Printout

Tuesday, January 09, 2018 9:44 AM

2. TWO-VARIABLE REGRESSION ANALYSIS

In order to understand two-variable regression, consider the data given in Table 2.1.
The data in the below table refer to a total **Population** of 42 families with their weekly income (X) and weekly consumption expenditure (Y).

At $X_i = 500$

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

$360 = 350 + u_i$

$313 = 350 + u_i$

Table 2.1: Weekly family Expenditure (Y), Baht and Income (X), Baht

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

Y_i

$E(Y|X_i)$

Notes -

① As X increases, E(Y|X) also increases.

② At a given level of income, some families spend higher than its own conditional mean of Y [E(Y|X)] and some spend less than E(Y|X).

Table 2.2: 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
	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
Y= Weekly	1/8	1/7	1/6	1/7	1/6	1/8
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	-	1/7	-	1/8
	1/8	-	-	-	-	1/8
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 , $E(Y|X)$

$E(Y|X=500) = 350$

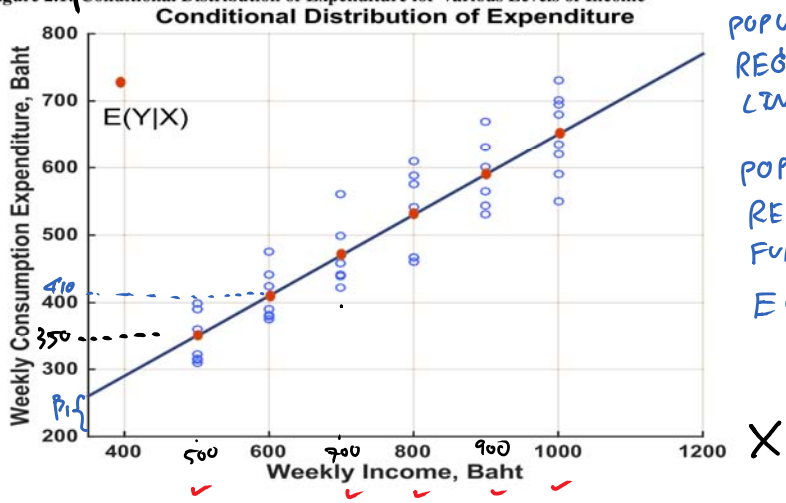
$E(Y|X=1000) = 650.$

Unconditional expected value ,$E(Y)$ → gives us "average spending" of 42 families regardless of income level.

$$E(Y) = \frac{(\ldots + \ldots + \ldots + \ldots)}{42}$$

coefficient

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



POPULATION REGRESSION LINE (PRL)

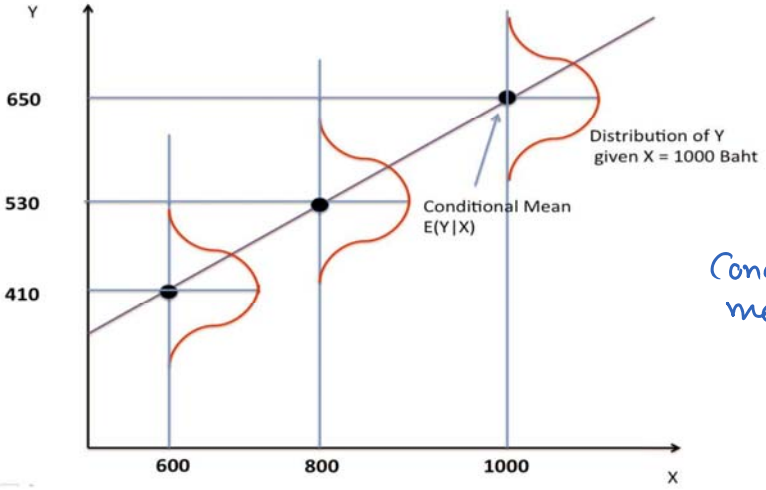
POPULATION REGRESSION FUNCTION:

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

$= \beta_1 + \beta_2 X$.

intercept coefficient — slope coefficient

Figure 2.2: Population Regression Line (PRL)



Conditional mean of Y

2.1 The Concept of Population Regression Function (PRF)

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

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

2.1.1 What form does the function f(X) assume?

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

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

$\frac{\Delta E(Y|X_i)}{\Delta X_i} = \beta_2$ → reports "marginal effect of X_i on $E(Y|X_i)$"

2.1.2 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 TO THE POWER OF TWO

$E(Y|X_i) = \beta_1 + \beta_2 \sqrt{X_i}$ NOT LINEARITY IN VARIABLE

LINEARITY in the parameters

$E(Y|X_i) = \beta_1 + \beta_2^2 X_i$

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

$E(Y|X_i) = \beta_1 + \beta_2 \cdot \beta_3 X_i$

} Not linearity in "PARAMETERS"

The linear regression model refers to a model that is "linear in parameter". It may not be linear in variable.

linear in parameters \ linear in variables	Y	N
Y	LRM ✓	LRM ✓
N	NLRM	NLRM

□

LRM = Linear regression model
NLRM = Non linear regression model

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

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

} LRM

2.2 Stochastic Specification of PRF

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

At $X = 500$,

Ex: $Y_i = \overbrace{E(Y|X=500)}^{350} + u_i$

$390 = 350 + u_i$

$u_i = 390 - E(Y|X=500)$

$u_i = 390 - 350$

$u_i = +40$

In general, $\boxed{u_i = Y_i - E(Y|X_i)}$

For a given income level (X_i), an individual family expenditure (Y_i) can be decomposed into 2 components

$Y_i = E(Y|X_i) + u_i$ → Random component or Stochastic component.

$E(Y|X_i)$: mean consumption expenditure w/ the same income level

This part is so called "systematic component" or "deterministic component"

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

2.2.1 The roles of the stochastic disturbance term (u_i)

1. Vagueness of theory

$Y_i = f(X_i)$ → Keysian consumption expenditure.

u_i acts as a representative of the omitted variables

2. Unavailability of data

$Y_i = f(X_1, X_2, X_3, X_4, X_5, \dots X_n)$

3. Core variables versus peripheral variables

4. Intrinsic randomness in human behavior

5. Poor proxy variable

Milton Friedman : Permanent Consumption = f(permanent income)

When poor proxy variables are being used, error in measurement may arise and it will be reflected by the size of u_i

6. Principle of parsimony

"Keep your regression model as simple as possible"

7. Wrong functional form

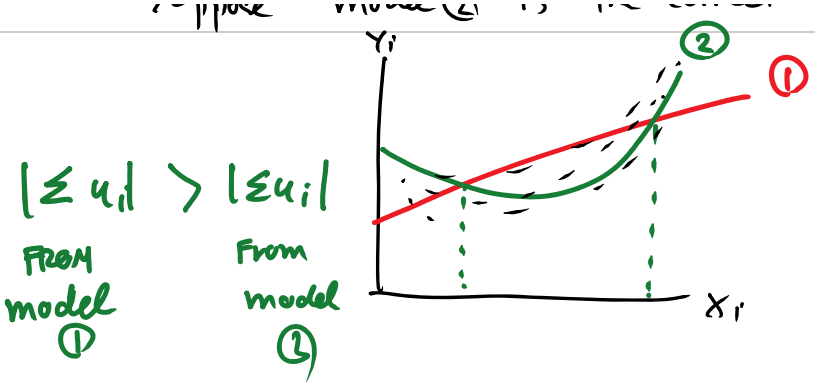
$Y_i = \beta_1 + \beta_2 X_i + u_i$ —(1) → linear

$Y_i = \beta_2 + \beta_2 X_i + \beta_3 X_i^2 + u_i$ —(2) nonlinear

Y_i = cost of production in the short run

X_i = output

Suppose model (2) is the correct functional form ...


