

統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 14(1): Simple Linear Regression

上課時間地點: 二 D56, 資訊 140306

授課教師: 吳漢銘 (國立政治大學統計學系副教授)

教學網站: <http://www.hmwu.idv.tw>

系級: _____ 學號: _____ 姓名: _____

Overview

1. Managerial decisions often are based on the relationship between two or more variables.
2. **Examples**
 - (a) After considering the relationship between advertising expenditures and sales, a marketing manager might attempt to _____ sales for a given level of advertising expenditures.
 - (b) A public utility might use the relationship between the daily high temperature and the demand for electricity to _____ electricity usage on the basis of next month's anticipated daily high temperatures.
3. Regression analysis can be used to develop _____ showing how the variables are related.
4. In regression terminology, the variable being predicted is called the _____ variable (denoted by _____). The variable or variables being used to predict the value of the dependent variable are called the _____ variables (denoted by _____).

5. **Simple linear regression:** the simplest type of regression analysis involving _____ independent variable and _____ dependent variable in which the relationship between the variables is approximated by a _____.
6. Regression analysis involving two or more _____ variables is called _____ regression analysis.
7. Multiple regression and cases involving _____ relationships are covered in Chapters 15 and 16.

14.1 Simple Linear Regression Model

1. **Example** Armand's Pizza Parlors
 - (a) Armand's Pizza Parlors is a chain of Italian-food restaurants located in a five-state area. Armand's most successful locations are near college campuses.
 - (b) The managers believe that _____ for these restaurants (denoted by _____) are related positively to the _____ population (denoted by _____);
 - (c) Restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population.
2. Using regression analysis, we can develop an equation showing how the dependent variable y is related to the independent variable x .

Regression Model and Regression Equation

1. (**population**) In the Armand's Pizza Parlors example, the population consists of all the Armand's restaurants. For every restaurant in the population, there is a value of _____ (student population) and a corresponding value of _____ (quarterly sales).

2. (**regression model**) The _____ that describes how y is related to x and an _____ is called the regression model.

3. Simple Linear Regression Model

$$\text{_____} \quad (14.1)$$

β_0 and β_1 are referred to as the _____ of the model, and ϵ (the Greek letter epsilon) is a _____ referred to as the _____.

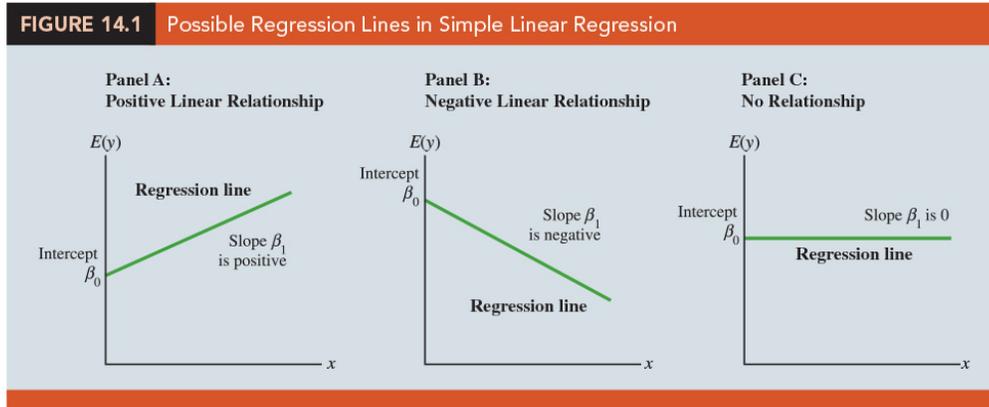
4. The error term accounts for the _____ that cannot be explained by the _____ between x and y .
5. The population of all Armand's restaurants can also be viewed as a collection of _____, one for each distinct value of _____.
- (a) For example, one subpopulation consists of all Armand's restaurants located near college campuses with _____; another subpopulation consists of all Armand's restaurants located near college campuses with _____; and so on.
- (b) Each subpopulation has a corresponding _____. Thus, a distribution of y values is associated with restaurants located near campuses with 8000 students; a distribution of y values is associated with restaurants located near campuses with 9000 students; and so on.
6. (**regression equation**) Each distribution of y values has its own _____ or _____. The equation that describes how the expected value of y , denoted $E(y)$, is related to x is called the _____.

7. Simple Linear Regression Equation

$$\text{_____} \quad (14.2)$$

The graph of the simple linear regression equation is a straight line; β_0 is the _____ of the regression line, β_1 is the _____, and $E(y)$ is the mean or expected value of y for a given value of x .

