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).

Table 2.1: 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 -

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 , $E(Y|X)$

Unconditional expected value ,$E(Y)$

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

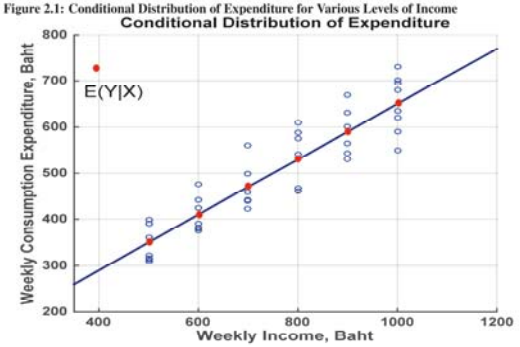
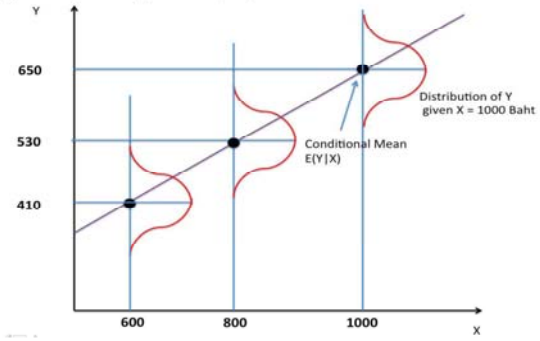


Figure 2.2: Population Regression Line (PRL)



2.1 The Concept of Population Regression Function (PRF)

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

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

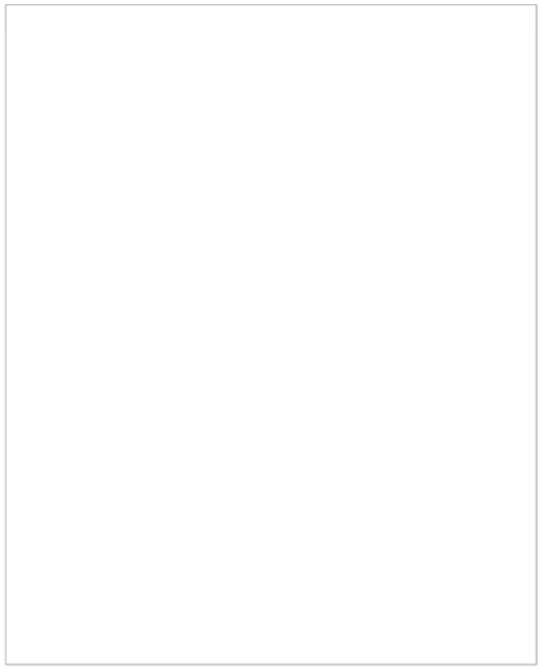
2.1.2 What is the meaning of the term LINEAR?

LINEARITY in the variables

LINEARITY in the parameters

2.2 Stochastic Specification of PRF

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



2.2.1 The roles of the stochastic disturbance term

1. Vagueness of theory

2. Unavailability of data

3. Core variables versus peripheral variables

4. Intrinsic randomness in human behavior

5. Poor proxy variable

6. Principle of parsimony

7.Wrong functional form

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

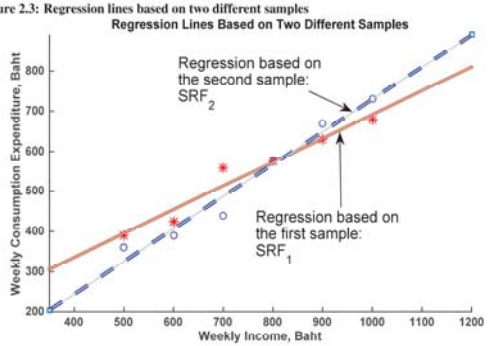
Table 2.3: Random Sample From the Population

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

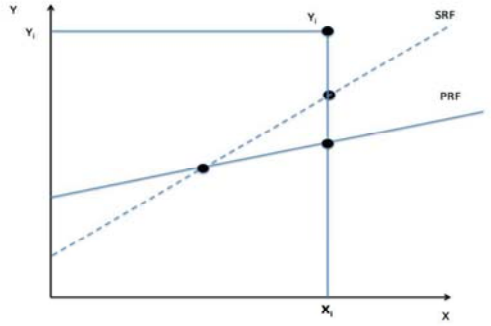
We can express the SRF in its stochastic form as follows:

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

In sum, our ultimate goal is to estimate
the PRF

on the basis of
the SRF

Figure 2.4: Sample and Population Regression Lines



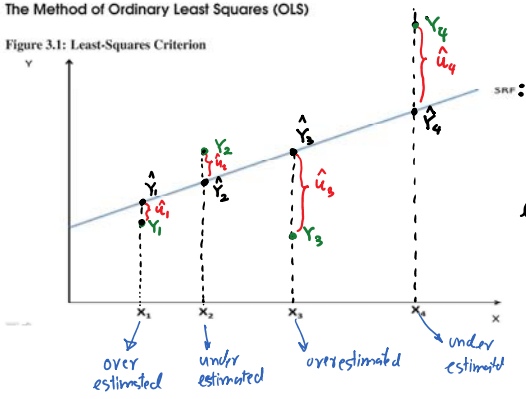
3. REGRESSION:THE PROBLEM OF ESTIMATION

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 the method of estimation:Ordinary Least Squares (OLS)

3.1 The Method of Ordinary Least Squares (OLS)

Figure 3.1: Least-Squares Criterion



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

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

ex:

$Y_1 = \hat{Y}_1 + \hat{u}_1$

$Y_2 = \hat{Y}_2 + \hat{u}_2$

$Y_3 = \hat{Y}_3 + \hat{u}_3$

$Y_4 = \hat{Y}_4 + \hat{u}_4$

actual estimated error

Goal: To find the SRF that is best fitted.

