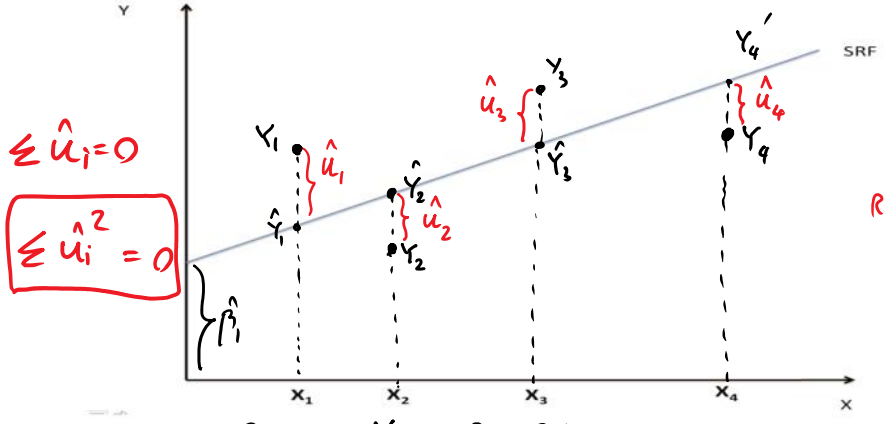
3. REGRESSION:THE PROBLEM OF ESTIMATION

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



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

$\sum \hat{u}_i^2 = 0$

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

ACTUAL Y

ESTIMATED Y

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

RESIDUALS OR ERROR TERM

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

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

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

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

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

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 $\sum_{i=1}^{n} \hat{u}_i^2$ IS AS LEAST AS POSSIBLE.

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

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

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

F.O.C $\frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_1} = 2 \sum_{i=1}^{n} (\overbrace{Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}^{\hat{u}_i})(-1) = 0 \Rightarrow \sum \hat{u}_i = 0$ — (1)

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

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

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

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

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

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

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

$\boxed{\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}}$ — (3)

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

FROM (2)

$$\sum_{i=1}^{n}(Y_i-\hat{\beta}_1-\hat{\beta}_2X_i)(X_i)=0$$

$$\sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}\hat{\beta}_1X_i-\sum_{i=1}^{n}\hat{\beta}_2X_i^2=0$$

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

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

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

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

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

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

$$\hat{\beta}_2=\frac{n\sum_{i=1}^{n}X_iY_i-\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}\qquad -(4)$$

From the SRF:

$$Y_i=\hat{\beta}_1+\hat{\beta}_2X_i+\hat{u}_i$$

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

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

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

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

$$\begin{aligned}x_i &= (X_i-\bar{X})\\ y_i &= (Y_i-\bar{Y})\end{aligned}\qquad(3.3)$$

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

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum x_iy_i}{\sum x_i^2}\\ &= \frac{\sum x_iY_i}{\sum X_i^2-n\bar{X}^2}\\ &= \frac{\sum X_iy_i}{\sum X_i^2-n\bar{X}^2}\end{aligned}$$

$$= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2}$$
$$= \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2} \quad (3.4)$$

Show that

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

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

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

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

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

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

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

EXAMPLE

Table 3.1: A Random Sample From the Population

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

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			
	$\sum Y_i$	$\sum X_i$	$\sum X_iY_i$	$\sum X_i^2$				$\sum x_iy_i$			
Sum	3,259	4,500	2,548,000	3,550,000	0	0	175,000	103,750			
	$\bar{Y}$	$\bar{X}$									
Mean	543.17	750	424,666.67	591,666.670	0	0	29,166.67	17,291.67			