8. (Figure 14.1) Possible regression lines



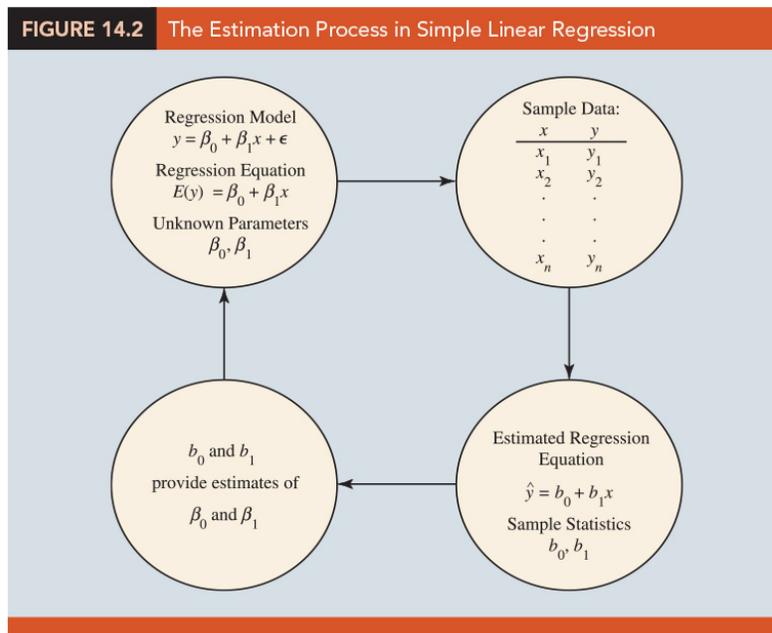
Estimated Regression Equation

1. If the values of the population parameters _____ and _____ were known, we could use equation (14.2) to compute the mean value of y for a given value of x .
2. In practice, the parameter values are not known and must be estimated using _____. Sample statistics (denoted _____ and _____) are computed as estimates of the population parameters β_0 and β_1 . Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1 in the regression equation, we obtain _____.

3. Estimated Simple Linear Regression Equation

$$\text{_____} \quad (14.3)$$

4. (**the estimated regression line**) The graph of the estimated simple linear regression equation is called the estimated regression line; b_0 is the y -intercept and b_1 is the slope.
5. In general, _____ is the point estimator of $E(y)$, the mean value of y for a given value of x .
6. (Figure 14.2) A summary of the estimation process for simple linear regression.

7. **Example** Armand's Pizza Parlors

- To estimate the mean or expected value of quarterly sales for all restaurants located near campuses with 10,000 students, Armand's would substitute the value of 10,000 for x in equation (14.3).
 - In some cases, however, Armand's may be more interested in predicting sales for one particular restaurant.
 - For example, suppose Armand's would like to predict quarterly sales for the restaurant they are considering building near Talbot College, a school with 10,000 students. As it turns out, the best predictor of y for a given value of x is also provided by _____.
 - Thus, to predict quarterly sales for the restaurant located near Talbot College, Armand's would also substitute the value of 10,000 for x in equation (14.3).
8. The value of \hat{y} provides both a _____ of $E(y)$ for a given value of x and a _____ of an individual value of y for a given value of x .

Notes + Comments

- Regression analysis cannot be interpreted as a procedure for establishing a _____ relationship between variables. It can only indicate how or to what extent variables

are _____ with each other.

- Any conclusions about cause and effect must be based upon the _____ of those individuals most knowledgeable about the application.
- The regression equation in simple linear regression is $E(y) = \beta_0 + \beta_1 x$. More advanced texts in regression analysis often write the regression equation as

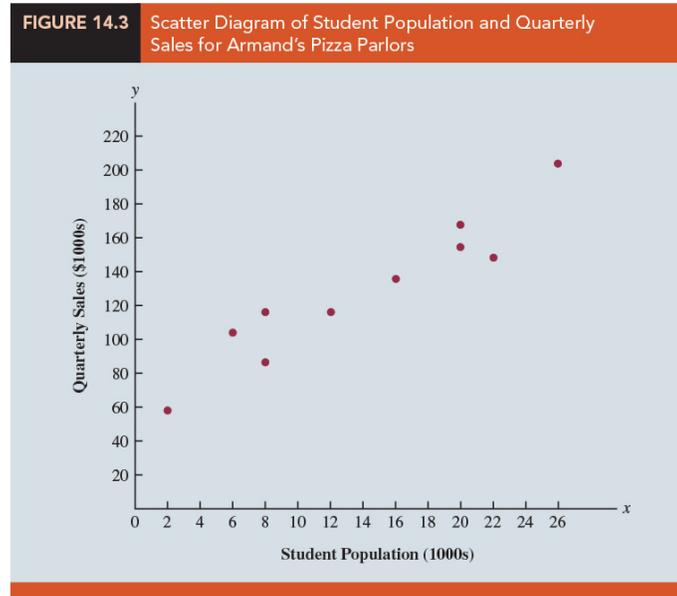
to emphasize that the regression equation provides the mean value of y for a given value of x .