Principle of Ordinary Least Squares (OLS)

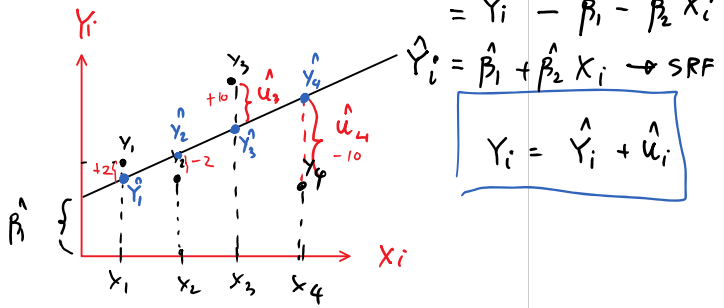
Consider SRF: $Y_i = \hat{Y}_i + \hat{u}_i$. (actual = estimated or predicted)

So $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$

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

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

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



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

OPTION 1 $\sum_{i=1}^{n} \hat{u}_i = 0 \rightarrow \hat{u}_1 + \hat{u}_2 + \hat{u}_3 - \dots + \hat{u}_n = 0$

(sum of errors is minimized) ✗ not a good idea

OPTION 2 $\sum_{i=1}^{n} \hat{u}_i^2 = 0 \rightarrow \hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \dots + \hat{u}_n^2 = 0$ ☺

We should choose the SRF such that $\sum \hat{u}_i^2$ is as smallest as possible.

$\sum \hat{u}_i^2$: sum of squared errors or sum of errors squared.

3.1 The Method of Ordinary Least Squares (OLS) 53

3.1.1 The Method to Find Out the Least-Squares Estimators β_1 and β_2

objective function

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

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

$= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$ — (1)

Goal: Choose $\hat{\beta}_1$ and $\hat{\beta}_2$ such that for a given sample

$\hat{\beta}_1, \hat{\beta}_2$ $\quad \sum_{i=1} \quad \sum_{i=1}$ $\qquad Y_i = \beta_1 + \beta_2 X_i$

$$= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad \text{---} (1)$$

Goal: Choose $\hat{\beta}_1$ and $\hat{\beta}_2$ such that for a given sample or set of data, $\sum \hat{u}_i^2$ is as smallest as possible.

F.O.C:

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

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

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

$$= -2 \sum_{i=1}^n \hat{u}_i = 0$$

$$= \sum_{i=1}^n \hat{u}_i = 0 \quad \text{---} (4)$$

$$\frac{\partial \sum_{i=1}^n \hat{u}_i^2}{\partial \hat{\beta}_2} = \frac{\partial \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2}{\partial \hat{\beta}_2} = 2 \sum_{i=1}^n (\underbrace{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}_{\hat{u}_i})((-X_i)) = 0$$

$$= \sum_{i=1}^n \hat{u}_i X_i = 0 \quad \text{---} (5)$$

From (4) $\quad \sum_{i=0}^n \hat{u}_i = 0$

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

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

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

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

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

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

54 Chapter 3. REGRESSION:THE PROBLEM OF ESTIMATION

From the SRF:

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

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

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \bar{Y} - \hat{\beta}_2 \bar{X} \qquad (3.1)$$

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \qquad (3.2)$$

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

(3.2)

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

$$x_i = (X_i - \bar{X}) \quad \rightarrow \text{called "Deviation form of X"}$$
$$y_i = (Y_i - \bar{Y}) \quad \rightarrow \text{called "Deviation form of Y"}$$

(3.3)

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

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2}$$

(3.4)

Show that

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

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

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

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

$$= \frac{\sum X_i Y_i - \frac{\sum Y_i}{n}\cdot\sum X_i}{\sum X_i^2 - \frac{2}{n}(\sum X_i)^2 + \frac{1}{n}(\sum X_i)^2}$$

$$= \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2} \quad \times\frac{n}{n} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$ #

EXAMPLE

Table 3.1: Random Sample From the Population

X	Y

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

From (5)

$$\sum_{i=1}^{n} \hat{u}_i X_i = 0$$

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

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

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

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

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

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

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

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

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

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

EXAMPLE

Table 3.1: Random Sample From the Population

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

$\hat{\beta}_1 = ?$

$\hat{\beta}_2 = ?$

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

$$\hat{Y}_i = 98.5238 + 0.5929 X_i$$ → SRF from our sample in Table 3.1

(98.5238: estimated; 0.5929: estimated)

Prediction:

- IF $X_i = 0$, $\hat{Y}_i = 98.5238$ (Autonomous spending)
- MPC = $\hat{\beta}_2 = 0.5929$

IF income rises by 1 extra baht, on average, a family would spend about 0.5929 baht.

56 Chapter 3. REGRESSION:THE PROBLEM OF ESTIMATION

Table 3.2: Raw Data Based on the Sample Data on Table 3.1

	(1) Y_i	(2) X_i	(3) Y_iX_i	(4) X_i^2	(5) $x_i = X_i - \bar{X}$	(6) $y_i = Y_i - \bar{Y}$	(7) x_i^2	(8) x_iy_i	(9) $\hat{Y}_i$	(10) $\hat{u}_i = Y_i - \hat{Y}_i$	(11) $\hat{Y}_i\hat{u}_i$
	390	500	195,000	250,000	−250	−153.17	62,500	38,291.67			
	425	600	255,000	360,000	−150	−118.17	22,500	17,725			
	560	700	392,000	490,000	−50	16.83	2,500	−841.67			
	575	800	460,000	640,000	50	31.83	2,500	1,591.67			
	630	900	567,000	810,000	150	86.83	22,500	13,025			
	679	1,000	679,000	1,000,000	250	135.83	62,500	33,958.33			
Sum	3,259	4,500	2,548,000	3,550,000	0	0	175,000	103,750			
Mean	543.17	750	424,666.67	591,666.670	0	0	29,166.67	17,291.67			

$\bar{Y}$ $\bar{X}$

0!!! $\sum \hat{u}_i = 0$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{103,750}{175,000} = 0.5929 \#$$

