Chapter 14.1 Simple Linear Regression Model (迴歸模型基本概念介紹)

- Managerial decisions often are based on the relationship between two or more variables.
- Regression analysis can be used to develop an equation showing how the variables are related.
- The variable being predicted is called the dependent variable and is denoted by *y*.
- The variables being used to predict the value of the dependent variable are called the independent variables and are denoted by *x*.
- For example, the effect of advertising expenditures on sales…

- The equation that describes how *y* is related to *x* and an error term is called the regression model.

- The simple linear regression model is: $y = \beta_0 + \beta_1 x + \varepsilon$

The estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x$$

- The graph is called the estimated regression line.
- b_0 is the y intercept of the line.
- b_1 is the slope of the line.
- $\hat{y}$ is the estimated value of y for a given x value.

FIGURE 14.1 POSSIBLE REGRESSION LINES IN SIMPLE LINEAR REGRESSION

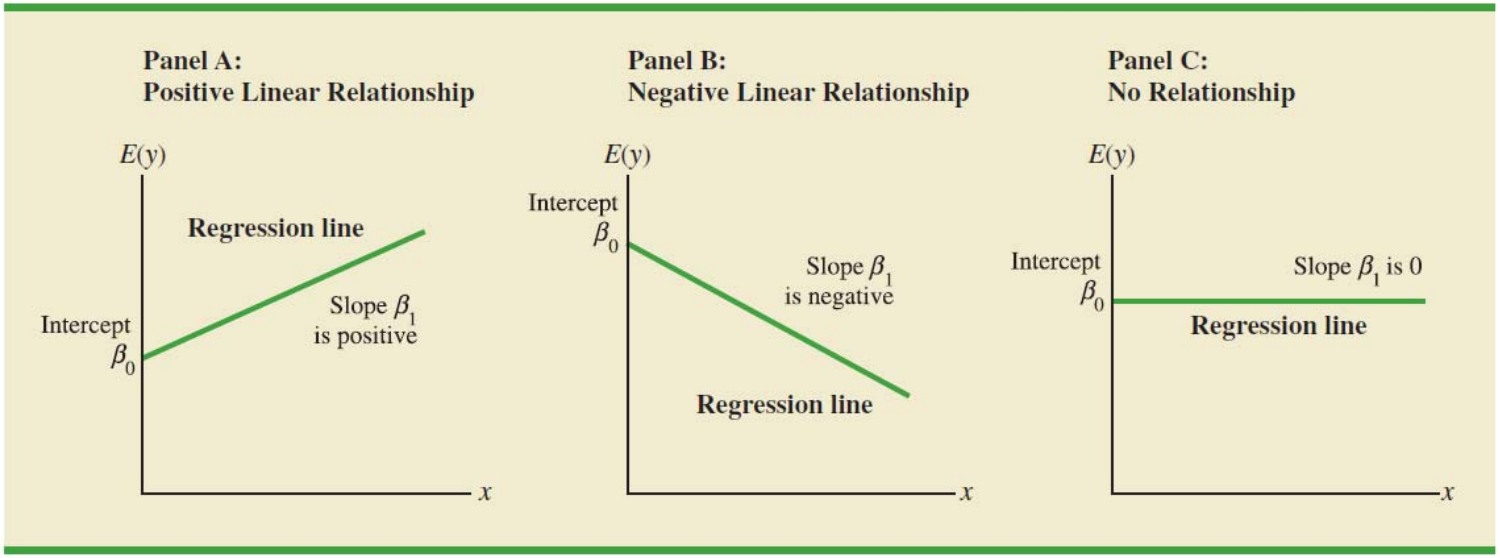
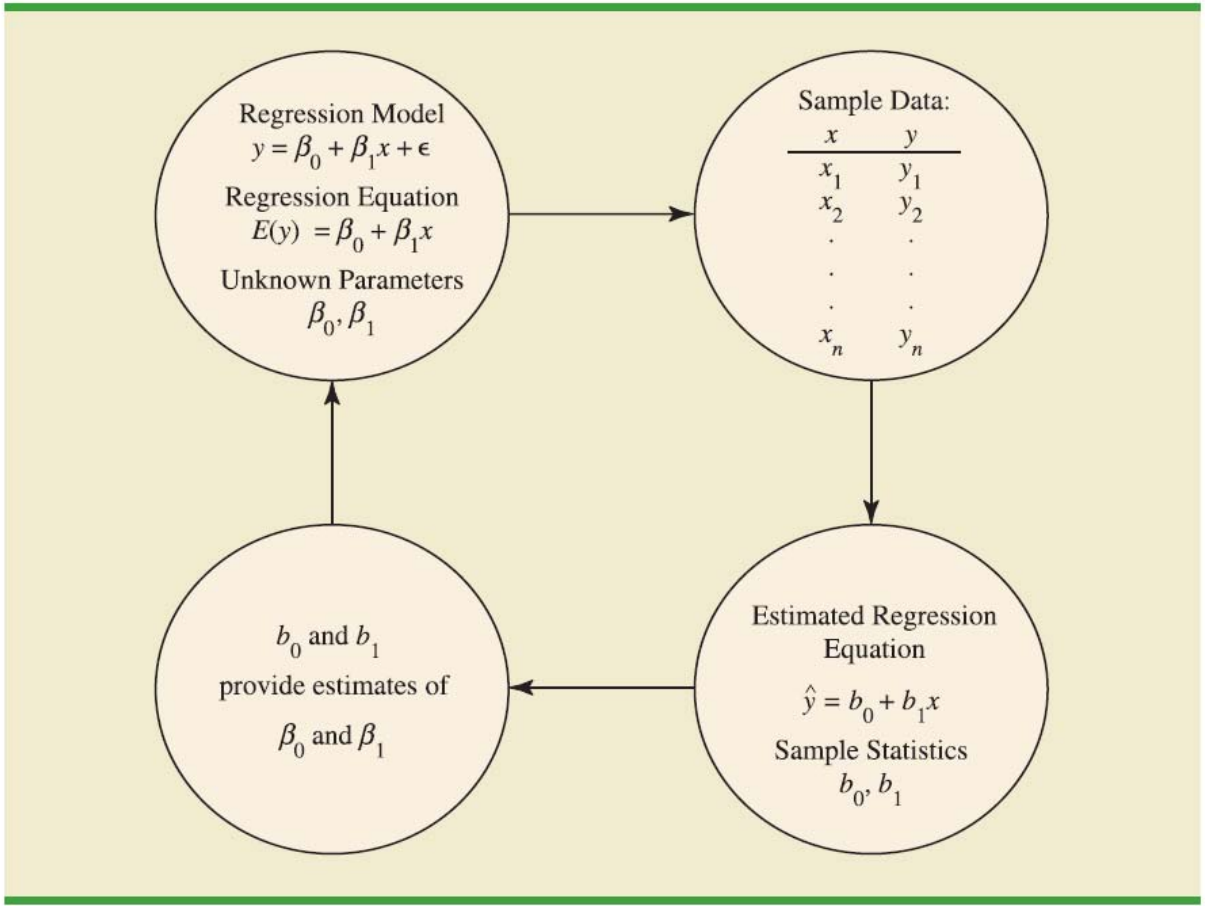


