CHAPTER 6

Simple Regression Analysis

6.1 THE TWO-VARIABLE LINEAR MODEL

The two-variable linear model, or *simple regression analysis*, is used for testing hypotheses about the relationship between a dependent variable Y and an independent or explanatory variable X and for prediction. Simple *linear* regression analysis usually begins by plotting the set of XY values on a *scatter diagram* and determining by inspection if there exists an approximate linear relationship:

$$Y_i = b_0 + b_1 X_i \tag{6.1}$$

Since the points are unlikely to fall precisely on the line, the exact linear relationship in Eq. (*6.1*) must be modified to include a *random disturbance*, *error*, or *stochastic term*, u_i (see Sec. 1.2 and Prob. 1.8):

$$Y_i = b_0 + b_1 X_i + u_i \tag{6.2}$$

The error term is assumed to be (1) normally distributed, with (2) zero expected value or mean, and (3) constant variance, and it is further assumed (4) that the error terms are uncorrelated or unrelated to each other, and (5) that the explanatory variable assumes fixed values in repeated sampling (so that X_i and u_i are also uncorrelated).

EXAMPLE 1. Table 6.1 gives the bushels of corn per acre, Y, resulting from the use of various amounts of fertilizer in pounds per acre, X, produced on a farm in each of 10 years from 1971 to 1980. These are plotted in the scatter diagram of Fig. 6-1. The relationship between X and Y in Fig. 6-1 is approximately linear (i.e., the points would fall on or near a straight line).

6.2 THE ORDINARY LEAST-SQUARES METHOD

The *ordinary least-squares method* (OLS) is a technique for fitting the "best" straight line to the sample of XY observations. It involves minimizing the sum of the squared (vertical) deviations of points from the line:

$$\text{Min} \sum (Y_i - \hat{Y}_i)^2 \tag{6.3}$$

where Y_i refers to the actual observations, and $\hat{Y}_i$ refers to the corresponding *fitted* values, so that $Y_i - \hat{Y}_i = e_i$, the *residual*. This gives the following two *normal equations* (see Prob. 6.6):

Table 6.1 Corn Produced with Fertilizer Used

Year	n	Y_i	X_i
1971	1	40	6
1972	2	44	10
1973	3	46	12
1974	4	48	14
1975	5	52	16
1976	6	58	18
1977	7	60	22
1978	8	68	24
1979	9	74	26
1980	10	80	32

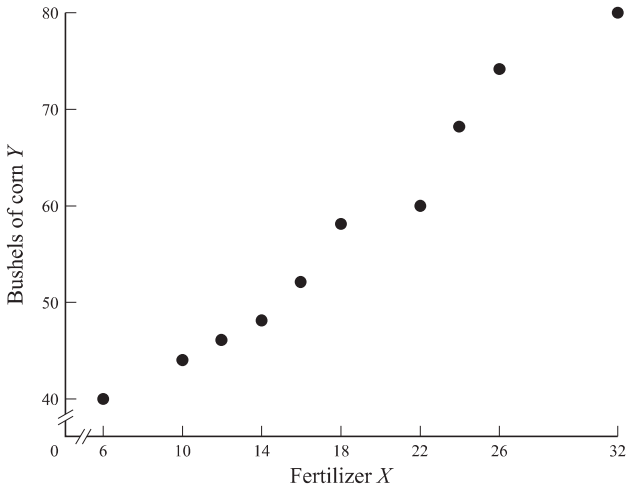

Fig. 6-1

$$\sum Y_i = nb_0 + \hat{b}_1 \sum X_i \tag{6.4}$$

$$\sum X_i Y_i = \hat{b}_0 \sum X_i + \hat{b}_1 \sum X_i^2 \tag{6.5}$$

where n is the number of observations and $\hat{b}_0$ and $\hat{b}_1$ are estimators of the true parameters b_0 and b_1.