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

$$= 543.17 - (0.5929)(750)$$

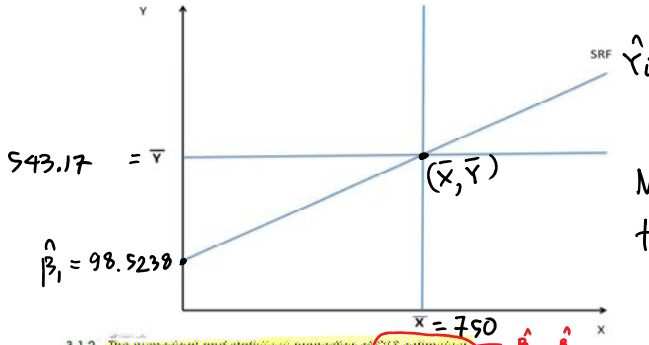
$$= 98.5238. \#$$

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

$$= \frac{6(2,548,000) - (4500)(3259)}{6(3,550,000) - (4500)^2}$$

$$= \frac{622,500}{1,050,000} = 0.5929 \#$$

Figure 3.2: Sample Regression Line Based on the Data of Table 3.2



$\hat{Y}_i = 98.5238 + 0.5928 \hat{X}_i$

$\hat{\beta}_1$ $\hat{\beta}_2$

Notice that our SRF will pass through (or will hit) $(\bar{X}, \bar{Y})$

i.e., $\hat{Y}_i = 98.5238 + 0.5929(750)$

$\hat{Y}_i = 543.1988 \simeq 543.17$ $(\bar{Y})$

In other words, $(\bar{X}, \bar{Y})$ is a member of SRF family.

3.1.2 The numerical and statistical properties of OLS estimators $\hat{\beta}_1$, $\hat{\beta}_2$

1. The OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are expressed solely in terms of the observable (Sample size) and quantities (i.e X and Y).

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \bar{Y} - \hat{\beta}_2 \bar{X} \quad (3.5)$$

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \quad \text{or} \quad \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (3.6)$$

$x_i = X_i - \bar{X}$

$y_i = Y_i - \bar{Y}$

2. They are **point estimators**.

SRF

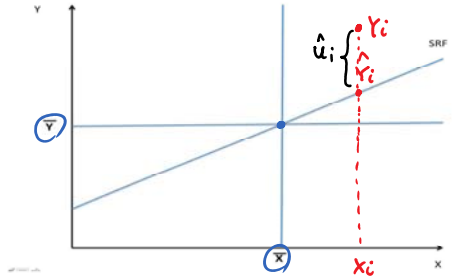
3. The regression line has the following properties.

3.1 The sample regression function (SRF) passes through the sample means of Y and X ($\bar{Y}$ and $\bar{X}$).

SAMPLE1 → $(\hat{\beta}_1, \hat{\beta}_2)$

SAMPLE2 → $(\hat{\beta}_1, \hat{\beta}_2)$

Figure 3.3: The Sample regression Line Passes through the Sample Mean Values of Y and X



Proof:

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

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

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

$$\sum Y_i = \sum \hat{\beta}_1 + \hat{\beta}_2 \sum X_i + \sum \hat{u}_i \quad (\sum \hat{u}_i = 0$$ (From Eq(4) of F.O.C)

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

$$\frac{\sum Y_i}{n} = \frac{n \cdot \hat{\beta}_1 + \hat{\beta}_2 \sum X_i}{n}$$

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}. \quad \#$$ ☺

3.2 The mean value of the estimated $Y = \hat{Y}_i$ is equal to the mean value of the actual Y, i.e. $\bar{\hat{Y}}_i = \bar{Y}$.

Proof:

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

$$\hat{Y}_i = (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i$$

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_2 (X_i - \bar{X})$$

$$\bar{\hat{Y}}_i = \frac{\sum \hat{Y}_i}{n} = \frac{\sum \bar{Y}}{n} + \hat{\beta}_2 \frac{\sum (X_i - \bar{X})}{n} = \bar{Y}$$

($\sum (X_i - \bar{X}) = 0$)

$$\sum (X_i - \bar{X}) = \sum X_i - \sum \bar{X}$$
$$= \sum X_i - n \cdot \bar{X}$$
$$= \sum X_i - n \cdot \frac{\sum X_i}{n}$$
$$= \sum X_i - \sum X_i$$
$$= 0 . \#$$

OR $\sum (X_i - \bar{X}) = \sum x_i = 0$.

3.3. The mean value of the residuals $\hat{u}_i$ is zero.

$\frac{\sum \hat{u}_i}{n} = 0$

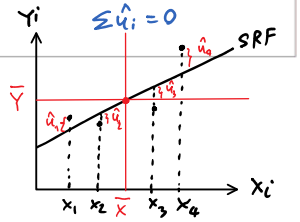
Proof; From the F.O.C:

$$-2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

($\hat{u}_i$)

$\sum \hat{u}_i = 0$

Meaning ?



$\sum \hat{u}_i = 0$

60 Chapter 3. REGRESSION:THE PROBLEM OF ESTIMATION

3.4 The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i.

i.e., $\sum \hat{y}_i \hat{u}_i = 0$

Proof: Recall that $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$ —①

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$ —②

②-①: $Y_i - \bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i - \hat{\beta}_1 - \hat{\beta}_2 \bar{X}$

$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + \hat{u}_i$

$y_i = \hat{\beta}_2 x_i + \hat{u}_i$ where $Y_i - \bar{Y} = y_i$, $X_i - \bar{X} = x_i$

$y_i = \hat{y}_i + \hat{u}_i$ where $\hat{y}_i = \hat{\beta}_2 x_i$

From $\hat{y}_i = \hat{\beta}_2 x_i$

$\hat{y}_i \cdot \hat{u}_i = \hat{\beta}_2 x_i \cdot \hat{u}_i$

$\sum \hat{y}_i \hat{u}_i = \hat{\beta}_2 \sum x_i \hat{u}_i$

$\sum \hat{y}_i \hat{u}_i = \hat{\beta}_2 \sum x_i (y_i - \hat{\beta}_2 x_i)$

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

$= \hat{\beta}_2 (\hat{\beta}_2 \sum x_i^2) - \hat{\beta}_2^2 \sum x_i^2$

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

$\boxed{\sum \hat{y}_i \cdot \hat{u}_i = 0}$ #

No any connection between $\hat{y}_i$ and $\hat{u}_i$!