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

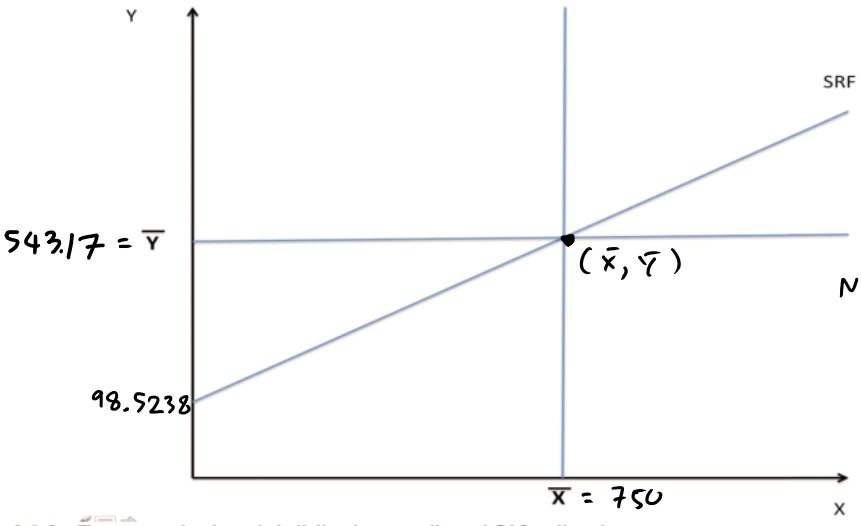
$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = 543.17 - 0.5929 \cdot 750 = 98.5238. \#$$

$$\hat{Y}_i = 98.5238 + 0.5929X_i \rightarrow \text{SRF}$$

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

$$\hat{\beta}_2 = \frac{n\sum X_iY_i - \sum X_i\sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{6 \cdot 2,548,000 - (4500)(3259)}{6 \cdot 3,550,000 - (4500)^2} = \frac{622,500}{1,050,000} = 0.5929$$

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



$\hat{Y}_i = 98.5238 + 0.5929 X_i$

NOTICE THAT THE SRF PASSES THROUGH $(\bar{X}, \bar{Y})$

$\bar{X}$: SAMPLE MEAN OF X

$\bar{Y}$: SAMPLE MEAN OF Y

3.1.2 The numerical and statistical properties of OLS estimators

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).

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X}\end{aligned} \tag{3.5}$$

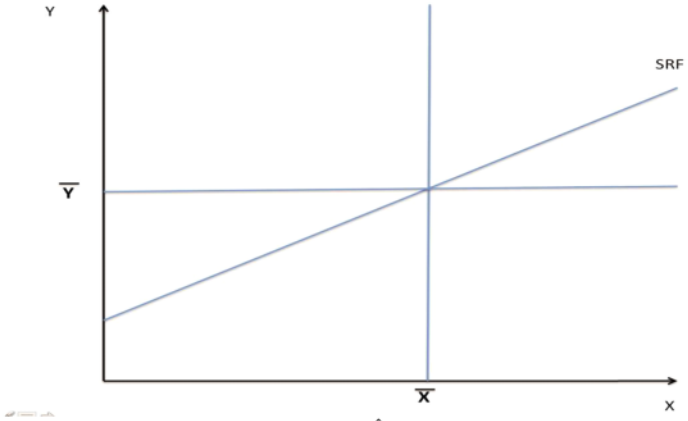
$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum Xi \sum Yi}{n\sum X_i^2 - (\sum X_i)^2} \tag{3.6}$$

2. They are **point estimators**.

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 $\bar{X}$).

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



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

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

PROOF:

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

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

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

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

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

$$= \sum X_i - n\bar{X}$$

$$= \sum X_i - n \cdot \frac{\sum X_i}{n}$$

$$= 0$$

OR IF WE DEFINE

$$x_i = X_i - \bar{X}$$

$$\sum x_i = 0.$$

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

FROM LAGRANGIAN:

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

SO $\frac{\sum \hat{u}_i}{n}$ MUST BE EQUAL TO ZERO.