14.2 Least Squares Method

- The _____ is a procedure for using sample data to find the estimated regression equation.
- (Table 14.1) **Example** Armand's Pizza Parlor
Suppose data were collected from a sample of 10 Armand's Pizza Parlor restaurants located near college campuses. For the i th observation or restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars).

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

3. (Figure 14.3) Scatter diagrams for regression analysis are constructed with the independent variable x (student population) on the horizontal axis and the dependent variable y (quarterly sales) on the vertical axis.



- (a) The _____ enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.
- (b) Quarterly sales appear to be higher at campuses with larger student populations.
- (c) In addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a _____.
- (d) A _____ relationship is indicated between x and y .
- (e) We therefore choose the _____ model to represent the relationship between quarterly sales and student population.
- (f) Next task is to use the sample data in Table 14.1 to determine the values of b_0 and b_1 in the estimated simple linear regression equation.
4. For the i th restaurant, the estimated regression equation provides

$$\text{_____} \quad (14.4)$$

where

- \hat{y}_i : predicted value of quarterly sales (\$1000s) for the i th restaurant
 - b_0 : the _____ of the estimated regression line
 - b_1 : the _____ of the estimated regression line
 - x_i : size of the student population (1000s) for the i th restaurant
5. In simple linear regression, each observation _____ consists of two values: one for the independent variable and one for the dependent variable.
 6. Every restaurant in the sample will have an observed value of sales y_i and a predicted value of sales \hat{y}_i .
 7. For the estimated regression line to provide a good fit to the data, we want the differences between the observed sales values and the predicted sales values _____.
 8. **(the least squares method)** The least squares method uses the sample data to provide the values of b_0 and b_1 that _____ the _____ of the _____ between the observed values of the dependent variable y_i and the predicted values of the dependent variable \hat{y}_i .

9. Least Squares Criterion

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (14.5)$$

where

- y_i : _____ of the dependent variable for the i th observation
- \hat{y}_i : _____ of the dependent variable for the i th observation

10. **Slope and Y-Intercept for the Estimated Regression Equation** Differential calculus can be used to show (see Appendix 14.1) that the values of b_0 and b_1 that minimize expression (14.5) can be found by:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \quad (14.6)$$

$$b_0 = \bar{y} - b_1 \bar{x} \quad (14.7)$$

where

- x_i : value of the independent variable for the i th observation
- y_i : value of the dependent variable for the i th observation
- \bar{x} : mean value for the independent variable
- \bar{y} : mean value for the dependent variable
- n : total number of observations

補充說明：

 Question (p660)

Using data in Table 14.2 to calculate the slope and intercept of the estimated regression equation for Armand's Pizza Parlors example.

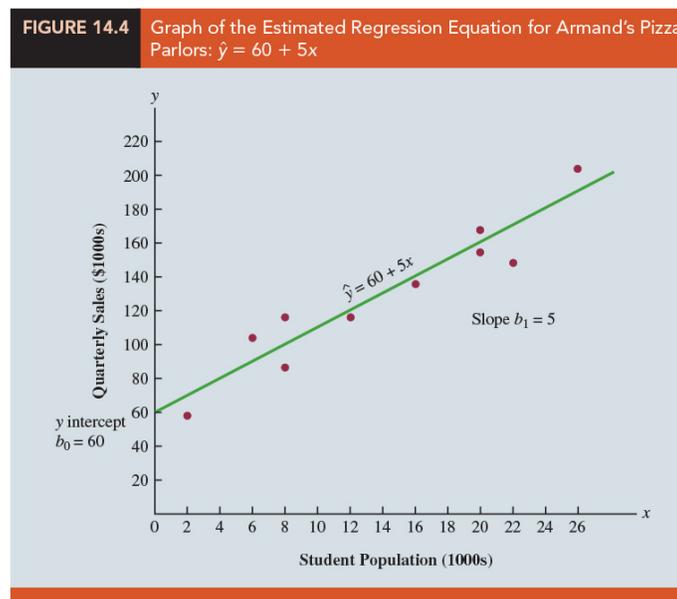
sol:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{140}{10} = 14, & \bar{y} &= \frac{\sum y_i}{n} = \frac{1300}{10} = 130 \\ b_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{2840}{568} = 5 \\ b_0 &= \bar{y} - b_1 \bar{x} = 130 - 5(14) = 60\end{aligned}$$

Thus, the estimated regression equation is _____.

