REGRESSION ANALYSIS

- **Ordinary Least Squares Method (*OLS*)**

Recall the two-variable PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.4.2}$$

estimate it from the SRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

$$= \hat{Y}_i + \hat{u}_i \tag{2.6.3}$$

where $\hat{Y}_i$ is the estimated (conditional mean) value of Y_i.

But how is the SRF itself determined? To see this, let us proceed as follows. First, express (2.6.3) as

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \end{aligned} \tag{3.1.1}$$

which shows that the $\hat{u}_i$ (the residuals) are simply the differences between the actual and estimated Y values.

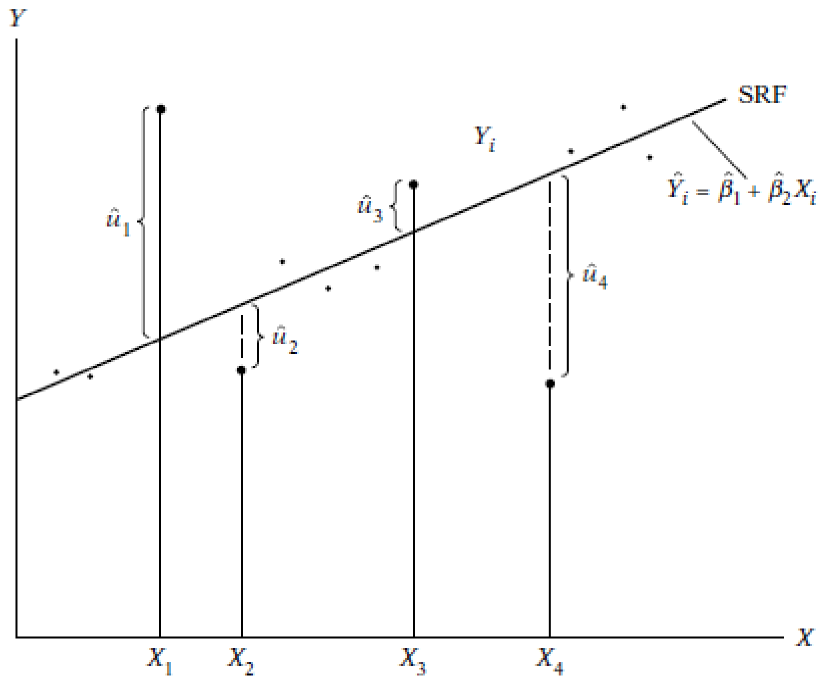

FIGURE 3.1 Least-squares criterion.

adopt the *least-squares criterion,* which states that the SRF can be fixed in such a way that

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad (3.1.2)$$

is as small as possible, where $\hat{u}_i^2$ are the squared residuals. By squaring $\hat{u}_i$, this method gives more weight to residuals such as $\hat{u}_1$ and $\hat{u}_4$ in Figure 3.1 than the residuals $\hat{u}_2$ and $\hat{u}_3$. As noted previously, under the minimum $\sum \hat{u}_i$

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \quad (3.1.4)$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \quad (3.1.5)$$

where n is the sample size. These simultaneous equations are known as the **normal equations.**

Solving the normal equations simultaneously, we obtain

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2} \quad (3.1.6)$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y and where we define $x_i = (X_i - \bar{X})$ and $y_i = (Y_i - \bar{Y})$. *Henceforth we adopt the convention of letting the lowercase letters denote deviations from mean values.*

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} = \bar{Y} - \hat{\beta}_2 \bar{X} \quad (3.1.7)$$

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} \qquad \textbf{(3.1.7)}$$
$$= \bar{Y} - \hat{\beta}_2 \bar{X}$$