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

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

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

② − ①: $Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + \hat{u}_i$ —③

$y_i = Y_i - \bar{Y}$ AND $x_i = X_i - \bar{X}$

$y_i = \hat{\beta}_2 x_i + \hat{u}_i$ —④ CALLED "DEVIATION FORM"

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

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

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

(FROM PAGE 49)

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

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

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

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

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

$= \sum \hat{\beta}_2 x_i \cdot (y_i - \hat{\beta}_2 x_i)$

$= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 = 0$!

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

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

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

↳ NO CORRELATION BETWEEN $\hat{y}_i$ AND $\hat{u}_i$.

3.1.3 The Assumptions Underlying the Method of Least Squares

Assumption 1: Linear regression model

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

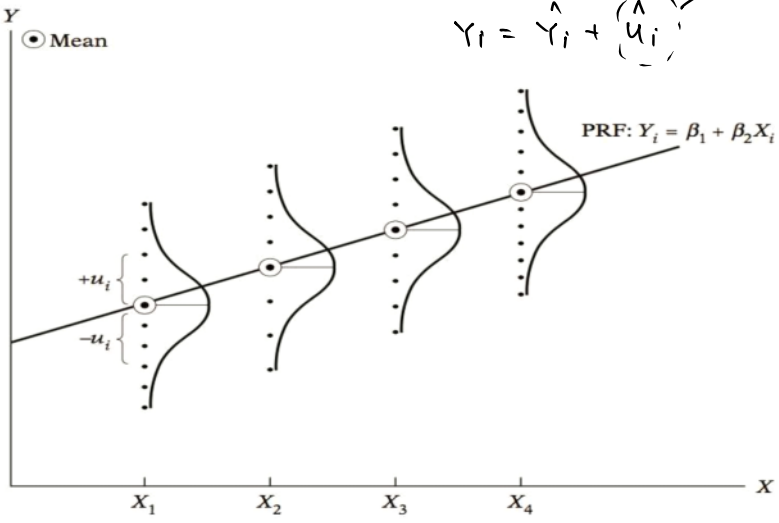
Assumption 2: X values are fixed in repeated sampling

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

EQUAL SPREAD

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

$$= E\left[u_i^2 \mid X_i\right] = \sigma^2$$

Figure 3.5: Homoscedasticity

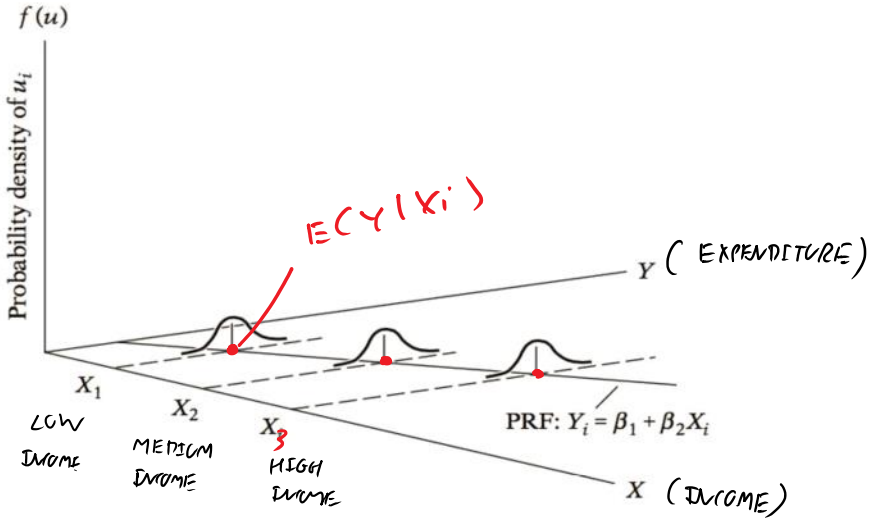
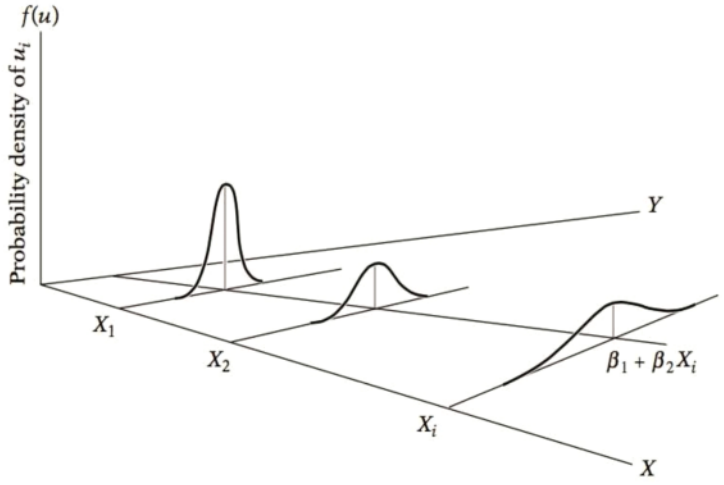