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

So, $\sum x_i y_i = \hat{\beta}_2 \cdot \sum x_i^2$

3.5 The residuals $\hat{u}_i$ are uncorrelated with X_i.

$\sum \hat{u}_i x_i = 0$. Proof: Look at F.O.C. equation(5)

$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_2} = 0$

3.1.3 The Assumptions Underlying the Method of Least Squares

Assumption 1: Linear regression model

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

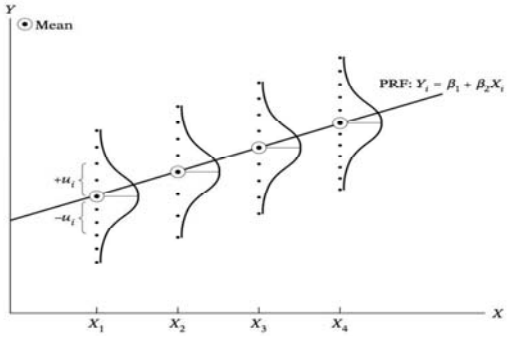
Assumption 2: X values are fixed in repeated sampling

X is assumed to be nonstochastic.

Assumption 3: Zero mean value of disturbance u_i

$$E(u_i|Xi) = 0$$

Figure 3.4: Conditional Distribution of the Disturbances u_i



Assumption 4: Homoscedasticity or Equal Variance of u_i

$$\text{var}(u_i|X_i) = E\left[u_i - E(u_i|X_i)\right]^2$$

$$= E[u_i^2|X_i] = \sigma^2.$$

In short, variance of u_i is constant. (Homoscedastic variance)

= Equal

Figure 3.5 Homoscedasticity

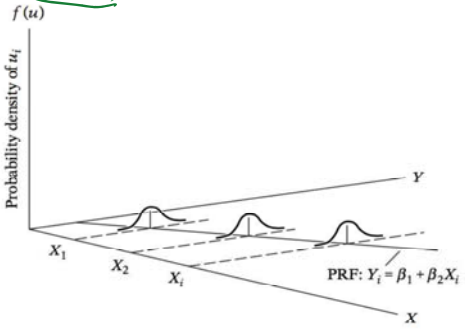
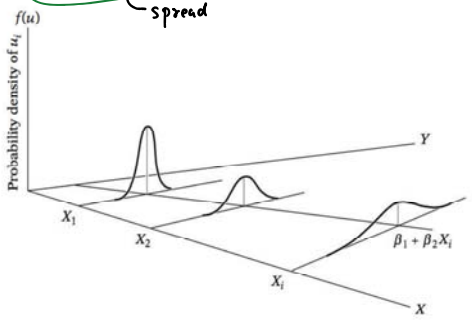


Figure 3.6: Heteroscedasticity = unequal variance

spread

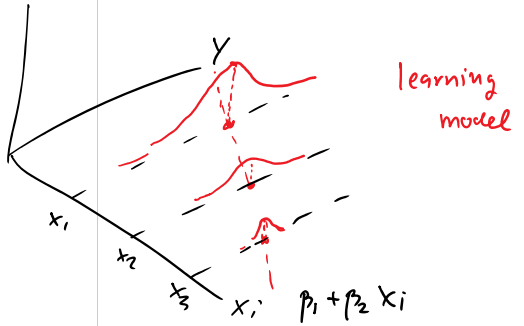


Assumption 5: No Autocorrelation Between the Disturbances

Given X_i and X_j where $i \neq j$

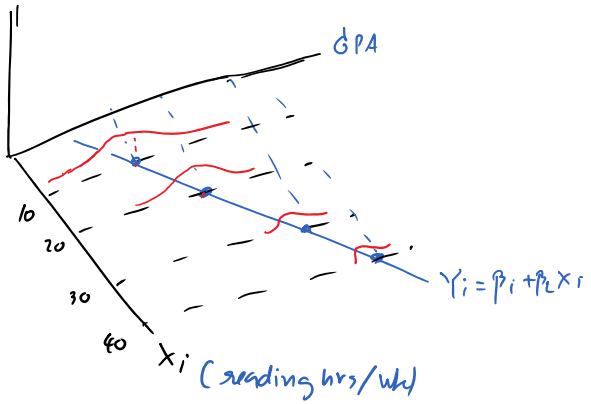
$$\text{cov}(u_i, u_j \mid X_i, X_j) = E\left[(u_i - E(u_i|X_i)) \cdot (u_j - E(u_j|X_j))\right]$$

$$= E\left[(u_i|X_i) \cdot (u_j|X_j)\right] = 0$$



learning model

- As X_i ↑, $E(Y_i|X_i)$ ↓



Assumption 6: Zero Covariance Between u_i and X_i

$$\text{Cov}(u_i, X_i) = E[(u_i - E(u_i)) \cdot (X_i - E(X_i))]$$

$$= E(u_i X_i) - E(X_i)\underbrace{E(u_i)}_{=0}$$

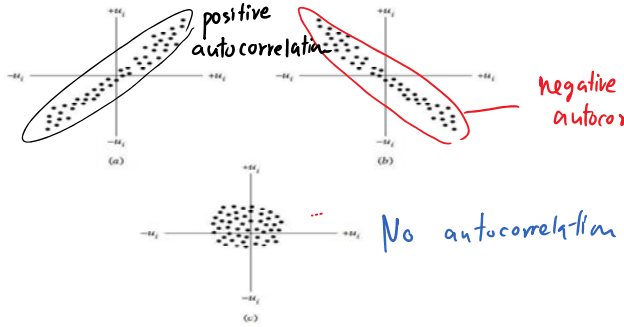
$$= E(u_i X_i) = 0.$$

No any association between u_i and X_i

: knowledge about X_i you have tells you nothing about how u_i behaves.

Figure 3.7: Patterns of Correlation Among the disturbances (Related to assumption #5)

u_i u_j



Assumption 7: The number of observations n must be greater than the number of parameters to be estimated.

$n > k$

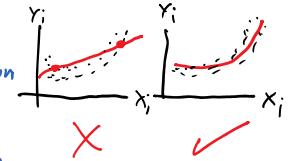
Assumption 8: Variability in X values.

in short, to predict spending behavior, you need to interview different groups of income level, not only one group!

Assumption 9: The regression model is correctly specified.

It means that when we use OLS, we need to perform a test to check if we use the correct functional form.

