

---

## Regression Analysis: Terminology and Notation

Consider the generic version of the *simple (two-variable) linear regression model*.

It is represented by the following **population regression equation** (called the **PRE** for short):

$$Y_i = f(X_i) + u_i = \beta_0 + \beta_1 X_i + u_i$$

- **The PRF (population regression function):**

$$f(X_i) = \beta_0 + \beta_1 X_i$$

= the *i*-th value of the population regression function (PRF).

- **Observable Variables:**

$Y_i \equiv$  the *i*-th value of the dependent variable *Y*

$X_i \equiv$  the *i*-th value of the independent variable *X*

- **Unobservable Variable:**

$u_i \equiv$  the random error term for the *i*-th member of the population

- **Unknown Parameters:** the regression coefficients

$\beta_0 =$  the *intercept coefficient*

$\beta_1 =$  the *slope coefficient on  $X_i$*

The **true population values** of the regression coefficients  $\beta_0$  and  $\beta_1$  are ***unknown***.

---

## Variables and Parameters

**PRE:**  $Y_i = f(X_i) + u_i = \beta_0 + \beta_1 X_i + u_i$

- The **variables** of the regression model are  $Y_i$ ,  $X_i$ , and  $u_i$ .

$Y_i$  and  $X_i$  are the **observable variables**; their values can be observed or measured.

$Y_i$  is called any of the following: (1) the **dependent variable**  
(2) the **regressand**  
(3) the **explained variable**.

$X_i$  is called any of the following: (1) the **independent variable**  
(2) the **regressor**  
(3) the **explanatory variable**.

- $u_i$  is an **unobservable random variable**; its value cannot be observed or measured. It is called a **random error term**.
- $\beta_0$  and  $\beta_1$  are the **parameters** of the regression model, together with any unknown parameters of the probability distribution of the random error term  $u_i$ .

$\beta_0$  and  $\beta_1$  are called **regression coefficients**; in particular,

$\beta_0 \equiv$  the **intercept coefficient**,

and

$\beta_1 \equiv$  the **slope coefficient** of  $X$ .

The **true population values** of the regression coefficients  $\beta_0$  and  $\beta_1$  are **unknown**.

## Simple Regression versus Multiple Regression

- A **simple** regression model has *only two* observable variables:
  - (1) **one dependent** variable or *regressand*  $Y_i$ ;
  - (2) **one independent** variable or *regressor*  $X_i$ .
- A **multiple** regression model has *three or more* observable variables:
  - (1) **one dependent** variable or *regressand*  $Y_i$ ;
  - (2) **two or more independent** variables or *regressors*  $X_{1i}, X_{2i}, \dots, X_{ki}$ , where
$$X_{ji} \equiv \text{the } i\text{-th value of the } j\text{-th regressor } X_j \text{ (} j = 1, 2, \dots, k\text{)}.$$

## The Simple Linear Regression Model

- The **PRE (population regression equation)** for the simple linear regression model:

$$Y_i = f(X_i) + u_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{PRF}} + \underbrace{u_i}_{\text{random error}} \quad (1a)$$

$$f(X_i) = \beta_0 + \beta_1 X_i$$

= the **PRF** (population regression function) for the i-th population member

$$u_i = Y_i - f(X_i) = Y_i - \beta_0 - \beta_1 X_i$$

= the **random error** for the i-th population member

$\beta_0, \beta_1$  = the unknown regression coefficients  $\beta_0$  and  $\beta_1$

Number of regression coefficients =  $K = 2$ .

Number of slope coefficients =  $K - 1 = 2 - 1 = 1$ .

- Sample Data:** A *random sample of N members of the population* for which the observed values of Y and X are measured. Each sample observation is of the form

$$(Y_i, X_i), \quad i = 1, \dots, N$$

- The **SRE (sample regression equation)** for the simple linear regression model:

$$Y_i = \hat{f}(X_i) + \hat{u}_i = \hat{Y}_i + \hat{u}_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\text{SRF}} + \underbrace{\hat{u}_i}_{\text{residual}} \quad (1b)$$

$$\hat{f}(X_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

= the **SRF (sample regression function)** for sample observation i

$$\hat{u}_i = Y_i - \hat{f}(X_i) = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

= the **residual** for sample observation i

$\hat{\beta}_0, \hat{\beta}_1$  = *estimators* or *estimates* of the regression coefficients  $\beta_0$  and  $\beta_1$

## The Multiple Linear Regression Model

- The **PRE (population regression equation)** for the *multiple linear regression model* is:

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i = \underbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}}_{\text{PRF}} + \underbrace{u_i}_{\text{random error}} \quad (2a)$$

$$f(X_{1i}, X_{2i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$$

= the **PRF** (population regression function) for the *i*-th population member

$$u_i = Y_i - f(X_{1i}, X_{2i}, \dots, X_{ki}) = Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \dots - \beta_k X_{ki}$$

= the **random error** for the *i*-th population member

$$\beta_0, \beta_1, \beta_2, \dots, \beta_k = \text{the unknown regression coefficients } \beta_0, \beta_1, \beta_2, \dots, \beta_k$$

Number of regression coefficients =  $K$ .

Number of slope coefficients =  $k = K - 1$ .