Solving simultaneously Eqs. (*6.4*) and (*6.5*), we get [see Prob. 6.7(*a*)]

$$\hat{b}_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \tag{6.6}$$

The value of $\hat{b}_0$ is then given by [see Prob. 6.7(*b*)]

$$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X} \tag{6.7}$$

It is often useful to use an equivalent formula for estimating $\hat{b}_1$ [see Prob. 6.10(*a*)]:

$$\hat{b}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\text{cov}(X, Y)}{\sigma_X^2} \tag{6.8}$$

where $x_i = X_i - \overline{X}$, and $y_i = Y_i - \overline{Y}$. The estimated least-squares regression (OLS) equation is then

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i \tag{6.9}$$

EXAMPLE 2. Table 6.2 shows the calculations to estimate the regression equation for the corn-fertilizer problem in Table 6.1. Using Eq. (*6.8*),

Table 6.2 Corn Produced with Fertilizer Used: Calculations

n	Y_i (Corn)	X_i (Fertilizer)	y_i	x_i	$x_i y_i$	x_i^2
1	40	6	−17	−12	204	144
2	44	10	−13	−8	104	64
3	46	12	−11	−6	66	36
4	48	14	−9	−4	36	16
5	52	16	−5	−2	10	4
6	58	18	1	0	0	0
7	60	22	3	4	12	16
8	68	24	11	6	66	36
9	74	26	17	8	136	64
10	80	32	23	14	322	196
$n = 10$	$\sum Y_i = 570$ $\overline{Y} = 57$	$\sum X_i = 180$ $\overline{X} = 18$	$\sum y_i = 0$	$\sum x_i = 0$	$\sum x_i y_i = 956$	$\sum x_i^2 = 576$

$$\hat{b}_i = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{956}{576} = 1.66 \qquad \text{(the slope of the estimated regression line)}$$

$$\hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X} \cong 57 - (1.66)(18) \cong 57 - 29.88 \cong 27.12 \qquad \text{(the } Y \text{ intercept)}$$

$$\hat{Y}_i = 27.12 + 1.66 X_i \qquad \text{(the estimated regression equation)}$$

Thus, when $X_i = 0$, $\hat{Y} = 27.12 = \hat{b}_0$. When $X_i = 18 = \overline{X}$, $\hat{Y} = 27.12 + 1.66(18) = 57 = \overline{Y}$. As a result, the regression line passes through point $\overline{XY}$ (see Fig. 6-2).

6.3 TESTS OF SIGNIFICANCE OF PARAMETER ESTIMATES

In order to test for the statistical significance of the parameter estimates of the regression, the variance of $\hat{b}_0$ and $\hat{b}_1$ is required (see Probs. 6.14 and 6.15):

$$\operatorname{Var} \hat{b}_0 = \sigma_u^2 \frac{\sum X_i^2}{n \sum x_i^2} \tag{6.10}$$

$$\operatorname{Var} \hat{b}_1 = \sigma_u^2 \frac{1}{\sum x_i^2} \tag{6.11}$$

Since σ_u^2 is unknown, the *residual variance* s^2 is used as an (unbiased) estimate of σ_u^2:

$$s^2 = \hat{\sigma}_u^2 = \frac{\sum e_i^2}{n - k} \tag{6.12}$$

where k represents the number of parameter estimates.

Unbiased estimates of the variance of $\hat{b}_0$ and $\hat{b}_1$ are then given by