Figure 3.6: Heteroscedasticity



Assumption 5: No Autocorrelation Between the Disturbances

GIVEN X_i AND X_j WHERE $i \neq j$

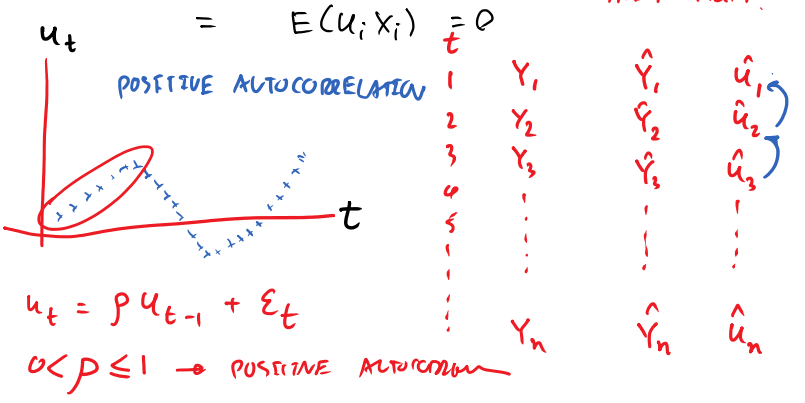
$$\text{COV}[u_i, u_j \mid X_i, X_j]$$

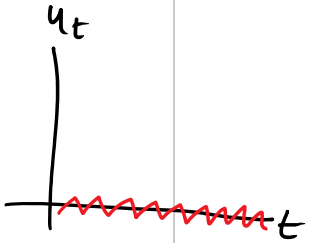
$$= E\left[\left(u_i - E(u_i \mid x_i)\right)\cdot\left(u_j - E(u_j \mid x_j)\right)\right]$$

$$= E(u_i x_i) - E(x_i)\,E(u_i) \quad (E(u_i) = 0)$$

CONSTANT TERM

$$= E(u_i x_i) = 0$$





$u_t = \rho u_{t-1} + \varepsilon_t$

$0 < \rho \leq 1$ → POSITIVE AUTOCORRELATION

$-1 \leq \rho < 0$ → NEGATIVE AUTOCORRELATION $-1 \leq \rho \leq 1$

$\rho = 0$ → NO AUTOCORRELATION

Assumption 6: Zero Covariance Between u_i and X_i

$$\text{COV}(u_i, X_i) = 0$$

VARIATION IN X_i IS NOT RELATED WITH VARIATION IN u_i.