(Model Specification Test)



X ✓

(available in any software package (EViews, Stata))

Assumption 10: There is no perfect multicollinearity.

Independent variables you use to predict a dependent variables must not be perfectly correlated!

Ex: $Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u_i$ (SPENDING, INCOME, WEALTH)

If $X_2 = 2X_3$ then...

$Y_i = \beta_1 + \beta_2(2X_3) + \beta_3 X_3 + u_i$

$= \beta_1 + (2\beta_2 + \beta_3) X_3 + u_i$

they are mixed now ☹

⇒ You cannot estimate a separate effect of X_2 on Y and effect of X_3 on Y!

66 Chapter 3. REGRESSION:THE PROBLEM OF ESTIMATION

3.1.4 Standard Errors of Least-Squares Estimates → $var(\hat{\beta}_2) = ?$ $var(\hat{\beta}_1) = ?$

The standard errors of the OLS estimates can be obtained as follows: $se(\hat{\beta}_2) = ?$ $se(\hat{\beta}_2) = ?$

We know that

(standard deviation or standard errors)

$$\hat{\beta}_2 = \frac{\sum k_i Y_i}{\sum k_i} = \sum k_i Y_i$$

where

$$k_i = \frac{x_i}{\sum x_i^2}$$

The properties of the weights k_i

1. The k_i are nonstochastic.
2. $\sum k_i = 0$
3. $\sum k_i^2 = \frac{1}{\sum x_i^2}$
4. $\sum k_i x_i = \sum k_i X_i = 1$

Since

$$var(\hat{\beta}_2) = E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2$$

First Step

Find the $E(\hat{\beta}_2)$

sample 1 → ols → $\hat{\beta}_1, \hat{\beta}_2$

sample 2 → ols → $\hat{\beta}_1, \hat{\beta}_2$

⋮

sample n → ols → $\hat{\beta}_1, \hat{\beta}_2$

From $\hat{\beta}_2 = \sum_{i=1}^{n} k_i Y_i$

$= \sum_{i=1}^{n} k_i (\beta_1 + \beta_2 X_i + u_i)$

$= \sum_{i=1}^{n} [k_i \beta_1 + \beta_2 k_i X_i + k_i u_i]$

$= \beta_1 \underbrace{\sum_{i=1}^{n} k_i}_{=0} + \beta_2 \underbrace{\sum_{i=1}^{n} k_i X_i}_{=1} + \sum_{i=1}^{n} k_i u_i$

$\hat{\beta}_2 = \beta_2 + \sum_{i=1}^{n} k_i u_i$

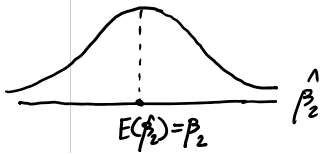
$E(\hat{\beta}_2) = E(\beta_2) + E(\sum_{i=1}^{n} k_i u_i)$

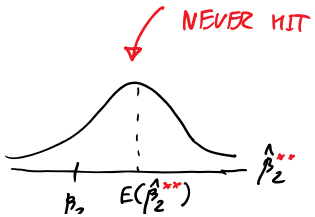
$\sum_{i=1}^{n} \underbrace{E(k_i u_i)}_{=0}$

$$\boxed{E(\hat{\beta}_2) = \beta_2}$$!!!

Therefore, $\hat{\beta}_2$ is an unbiased estimator of the true β_2 !!! ☺

3.1 The Method of Ordinary Least Squares (OLS) 67





so, $\hat{\beta}_2^{**}$ is an biased estimator.

N = 42 families (as population), 6 income groups

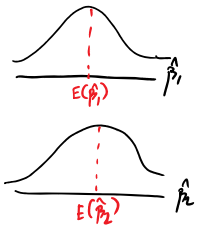
sample 1 → ols → $(\hat{\beta}_1, \hat{\beta}_2)$ SET 1

sample 2 → ols → $(\hat{\beta}_1, \hat{\beta}_2)$ SET 2

sample 3 → ols → $(\hat{\beta}_1, \hat{\beta}_2)$ SET 3

⋮

sample n → $(\hat{\beta}_1, \hat{\beta}_2)$ SET n



Solution: • $Var(\hat{\beta}_2) = \frac{\sigma_u^2}{\sum x_i^2}$ • $Var(\hat{\beta}_1) = \frac{\sigma_u^2 \sum X_i^2}{n \sum x_i^2}$

• $se(\hat{\beta}_2) = \frac{\sigma_u}{\sqrt{\sum x_i^2}}$ • $se(\hat{\beta}_1) = \sigma_u \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}$

• $cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\bar{X} \sigma_u^2}{\sum x_i^2} = -\bar{X} \cdot var(\hat{\beta}_2) = -\bar{X} \cdot \frac{\sigma_u^2}{\sum x_i^2}$

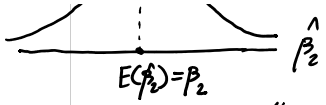
⇒ σ_u^2 is so called "variance of the disturbance term"

or "error variance"

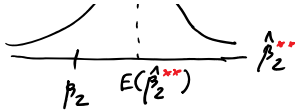
or "variance of errors"

⇒ Unfornately, σ_u^2 is unknown (since we don't have N = 42 families)

and so, we have to find a way to "estimate" it!!!



"unbiased property"



TRUE PARAMETER IS HERE, SUPPOSED

So, $\hat{\beta}_2$ is an biased estimator.

D-I-Y VERIFY THAT $E(\hat{\beta}_1) = \beta_1$.

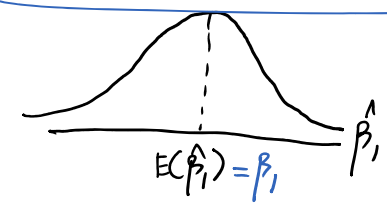
From $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$

$= \bar{Y} - \sum_{i=1}^{n} k_i Y_i \cdot \bar{X}$

$= \frac{\sum_{i=1}^{n} Y_i}{n} - \bar{X} \sum_{i=1}^{n} k_i Y_i$

$= \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{X} \cdot k_i \right) Y_i$

$= \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{X} k_i \right) (\beta_1 + \beta_2 X_i + u_i)$



So, $\hat{\beta}_1$ is an unbiased estimator.

continue

68 Chapter 3. REGRESSION:THE PROBLEM OF ESTIMATION

Second Step

Using the definition of variance

$$var(\hat{\beta}_2) = E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2$$

Solution for $var(\hat{\beta}_2)$:

$$var(\hat{\beta}_2) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$$

we can estimate him by $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$.