- Sample Data:** A *random sample of N members of the population* for which the observed values of  $Y$  and  $X_1, X_2, \dots, X_k$  are measured. Each sample observation is of the form

$$(Y_i, X_{1i}, X_{2i}, \dots, X_{ki}), \quad i = 1, \dots, N$$

- The **SRE (sample regression equation)** for the multiple linear regression model:

$$Y_i = \hat{f}(X_{1i}, X_{2i}, \dots, X_{ki}) + \hat{u}_i = \hat{Y}_i + \hat{u}_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}}_{\text{SRF}} + \underbrace{\hat{u}_i}_{\text{residual}} \quad (2b)$$

$$\hat{f}(X_{1i}, X_{2i}, \dots, X_{ki}) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki}$$

= the **SRF (sample regression function)** for sample observation *i*

$$\hat{u}_i = Y_i - \hat{f}(X_{1i}, X_{2i}, \dots, X_{ki}) = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \dots - \hat{\beta}_k X_{ki}$$

= the **residual** for sample observation *i*

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k = \text{estimators or estimates of the regression coefficients}$$

- Examples of **multiple regression models**

- ♦ A ***three-variable* linear regression model** has **two** regressors; its PRE is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i.$$

Total number of regression coefficients =  $K = 3$

Number of *slope* coefficients =  $k = K - 1 = 3 - 1 = 2$

- ♦ A ***four-variable* linear regression model** has **three** regressors; its PRE is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i.$$

Total number of regression coefficients =  $K = 4$

Number of *slope* coefficients =  $k = K - 1 = 4 - 1 = 3$

- ♦ The ***general* multiple linear regression model** has  **$K - 1$**  regressors; its PRE is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i.$$

Total number of regression coefficients =  $K$ .

Number of *slope* coefficients =  $k = K - 1$ .

## **Regression Analysis: A Hypothetical Numerical Example**

**Reference:** D. Gujarati (1995), Chapter 2, pp. 32-36.

**Purpose:** To illustrate some of the **basic ideas of linear regression analysis**.

**The Model:** A simple consumption function representing the relationship between

$Y_i$   $\equiv$  the weekly consumption expenditure of family  $i$  (\$ per week);

$X_i$   $\equiv$  the weekly disposable (after-tax) income of family  $i$  (\$ per week);

The **PRE (population regression equation)** for this model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (1)$$

**The Population:** consists entirely of **60 families**.

We assume that the weekly disposable incomes of these families take only **10 distinct values** -- i.e.,  $X$  takes only the 10 distinct values

**$X_i = 80, 100, 120, 140, 160, 180, 200, 220, 240, 260$ .**

We further assume that we can observe the entire population of 60 families.

The **data for the complete population** is given in **Table 2.1**.

**Table 2.1: Population data points ( $Y_i, X_i$ ) for the population of 60 families.**

$X_i$ values →	80	100	120	140	160	180	200	220	240	260
$Y_i$ values ↓	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	--	88	--	113	125	140	--	160	189	185
	--	--	--	115	--	--	--	162	--	191
<b>Sum <math>Y_i</math> values</b>	<b>325</b>	<b>462</b>	<b>445</b>	<b>707</b>	<b>678</b>	<b>750</b>	<b>685</b>	<b>1043</b>	<b>966</b>	<b>1211</b>
<b>Number of <math>Y_i</math></b>	<b>5</b>	<b>6</b>	<b>5</b>	<b>7</b>	<b>6</b>	<b>6</b>	<b>5</b>	<b>7</b>	<b>6</b>	<b>7</b>

- **Interpretation of Table 2.1:**

**Each column of Table 2.1** represents the **population conditional distribution of Y** (families' weekly consumption expenditure) **for the corresponding value of X** (families' weekly disposable income).

- ♦ The first column gives the conditional distribution of Y for  $X_i = 80$ ; five families in the population have weekly disposable income equal to 80 dollars.
- ♦ The fifth column gives the conditional distribution of Y for  $X_i = 160$ ; six families in the population have weekly disposable income equal to 160 dollars.
- ♦ The tenth (last) column gives the conditional distribution of Y for  $X_i = 260$ ; seven families in the population have weekly disposable income equal to 260 dollars.



**Table 2.2: Population conditional probabilities of Y for each population value of X.**

• **Notation:**

$p(Y|X_i) = p(Y_j|X_i)$  = the conditional probability of Y for  $X = X_i$   
 = the probability that the random variable Y takes the numerical value  $Y_j$  given that the variable X is equal to the numerical value  $X_i$ .

Conditional probabilities  $p(Y|X_i)$  for the population data in Table 2.1

<b><math>X_i</math> values →</b>	<b>80</b>	<b>100</b>	<b>120</b>	<b>140</b>	<b>160</b>	<b>180</b>	<b>200</b>	<b>220</b>	<b>240</b>	<b>260</b>
<b><math>p(Y X_i)</math> ↓</b>	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	--	1/6	--	1/7	1/6	1/6	--	1/7	1/6	1/7
	--	--	--	1/7	--	--	--	1/7	--	1/7
<b>Sum <math>Y_i</math> values</b>	<b>325</b>	<b>462</b>	<b>445</b>	<b>707</b>	<b>678</b>	<b>750</b>	<b>685</b>	<b>1043</b>	<b>966</b>	<b>1211</b>
<b>Number of <math>Y_i</math></b>	<b>5</b>	<b>6</b>	<b>5</b>	<b>7</b>	<b>6</b>	<b>6</b>	<b>5</b>	<b>7</b>	<b>6</b>	<b>7</b>
<b><math>E(Y X_i)</math></b>	<b>65</b>	<b>77</b>	<b>89</b>	<b>101</b>	<b>113</b>	<b>125</b>	<b>137</b>	<b>149</b>	<b>161</b>	<b>173</b>