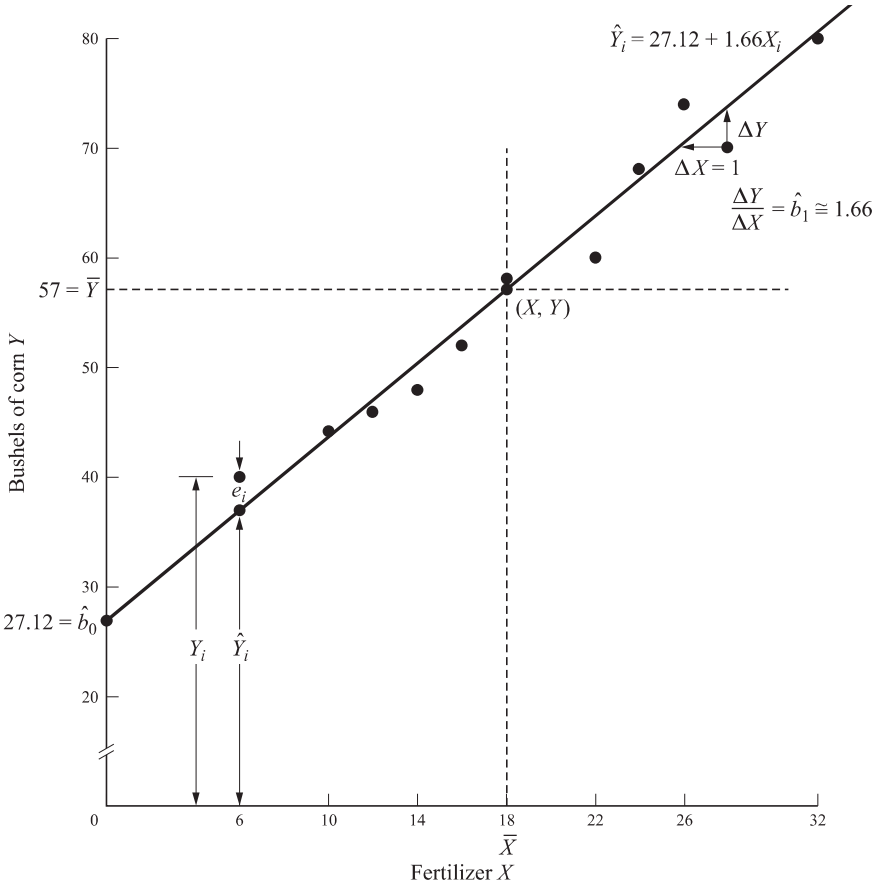

Fig. 6-2

$$s^2_{\hat{b}_0} = \frac{\sum e_i^2}{n-k}\frac{\sum X_i^2}{n\sum x_i^2} \tag{6.13}$$

$$s^2_{\hat{b}_1} = \frac{\sum e_i^2}{n-k}\frac{1}{\sum x_i^2} \tag{6.14}$$

so that $s_{\hat{b}_0}$ and $s_{\hat{b}_1}$ are the *standard errors of the estimates*. Since u_i is normally distributed, Y_i and therefore $\hat{b}_0$ and $\hat{b}_1$ are also normally distributed, so that we can use the t distribution with $n-k$ degrees of freedom, to test hypotheses about and construct confidence intervals for $\hat{b}_0$ and $\hat{b}_1$ (see Secs. 4.4 and 5.2).

EXAMPLE 3. Table 6.3 (an extension of Table 6.2) shows the calculations required to test the statistical significance of $\hat{b}_0$ and $\hat{b}_1$. The values of $\hat{Y}_i$ in Table 6.3 are obtained by substituting the values of X_i into the estimated regression equation found in Example 2. (The values of y_i^2 are obtained by squaring y_i from Table 6.2 and are to be used in Sec. 6.4.)

$$s^2_{\hat{b}_0} = \frac{\sum e_i^2}{n-k}\frac{\sum X_i^2}{n\sum x_i^2} \cong \frac{47.3056}{10-2}\frac{3816}{10(576)} \cong 3.92 \qquad \text{and} \qquad s_{\hat{b}_0} = \sqrt{3.92} \cong 1.98$$

$$s^2_{\hat{b}_1} = \frac{\sum e_i^2}{(n-k)\sum x_i^2} \cong \frac{47.3056}{(10-2)576} \cong 0.01 \qquad \text{and} \qquad s_{\hat{b}_1} \cong \sqrt{0.01} \cong 0.1$$

Therefore $$t_0 = \frac{\hat{b}_0 - b_0}{s_{\hat{b}_0}} \cong \frac{27.12-0}{1.98} \cong 13.7 \qquad \text{and} \qquad t_1 = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \cong \frac{1.66}{0.1} \cong 16.6$$

Table 6.3 Corn-Fertilizer Calculations to Test Significance of Parameters

Year	Y_i	X_i	$\hat{Y}_i$	e_i	e_i^2	X_i^2	x_i^2	y_i^2
1	40	6	37.08	2.92	8.5264	36	144	289
2	44	10	43.72	0.28	0.0784	100	64	169
3	46	12	47.04	−1.04	1.0816	144	36	121
4	48	14	50.36	−2.36	5.5696	196	16	81
5	52	16	53.68	−1.68	2.8224	256	4	25
6	58	18	57.00	1.00	1.0000	324	0	1
7	60	22	63.64	−3.64	13.2496	484	16	9
8	68	24	66.96	1.04	1.0816	576	36	121
9	74	26	70.28	3.72	13.8384	676	64	289
10	80	32	80.24	−0.24	0.0576	1024	196	529
$n = 10$				$\sum e_i = 0$	$\sum e_i^2 = 47.3056$	$\sum X_i^2 = 3816$	$\sum x_i^2 = 576$	$\sum y_i^2 = 1634$

Since both t_0 and t_1 exceed $t = 2.306$ with 8 df at the 5% level of significance (from App. 5), we conclude that both b_0 and b_1 are statistically significant at the 5% level.