[see Appendix 3A.3 for the proof]

$\sum_{i=1}^{n} x_i$

[see Appendix 3A.3 for the proof]

The covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$

Solution for $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$:

$$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\bar{X} \cdot \sigma_u^2}{\sum_{i=1}^{n} x_i^2}$$

estimated by $\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}$.

$\frac{}{n-2}$.

3.1.5 The Least-Square Estimator of σ^2

$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

Solution: $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-2}$ is used to "estimate" true but unknown σ_u^2

$\hat{\sigma}_u^2$ is an UNBIASED ESTIMATOR OF σ_u^2.

true but unknown.

Remark: look at Appendix 3A.5 (in Chapter 3) Gujarati

In sum, the standard errors of the OLS estimators can be obtained as follow:

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum x_i^2} \\ \text{se}(\hat{\beta}_2) &= \frac{\sigma}{\sqrt{\sum x_i^2}} \end{aligned} \tag{3.7}$$

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2 \\ \text{se}(\hat{\beta}_1) &= \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}\sigma \end{aligned} \tag{3.8}$$

We can estimate the σ^2 from the data where the formula for the estimated σ^2 is following :

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

where

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2$$

The alternative expression for computing $\sum \hat{u}_i^2$ is

$$\sum \hat{u}_i^2 = \sum y_i^2 - \frac{(\sum x_i y_i)^2}{\sum x_i^2}$$

The covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ is:

$$\begin{aligned} \text{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X}\,\text{var}(\hat{\beta}_2) \\ &= -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right) \end{aligned} \tag{3.9}$$

3.1.6 Properties of Least-Squares Estimators: The Gauss-Markov Theorem

Given the assumptions of the classical linear regression model, the least-square estimators are satisfied the optimum properties which is known as "The Gauss- Markov Theorem." To understand this theorem, we need to know the small-sample properties of an estimator first.

estimated by

θ (TRUE BUT UNKNOWN) $\rightarrow$ $\hat{\theta}$ (OUR ESTIMATOR)

The Small-Sample Properties of An Estimator

1. Unbiasedness

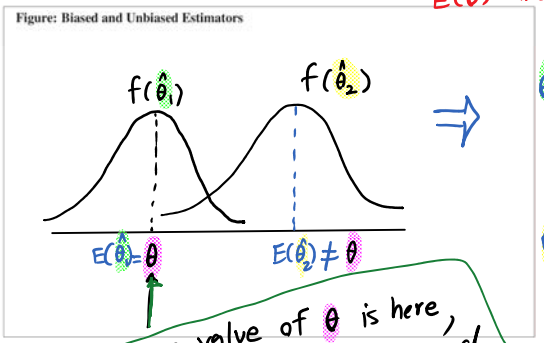
An estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if the expected value of $\hat{\theta}$ is equal to the true θ

$E(\hat{\theta}) = \theta$ in repeated samplings.

Therefore, if the expected value of $\hat{\theta}$ is not equal to the true θ, then the estimator is said to be biased. We can calculate the biased as:

$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$ — the difference between $E(\hat{\theta})$ and its true value.

Figure: Biased and Unbiased Estimators



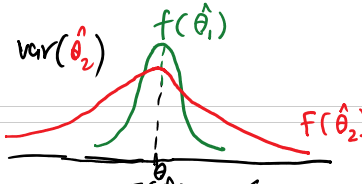
$\hat{\theta}_1$ is an unbiased estimator for θ.
(notice that $E(\hat{\theta}_1) - \theta = 0$.)
$\text{Bias}(\hat{\theta}_1) = 0$

$\hat{\theta}_2$ is a biased estimator for θ
(notice that $E(\hat{\theta}_2) - \theta \neq 0$)

JARGON TERMS

① Mean of $\hat{\theta} \rightarrow E(\hat{\theta})$
(average value of $\hat{\theta}$)
(expected value of $\hat{\theta}$)

$\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$



- $E(\hat{\theta}_1) = \theta$ ✓
- $E(\hat{\theta}_2) = \theta$ ✓

③ variance of $\hat{\theta} \rightarrow E\left[\hat{\theta} - E(\hat{\theta})\right]^2$

③ sampling error $\rightarrow \hat{\theta} - \theta$ (Estimated, true)

2. Minimum Variance

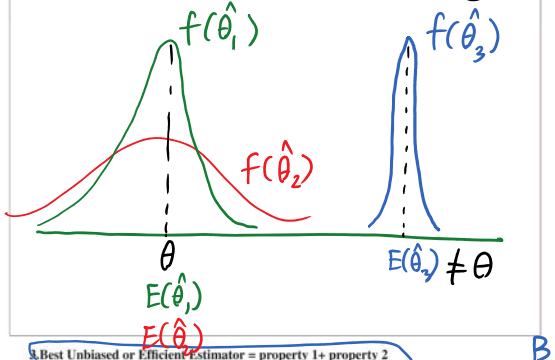
$\hat{\theta}_1$ is said to be a minimum variance estimator of θ if the variance of $\hat{\theta}_1$ is smaller that or at most equal to the variance of $\hat{\theta}_2$, which is any other estimator of θ

④ Bias $\rightarrow E(\hat{\theta}) - \theta$

Figure: Minimum Variance

to the variance of θ_2, which is any other estimator of θ

Figure: Minimum Variance



④

Bias $\longrightarrow$ $E(\hat{\theta}) - \theta$

To select your estimator for θ,

unbiasness comes first!

3.Best Unbiased or Efficient Estimator = property 1+ property 2

BUE

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of θ and the variance of $\hat{\theta}_1$ is smaller that or at most equal to the variance of $\hat{\theta}_2$, then $\hat{\theta}_1$ is **a minimum-variance unbiased estimator or best unbiased estimator.**

4. Linearity

$\hat{\beta}_2$ is a linear function of Y_i : $\hat{\beta}_2 = \sum_{i=1}^{n} k_i Y_i$

An estimator $\hat{\theta}$ is said to be a linear estimator of θ if it is a linear function of the sample observations.