• **Interpretation of Table 2.2:**

**Each column of Table 2.2 contains the population conditional probabilities of Y (families' weekly consumption expenditure) for the corresponding value of X (families' weekly disposable income).**

---

***Examples: Computing the Conditional Probabilities of Individual Y Values***

1. Consider the column corresponding to  $X_i = 80$ .

There are *five* different values of Y for  $X_i = 80$ : .

$$Y | X_i = 80: 55, 60, 65, 70, 75.$$

The probability of observing any one family whose weekly disposable income is  $X_i = 80$  equals 1/5: e.g.,

$$p(Y = 55 | X_i = 80) = \frac{1}{5}.$$

$$p(Y = 60 | X_i = 80) = \frac{1}{5}.$$

$$p(Y = 65 | X_i = 80) = \frac{1}{5}.$$

$$p(Y = 70 | X_i = 80) = \frac{1}{5}.$$

$$p(Y = 75 | X_i = 80) = \frac{1}{5}.$$

2. Consider the column corresponding to  $X_i = 160$ .

There are *six* different values of Y for  $X_i = 160$ : .

$$Y | X_i = 160: 102, 107, 110, 116, 118, 125.$$

The probability of observing any one family whose weekly disposable income is  $X_i = 160$  equals 1/6: e.g.,

$$p(Y = 102 | X_i = 160) = \frac{1}{6}.$$

$$p(Y = 110 | X_i = 160) = \frac{1}{6}.$$

## Population Conditional Means of Y

For each of the 10 population values of  $X_i$ , we can compute from Tables 2.1 and 2.2 the corresponding *conditional mean value* of the **population values of Y**.

For each of the values  $X_i$  of  $X$ , the population mean value of  $Y$  is called

(1) the **population conditional mean of Y**

or

(2) the **population conditional expectation of Y**.

- **Notation:**

$$\begin{aligned} E(Y | X_i) &= E(Y | X = X_i) \\ &= \text{the } \mathbf{population\ conditional\ mean\ of\ Y\ for\ X = X_i} \\ &= \text{the "expected value of Y given that X takes the specific value } X_i\text{"} \end{aligned}$$

- **Definition:**

$$E(Y | X_i) = E(Y | X = X_i) = \sum_{X=X_i} p(Y | X_i) Y$$

where

$$p(Y | X_i) = \text{the conditional probability of Y when } X = X_i;$$

$$p(Y | X_i) Y = \text{the product of each population value of Y and its corresponding conditional probability for } X = X_i.$$

In words, the above formula for  $E(Y | X_i) = E(Y | X = X_i)$  says that for the value  $X_i$  of  $X$ ,

- (1) **multiply** each population value of  $Y$  by its associated conditional probability  $p(Y | X_i)$  to get the product  $p(Y | X_i) Y$
- (2) then **sum these products**  $p(Y | X_i) Y$  over all the population values of  $Y$  corresponding to  $X = X_i$ .

---

- **Illustrative Calculations of  $E(Y | X_i)$ :**

1. For  $X_i = 80$ ,  $p(Y | X_i) = 1/5$ :

$$\begin{aligned} E(Y | X_i = 80) &= \frac{1}{5}55 + \frac{1}{5}60 + \frac{1}{5}65 + \frac{1}{5}70 + \frac{1}{5}75 \\ &= \frac{55 + 60 + 65 + 70 + 75}{5} \\ &= \frac{325}{5} \\ &= 65 \end{aligned}$$

2. For  $X_i = 160$ ,  $p(Y | X_i) = 1/6$ :

$$\begin{aligned} E(Y | X_i = 160) &= \frac{1}{6}102 + \frac{1}{6}107 + \frac{1}{6}110 + \frac{1}{6}116 + \frac{1}{6}118 + \frac{1}{6}125 \\ &= \frac{102 + 107 + 110 + 116 + 118 + 125}{6} \\ &= \frac{678}{6} \\ &= 113 \end{aligned}$$

3. For  $X_i = 260$ ,  $p(Y | X_i) = 1/7$ :

$$\begin{aligned} E(Y | X_i = 260) &= \frac{1}{7}150 + \frac{1}{7}152 + \frac{1}{7}175 + \frac{1}{7}178 + \frac{1}{7}180 + \frac{1}{7}185 + \frac{1}{7}191 \\ &= \frac{150 + 152 + 175 + 178 + 180 + 185 + 191}{7} \\ &= \frac{1211}{7} \\ &= 173 \end{aligned}$$

**Table 2.3: Population Conditional Means of Y**

**Table 2.3**

$X_i$	$E(Y X_i)$
80	65
100	77
120	89
140	101
160	113
180	125
200	137
220	149
240	161
260	173

- **Interpretation of Table 2.3:**

Table 2.3 tabulates the relationship between  $E(Y|X_i)$  and  $X_i$  for this particular population of 60 families.

This population relationship between  $E(Y|X_i)$  and  $X_i$  is called either

(1) the **population regression function**, or **PRF**.

or

(2) the **population conditional mean function**, or **population CMF**

So **Table 2.3** is a *tabular representation of the PRF* for the population of 60 families.

