2. TWO-VARIABLE REGRESSION ANALYSIS

Y f(X) — INDEPENDENT VARIABLE
DEPENDENT VARIABLE

MATHS STATISTICS ECONOMETRICS

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).

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	425	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
	405	-	-	-	-	635
Total	2800	2870	2820	3710	3540	5200
Conditional means of Y, $E(Y\|X)$	350	410	470	530	590	650
Notes -						

POPULATION SAMPLE
STATISTICAL INFERENCE

$Y_i = E(Y|X=500) + u_i$

$360 = 350 + u_1 = \beta_1 + \beta_2(500) + u_1$
$313 = 350 + u_2 = \beta_1 + \beta_2(500) + u_2$
$322 = 350 + u_3 = \beta_1 + \beta_2(500) + u_3$
$310 = 350 + u_4 = \beta_1 + \beta_2(500) + u_4$
$390 = 350 + u_5 = \beta_1 + \beta_2(500) + u_5$
$315 = 350 + u_6 = \beta_1 + \beta_2(500) + u_6$
$390 = 350 + u_7 = \beta_1 + \beta_2(500) + u_7$
$405 = (350) + u_8 = \beta_1 + \beta_2(500) + u_8$

$= \beta_1 + \beta_2(500)$

$E(u_i|X=500) = 0$,
(SEE THE PROOF IN GUJARATI.)

$E(Y) = ?$ $E(Y|X=500) = 350$
$E(Y)|X=900) = 590$
$= 498.5714$

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

② AT A GIVE LEVEL OF INCOME, SOME FAMILIES SPEND HIGHER THAT ITS OWN CONDITIONAL MEAN OF Y ($E(Y|X)$), SOME SPEND LESS THAN $E(Y|X)$

39

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
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 -						

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

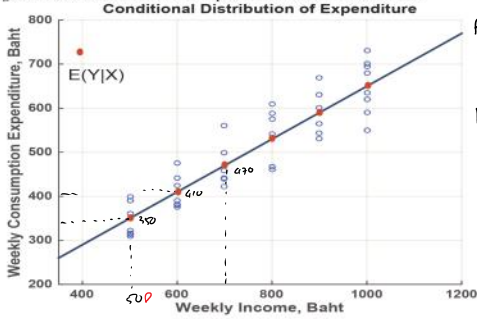
$E(Y|X=500) = 350$

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

Unconditional expected value ,$E(Y)$

$E(Y) =$

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

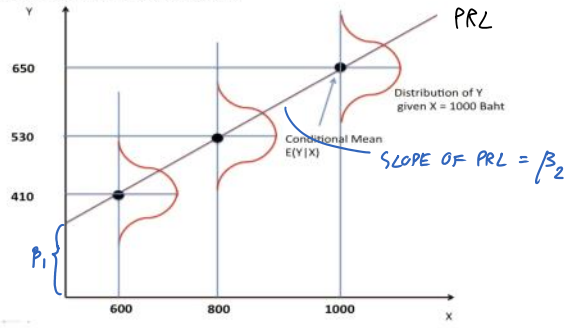


POPULATION REGRESSION LINE (PRL):

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

X_i

Figure 2.2: Population Regression Line (PRL)



PRL

SLOPE OF PRL = β_2

β_1

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

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

β_1 = INTERCEPT Y-AXIS OF PRL

β_2 = SLOPE OF THE PRL

$\frac{\Delta E(Y|X_i)}{\Delta X_i} = \beta_2$ = 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

Ex: $E(Y|X_i) = \beta_1 + \beta_2 X_i$ → 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 2

SUMMARY TO HAVE LINEARITY IN VARIABLES, X_i MUST BE RAISED TO THE POWER OF 1 ONLY.

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

LINEARITY in the parameters

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

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

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

} NOT LINEARITY IN PARAMETERS.

SUMMARY THE TERM "LINEAR REGRESSION MODEL" REFERS TO A MODEL THAT IS "LINEAR IN PARAMETER". IT MAY BE OR 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

- $Y = \beta_1 + \beta_2 X + \beta_3 X^2$ → QUADRATIC FORM
- $Y = e^{\beta_1 + \beta_2 X}$ → EXPONENTIAL FORM
- $Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3$ → CUBIC FORM

THESE ARE LINEAR-IN-PARAMETER FUNCTIONS!