6.4 TEST OF GOODNESS OF FIT AND CORRELATION

The closer the observations fall to the regression line (i.e., the smaller the residuals), the greater is the variation in Y "explained" by the estimated regression equation. The total variation in Y is equal to the explained plus the residual variation:

$$\begin{array}{ccccc} \sum(Y_i - \overline{Y})^2 & = & \sum(\hat{Y}_i - \overline{Y})^2 & + & \sum(Y_i - \hat{Y}_i)^2 \\ \text{Total variation in } Y \text{ [or total sum of squares (TSS)]} & = & \text{Explained variation in } Y \text{ [or regression sum of squares (RSS)]} & + & \text{Residual variation in } Y \text{ [or error sum of squares (ESS)]} \end{array} \tag{6.15}$$

Dividing both sides by TSS gives

$$1 = \frac{\text{RSS}}{\text{TSS}} + \frac{\text{ESS}}{\text{TSS}}$$

The *coefficient of determination*, or R^2, is then defined as the proportion of the total variation in Y "explained" by the regression of Y on X:

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}} \tag{6.16}$$

R^2 can be calculated by

$$R^2 = \frac{\sum \hat{y}^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2} \tag{6.17}$$

where $$\sum \hat{y}_i^2 = \sum(\hat{Y}_i - \overline{Y}_i)^2$$

R^2 ranges in value from 0 (when the estimated regression equation explains none of the variation in Y) to 1 (when all points lie on the regression line).

The *correlation coefficient* r is given by (see Prob. 6.22)

$$r = \sqrt{R^2} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \sqrt{\hat{b}_1 \frac{\sum x_i y_i}{\sum y_i^2}} \tag{6.18}$$

r ranges in value from -1 (for perfect negative linear correlation) to $+1$ (for perfect positive linear correlation) and does not imply causality or dependence. With qualitative data, the *rank* or (the *Spearman*) *correlation coefficient* r' (see Prob. 6.25) can be used.

EXAMPLE 4. The coefficient of determination for the corn-fertilizer example can be found from Table 6.3:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} \cong 1 - \frac{47.31}{1634} \cong 1 - 0.0290 \cong 0.9710, \text{ or } 97.10\%$$

Thus the regression equation explains about 97% of the total variation in corn output. The remaining 3% is attributed to factors included in the error term. Then $r = \sqrt{R^2} \cong \sqrt{0.9710} \cong 0.9854$, or 98.54%, and is positive because $\hat{b}_1$ is positive. Figure 6-3 shows the total, the explained, and the residual variation of Y.