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			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

11. (Figure 14.4) The graph of this equation on the scatter diagram.



- (a) The slope of the estimated regression equation _____, implying that as student population increases, sales increase.
- (b) We can conclude (based on sales measured in \$1000s and student population in 1000s) that an _____ in the student population of 1000 is associated with an _____ of \$5000 in _____ sales; that is, quarterly sales are expected to increase by \$5 per student.

12. If we wanted to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\hat{y} = \underline{\hspace{2cm}}$$

Hence, we would predict quarterly sales of \$140,000 for this restaurant.

13. This least squares criterion is used to choose the equation that provides the _____.
14. If some other criterion were used, such as minimizing the sum of the _____ between y_i and \hat{y}_i , a different equation would be obtained. In practice, the least squares method is the _____.

☺ EXERCISES 14.2: 1, 5, 6

14.3 Coefficient of Determination

1. How well does the estimated regression equation fit the data? The _____ provides a measure of the _____ for the estimated regression equation.
2. (**residual**) For the i th observation, the difference between the observed value of the dependent variable, _____, and the predicted value of the dependent variable, _____, is called the i th residual. The i th residual represents the error in using \hat{y}_i to estimate y_i . Thus, for the i th observation, the residual is _____.
3. (**Sum of Squares Due to Error**) The sum of squares of these residuals or errors is the quantity that is minimized by the least squares method. This quantity, also known as the sum of squares due to error, is denoted by _____:

$$SSE = \underline{\hspace{2cm}} \quad (14.8)$$

4. (Table 14.3) The value of SSE is a measure of the error in using the estimated regression equation to predict the values of the dependent variable in the sample. $SSE = 1530$ measures the error in using the estimated regression equation $\hat{y} = 60 + 5x$ to predict sales.

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
					$SSE = 1530$

5. Now suppose we are asked to develop an estimate of quarterly sales _____ knowledge of the size of the student population. Without knowledge of any related variables, we would use the _____ as an estimate of quarterly sales at any given restaurant.
6. (Table 14.4) (**Total Sum of Squares**) We show the sum of squared deviations obtained by using the _____ to predict the value of quarterly sales for each restaurant in the sample. For the i th restaurant in the sample, the difference _____ provides a measure of the error involved in using \bar{y} to predict sales. The corresponding sum of squares, called the total sum of squares, is denoted _____.

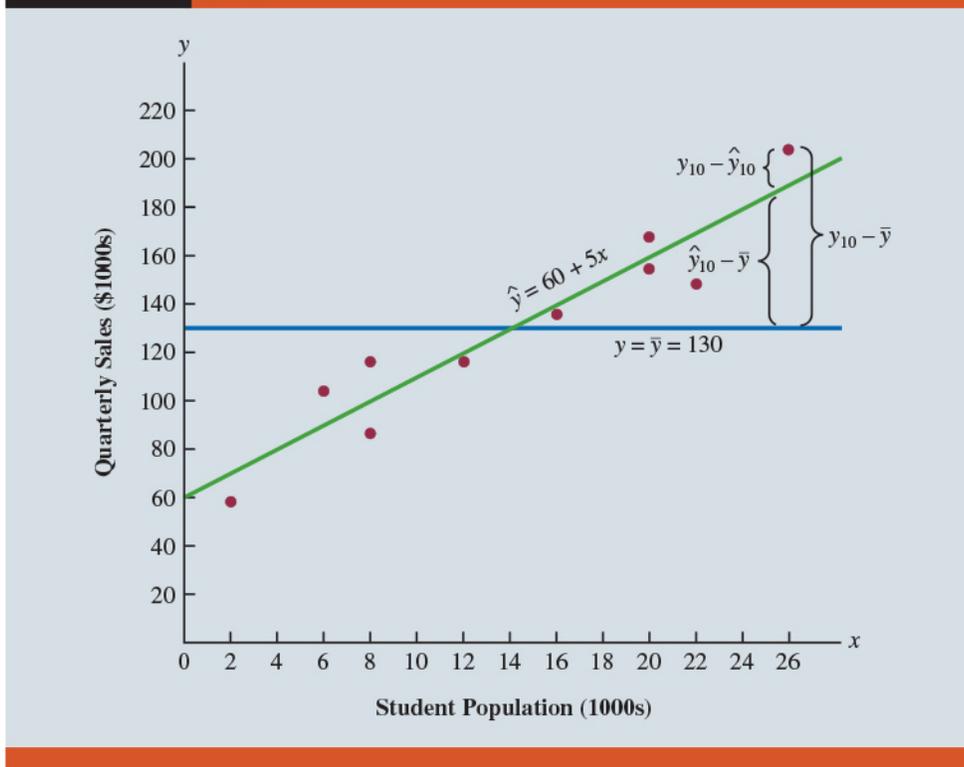
$$SST = \underline{\hspace{2cm}} \quad (14.9)$$

