{\rho} = 0 &\Rightarrow DW = 2 \\ \hat{\rho} > 0 &\Rightarrow DW < 2 \\ \hat{\rho} < 0 &\Rightarrow DW > 2\end{aligned}$$

$H_o$  : no positive autocorrelation, serial-correlation

$H_a$  : no negative serial correlation

To perform the test:

1. Estimate  $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + u_t$
2. Obtain  $\hat{u}_t, \hat{u}_{t-1} ; \forall t = 1, 2, \dots, n$
3. Calculate  $DW$  from eq.(2)
4. Find the critical  $d_L$  and  $d_u$  values (say, at the 5% level of significance) for the given sample size and # of regressors.
5. Follow the decision rule in the picture.

Example:

Suppose the calculated value of  $DW = 0.80$ ,  $n = 45$ ,  $k = 4$ .

From this, we get  $d_L = \underline{\hspace{2cm}}$  and  $d_u = \underline{\hspace{2cm}}$

### 7.3 Testing for AR(1) serial correlation "without" strictly exogenous regressors

7.4 *Testing for AR(q) serial correlation "without" strictly exogenous regressors*

## 8 Correcting for serial correlation

### 8.1 *Passive way*

Use the type of standard error that is robust to the serial correlation, autocorrelation problem

### 8.2 *Active way –*



# Multicolinearity

## 1 The Nature of Multicolinearity

- 

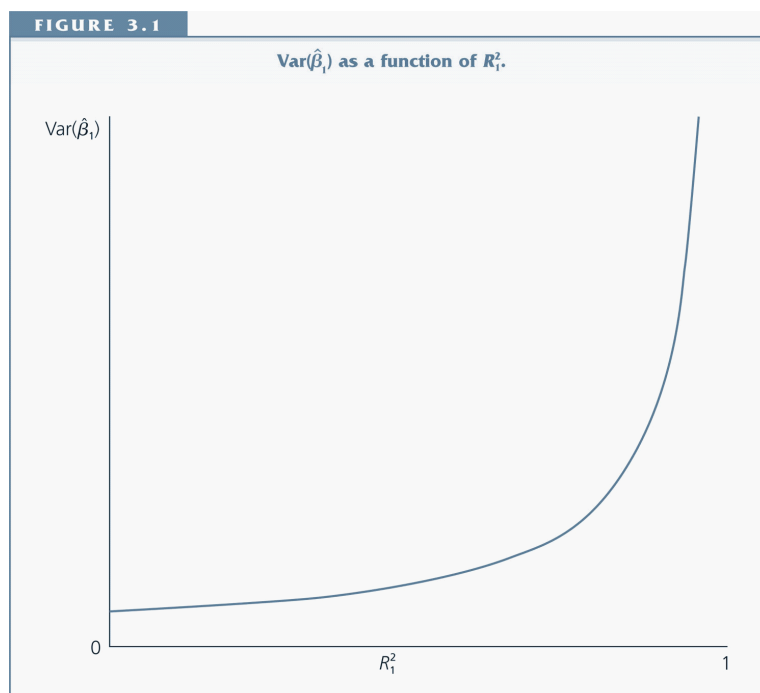
observation	$x_1$	$x_2$	$3x_1 - x_2$
1	6	18	0
2	12	36	0
3	7	21	0
4	-5	-15	0

observation	$x_1$	$x_2$	$3x_1 - x_2$
1	6	16	-2
2	12	45	9
3	7	18	-3
4	-5	-12	3

## 2 Consequences of Multicolinearity

2.1 *The OLS estimator will still be BLUE.*

2.2 *The variances and covariances will be very large. This makes precise estimation difficult.*



### 3 Detection of multicollinearity

1. There is conflicting test between t- and F-test: if we find that the conclusion derived from the two tests are inconsistent, specifically  $R^2$  is high and F-test results in statistical overall significance; whereas, at least, one null hypothesis of some t-tests cannot be rejected, it is reasonable to suspect the multicollinearity problem.
2. Correlation of regressors is greater than 0.8: the higher the correlation, the higher the variance of estimators.
3. Variance inflation factor (VIF) is greater than 10: when the regressors face the multicollinearity problem, the value of VIF might be so high that the resulting high variance of estimators adversely affects the regression analysis.
  - The VIF (variance inflation factor) to detect high multicollinearity:

4. Scatter plot of two regressors is relatively linear: when we plot the value of one regressor against another and we find that both of them tend to change in the same way, this fact might suggest the existence of multicollinearity.

## 4 Remedial Measures

1. Do nothing
  
2. Apply prior relationship among explanatory variables -
  
3. Discard some explanatory variables - the removal of the variables could mitigate the problem; but, another problem, namely specification bias problem, might occur instead. For example, suppose we want to construct the model where the production is the explained variables; and labor and capital are the explanatory ones. If there is linear relationship between labor and capital, the elimination of one variable might assuage the multicollinearity problem, but might be contrary to economic reasoning. Hence, the decision of which variables will be disposed of should be based on economic theory.
  
4. Collect more observations - this practice will increase  $\frac{1}{n}$ , which is the component of the variances. As a result, the variances will be lower despite high correlation among explanatory variables.
  
5. Transform the variables - although there is linear relationship among explanatory variables, it is not necessary that the first difference or ratio transformation of the variables will have that relationship