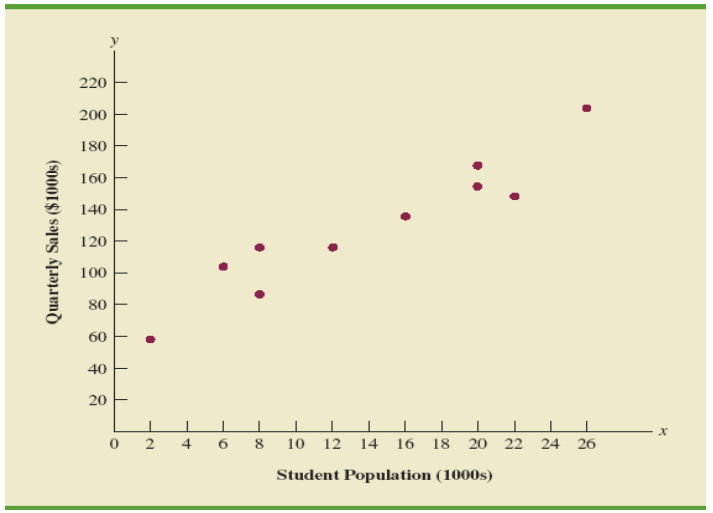
FIGURE 14.2 THE ESTIMATION PROCESS IN SIMPLE LINEAR REGRESSION



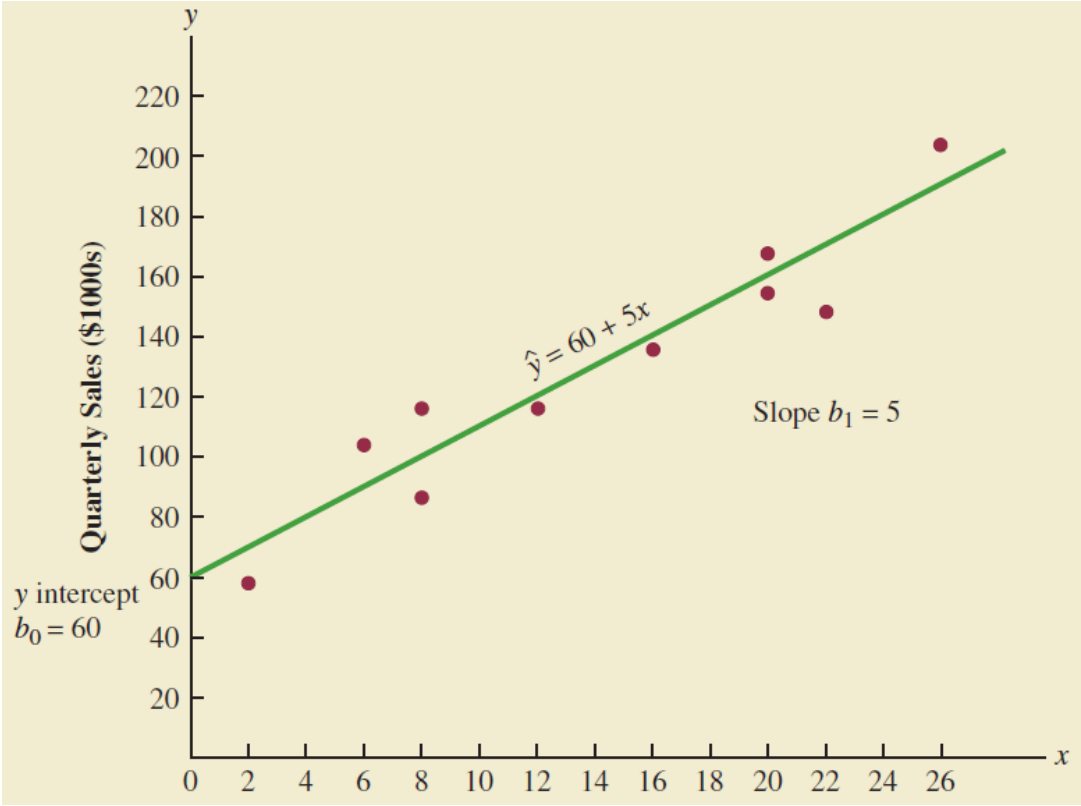
Chapter 14.2 Least Squares Methods (最小平方法)

- The least squares method is a procedure for using sample data to find the estimated regression equation.
- For example, suppose data were collected from a sample of 10 Armand's Pizza Parlor restaurants located near college campuses.