$\rho = 0$ → NO AUTOCORRELATION

Assumption 6: Zero Covariance Between u_i and X_i

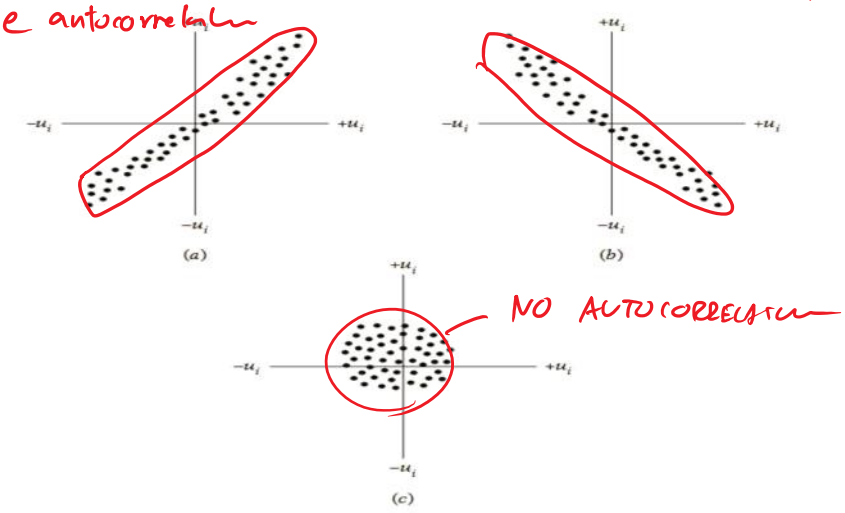
$COV(u_i, X_i) = 0$

VARIATION IN X_i IS NOT RELATED WITH VARIATION IN u_i.

Figure 3.7: Patterns of Correlation Among the disturbances

+ve autocorrelation

-ve autocorrelation



NO AUTOCORRELATION

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

Assumption 8:Variability in X values.

X_i 500 600 700 ... 1000

$$\text{var}(X) = \frac{\sum (X_i - \bar{X})^2}{n-1} \neq 0.$$

Assumption 9: The regression model is correctly specified.

Assumption 10: There is no perfect multicollinearity.

EX: $Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3$ (CONSUMPTION, INCOME, WEALTH)

HOWEVER, IF $X_3 = 2X_2$, THEN WE HAVE "PERFECT MULTICOLLINEARITY"

RESULT 9) $\Rightarrow$ $\hat{\beta}_2$ AND $\hat{\beta}_3$ CANNOT BE ESTIMATED.

HOWEVER IF $X_3 = 2X_2 + v_i$ WHERE $v_i \neq 0$

i	X_2	X_3	v_i
1	-	-	-1
2	-	-	100
3	-	-	-20
4	-	-	-50
:	-	-	+50
:	-	-	0

"IMPERFECT MULTICOLLINEARITY"

OLS CAN GIVE AN ESTIMATION OF $\hat{\beta}_2$ AND $\hat{\beta}_3$.

H A M

REPEATED SAMPLING

3.1.4 Standard Errors of Least-Squares Estimates

The standard errors of the OLS estimates can be obtained as follows:
We know that

OBSERVE THAT $\hat{\beta}_2 = f(Y_i)$

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

$\hat{\beta}_1, \hat{\beta}_2$ — (1) (2) (3) (4) (5) (6) ...

where

$$k_i = \frac{x_i}{\sum x_i^2} \qquad x_i = X_i - \bar{X}, \quad \sum_{i=1}^{n} (x_i - \bar{X})^2$$

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. $\sum k_i x_i = \sum k_i X_i = 1$

PROOF (2)

$$\sum k_i = 0$$

$$\sum \frac{x_i}{\sum x_i^2} = \frac{\sum (X_i - \bar{X})}{\sum x_i^2} = \frac{\sum X_i - n\bar{X}}{\sum x_i^2} = \frac{\sum X_i - n\frac{\sum X_i}{n}}{\sum x_i^2}$$

Since

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

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

PROOF (3)

$$\sum k_i^2 = \frac{1}{\sum x_i^2} = 0. \#$$

$$\sum \left(\frac{x_i}{\sum x_i^2}\right)^2 = \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{1}{\sum x_i^2}.$$