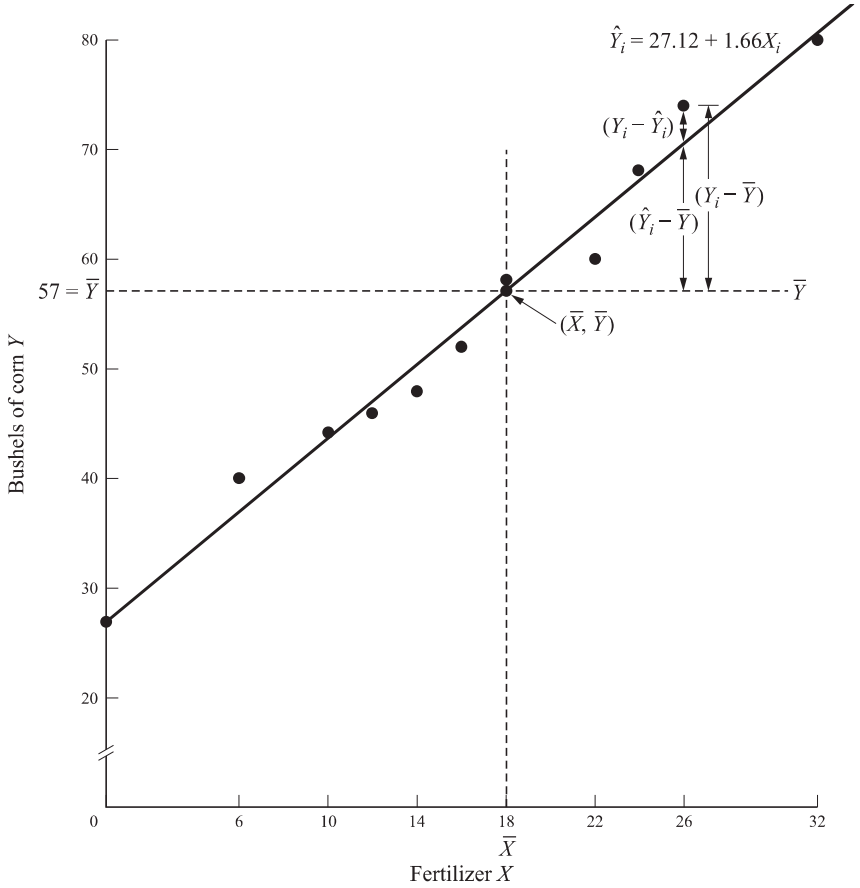

Fig. 6-3

6.5 PROPERTIES OF ORDINARY LEAST-SQUARES ESTIMATORS

Ordinary least-squares (OLS) estimators are *best linear unbiased estimators* (BLUE). Lack of bias means

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

so that

$$\text{Bias} = E(\hat{b}) - b$$

Best unbiased or efficient means smallest variance. Thus OLS estimators are the best among all unbiased linear estimators [see Probs. 6.14(*a*) and 6.15(*b*)]. This is known as the *Gauss-Markov theorem* and represents the most important justification for using OLS.

Sometimes, a researcher may want to trade off some bias for a possibly smaller variance and minimize the mean square error, MSE (see Prob. 6.29):

$$\text{MSE}(\hat{b}) = E(\hat{b} - b)^2 = \text{var}(\hat{b}) + (\text{bias}\,\hat{b})^2$$

An estimator is *consistent* if, as the sample size approaches infinity in the limit, its value approaches the true parameter (i.e., it is asymptotically unbiased) and its distribution collapses on the true parameter (see Prob. 6.30).

EXAMPLE 5. OLS estimators $\hat{b}_0$ and $\hat{b}_1$ found in Example 2 are unbiased linear estimators of b_0 and b_1 because

$$E(\hat{b}_0) = b_0 \qquad \text{and} \qquad E(\hat{b}_1) = b_1$$

$\text{Var}\,\hat{b}_0$ and $\text{var}\,\hat{b}_1$ found in Example 3 are also lower than for any other linear unbiased estimators. Therefore $\hat{b}_0$ and $\hat{b}_1$ are BLUE.

Solved Problems

THE TWO-VARIABLE LINEAR MODEL

6.1 What is meant by and what is the function of (*a*) Simple regression analysis? (*b*) Linear regression analysis? (*c*) A scatter diagram? (*d*) An error term?

(*a*) *Simple regression* is used for testing hypotheses about the relationship between a dependent variable Y and an independent or explanatory variable X and for prediction. This is to be contrasted with *multiple regression* analysis, in which there are not one, but two or more independent or explanatory variables. Multiple regression analysis is discussed in Chap. 7.

(*b*) *Linear regression analysis* assumes that there is an approximate linear relationship between X and Y (i.e., the set of random sample values of X and Y fall on or near a straight line). This is to be contrasted with *nonlinear regression analysis* (discussed in Sec. 8.1).

(*c*) A *scatter diagram* is a figure in which each pair of independent-dependent observations is plotted as a point in the XY plane. Its purpose is to determine (by inspection) if there exists an approximate linear relationship between the dependent variable Y and the independent or explanatory variable X.

(*d*) The *error term* (also known as the *disturbance* or *stochastic term*) measures the deviation of each observed Y value from the true (but unobserved) regression line. These error terms, designated by u_i and e_i, arise because of (1) numerous explanatory variables with only slight and irregular effects on Y that are omitted from the exact linear relationship given by Eq. (*6.1*), (2) possible errors of measurement in Y, and (3) random human behavior (see Prob. 1.8).

6.2 The data in Table 6.4 reports the aggregate consumption (Y, in billions of U.S. dollars) and disposable income (X, also in billions of U.S. dollars) for a developing economy for the 12 years from 1988 to 1999. Draw a scatter diagram for the data and determine by inspection if there exists an approximate linear relationship between Y and X.

From Fig. 6-4 it can be seen that the relationship between consumption expenditures Y and disposable income X is approximately linear, as required by the linear regression model.

6.3 State the general relationship between consumption Y and disposable income X in (*a*) exact linear form and (*b*) stochastic form. (*c*) Why would you expect most observed values of Y not to fall exactly on a straight line?

(*a*) The exact or deterministic general relationship between aggregate consumption expenditures Y and aggregate disposable income X can be written as