Lecture 5
CHAPTER 2: TWO-VARIABLE REGRESSION ANALYSIS

2.1 Example

In order to understand two-variable regression, consider the data given in Table 1.

The data in the below table refer to a total Population of 42 families with their weekly income (X) and weekly consumption expenditure (Y).

Table 1. Weekly family Expenditure (Y), Baht and Income (X), Baht

	X=Weekly family Income, Baht					
	500	600	700	800	900	1000
	360	376	458	610	600	700
	313	475	422	468	531	679
	322	380	498	575	670	730
Y= Weeky	310	382	560	542	630	591
Family Expenditure	390	390	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
Notes -						

Conditional expected value of weekly consumption expenditure given the income level =X , $E(Y|X)$

$E(Y|X=500) = 350$, $E(Y|X=900) = 590$

Unconditional expected value ,$E(Y)$ $= \dfrac{2800+2870+2820+3710+3540+5200}{42}$

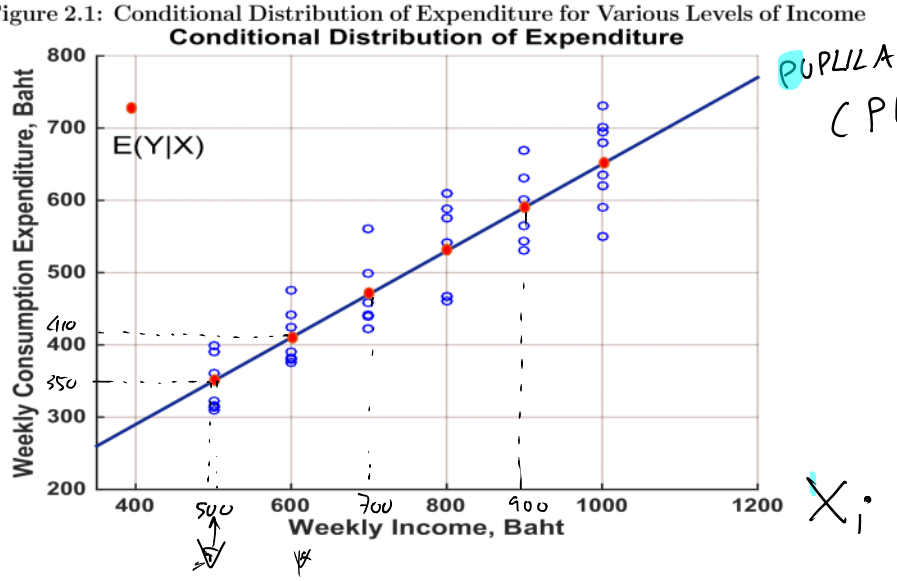
$= 498.5714$

Table 2. Conditional Probabilities $p(Y|X_i)$ for the Weekly Family Income (X) and Expenditure (Y)

	X=Weekly family Income, Baht					
	500	600	700	800	900	1000
Y= Weekly Family Expenditure	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

Notes -

Figure 2.1: Conditional Distribution of Expenditure for Various Levels of Income



Yi

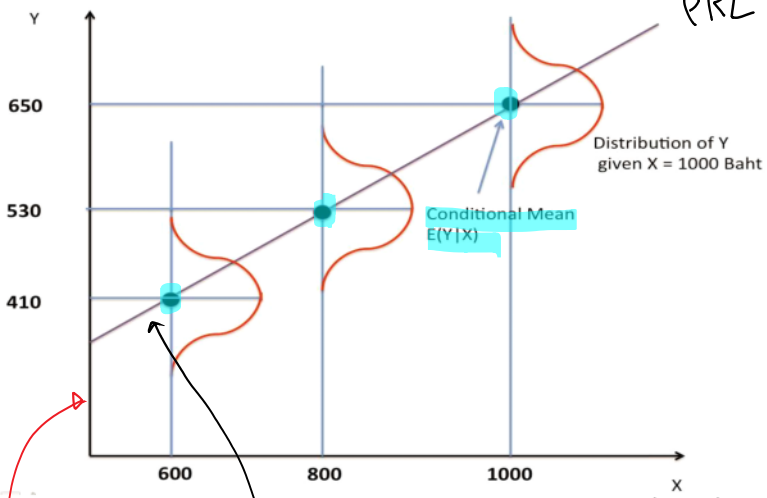
POPULATION REGRESSION LINE (PRL)

Xi

① AS INCOME INCREASES, E(Y|X) INCREASES.