PRF IN STOCHASTIC FORM

$$\hat{\beta}_2 = \sum k_i Y_i$$
$$= \sum k_i (\beta_1 + \beta_2 X_i + u_i)$$
$$= \sum [k_i \beta_1 + k_i \beta_2 X_i + k_i u_i]$$
$$= \beta_1 \sum k_i + \beta_2 \sum k_i X_i + \sum k_i u_i \quad (\sum k_i = 0,\ \sum k_i X_i = 1)$$
$$= 0 + \beta_2 + \sum k_i u_i$$

PROOF (4)

$$\sum k_i x_i = \sum k_i X_i = 1$$

$$\sum k_i (X_i - \bar{X}) = \sum k_i X_i - \sum k_i \bar{X} \quad (?)$$

$$\sum k_i x_i = \sum k_i X_i. \#$$

$$\sum k_i x_i = \sum \frac{x_i}{\sum x_i^2} \cdot x_i = \frac{\sum x_i^2}{\sum x_i^2} = 1$$

TAKING EXPECTATION BOTH SIDE GIVES

$$E(\hat{\beta}_2) = \beta_2 + E(\sum k_i u_i)$$
$$= \beta_2 + \sum k_i E(u_i) \quad (E(u_i) = 0)$$
$$= \beta_2 + \sum E(k_i u_i) \quad (= 0)$$

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

$E(\hat{\beta}_2) = \beta_2$ — $\hat{\beta}_2$ IS AN UNBIASED ESTIMATOR.

Second Step
Using the definition of variance

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

$\hat{\beta}_2^*$ IS A BIASED ESTIMATOR AS IT NEVER HITS THE TRUE β_2.

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

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

VAR (D) = $E[D - E(D)]^2$

$$var(\hat{\beta}_2) = E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2$$
$$= E[\hat{\beta}_2 - \beta_2]^2$$

HITS THE TRUE β_2.

$$= E\left[\hat{\beta}_2 - \beta_2\right]^2$$

$$= E\left[\sum k_i u_i\right]^2$$

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

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

SINCE $\hat{\beta}_2 = \beta_2 + \sum k_i u_i$.

$$= E\left[k_1^2 u_1^2 + k_2^2 u_2^2 + \cdots + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + \cdots + 2k_{n-1} k_n u_{n-1} u_n\right]$$