Restaurant i	Student Population (1000s) x_i	Quarterly Sales ($1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202



透過最小平方法，將這十筆樣本資料建構出一條迴歸方程式(綠色線)



最小平方法的核心觀念(p.659)

Least Squares Criterion

$$min \sum (y_i - \hat{y}_i)^2$$

where:

y_i = the observed value of the dependent variable for the i^{th} observation

$\hat{y}_i$ = the estimated value of the dependent variable for the i^{th} observation

Step 1: 先算出斜率(b_1)

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

where:

x_i = the value of the independent variable for the i^{th} observation

y_i = the value of the dependent variable for the i^{th} observation

$\bar{x}$ = the mean value for the independent variable

$\bar{y}$ = the mean value for the dependent variable

Step 2: 算出截距(b_0)

$$b_0 = \bar{y} - b_1\bar{x}$$

TABLE 14.2 CALCULATIONS FOR THE LEAST SQUARES ESTIMATED REGRESSION EQUATION FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

The calculation of the y intercept (b_0) follows.

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

Thus, the estimated regression equation is

$$\hat{y} = 60 + 5x$$

If we believe the least squares estimated regression equation adequately describes the relationship between *x* and *y*, it would seem reasonable to use the estimated regression equation to predict the value of *y* for a given value of *x*. For example, if we wanted to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\hat{y} = 60 + 5(16) = 140$$

TABLE 14.2 CALCULATIONS FOR THE LEAST SQUARES ESTIMATED REGRESSION EQUATION FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

The calculation of the y intercept (b_0) follows.

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

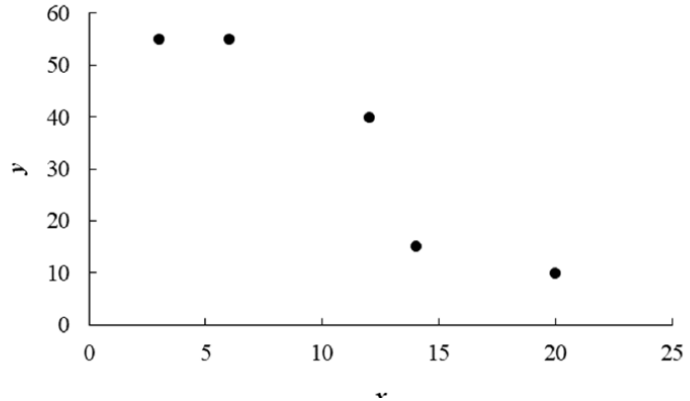
一起練習 Exercise 2 (p.662)

2. Given are five observations for two variables, x and y.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. Develop a scatter diagram for these data.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Try to approximate the relationship between x and y by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1 using equations (14.6) and (14.7).
e. Use the estimated regression equation to predict the value of y when $x = 10$.

a.



b. There appears to be a negative linear relationship between x and y.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between x *and* y; in part d we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

d.

	Xi	Yi	Xi-Xmean	Yi-Ymean	(Xi-Xmean)(Yi-Ymean)	(Xi-Xmean)(Xi-Xmean)
	3	55	-8	20	-160	64
	12	40	1	5	5	1
	6	55	-5	20	-100	25
	20	10	9	-25	-225	81
	14	15	3	-20	-60	9
Mean	11	35		Total	-540	180

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{55}{5} = 11 \qquad \bar{y} = \frac{\Sigma y_i}{n} = \frac{175}{5} = 35$$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = -540 \qquad \Sigma(x_i - \bar{x})^2 = 180$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{-540}{180} = -3$$

$$b_0 = \bar{y} - b_1\bar{x} = 35 - (-3)(11) = 68$$

$$\hat{y} = 68 - 3x$$

e. $\hat{y} = 68 - 3(10) = 38$

Chapter 14.3 Coefficient of Determination

- According to 14.2, we developed the estimated regression equation to appropriate the linear relationship between the size of the student population *x* and quarterly sales *y*.
- Now, we have a further question: How well does the estimated regression equation fit the data?
- The difference between the observed value of the dependent, y_i, and the predicted value of the dependent variable, $\hat{y}_i$, is called the *i*th residual.(殘差)