② AT A GIVEN LEVEL INCOME, SOME SPEND HIGHER THAN ITS E(Y|X) AND SOME SPEND LOWER THAN ITS E(Y|X)

Figure 2.2: Population Regression Line (PRL)



2.2 The Concept of Population Regression Function (PRF)

The population regression function (PRF) can be written as the function of X_i:

$$E(Y|X_i) = f(X_i)$$

→ CONDITIONAL EXPECTATION FUNCTION (CEF) OR POPULATION REGRESSION FUNCTION (PRF)

What form does the function $f(X_i)$ assume?

If we assume the PRF $E(Y|X_i)$ is a linear function of X_i, we get

SCOPE OF PRF

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

INTERCEPT Y-AXIS

2.3 What is the meaning of the term LINEAR?

LINEARITY in the variables

$E(Y|X_i) = \beta_1 + \beta_2 X_i$ → LINEARITY IN VARIABLE

$E(Y|X_i) = \beta_1 + \beta_2 X_i^2$ → NOT LINEARITY IN VARIABLE SINCE X_i IS RAISED BY POWER OF 2!

LINEARITY in the parameters

$E(Y|X_i) = \beta_1 + \beta_2^2 X_i$ → NOT LINEARITY IN PARAMETERS AS WE OBSERVE β_2^2.

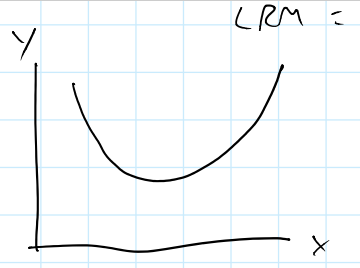
THE TERM "LINEAR REGRESSION MODEL (LRM)" REFERS TO A MODEL THAT IS "LINEAR IN PARAMETER". IT MAY BE / MAY BE NOT LINEAR IN VARIABLE.

SUMMARY

MODEL IS LINEAR IN PARAMETERS	MODEL IN YES	LINEAR IN VARIABLE YES	LINEAR IN VARIABLE NO
YES	LRM ☺	LRM ☺	LRM ☺
NO	NLRM	NLRM	NLRM

NLRM = NON LINEAR REGRESSION MODEL

LRM = LINEAR REGRESSION MODEL

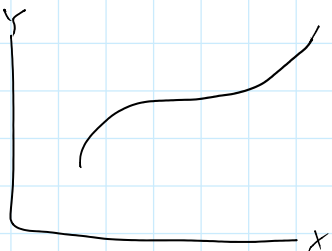


(a) QUADRATIC FORM

$Y = \beta_1 + \beta_2 X + \beta X^2$

(b) EXPONENTIAL FORM

$Y = e^{\beta_1 + \beta_2 X}$



(c) CUBIC FORM

$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3$

THESE ARE LINEAR-IN-PARAMETER FUNCTIONS

2.4 Stochastic Specification of PRF

We can write the **deviation** of an individual Y_i around it expected value as follows:

TAKE A LOOK AT 42 FAMILIES AGAIN... (TREATED AS POPULATION)

FOR AN INDIVIDUAL FAMILY, WE OBSERVE A DEVIATION FROM ITS CONDITIONAL MEAN, $E(Y|X_i)$

SO LET'S CALL THE DEVIATION FOR A FAMILY AS u_i

EX: AT $X = 500$, FOR $Y_i = 390$, WE CAN WRITE:

$Y_i \quad E(Y|X=500)$

$390 = 350 + u_i$

$Y_i \quad E(Y|X_i)$

OR $u_i = 390 - 350 = 40$

IN GENERAL,

$$u_i = Y_i - E(Y|X_i)$$

OR

$$\boxed{Y_i = E(Y|X_i) + u_i}$$

GIVEN AN INCOME LEVEL, A FAMILY'S WEEKLY EXPENDITURE (Y_i) COMPOSES OF 2 COMPONENTS, i.e.,

(1) $E(Y|X_i)$: MEAN CONSUMPTION EXPENDITURE OF ALL FAMILIES W/ THE SAME LEVEL OF INCOME

THE FIRST COMPONENT IS KNOWN AS "SYSTEMATIC" COMPONENT OR "DETERMINISTIC" COMPONENT

AND

(2) u_i: RANDOM COMPONENT OR STOCHASTIC, OR NONSYSTEMATIC COMPONENT

2.5 The roles of the stochastic disturbance term (u_i)

1. Vagueness of theory

CONSUMPTION EXPENDITURE / INCOME

BASE ON THEORY OF CONSUMPTION, Y_i DEPENDS ON X_i. HOWEVER, THE THEORY ABOVE IS INCOMPLETE. THEY ARE MANY OTHER VARIABLES THAT MIGHT INFLUENCE Y_i BUT WE EXCLUDE THEM FROM THIS MODEL.

2. Unavailability of data

SUPPOSE YOU KNOW THAT WE SHOULD INCLUDE "FAMILY WEALTH" INTO THE MODEL UNFORTUNATELY, WE USUALLY DON'T HAVE THIS DATA. ☹

3. Core variables versus peripheral variables

THERE ARE SO MANY VARIABLES THAT MIGHT AFFECT Y_i. HOWEVER, IT IS BETTER TO SELECT THE MOST IMPORTANT VARIABLES AND W/ HOPE, WHAT WE EXCLUDE FROM THE MODEL, THEIR EFFECT ARE CANCELLED OUT.

4. Intrinsic randomness in human behavior

EVEN IF YOU CAN INCLUDE ALL VARIABLE, BUT, SOME PART WILL BE LEFT UNEXPLAINED DUE TO "INTRINSIC RANDOMNESS OF HUMAN BEHAVIOR."

5. Poor proxy variable

EX: PERMANENT CONSUMPTION IS DETERMINED BY PERMANENT INCOME EXPENDITURE (MILTON FRIEDMAN'S THEORY OF CONSUMPTION) → ERRORS OF MEASUREMENT → SO u_i

6. Principle of parsimony

"KEEP YOUR REGRESSION AS SIMPLE AS POSSIBLE"

i.e., IF INCOME SUBSTANTIALLY EXPLAIN CONSUMPTION & YOU DON'T HAVE STRONG THEORETICAL FOUNDATION TO INCLUDE OTHER VARIABLES, THEN WHY WE HAVE TO INTRODUCE THEM!

7. Wrong functional form

EX: $Y_i = \beta_1 + \beta_2 X_i + u_i$ —(1) LINEAR

$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$ —(2) NON LINEAR → SUPPOSE THIS ONE IS A CORRECT FUNCTIONAL FORM,

Lecture 6

2.6 The Sample Regression Function (SRF)

As mentioned, in the real situation, we cannot find out all the population of Y values corresponding to the fixed X's. We only have a sample of Y values corresponding to some fixed X's.

Therefore, our goal in this section is to estimate the population regression line (PRF) on the basis of the **SAMPLE INFORMATION.**

As a result, for the fixed X's as given in table 1, we only have a randomly selected sample of Y values. For example, table 5 and table 19 show a random sample from the population of table 1