$$= k^2 E(u_1^2) + k_2^2 E(u_2^2) + \cdots + k_n^2 E(u_n^2) + 0.$$

CROSS TERMS : $2k_i k_j \cdot u_i u_j$ SINCE $E(u_i u_j) = 0$

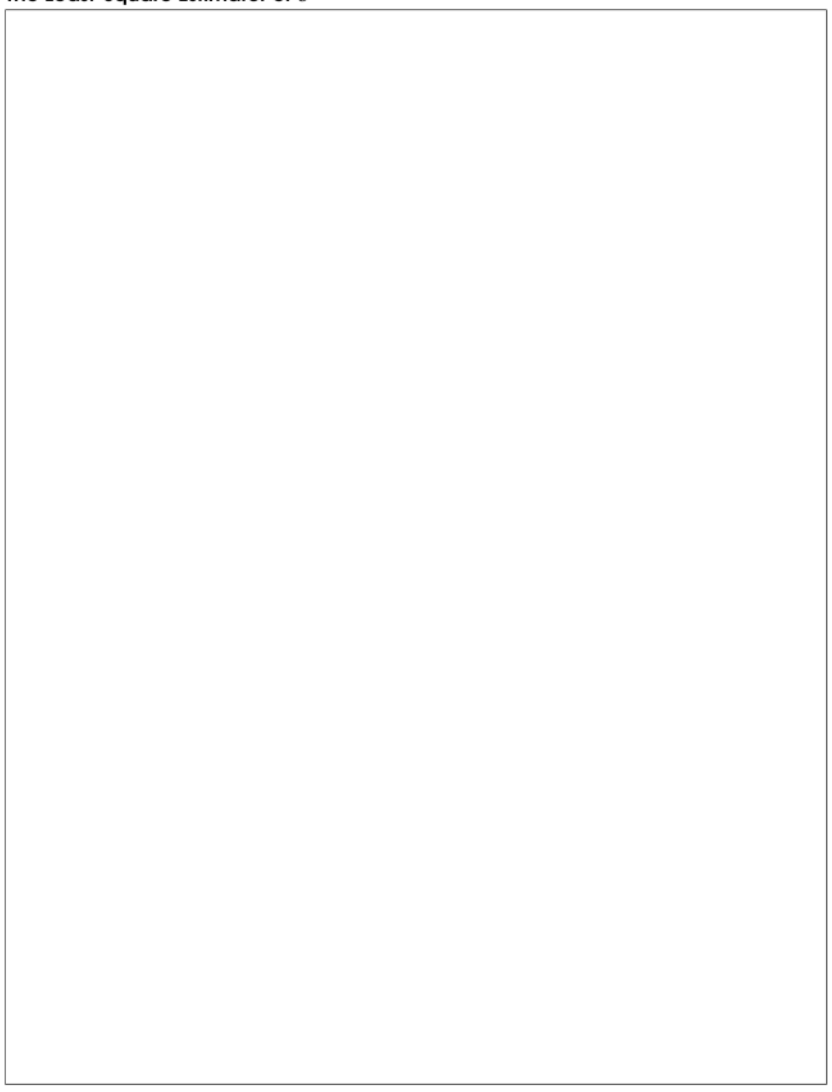
$$\text{var}(\hat{\beta}_2) = \sigma_u^2 \cdot \sum_{i=1}^{n} k_i^2 = \sigma_u^2 \frac{1}{\sum_{i=1}^{n} x_i^2}$$

IS VARIANCE of DISTURBANCE TERM WHICH IS UNKNOWN.

SO, IF WE WANT TO KNOW $\text{var}(\hat{\beta}_2)$, WE HAVE TO ESTIMATE σ_u^2 TOO.

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

3.1.5 The Least-Square Estimator of σ^2



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.

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

$$E(\hat{\theta}) = \theta$$

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:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Figure: Biased and Unbiased Estimators

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

Figure: Minimum Variance

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

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

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.

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

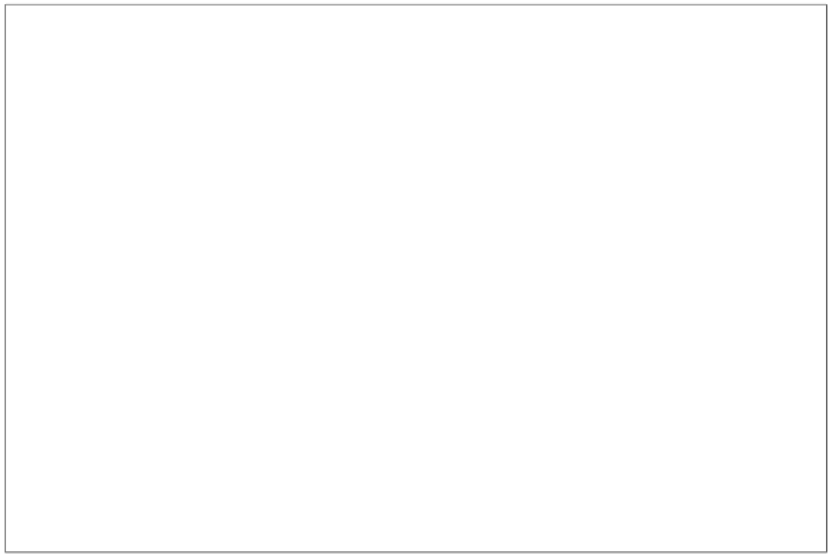
The MSE measures dispersion around the true value of the parameter. It is defined as:

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

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



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

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

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

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

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).