**Properties of the Population Regression Function, or PRF:****Table 2.3**

$X_i$	$E(Y X_i)$
80	65
100	77
120	89
140	101
160	113
180	125
200	137
220	149
240	161
260	173

1.  $E(Y|X_i)$  is a **function of  $X_i$** : i.e.,  $E(Y|X_i) = f(X_i)$ .

2.  $E(Y|X_i)$  is an **increasing function of  $X_i$** : i.e.,

$$\Delta X_i > 0 \Rightarrow \Delta E(Y|X_i) > 0 \quad \text{and} \quad \Delta X_i < 0 \Rightarrow \Delta E(Y|X_i) < 0.$$

3.  $E(Y|X_i)$  is a **linear function of  $X_i$** : i.e.,

- A plot of the 10 points in Table 2.3 lie on a straight line.
- Each 20-dollar increase in  $X$  induces a constant 12-dollar increase in  $E(Y|X_i)$ : i.e.,

$$\Delta X_i = 20 \Rightarrow \Delta E(Y|X_i) = 12 \Rightarrow \frac{\Delta E(Y|X_i)}{\Delta X_i} = \frac{12}{20} = \mathbf{0.60}.$$

4. The **population regression function (PRF)** -- also called the **population conditional mean function** -- takes the general linear form

$$E(Y | X_i) = \beta_0 + \beta_1 X_i.$$

5. The **population values** of the **regression coefficients**  $\beta_1$  and  $\beta_2$  for this hypothetical population of 60 families are:

$$\beta_0 = 17 \quad \text{and} \quad \beta_1 = 0.60.$$

6. The **population regression function, or PRF**, for this particular population of 60 families is therefore

$$E(Y | X_i) = \beta_0 + \beta_1 X_i = 17 + 0.60 X_i.$$

□ **Summary -- The Population Regression Function (PRF)**

The **PRF**, or **population regression function**, for this hypothetical population of 60 families is a **linear function of the population values**  $X_i$  of the regressor  $X$ ; it takes the form

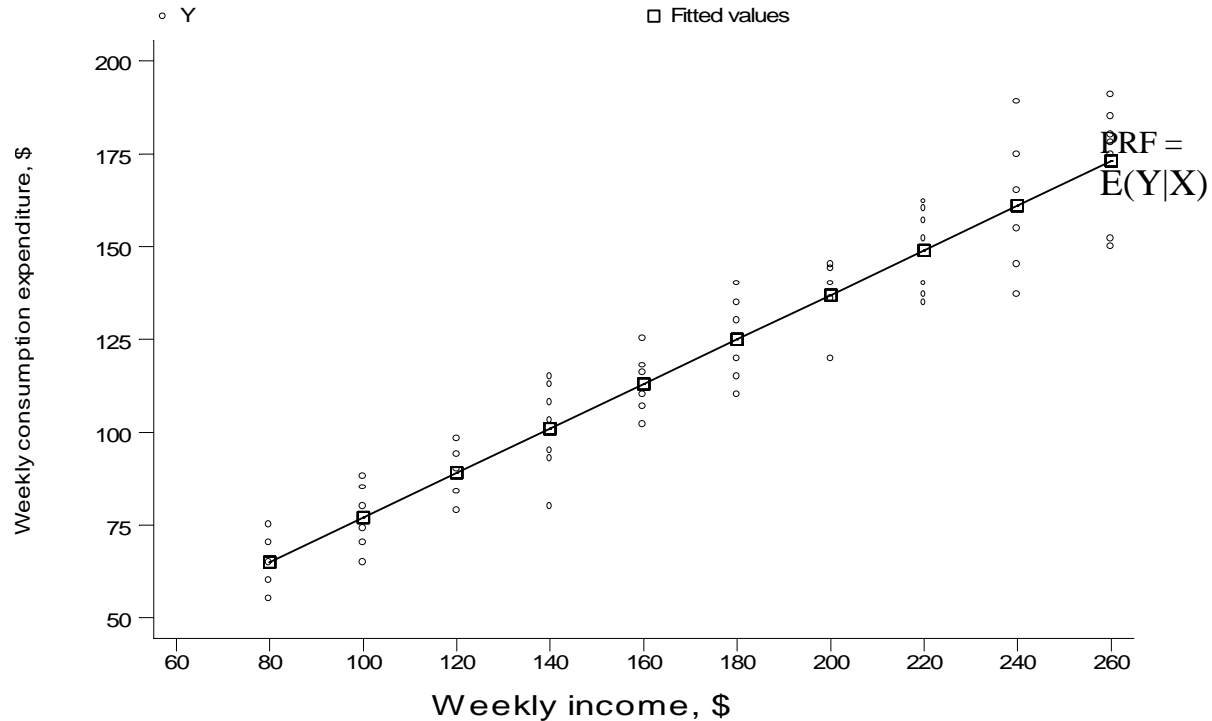
$$f(X_i) = E(Y | X_i) = \beta_0 + \beta_1 X_i = 17 + 0.60 X_i.$$

where

$\beta_0 = 17$  is the population value of the *intercept* coefficient

$\beta_1 = 0.60$  is the population value of the *slope* coefficient of  $X_i$ .

- **Figure 2.1** Plot of Population Data Points, Conditional Means  $E(Y|X)$ , and the Population Regression Function PRF



1. The *small dots* in Figure 2.1 constitute a **scatterplot** of the **population values of Y and X** for the population of 60 families:

Each small dot corresponds to a **single population data point** of the form  $(Y_i, X_i)$   $i = 1, 2, \dots, 60$ .