For example:

$$\bar{X} = \frac{1}{n}\sum X_i = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

Thus, $\bar{X}$ is a linear estimator because it is a linear function of the X values.

$\hat{\beta}_1, \hat{\beta}_2 \rightarrow$ are BLUE!

Best Linear Unbiased Estimators : BLUE

The estimator $\hat{\theta}$ is called as the Best Linear Unbiased Estimator **BLUE** if it is satisfied the properties 1,2,4 that is $\hat{\theta}$ is linear, is unbiased, and has the minimum variance in the class of all linear unbiased estimators of θ.

Minimum Mean-Square-Error (MSE) Estimator

The MSE measures dispersion around the true value of the parameter. It is defined as:

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

However, the variance of $\hat{\theta}$ measures the dispersion of the distribution of the distribution of $\hat{\theta}$ around its mean or expected value.

$$\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$$

The relationship between the $\text{MSE}(\hat{\theta})$ and the $\text{var}(\hat{\theta})$ is as follows:

$$\text{MSE}(\hat{\theta}) = E[\hat{\theta} - \theta]^2$$

$$= E[\{\hat{\theta} - E(\hat{\theta})\} + \{E(\hat{\theta}) - \theta\}]^2 \qquad (a+b)^2 = a^2 + 2ab + b^2.$$

$$= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 + 2E[\hat{\theta} - E(\hat{\theta})]E[E(\hat{\theta}) - \theta]$$

$E[E(\hat{\theta}) - \theta] = 0$ as $E(\hat{\theta}) = \theta$.

(I) (II) (III)

$$\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + [\text{BIAS}(\hat{\theta})]^2$$

Then IF BIAS $= 0$, then $\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta})$.

Message : If our estimator is an unbiased estimator, we can use $\text{var}(\hat{\theta})$ as a representative term of $\text{MSE}(\hat{\theta})$.

In other words, when BIAS $= 0$, i.e., $E(\hat{\theta}) = \theta$, one can use $\text{var}(\hat{\theta})$ "to estimate" $\text{MSE}(\hat{\theta})$.

we can use $\text{var}(\hat{\theta})$ as a representative term of $MSE(\theta)$.

In other words, when BIAS $= 0$, i.e., $E(\hat{\theta}) = \theta$, one can use $\text{var}(\hat{\theta})$ "to estimate" $MSE(\hat{\theta})$.

An estimator $\hat{\beta}_2$ is said to be a best linear unbiased estimator (BLUE) of β_2 if the following hold:

B-L-U-E ESTIMATOR (UNBIASED, LINEAR, BEST)

♣ **It is linear.** It is the linear function of a random variable.

Recall that this property is ...

F.O.C $\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\beta}_1} = 0$ $\frac{\partial \sum \hat{u}^2}{\partial \hat{\beta}_2}$

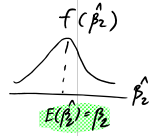
$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{or} = \sum k_i Y_i$$

$\Rightarrow \hat{\beta}_2$ is a linear function of Y_i. #

♣ **It is unbiased.** That is $E(\hat{\beta}_2)$ is equal to the true value, β_2

We have already proved this: $\hat{\beta}_2 = \beta_2 + \sum k_i u_i$

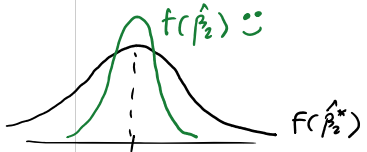
So $E(\hat{\beta}_2) = \beta_2$. #



♣ **It has the minimum variance in the class of all such linear unbiased estimators.**

Ex. $\hat{\beta}_2 = \sum k Y_i$ vs. $\hat{\beta}_2^* = \sum w_i Y_i$

$\text{var}(\hat{\beta}_2) \leq \text{var}(\hat{\beta}_2^*)$



$\beta_2 \rightarrow E(\hat{\beta}_2) = \beta_2$

$E(\hat{\beta}_2^*) = \beta_2$

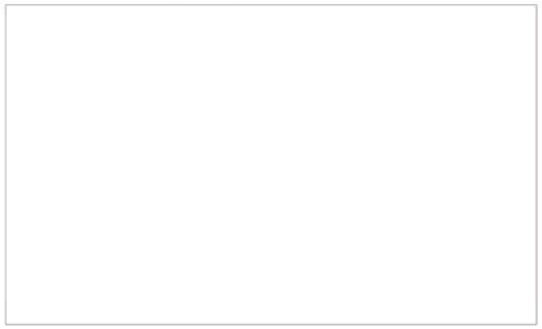
OLS OTHER METHODS

$\hat{\beta}_2$ has a minimum variance Among all other linear unbissed estimators.

OLS OTHER METHODS

$\hat{\beta}_2$ has a minimum variance Among all other linear unbissed estimators.

$E(\hat{\beta}_2) = \beta_2$



Gauss-Markov Theorem: Given the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is they are BLUE.

3.1.7 A measure of goodness of fit: r^2

In this section, we are going to study the goodness of fit of the fitted regression line to a set of data. Let us consider the following example:

Suppose we were to estimate the family expenditure (Y) based on our information from a random sample (as in Table 3.2).

What will happen if we set the estimated Y to be $\bar{Y}$?

Table 3.3: Estimating the expenditure of the household

Family Number (i)	Actual Y_i	Estimate $\hat{Y}_i = \bar{Y}$	Error in Estimation $Y_i - \hat{Y}$	Errors Squared $(Y_i - \hat{Y})^2$
1	390	543	-153	23460.03
2	425	543	-118	13963.36
3	560	543	17	283.36
4	575	543	32	1013.36
5	630	543	87	7540.03
6	679	543	136	18450.69
Sum	3259	3259	0	64710.83

$\bar{Y} = 543$

We can see all this graphically:

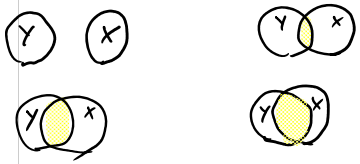
F1 F2 F3 F4 F5 F6

$Y = f(X)$

Q: how well X could be used to explain behavior/pattern/movement of Y?