Table 3. A Random Sample From the Population

X	Y
500	390
600	425
700	560
800	575
900	630
1000	679

Table 4. Another Random Sample From the Population

X	Y
500	360
600	390
700	440
800	575
900	670
1000	730

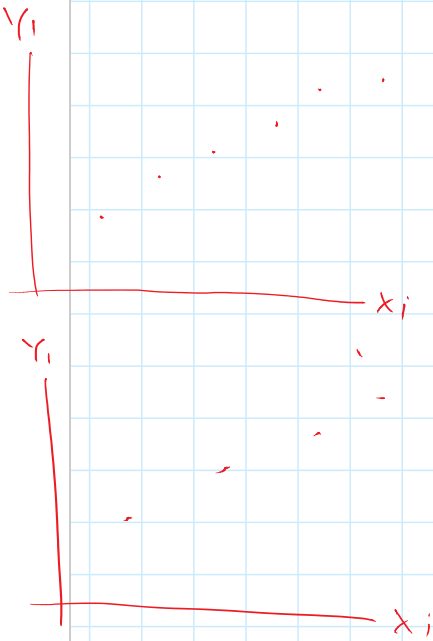
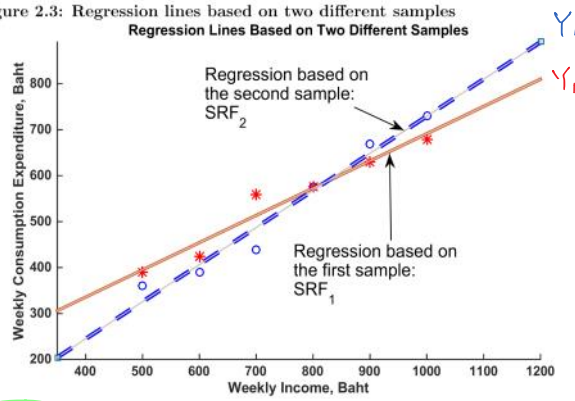


Figure 2.3: Regression lines based on two different samples



$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

The sample **regression function** (SRF) can be written as:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

where $\hat{Y}$ is read as "Y-hat"

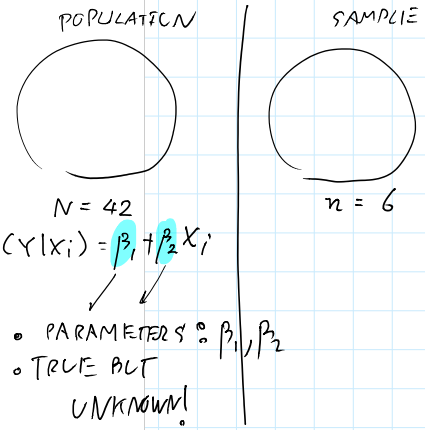
$\hat{Y}_i$ = estimator of $E(Y|X_i)$

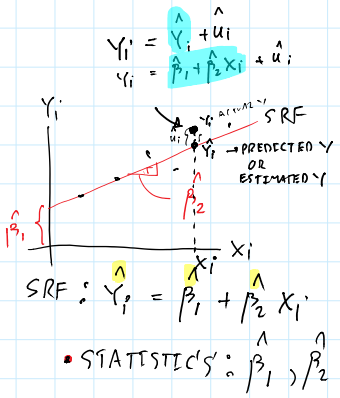
$\hat{\beta}_1$=estimator of β_1

$\hat{\beta}_2$=estimator of β_2

We can express the SRF in its stochastic form as follows:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$





Lecture Note: EE 325-2/2015: Introductory Econometrics—page—43

In sum, our ultimate goal is to estimate the PRF

$(N = 42)$ $E(Y|X_i) = \beta_1 + \beta_2 X_i$

on the basis of the SRF

$(n = 6)$ $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$

STOCHASTIC FORM OF PRF

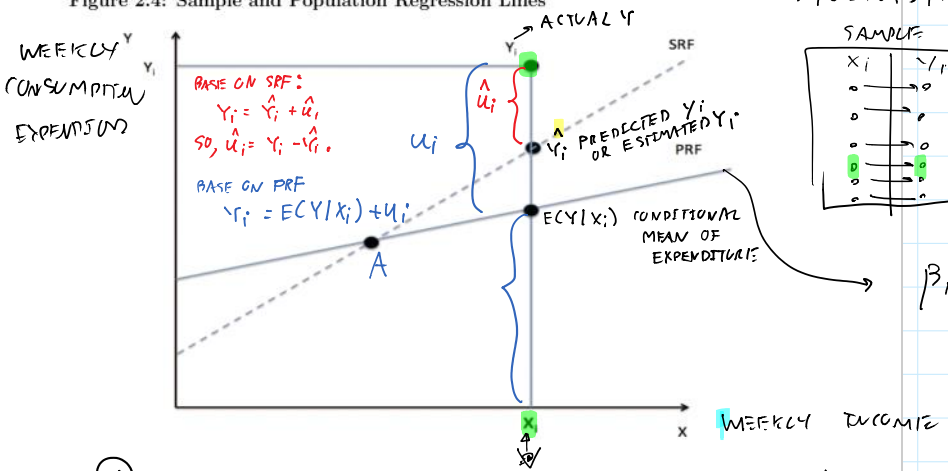
$Y_i = E(Y|X_i) + u_i$

$Y_i = \beta_1 + \beta_2 X_i + u_i$

PRF (STOCHASTIC FORM)