2. The *solid line* in Figure 2.1 is the **population regression line** for the population of 60 families.

Each pair of population values of  $(E(Y|X_i), X_i)$ , is represented by a **large square dot** in Figure 2.1.

This population regression line is the locus of the 10 points in Table 2.3 -- i.e., it connects the 10 points of the form  $(E(Y|X_i), X_i)$ ,  $i = 1, \dots, 10$ .



## The Random Error Terms

- **Definition:** The *unobservable random error term* for the  $i$ -th population member is denoted as  $u_i$  and defined as

$$u_i = Y_i - E(Y | X_i) \quad \forall i.$$

**For each population member** -- for each of the 60 families in our hypothetical population -- the **random error term**  $u_i$  equals the deviation of that population member's individual  $Y_i$  value from the population conditional mean value of  $Y$  for the corresponding value  $X_i$  of  $X$ .

**Terminology:** The random error term  $u_i$  is also known as the stochastic error term, the random disturbance term, or the stochastic disturbance term

- **Implication 1:** By simple re-arrangement of the above definition of  $u_i$ , it is obvious that each individual population value  $Y_i$  of  $Y$  can be written as

$$\begin{aligned} Y_i &= E(Y | X_i) + u_i \\ &= \beta_0 + \beta_1 X_i + u_i \quad \text{since } E(Y | X_i) = \beta_0 + \beta_1 X_i. \end{aligned}$$

This equation is called the **population regression equation**, or **PRE**.

**Interpretation:** The **PRE** indicates that **each population value**  $Y_i$  of  $Y$  can be expressed as the **sum of two components**:

- (1)  $E(Y | X_i) = \beta_0 + \beta_1 X_i$ 
  - = **the population conditional mean of  $Y$  for  $X = X_i$**
  - = the mean weekly consumption expenditure for all families in the population who have weekly disposable income  $X = X_i$ .
- (2)  $u_i =$  **the random error term for the  $i$ -th population member**
  - =  $Y_i - E(Y | X_i)$
  - = the deviation of family  $i$ 's weekly consumption expenditure  $Y_i$  from the population mean value  $E(Y | X_i)$  of all families in the population that have the same weekly disposable income  $X = X_i$ .

---

**Implication 2:** The population conditional mean value of the random error terms for each population value  $X_i$  of  $X$  equals 0 -- i.e.,

$$E(u_i | X_i) = 0 \quad \forall i.$$

**Proof:**

1. Take the conditional expectation for  $X = X_i$  of both sides of the PRE:

$$\begin{aligned} E(Y_i | X_i) &= E[E(Y | X_i)] + E(u_i | X_i) \\ &= E(Y | X_i) + E(u_i | X_i) \quad \text{since } E(Y | X_i) \text{ is a constant.} \end{aligned}$$

2. Since  $E(Y_i | X_i) = E(Y | X_i)$ , the above equation implies that  $E(u_i | X_i) = 0$ .

• ***What do the Random Error Terms  $u_i$  Represent?***

**The random error terms represent all the *unknown and unobservable* variables *other than X* that determine the *individual population values*  $Y_i$  of the dependent variable  $Y$ .**

They arise from the following factors:

1. ***Omitted variables*** that determine the population  $Y_i$  values
2. ***Intrinsic randomness*** in individual behaviour

- **Random Errors for Hypothetical Population of 60 Families**

**Random Error Terms for  $X_i = 100$** 

$Y_i$	$E(Y   X_i = 100)$	$u_i = Y_i - E(Y   X_i = 100)$
65	77	-12
70	77	-7
74	77	-3
80	77	3
85	77	8
88	77	11
Sum = 462		Sum = 0
Mean = $462/6 = 77$		Mean = 0

**Random Error Terms for  $X_i = 180$** 

$Y_i$	$E(Y   X_i = 180)$	$u_i = Y_i - E(Y   X_i = 180)$
110	125	-15
115	125	-10
120	125	-5
130	125	5
135	125	10
140	125	15
Sum = 750		Sum = 0
Mean = $750/6 = 125$		Mean = 0

**Random Error Terms for  $X_i = 240$** 

$Y_i$	$E(Y   X_i = 240)$	$u_i = Y_i - E(Y   X_i = 240)$
137	161	-24
145	161	-16
155	161	-6
165	161	4
175	161	14
189	161	28
Sum = 966		Sum = 0
Mean = $966/6 = 161$		Mean = 0

## The Sample Regression Function

- **Important Point 1:** Since in practice we do not observe the entire relevant population, and never know the true PRF, **we must estimate the PRF from sample data.**
- **Objective of Regression Analysis:** To estimate the PRF (population regression function) from sample data consisting of  $N$  randomly selected observations  $(X_i, Y_i)$ ,  $i = 1, \dots, N$  taken from the population.
- **Form of the Sample Regression Function (SRF):** The sample regression function, or SRF, takes the general form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (i = 1, \dots, N)$$

where

$\hat{Y}_i$  = an *estimate of the PRF*,  $f(X_i) = E(Y_i | X_i) = \beta_0 + \beta_1 X_i$ ;

$\hat{\beta}_0$  = an *estimate of the intercept coefficient*  $\beta_0$ ;

$\hat{\beta}_1$  = an *estimate of the slope coefficient*  $\beta_1$ .