- **Properties of OLS estimators:**

1. It passes through the sample means of Y and X.

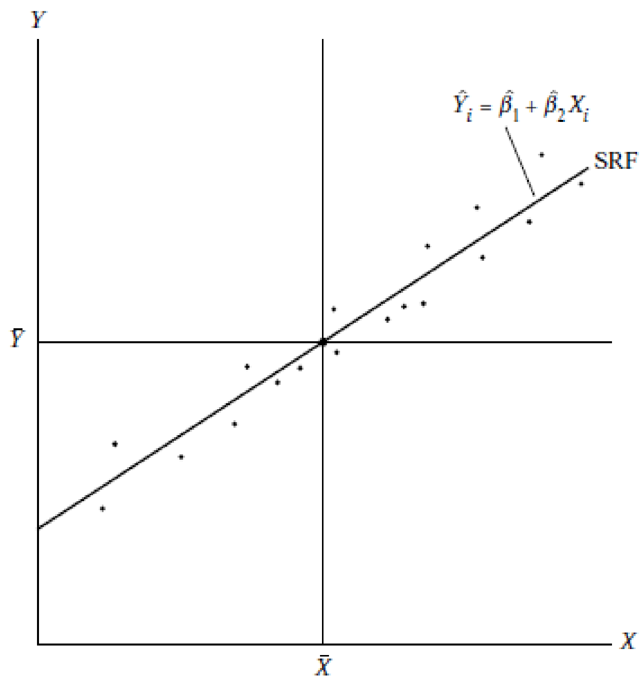

FIGURE 3.2 Diagram showing that the sample regression line passes through the sample mean values of Y and X.

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

$$\bar{\hat{Y}} = \bar{Y} \qquad \textbf{(3.1.10)}$$

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

$$E(u_i \mid X_i) = 0 \qquad \textbf{(3.2.1)}$$

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

5. The residuals $\hat{u}_i$ are uncorrelated with X_i; that is, $\sum \hat{u}_i X_i = 0$.

error (se).[17] Given the Gaussian assumptions, it is shown in Appendix 3A, Section 3A.3 that the standard errors of the OLS estimates can be obtained

as follows:

$$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \tag{3.3.1}$$

$$\operatorname{se}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} \tag{3.3.2}$$

$$\operatorname{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2}\sigma^2 \tag{3.3.3}$$

$$\operatorname{se}(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}\sigma \tag{3.3.4}$$

where var = variance and se = standard error and where σ^2 is the constant or homoscedastic variance of u_i of Assumption 4.

All the quantities entering into the preceding equations except σ^2 can be estimated from the data. As shown in Appendix 3A, Section 3A.5, σ^2 itself is estimated by the following formula:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \tag{3.3.5}$$

where $\hat{\sigma}^2$ is the OLS estimator of the true but unknown σ^2 and where the expression $n-2$ is known as the **number of degrees of freedom (df)**, $\sum \hat{u}_i^2$ being the sum of the residuals squared or the **residual sum of squares (RSS).**[18]

We now define r^2 as

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}} \tag{3.5.5}$$

or, alternatively, as

$$\begin{aligned} r^2 &= 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{RSS}}{\text{TSS}} \end{aligned} \tag{3.5.5a}$$

The quantity r^2 thus defined is known as the (sample) **coefficient of determination** and is the most commonly used measure of the goodness of fit of a regression line. Verbally, r^2 *measures the proportion or percentage of the total variation in Y explained by the regression model.*

Two properties of r^2 may be noted:

1. It is a nonnegative quantity. (Why?)

2. Its limits are $0 \le r^2 \le 1$. An r^2 of 1 means a perfect fit, that is, $\hat{Y}_i = Y_i$ for each i. On the other hand, an r^2 of zero means that there is no relationship between the regressand and the regressor whatsoever (i.e., $\hat{\beta}_2 = 0$). In this case, as (3.1.9) shows, $\hat{Y}_i = \hat{\beta}_1 = \bar{Y}$, that is, the best prediction of any Y value is simply its mean value. In this situation therefore the regression line will be horizontal to the X axis.