Sum of squares due to error (殘差平方和)➔ $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- 回到課本原本 Armand's Pizza Parlors 範例，我們來計算估計迴歸方程式 $\hat{y} = 60 + 5x$ 產生的 SSE

p. 668

TABLE 14.3 CALCULATION OF SSE FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

p. 669 Total Sum of Squares

TOTAL SUM OF SQUARES

$$SST = \Sigma(y_i - \bar{y})^2 \qquad (14.9)$$

TABLE 14.4 COMPUTATION OF THE TOTAL SUM OF SQUARES FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
				SST = 15,730

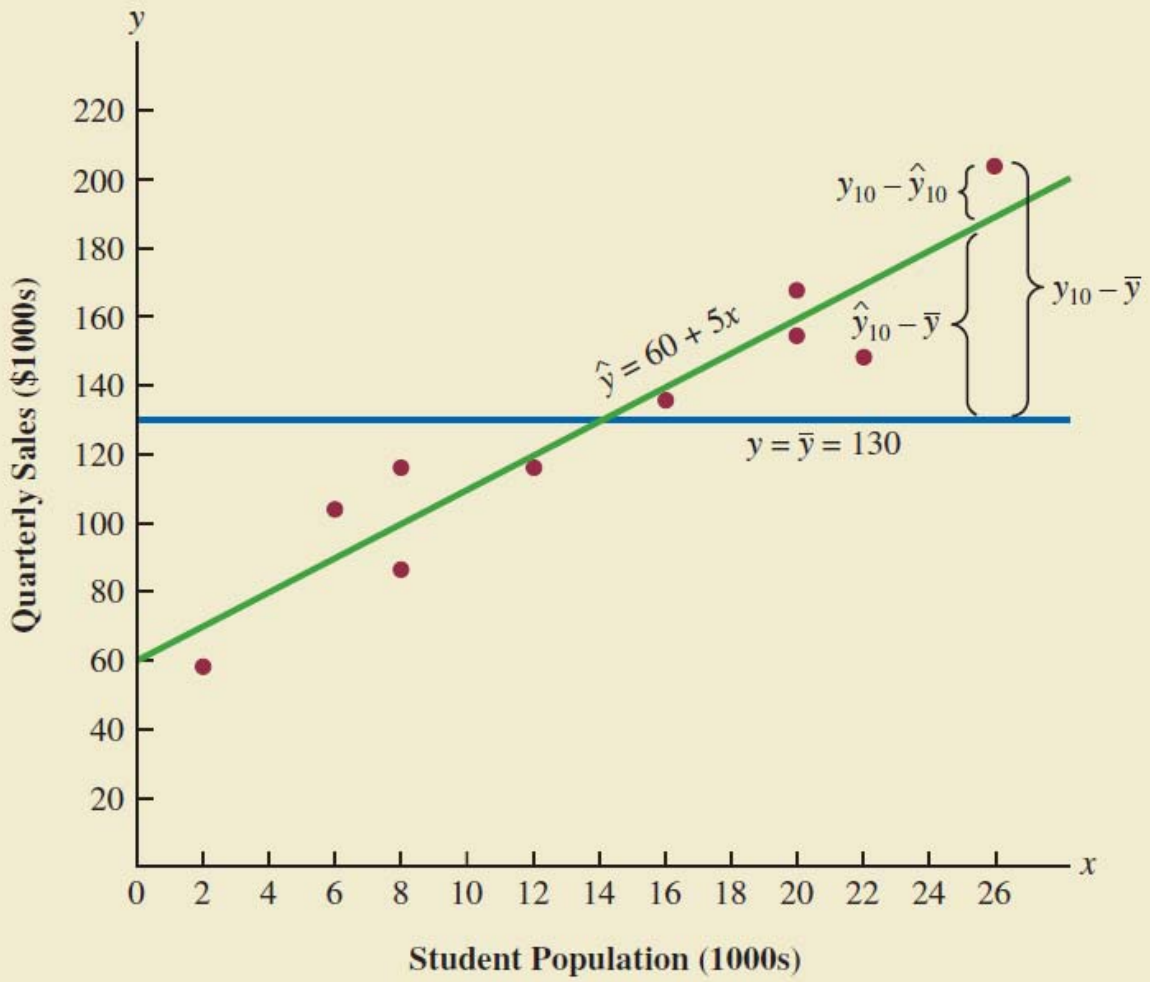
p. 669 Sum of Squares due to Regression

SUM OF SQUARES DUE TO REGRESSION

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 \qquad (14.10)$$

Predicted sales	Y mean	Deviation	Squared Deviation
70	130	-60	3600
90	130	-40	1600
100	130	-30	900
100	130	-30	900
120	130	-10	100
140	130	10	100
160	130	30	900
160	130	30	900
170	130	40	1600
190	130	60	3600
		SSR	14200

FIGURE 14.5 DEVIATIONS ABOUT THE ESTIMATED REGRESSION LINE AND THE LINE $y = \bar{y}$ FOR ARMAND'S PIZZA PARLORS



Relationship among SST, SSR, and SSE (p.670)

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

- Because SST=SSR+SSE, we see that for a perfect fit SSR must equal SST, and the ratio (SSR/SST) must equal one.

The coefficient of determination is:

$$r^2 = \frac{SSR}{SST}$$

where:

SSR = sum of squares due to regression
SST = total sum of squares

- For the Armand's Pizza Parlors example, the value of the coefficient of determination is:

  r^2 = SSR/SST = 14,200/15,730 = .9027