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

7. (Figure 14.5)

FIGURE 14.5 Deviations About the Estimated Regression Line and the Line $y = \bar{y}$ for Armand's Pizza Parlors



8. We can think of _____ as a measure of how well the observations cluster about

the _____ and SSE as a measure of how well the observations cluster about the _____.

9. (**Sum of Squares Due to Regression**) the sum of squares due to regression, is denoted _____, measures how much the \hat{y} values on the estimated regression line deviate from \bar{y} :

$$SSR = \text{_____} \quad (14.10)$$

10. (**Relationship Among SST , SSR , and SSE**) From the preceding discussion, we should expect that SST , SSR , and SSE are related.

$$\text{_____} \quad (14.11)$$

where

- SST : total sum of squares
- SSR : sum of squares due to regression
- SSE : sum of squares due to error

11. SSR can be thought of as the _____ portion of SST , and SSE can be thought of as the _____ portion of SST .

12. **Example** Armand's Pizza Parlors example
we already know that $SSE = 1530$ and $SST = 15,730$; therefore, solving for SSR in equation (14.11), we find that the sum of squares due to regression is

$$SSR = \text{_____} = 15,730 - 1530 = 14,200$$

13. How the three sums of squares, SST , SSR , and SSE , can be used to provide a measure of the goodness of fit for the estimated regression equation?

- (a) The estimated regression equation would provide a perfect fit if every value of the dependent variable y_i happened to lie on the estimated regression line.
- (b) In this case, _____ would be zero for each observation, resulting in _____.
- (c) Because $SST = SSR + SSE$, we see that for a perfect fit SSR must equal SST , and the ratio (_____) must equal one.

- (d) Poorer fits will result in larger values for SSE . Hence the poorest fit occurs when _____ and _____.
14. (**Coefficient of Determination**) The ratio SSR/SST , which will take values between zero and one, is used to evaluate the goodness of fit for the estimated regression equation. This ratio is called the coefficient of determination and is denoted by (_____) (Other textbook: _____).

$$r^2 = \frac{SSR}{SST} \quad (14.12)$$

15. When we express the coefficient of determination as a percentage, r^2 can be interpreted as the _____ of the total sum of squares that can be explained by using _____.
16. (**Example**) Armand's Pizza Parlors example

- (a) The value of the coefficient of determination is

$$r^2 = \frac{SSR}{SST} = \frac{14,200}{15,730} = 0.9027$$

- (b) For Armand's Pizza Parlors, we can conclude that 90.27% of the total sum of squares can be explained by using the estimated regression equation $\hat{y} = 60 + 5x$ to predict quantity.
- (c) In other words, _____ can be explained by the linear relationship between the size of the student population and sales. We should be pleased to find such a good fit for the estimated regression equation.

Correlation Coefficient

- In Chapter 3 we introduced the correlation coefficient as a descriptive measure of the strength of linear association between two variables, x and y . Values of the correlation coefficient are always between _____.
- A value of $+1$ indicates that the two variables x and y are _____ in a _____ linear sense. A value of -1 indicates that x and y are perfectly related in a _____ linear sense, with all data points on a straight line that has a negative slope. Values of the correlation coefficient close to zero indicate that x and y are _____.

3. (**Sample Correlation Coefficient**) If a regression analysis has already been performed and the coefficient of determination r^2 computed, the sample correlation coefficient can be computed:

$$r_{xy} = \frac{b_1}{s_y} = \frac{s_x}{s_y} r \quad (14.13)$$

where b_1 is the slope of the estimated regression equation $\hat{y} = b_0 + b_1x$

補充說明 : Show that the coefficient of determination of a simple linear regression is the square of the sample correlation coefficient of $(x_1, y_1), \dots, (x_n, y_n)$.