$$r^2 = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2} \tag{3.5.8}$$

TABLE 3.2 HYPOTHETICAL DATA ON WEEKLY FAMILY CONSUMPTION EXPENDITURE *Y* AND WEEKLY FAMILY INCOME *X*

Y, $	*X*, $
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

TABLE 3.3 RAW DATA BASED ON TABLE 3.2

	Y_i (1)	X_i (2)	Y_iX_i (3)	X_i^2 (4)	$x_i = X_i - \bar{X}$ (5)	$y_i = Y_i - \bar{Y}$ (6)	x_i^2 (7)	x_iy_i (8)	$\hat{Y}_i$ (9)	$\hat{u}_i = Y_i - \hat{Y}_i$ (10)	$\hat{Y}_i\hat{u}_i$ (11)
	70	80	5600	6400	−90	−41	8100	3690	65.1818	4.8181	314.0524
	65	100	6500	10000	−70	−46	4900	3220	75.3636	−10.3636	−781.0382
	90	120	10800	14400	−50	−21	2500	1050	85.5454	4.4545	381.0620
	95	140	13300	19600	−30	−16	900	480	95.7272	−0.7272	−69.6128
	110	160	17600	25600	−10	−1	100	10	105.9090	4.0909	433.2631
	115	180	20700	32400	10	4	100	40	116.0909	−1.0909	−126.6434
	120	200	24000	40000	30	9	900	270	125.2727	−6.2727	−792.0708
	140	220	30800	48400	50	29	2500	1450	136.4545	3.5454	483.7858
	155	240	37200	57600	70	44	4900	3080	145.6363	8.3636	1226.4073
	150	260	39000	67600	90	39	8100	3510	156.8181	−6.8181	−1069.2014
Sum	1110	1700	205500	322000	0	0	33000	16800	1109.9995 ≈ 1110.0	0	0.0040 ≈ 0.0
Mean	111	170	nc	nc	0	0	nc	nc	110	0	0

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = 16{,}800/33{,}000 = 0.5091$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = 111 - 0.5091(170) = 24.4545$$

Notes: ≈ symbolizes "approximately equal to"; nc means "not computed."

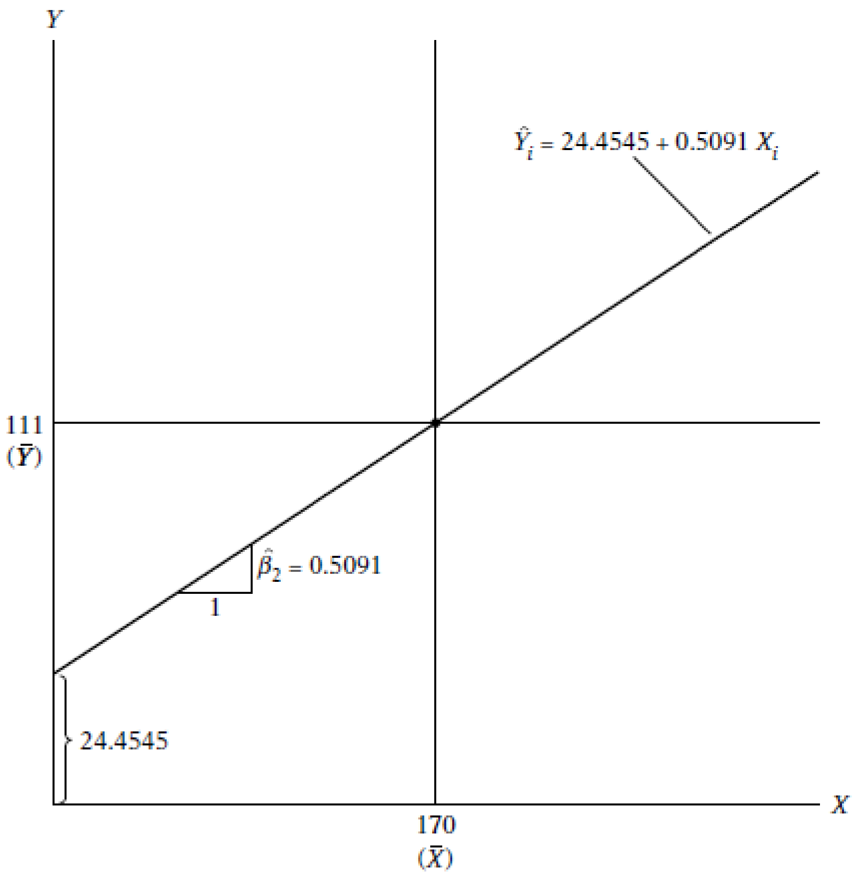

FIGURE 3.12 Sample regression line based on the data of Table 3.2.