SRF (STOCHASTIC FORM)

Figure 2.4: Sample and Population Regression Lines



SAMPLE

X_i	Y_i
D	d

β_1, β_2 $\hat{\beta}_1, \hat{\beta}_2$

(1) ∘ ON THE RIGHT HAND SIDE OF A, GIVEN OUR SAMPLE, WE OVERESTIMATE $E(Y|X_i)$

(2) ∘ ON THE LEFT HAND SIDE OF A, GIVEN OUR SAMPLE, WE UNDERESTIMATE $E(Y|X_i)$

Lecture Note: EE 325-2/2015: Introductory Econometrics—page—44

CHAPTER 3: TWO-VARIABLE REGRESSION MODEL: THE PROBLEM OF ESTIMATION

As mentioned in the previous chapter, our main objective is to estimate the population regression function (PRF) based on the basis of the sample regression function (SRF) as accurately as possible.

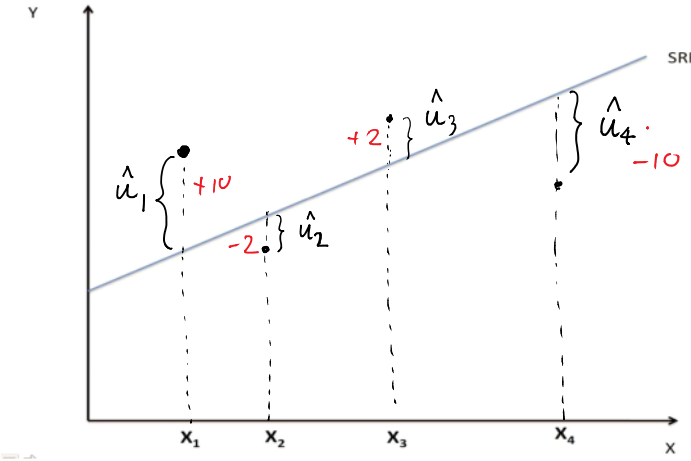
In this chapter, we are going to discuss two methods of estimation:

(1) Ordinary Least Squares (OLS) and

(2) Maximum Likelihood (ML).

3.1 The Method of Ordinary Least Squares (OLS)

Figure 3.1: Least-Squares Criterion



Lecture Note: EE 325-2/2015: Introductory Econometrics—page—45

SRF : $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

RECALL THAT

$Y_i = \hat{Y}_i + \hat{u}_i$

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

RESIDUALS OR ERRORS

PRF : $Y_i = \beta_1 + \beta_2 X_i + u_i$

SRF : $Y_i = \hat{Y}_i + \hat{u}_i$

$= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$

SO $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$.

CRITERIA TO CHOOSE THE BEST SRF THAT WE CAN USE TO ESTIMATE PRF ?

OPTION (1) $\sum_{i=1}^{n} \hat{u}_i = 0$. → $\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \hat{u}_4 + \dots + \hat{u}_n = 0$

OPTION (2) $\sum \hat{u}_i^2 = 0$ → $\hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \dots + \hat{u}_n^2 = 0$

WE SHOULD CHOOSE THE SRF SUCH THAT $\sum \hat{u}_i^2$ IS AS SMALLEST AS POSSIBLE.

i.e., SUM OF SQUARED RESIDUALS

The Method to Find Out the Least-Squares Estimators: $\hat{\beta}_1$ and $\hat{\beta}_2$

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

$$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \longrightarrow \text{OBJECTIVE FUNCTION}$$

$$\min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad \text{——— (1)}$$

THE METHOD OF LEAST SQUARES : CHOOSE $\hat{\beta}_1$ AND $\hat{\beta}_2$ SUCH THAT, FOR A GIVEN SAMPLE OR A SET OF DATA, $\sum_{i=1}^{n} \hat{u}_i^2$ IS AS SMALLEST AS POSSIBLE !