A: Let's look at r^2. (r-squared)

$r^2 = 0$



$\sum_{i=1}^{6} (Y_i - \bar{Y})^2$

= Total Sum of squared errors (TSS) refers to...

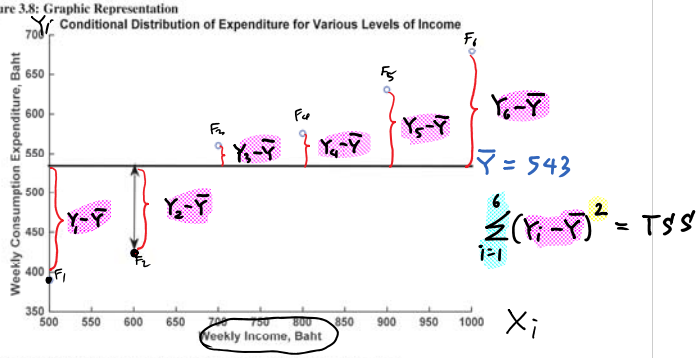
Total estimation errors this guy commits when he use just "sample mean" ($\bar{Y}$) as his predictor.

[we can also call TSS as "baseline errors"

$\rightarrow r^2 = 1$ (Perfect fit)

meaning: 100% variation in Y could be explained by variation in X.

Figure 3.8: Graphic Representation



$\bar{Y} = 543$

$\sum_{i=1}^{6}(Y_i - \bar{Y})^2 = TSS$

X_i

Question: Can we determine the total estimation error for this sample data?

Answer: Yes, we can calculate the total (combined) amount of estimation error for all observations in the sample when **using the mean as the estimate** as following:

$$TSS = \sum(Y_i - \bar{Y})^2$$

It is called the total sum of squares (TSS) which is the total variation of the actual Y values about their sample mean.

Since our objective in estimation is to minimize error (maximize precision), we need to cut down the amount of the estimation error (TSS).

We can achieve this by using information about other variables suspected to be strong predictors (strongly related to) the expenditure of the families.

We now can attempt to estimate the expenditure from the information on the income level of the family rather than from its own mean.

estimated — actual — errors

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

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

SRF, x_i

Table 3.4: Estimating the expenditure of the household with income (X_i)

Family (i)	Actual Y_i	Income X_i	x_i $X - \bar{X}$	y_i $Y - \bar{Y}$	$x_i y_i$ $(X-\bar{X})(Y-\bar{Y})$	x_i^2 $(X-\bar{X})^2$
1	390	500	-250	-153.17	38291.67	62500
2	425	600	-150	-118.17	17725.00	22500
3	560	700	-50	16.83	-841.67	2500
4	575	800	50	31.83	1591.67	2500

- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$
- $\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$

} Recipe for $\hat{\beta}_1$, $\hat{\beta}_2$

$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{103750}{175000} = 0.5928.$

Family (i)	Actual Y_i	Income X_i	x_i $X-\bar{X}$	y_i $Y-\bar{Y}$	$x_i y_i$ $(X-\bar{X})(Y-\bar{Y})$	x_i^2 $(X-\bar{X})^2$
1	390	500	-250	-153.17	38291.67	62500
2	425	600	-150	-118.17	17725.00	22500
3	560	700	-50	16.83	-841.67	2500
4	575	800	50	31.83	1591.67	2500
5	630	900	150	86.83	13025.00	22500
6	679	1000	250	135.83	33958.33	62500
Sum	3259	4500	0	0	103750	175000

$\bar{Y}$ = 543.16 ; $\bar{X}$ = 750 ; $\sum x_i y_i$; $\sum x_i^2$

From the table, we can calculate the simple regression as following:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{103750}{175000} = 0.5928.$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 543.16 - (0.5928)(750) = 98.524.$$

B/F

$\hat{Y}_i = \bar{Y}$

(use sample mean to predict $\hat{Y}_i$)

$TSS'_{OLD} = 64,710.83$

(B/F) $\hat{Y}_i = \bar{Y} = 543$

A/F

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

$\hat{Y}_i = 98.524 + 0.5928 X_i$ → SRF

(use income(X_i) to predict $\hat{Y}_i$)

(A/F) If $X_i = 500$, $\hat{Y}_i = ?$

$\hat{Y}_i = 98.524 + 0.5928(500)$

$\hat{Y}_i = 394.92.$

$\hat{\beta}_2 = 0.5928 \Rightarrow$ On average, if income rises by 1 baht, individual family would spend about 0.5928 baht, holding all other factors that might affect family spending constant.

Figure 3.9: Breakdown of the variation of Y_i into two components

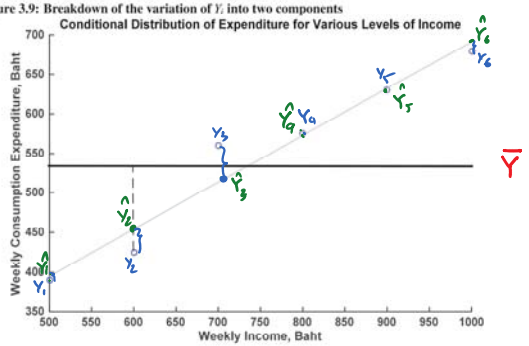


Table 3.5: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	Regression Estimate $\hat{Y}$	Residual $Y - \hat{Y}$	Residual squared $(Y - \hat{Y})^2$
1	390	500	394.95	-4.95	24.53
2	425	600	454.24	-29.24	854.87
3	560	700	513.52	46.48	2160.04
4	575	800	572.81	2.19	4.80
5	630	900	632.10	-2.10	4.39
6	679	1000	691.38	-12.38	153.29
Sum	3259	4500	0	0	3201.90

From the table 9, we can calculate the estimation error we have committed by using the regression line as:

$$RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{a}_i^2$$

where RSS stands for the residual sum of squares which is the unexplained variation of the Y values about the regression line.

B/F

TSS → RSS

64,XXX → 3201.90

64,XXX − 3201.90 = errors that is being eliminated by the regression

61508.93