- **Nature of the Sample Data:** A *sample* is a *randomly-selected subset of population members*.
  1. The sample observations  $\{(Y_i, X_i): i = 1, \dots, N\}$  are typically a small subset of the parent population of all population data points  $(Y_i, X_i)$ .  
  
Sample size  $N$  is much smaller than the number of population data points.
  2. **Each random sample** from a given population **yields one estimate of the PRF** -- i.e., one estimate of the numerical value of  $\beta_0$ , and one estimate of the numerical value of  $\beta_1$ .

- **Important Point 2: Each random sample** from the same population **yields a different SRF** -- i.e., a different numerical value of  $\hat{\beta}_0$ , and a different numerical value of  $\hat{\beta}_1$ .

**Example:** Consider **two random samples of 10 observations** from the population of 60 families. Each sample consists of one family for each of the 10 different population values of X.

**Tables 2.4 and 2.5**

Sample 1		Sample 2	
$X_i$	$Y_i$	$X_i$	$Y_i$
80	70	80	55
100	65	100	88
120	90	120	90
140	95	140	80
160	110	160	118
180	115	180	120
200	120	200	145
220	140	220	135
240	155	240	145
260	150	260	175

Because the two samples contain **different  $Y_i$  values** for the 10  $X_i$  values, they will yield **different SRFs** -- a different numerical value of  $\hat{\beta}_0$ , and a different numerical value of  $\hat{\beta}_1$ .

- **Sample 1 SRF (SRF<sub>1</sub>):**  $\hat{Y}_i = 24.46 + 0.5091X_i$ ,  
where the Sample 1 coefficient estimates are  $\hat{\beta}_0(1) = 24.46$  and  $\hat{\beta}_1(1) = 0.5091$
- **Sample 2 SRF (SRF<sub>2</sub>):**  $\hat{Y}_i = 17.17 + 0.5761X_i$ ,  
where the Sample 2 coefficient estimates are  $\hat{\beta}_0(2) = 17.17$  and  $\hat{\beta}_1(2) = 0.5761$

• **Figure 2.2** Plot of Sample Data Points and Sample Regression Functions for Random Samples 1 and 2

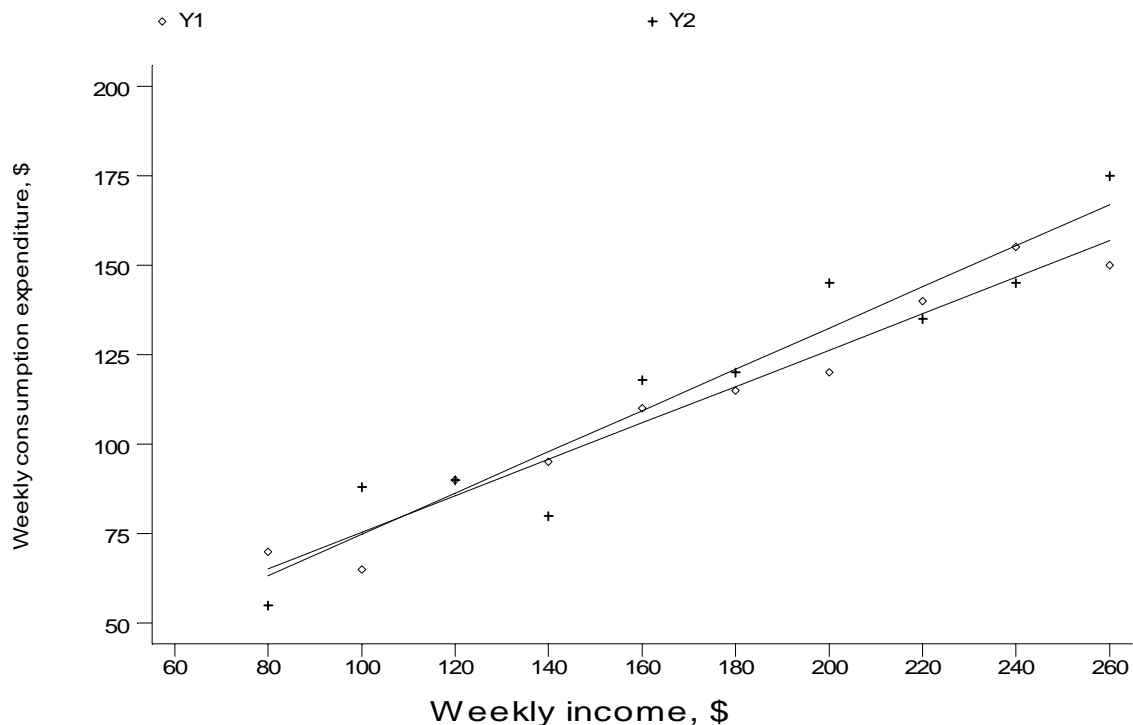
SRF<sub>1</sub> is the SRF based on Sample 1:  $\hat{Y}_i = 24.46 + 0.5091X_i$

SRF<sub>2</sub> is the SRF based on Sample 2:  $\hat{Y}_i = 17.17 + 0.5761X_i$

SRF<sub>1</sub> is the *flatter* regression line, SRF<sub>2</sub> is the *steeper* regression line.