F.O.C. $\hat{u}_i$

$$\frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-1) = 0 \rightarrow \sum_{i=1}^{n} \hat{u}_i = 0 \quad \text{——— (2)}$$

$$\frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_2} = 2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-X_i) = 0 \rightarrow \sum_{i=1}^{n} \hat{u}_i X_i = 0 \quad \text{——— (3)}$$

FROM (2) :

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_1 - \sum \hat{\beta}_2 X_i = 0$$

$$\sum_{i=1}^{n} Y_i - n \cdot \hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} X_i = 0$$

$$n \cdot \hat{\beta}_1 = \sum_{i=1}^{n} Y_i - \hat{\beta}_2 \sum_{i=1}^{n} X_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_2 \sum_{i=1}^{n} X_i}{n}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta}_2 \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\boxed{\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}} \quad \#$$

FROM (3) :

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(X_i) = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \hat{\beta}_1 X_i - \sum_{i=1}^{n} \hat{\beta}_2 X_i^2 = 0$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} X_i Y_i = (\bar{Y} - \hat{\beta}_2 \bar{X}) \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} X_i Y_i = \bar{Y} \sum_{i=1}^{n} X_i - \hat{\beta}_2 \bar{X} \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2$$

$$\hat{\beta}_2 \sum_{i=1}^{n} X_i^2 - \hat{\beta}_2 \bar{X} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i$$

Lecture Note: EE 325-2/2015: Introductory Econometrics—page—46

$$\sum_{i=1}^{n} x_i Y_i = \bar{Y}\sum_{i=1}^{n} x_i - \hat{\beta}_2 \bar{x}\sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_2 \sum_{i=1}^{n} x_i^2 - \hat{\beta}_2 \bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i Y_i - \bar{Y}\sum_{i=1}^{n} x_i$$

$$\hat{\beta}_2\left(\sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i}{n}\right) = \sum_{i=1}^{n} x_i Y_i - \frac{\sum_{i=1}^{n} Y_i \sum_{i=1}^{n} x_i}{n}$$

$$\frac{\hat{\beta}_2\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)}{n} = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n}$$

$$\hat{\beta}_2 = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \#$$

Page 12

Tuesday, February 02, 2016 8:54 AM

Lecture 7

From the SRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Now, we obtain the **least-squares estimators**:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X} \end{aligned} \tag{Eq.1}$$

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum Xi \sum Yi}{n\sum X_i^2 - (\sum X_i)^2} \tag{Eq.2}$$

If we define $\bar{X}$ and $\bar{Y}$ to be the sample means of X and Y. Then:

$$\begin{aligned} x_i &= (X_i - \bar{X}) \\ y_i &= (Y_i - \bar{Y}) \end{aligned} \tag{Eq.3}$$

We can have the alternative expressions for $\hat{\beta}_2$:

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} \\ &= \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2} \end{aligned} \tag{Eq.4}$$

$x_i = X_i - \bar{X}$

$y_i = Y_i - \bar{Y}$

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

Show that

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$= \frac{\sum (X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X}\bar{Y})}{\sum (X_i^2 - 2X_i \bar{X} + \bar{X}^2)}$$

$$= \frac{\sum X_i Y_i - \bar{Y}\sum X_i - \bar{X}\sum Y_i + \sum \bar{X}\bar{Y}}{\sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2}$$

$$= \frac{\sum X_i Y_i - \frac{\sum Y_i \sum X_i}{n} - \frac{\sum X_i \sum Y_i}{n} + n\frac{\sum X_i}{n}\frac{\sum Y_i}{n}}{\sum X_i^2 - 2\frac{\sum X_i \sum X_i}{n} + n\left(\frac{\sum X_i}{n}\right)^2}$$

$$= \frac{n\sum X_i Y_i - \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

EXAMPLE

Table 5. A Random Sample From the Population

X	Y
500	390
600	425
700	560
800	575
900	630
1000	679