- When we express the coefficient of determination as a percentage, r^2 can be interpreted as the percentage of the SST that can be explained by using the SSR.

- 90.27% of the variability in sales can be explained by the linear relationship between the size of the student population and sales.

課堂練習：Exercise 16 (延續 Exercise 2)

16. The data from exercise 2 follow.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

The estimated regression equation for these data is $\hat{y} = 68 - 3x$.

a. Compute SSE, SST, and SSR.

b. Compute the coefficient of determination r^2. Comment on the goodness of fit.

c. Compute the sample correlation coefficient.

	Xi	Yi	Predicted Yi	Deviation (Yi - Predicted Yi)	Sqared (Yi-Predicted Yi)	Yi-Ymean	Squared (Yi-Y mean)	Predicted Yi-Ymean	Squared (Predicted Yi-Ymean)
	3	55	59	-4	16	20	400	24	576
	12	40	32	8	64	5	25	-3	9
	6	55	50	5	25	20	400	15	225
	20	10	8	2	4	-25	625	-27	729
	14	15	26	-11	121	-20	400	-9	81
Mean	11	35			230		1850		1620
					SSE		SST		SSR

a. The estimated regression equation and the mean for the dependent variable are:

$\hat{y}_i = 68 - 3x \qquad \bar{y} = 35$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 = 230 \qquad \text{SST} = \Sigma(y_i - \bar{y})^2 = 1850$

Thus, SSR = SST – SSE = 1850 – 230 = 1620

b. $r^2 = \text{SSR/SST} = 1620/1850 = .876$

The least squares line provided an excellent fit; 87.6% of the variability in y has been explained by the estimated regression equation.

c. $r_{xy} = -\sqrt{.876} = -.936$

Note: the sign for r is negative because the slope of the estimated regression equation is negative.

($b_1 = -3$)