PAGE 43 STOCHASTIC FORM OF PRF

TAKE A LOOK AT 42 FAMILIES AGAIN ...
FOR AN INDIVIDUAL FAMILY, WE OBSERVE "A DEVIATION" FROM ITS CONDITIONAL MEAN, $E(Y|X_i)$.
LET US CALL "THE DEVIATION" FOR A FAMILY AS u_i.
EX: AT $X = 500$, FOR $Y_i = 390$, WE CAN WRITE:

Y_i

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

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

$= 390 - 350$

$= +40$

IN GENERAL, $u_i = Y_i - E(Y|X_i)$

FOR A GIVEN INCOME LEVEL (X_i), AN INDIVIDUAL FAMILY'S WEEKLY EXPENDITURE (Y_i) CAN BE DECOMPOSED INTO 2 COMPONENTS

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

$E(Y|X_i)$: MEAN EXPENDITURE OF ALL FAMILIES W/ THE SAME INCOME LEVEL — SYSTEMATIC OR DETERMINISTIC COMPONENT

u_i: RANDOM COMPONENT OR STOCHASTIC COMPONENT

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

2.2.1 The roles of the stochastic disturbance term

1. Vagueness of theory

2.2.1 The roles of the stochastic disturbance term

1. Vagueness of theory

$Y_i = f(X_i)$ → KEYSIAN CONSUMPTION FUNCTION

THE THEORY ABOVE IS NOT COMPLETE. THERE ARE MANY VARIABLES THAT MAY AFFECT Y_i BUT WE EXCLUDE THESE VARIABLES FROM THE MODEL. SO u_i IS ACTED LIKE A REPRESENTATIVE OF THESE VARIABLES

2. Unavailability of data

IT IS ABOUT DATA AVAILABILITY.

3. Core variables versus peripheral variables

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

CORE VARIABLES — PERIPHERAL VARIABLES

$Y_i = f(X_1, X_2, X_3)$

4. Intrinsic randomness in human behavior

NO MATTER HOW HARD TO EXPLAIN Y_i, YOU CANNOT FULLY EXPLAIN IT DUE TO INTRINSIC RANDOMNESS OF BEHAVIOR.

5. Poor proxy variable

EX: PERMANENT CONSUMPTION = f(PERMANENT INCOME)

↳ IF POOR PROXY VARIABLE IS BEING USED, ERROR IN MEASUREMENT MAY ARISE AND IT WILL REFLECTED BY THE SIZE OF u_i.

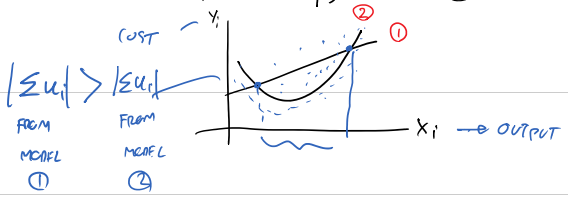
6. Principle of parsimony

" KEEP YOUR REGRESSION MODEL AS SIMPLE AS POSSIBLE "

7. Wrong functional form

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

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



$|\Sigma u_1| > |\Sigma u_2|$

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

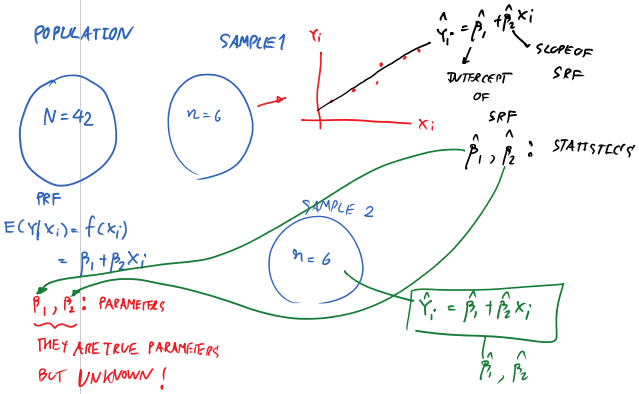
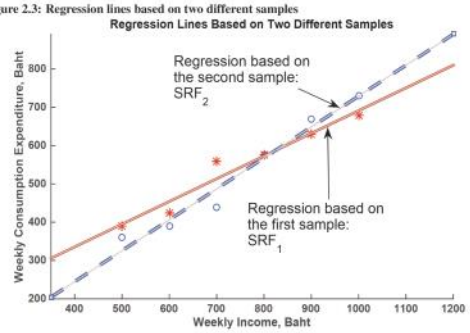


Figure 2.3: Regression lines based on two different samples



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

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

on the basis of the SRF

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

N = 42

n = 6

STOCHASTIC FORM OF PRF

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

$= \beta_1 + \beta_2 X_i + u_i$

SYSTEMATIC RANDOM

Figure 2.4: Sample and Population Regression Lines