**Important Points:**

- (1) Neither of these SRFs is identical to the true PRF. Each is merely an approximation to the true PRF.
- (2) How good an approximation any SRF provides to the true PRF depends on how the SRF is constructed from sample data -- i.e., on the properties of the coefficient estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .



### The Sample Regression Equation (SRE)

- The **sample regression equation (SRE)** is the sample counterpart of the population regression equation (PRE)

$$Y_i = f(X_i) + u_i = E(Y_i | X_i) + u_i = \beta_0 + \beta_1 X_i + u_i \quad \Leftarrow \text{the PRE}$$

- Form of the Sample Regression Equation (SRE):** The sample regression equation, or SRE, takes the general form

$$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i \quad (i = 1, \dots, N) \quad \Leftarrow \text{the SRE}$$

where

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \text{an } \textit{estimate of the PRF}, f(X_i) = E(Y_i | X_i) = \beta_0 + \beta_1 X_i;$$

$$\hat{\beta}_0 = \text{an } \textit{estimate of the intercept coefficient } \beta_0;$$

$$\hat{\beta}_1 = \text{an } \textit{estimate of the slope coefficient } \beta_1.$$

$$\hat{u}_i = \text{the } \textit{residual} \text{ for sample observation } i.$$

- Interpretation of the SRE:** The SRE represents each sample value of Y -- each  $Y_i$  value -- as the **sum of two components**:

(1) the **estimated (or predicted) value of Y** for each sample value  $X_i$  of X, i.e.,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (i = 1, \dots, N);$$

(2) the **residual** corresponding to the i-th sample observation, i.e.,

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \quad (i = 1, \dots, N).$$

$$\hat{u}_i = \text{the residual for the } i\text{-th sample observation}$$

$$= \text{the } \textit{observed} \text{ Y-value } (Y_i) - \text{the } \textit{estimated} \text{ Y-value } (\hat{Y}_i)$$

### Compare the Population and Sample Regression Equations: the PRE and SRE

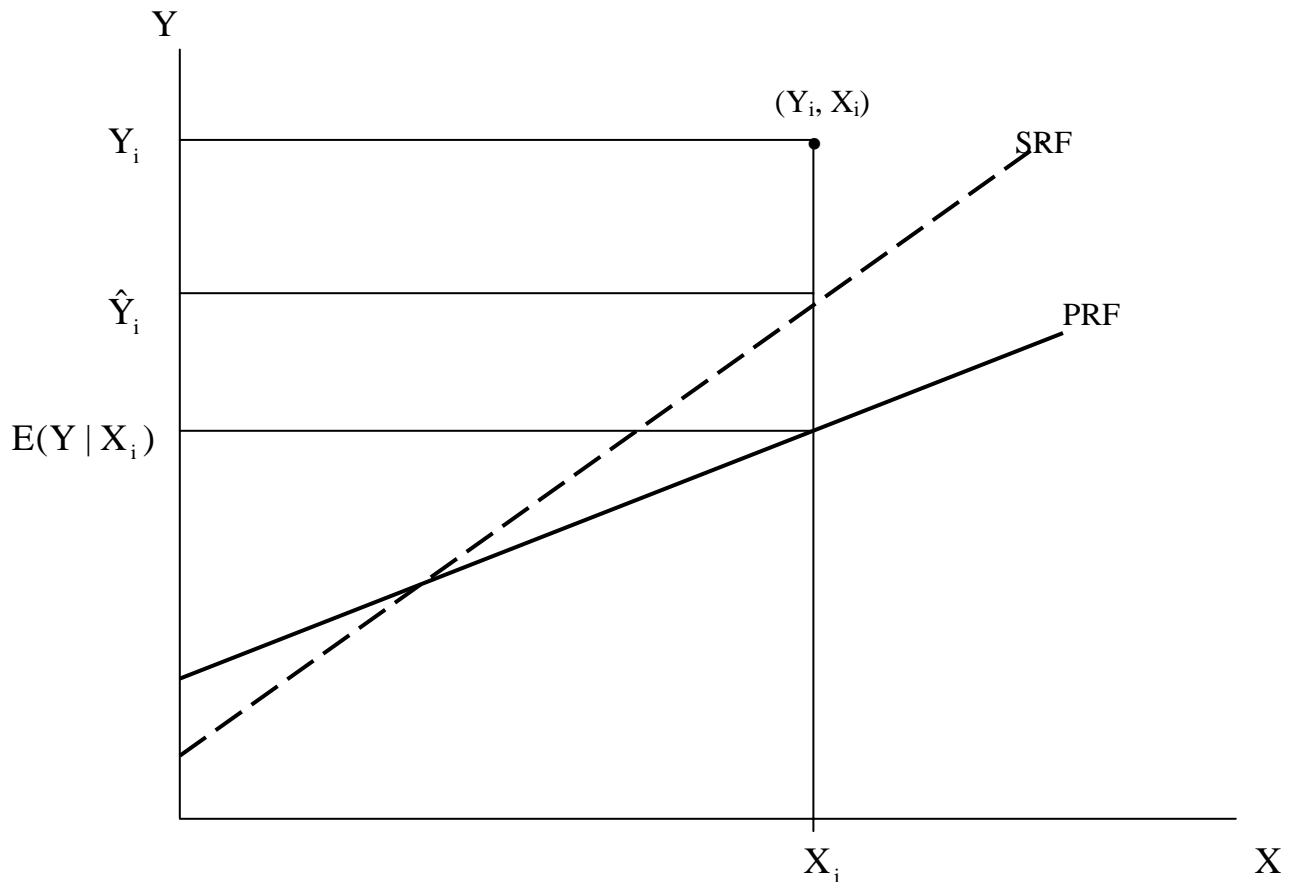
- The PRE for  $Y_i$  is:

$$Y_i = f(X_i) + u_i = E(Y_i | X_i) + u_i = \beta_0 + \beta_1 X_i + u_i$$

- The SRE for  $Y_i$  is:

$$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

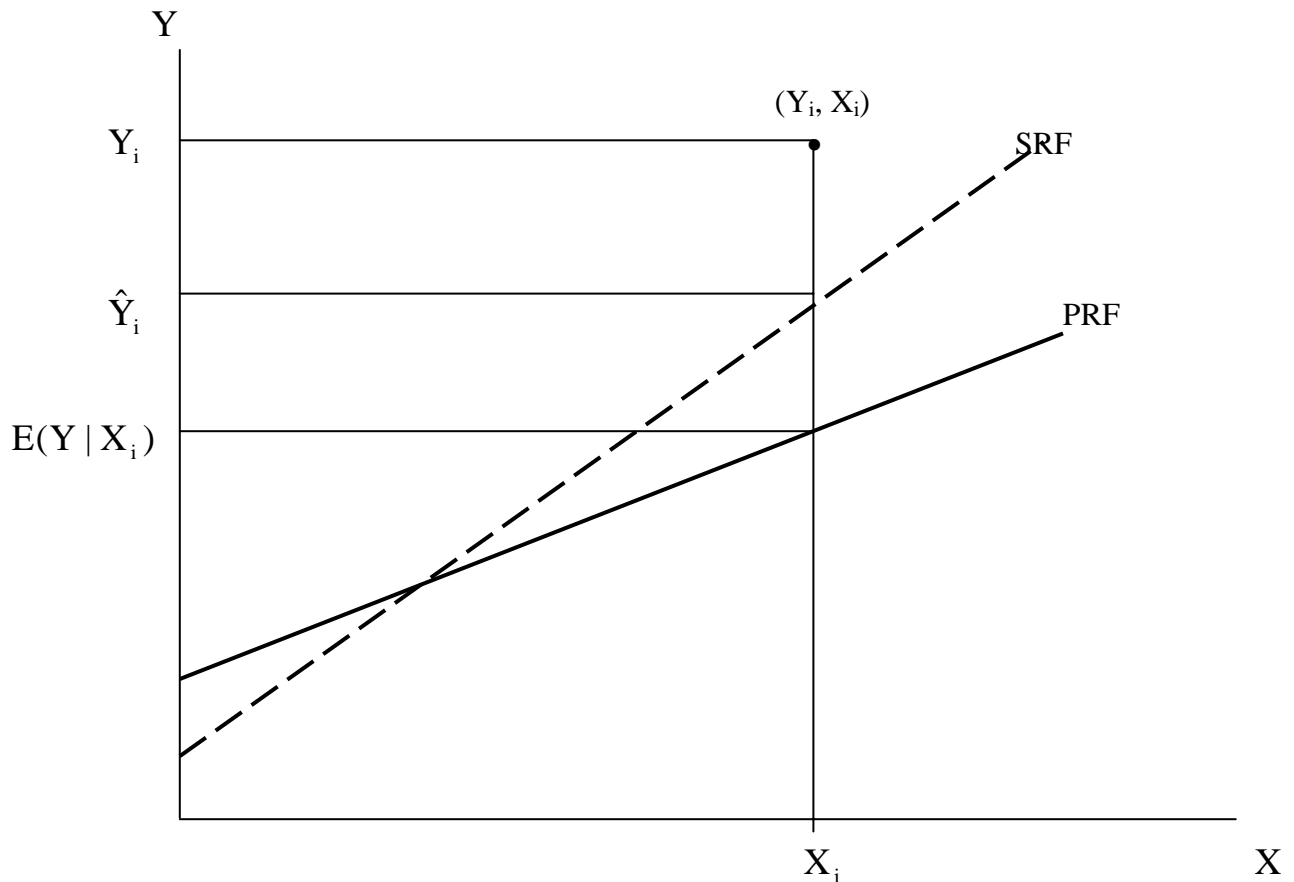
- **Figure 2.3: Comparison of Population and Sample Regression Lines**



- ♦ The *population regression line* is a plot of the PRF:  $E(Y_i | X_i) = \beta_0 + \beta_1 X_i$ .
- ♦ The *sample regression line* is a plot of the SRF:  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ .



- **Figure 2.3: Comparison of Population and Sample Regression Lines**



At  $X = X_i$ :

- ♦ The population regression equation (PRE) represents the population value  $Y_i$  of  $Y$  as the sum of two parts:

$$Y_i = E(Y_i | X_i) + u_i = \beta_0 + \beta_1 X_i + u_i, \text{ where } E(Y_i | X_i) = \beta_0 + \beta_1 X_i$$

$$u_i = Y_i - E(Y_i | X_i) = Y_i - \beta_0 - \beta_1 X_i = \text{distance between } Y_i \text{ and } E(Y_i | X_i)$$

- ♦ The sample regression equation (SRE) represents the population value  $Y_i$  of  $Y$  as the sum of two parts:

$$Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i, \text{ where } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = \text{distance between } Y_i \text{ and } \hat{Y}_i.$$