4. **Example** Armand's Pizza Parlor example
the value of the coefficient of determination corresponding to the estimated regression equation $\hat{y} = 60 + 5x$ is 0.9027. Because the slope of the estimated regression equation is positive, equation (14.13) shows that the sample correlation coefficient is $+\sqrt{0.9027} = +0.9501$. (a strong positive linear association exists between x and y .)
5. In the case of a _____ between two variables, both the coefficient of determination and the sample correlation coefficient provide measures of the strength of the relationship. The coefficient of determination provides a measure between zero and one, whereas the sample correlation coefficient provides a measure between -1 and $+1$.

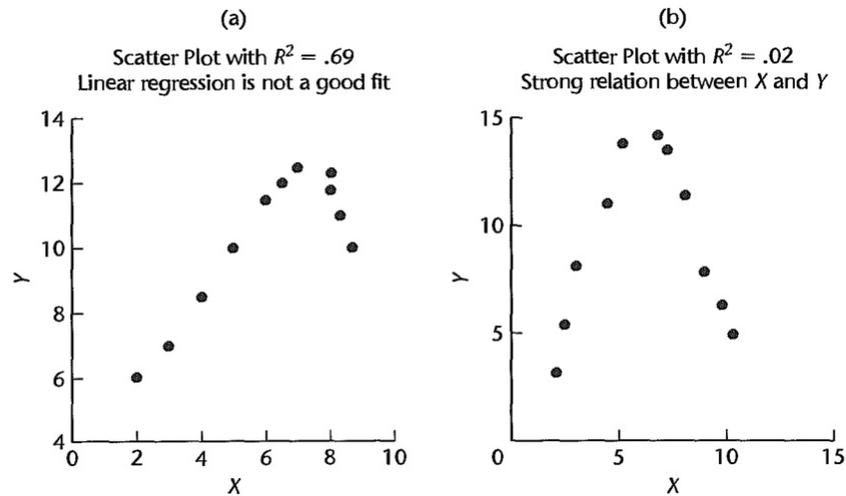
6. Although the sample correlation coefficient is restricted to a linear relationship between two variables, the coefficient of determination can be used for _____ relationships and for relationships that have _____ .
Thus, the coefficient of determination provides a wider range of applicability.

(補充) limitations of R^2 : three common misunderstandings

Source : Michael H. Kutner et al. (2019), Applied Linear Statistical Models: Applied Linear Regression Models, Mcgraw-Hill Inc., (5th edition)

1. **Misunderstanding 1:** A high R^2 indicates that _____ can be made. (not necessarily correct)
 - (a) (Toluca Company Example) the coefficient of determination was high ($R^2 = 0.82$). Yet the 90 percent prediction interval for the next lot, consisting of 100 units, was wide (332 to 507 hours) and not precise enough to permit management to schedule workers effectively.
 - (b) Misunderstanding 1 arises because R^2 measures only a _____ from SST and provides no information about absolute precision for estimating a mean response or predicting a new observation.
2. **Misunderstanding 2:** A high R^2 indicates that the estimated regression line is a _____. (not necessarily correct)
 - (a) (Figure 2.9a) a scatter plot where R^2 is high ($R^2 = 0.69$). Yet a linear regression function would not be a good fit since the regression relation is curvilinear.
3. **Misunderstanding 3:** A R^2 near zero indicates that X and Y are not related. (not necessarily correct).
 - (a) (Figure 2.9b) a scatter plot where R^2 between X and Y is $R^2 = 0.02$. Yet X and Y are strongly related; however, the relationship between the two variables is curvilinear.
 - (b) Misunderstandings 2 and 3 arise because R^2 measures the degree of _____ between X and Y , whereas the actual regression relation may be curvilinear.

FIGURE 2.9
Illustrations
of Two Misun-
derstandings
about
Coefficient of
Determination.



☺ EXERCISES 14.3: 15, 19, 20

14.4 Model Assumptions

1. In conducting a regression analysis, we begin by making an assumption about the appropriate model for the relationship between the dependent and independent variable(s).
2. For the case of simple linear regression, the assumed regression model is

3. Then the least squares method is used to develop values for b_0 and b_1 , the estimates of the model parameters β_0 and β_1 , respectively. The resulting estimated regression equation is

Even with a large value of r^2 , the estimated regression equation should not be used until further analysis of the appropriateness of the assumed model has been conducted.

4. An important step in determining whether the assumed model is appropriate involves _____ of the relationship. The tests of significance in regression analysis are based on the following assumptions about the error term ϵ .

5. **Assumptions About The Error Term ϵ in the Regression Model**

$$y = \beta_0 + \beta_1 x + \epsilon$$

- (a) The error term ϵ is a random variable with a mean or expected value of zero; that is, _____.