Y_i
2
1
X_i
$|\Sigma u_i| > |\Sigma u_i|$
FROM model 1
From model 2

2.3 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 2.1, we only have a randomly selected sample of Y values. For example, table 2.3 and table 2.4 show a random sample from the population of table 2.1

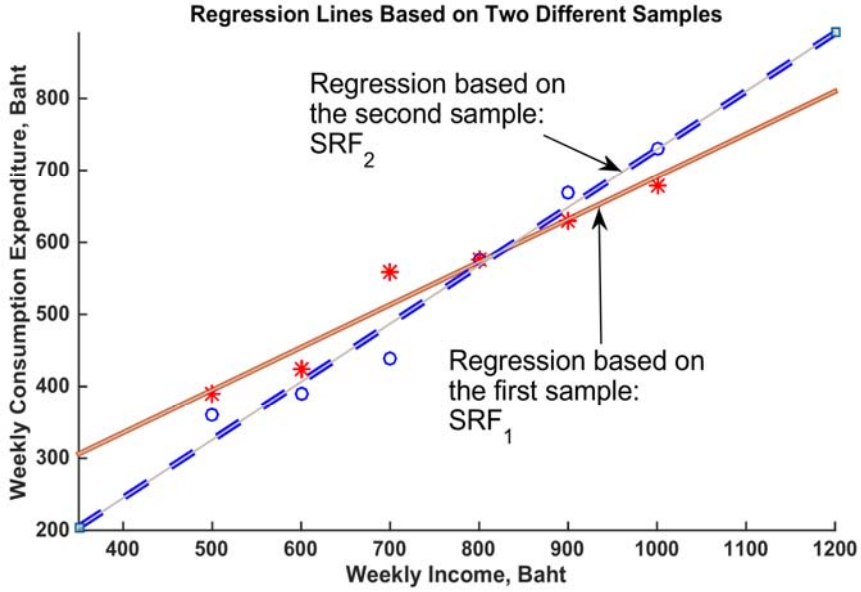
Table 2.3: A Random Sample From the Population

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

Table 2.4: 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



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:

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

In sum, our ultimate goal is to estimate
the PRF

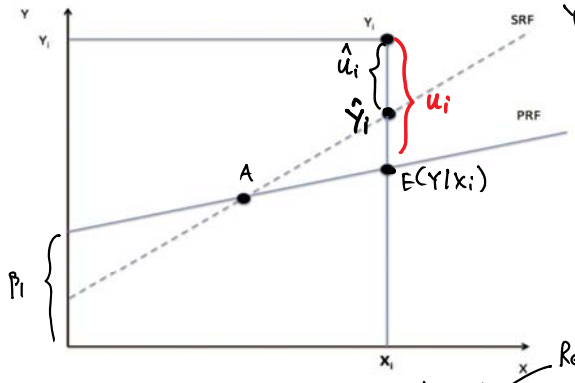
$N = 42$

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

on the basis of
the SRF

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

Figure 2.4: Sample and Population Regression Lines



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

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

Residuals

ON THE BASIS OF SRF, $Y_i = \hat{Y}_i + \hat{u}_i \rightarrow \hat{u}_i = Y_i - \hat{Y}_i$

observed Y; estimated Y; actual Y; estimated Y

ON THE BASIS OF PRF, $Y_i = E(Y|X_i) + u_i$

observed Y; conditional mean of Y; Disturbances.

Goal: get a clever method that can help us to get $\hat{\beta}_1$ as close as possible to β_1 and $\hat{\beta}_2$ "as close as" possible to β_2.

On the left hand side of A, we "underestimate" $E(Y|X_i)$ as $\hat{Y}_i < E(Y|X_i)$.

" — " Right " — ", we "overestimate" $E(Y_i|X_i)$ as $\hat{Y}_i > E(Y|X_i)$.

Note:

PRF	SRF
β_1	$\hat{\beta}_1$
β_2	$\hat{\beta}_2$
(population) parameters which are UNKNOWN!	(statistics) estimators.