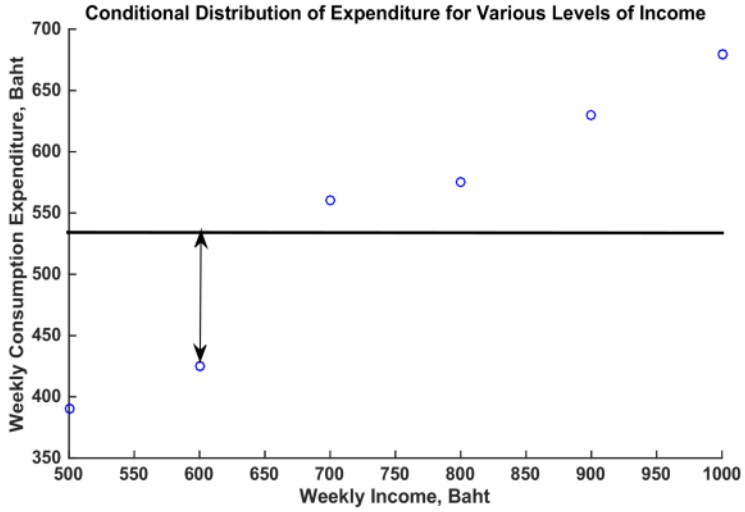
What will happen if we set the estimated Y to be $\bar{Y}$?

Table 3.3: Estimating the expenditure of the household

Family Number (i)	Actual Y_i	Estimate $\hat{Y}_i = \bar{Y}$	**Error in Estimation** $Y_i - \bar{Y}$	**Errors Squared** $(Y_i - \bar{Y})^2$
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

We can see all this graphically:

Figure 3.8: Graphic Representation



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.

Table 3.4: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
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

From the table 8, we can calculate the simple regression as following:

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

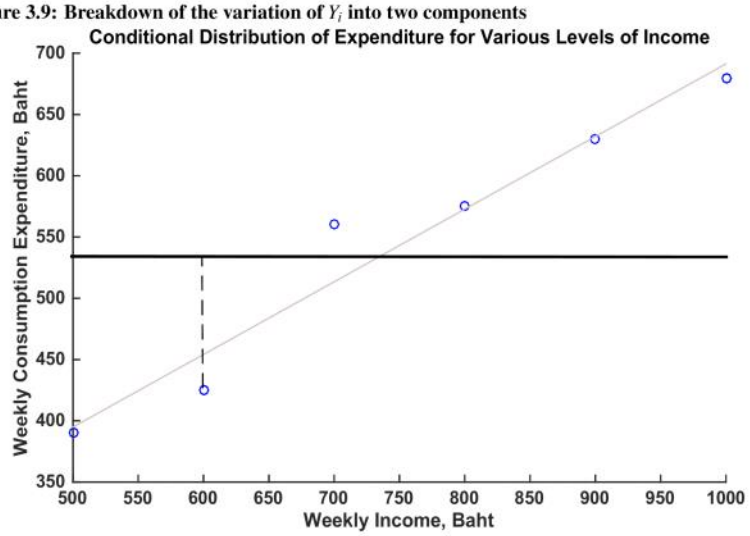


Table 3.5: Estimating the expenditure of the household with income

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:

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

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

Total Baseline Error using the mean (SS Total) =

New or Remaining Error (SS Error or SS Residual) =

QUESTION: How much of the original estimation error have we explained away (eliminated) by using the regression model (instead of the mean)?

ANS

QUESTION: What % of estimation error have we explained (eliminated by using the regression model?

ANS

QUESTION: What does the remaining% represent?

ANS
Percent of variation (differences) in expenditures that can be accounted for by: (a) all other potential predictors not included in the model, beyond income levels, and (b) unexplainable random/chance variations.

$$r^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

♣ r^2 is a measure of our success regarding accuracy of our estimation effort.

♣ r^2 = % of estimation error that we have been able to explain away by using the regression model, instead of using the mean.

♣ r^2 indicates how much better we can predict Y from information about Xs, rather than from using its own mean.

♣ r^2 = % of differences (variations) in Y values that is explained by (attributable to) differences in X values.