Implication: β_0 and β_1 are constants, thus, for a given value of x , the expected value of y is

$$\text{_____} \quad (14.14)$$

As we indicated previously, equation (14.14) is referred to as the regression equation.

- (b) The variance of ϵ , denoted by _____, is the same for all values of x .

Implication: The variance of y about the regression line equals σ^2 and is the same for _____.

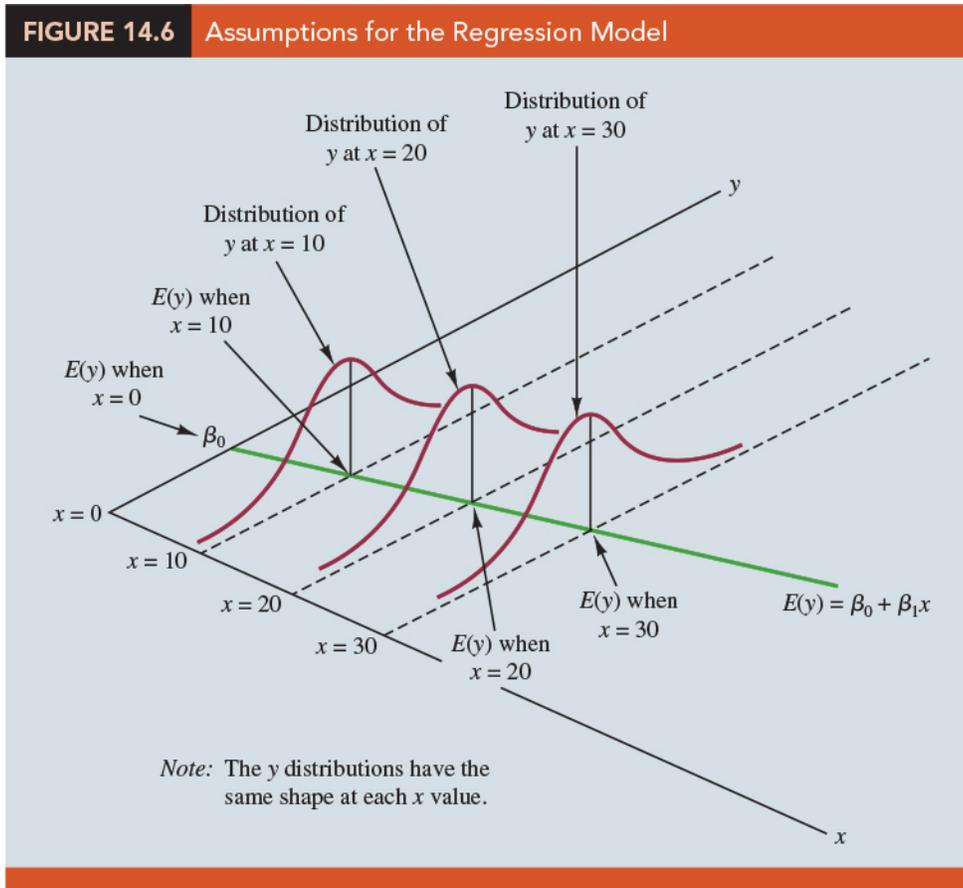
- (c) The values of ϵ are _____.

Implication: The value of ϵ for a particular value of x is not related to the value of ϵ for any other value of x ; thus, the value of y for a particular value of x is not related to the value of y for any other value of x .

- (d) The error term ϵ is a _____ r.v. for all values of x .

Implication: Because y is a linear function of ϵ , y is also a normally distributed random variable for all values of x .

6. Figure 14.6 illustrates the model assumptions and their implications; note that in this graphical interpretation, the value of _____ changes according to the specific value of x considered. However, regardless of the x value, the probability distribution of ϵ and hence the probability distributions of y are _____ distributed, each with the _____.



7. The specific value of the error ϵ at any particular point depends on whether the actual value of _____ is greater than or less than _____.
8. We assume that a straight line represented by _____ is the basis for the relationship between the variables.

14.5 Testing for Significance

1. In a simple linear regression equation, the mean or expected value of y is a linear function of x : $E(y) = \beta_0 + \beta_1x$. If the value of _____, the mean value of y does not depend on the value of x and hence we would conclude that x and y are _____.

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of _____.
- Two tests are commonly used. Both require an estimate of _____, the variance of ϵ in the regression model.

Estimate of σ^2

- From the regression model and its assumptions we can conclude that σ^2 , the variance of ϵ , also represents the variance of the y values about the regression line.
- Thus, _____, the sum of squared residuals, is a measure of the variability of the actual observations about the estimated regression line.

$$SSE = \underline{\hspace{2cm}} = \underline{\hspace{2cm}}$$

- Statisticians have shown that SSE has _____ degrees of freedom because two parameters (β_0 and β_1) must be estimated to compute SSE .
- The _____ provides the estimate of σ^2 ; it is SSE divided by its degrees of freedom.

5. Mean Square Error (Estimate of σ^2)

$$s^2 = MSE = \underline{\hspace{2cm}} \quad (14.15)$$

6. Standard Error of the Estimate

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} \quad (14.16)$$

- Example** Armand's Pizza Parlors example

$$s^2 = MSE = \frac{1530}{8} = 191.25$$

provides an unbiased estimate of σ^2 .

$$s = \sqrt{MSE} = \sqrt{191.25} = 13.829.$$

***t* Test**

1. The purpose of the *t* test is to see whether we can conclude that $\beta_1 \neq 0$. We will use the sample data to test the following hypotheses about the parameter β_1 .

2. If H_0 is rejected, we will conclude that $\beta_1 \neq 0$ and that a _____ relationship exists between the two variables.

3. If H_0 cannot be rejected, we will have _____ to conclude that a significant relationship exists.

4. The properties of the _____ of b_1 , the least squares estimator of b_1 , provide the basis for the hypothesis test.

5. Sampling Distribution of b_1

- Expected Value: _____
- Standard Deviation: _____
- Distribution Form: _____ (14.17)

6. Because we do not know the value of σ , we develop an estimate of σ_{b_1} , denoted s_{b_1} , by estimating σ with s in equation (14.17). Thus, we obtain the following estimate of σ_{b_1} .

7. Estimated Standard Deviation of b_1

$$s_{b_1} = \frac{s}{\sqrt{SSX}} \quad (14.18)$$

8. The standard deviation of b_1 is also referred to as the standard error of b_1 . Thus, s_{b_1} provides an estimate of the standard error of b_1 .

9. The *t* test for a significant relationship is based on the fact that the test statistic

follows a _____ distribution with _____ degrees of freedom. If the null hypothesis is true, then _____ and _____.

10. *t* Test for Significance in Simple Linear Regression

(a) Hypothesis:

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0$$

(b) Test Statistic: (14.19)

(c) Rejection Rule: _____

i. *p*-value approach: Reject H_0 if *p*-value $\leq \alpha$ ii. Critical value approach: Reject H_0 if _____ or if _____.where $t_{\alpha/2}$ is based on a *t* distribution with $n-2$ degrees of freedom.

 **Question** (p678)

Conduct this test of significance for Armand's Pizza Parlors at the $\alpha = 0.01$ level of significance.

sol:

1. Hypothesis: $H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$ 2. Level of significance $\alpha = 0.01.$ 3. Test statistic (under H_0): $t = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$

4. Decision rule

(a) Reject H_0 if *p*-value $< \alpha$ (b) Reject H_0 if $|t| \geq t_{\alpha/2, n-2} = 3.355$. (Table 2 of Appendix D, upper tail of the *t* distribution)

5. Decision:

(a) *p*-value (0.000) less than $2(0.005) = 0.01$. (Software), we reject H_0 .(b) $t = 8.62 > t_{\alpha/2, n-2} = 3.355$, we reject H_0 .6. Conclusion: We reject H_0 and conclude that β_1 is not equal to zero. This evidence is sufficient to conclude that a significant relationship exists between student population and quarterly sales.

Confidence Interval for β_1

1. The form of a confidence interval for β_1 is as follows:

2. The point estimator is _____ and the margin of error is _____.

3. Develop a 99% confidence interval estimate of b_1 for Armand's Pizza Parlors. From Table 2 of Appendix B we find $t_{0.005,8} = 3.355$. Thus, the 99% confidence interval estimate of b_1 is

$$b_1 \pm t_{\alpha/2, n-2} s_{b_1} = \underline{\hspace{2cm}} = 5 \pm 1.95$$

or 3.05 to 6.95.

4. At the $\alpha = 0.01$ level of significance, we can use the 99% confidence interval as an _____ for drawing the hypothesis testing conclusion for the Armand's data.
5. Because 0, the hypothesized value of b_1 , is _____ in the confidence interval (3.05 to 6.95), we can _____ and conclude that a significant statistical relationship exists between the size of the student population and quarterly sales.
6. In general, a confidence interval can be used to test any _____ about β_1 . If the hypothesized value of β_1 is _____ in the confidence interval, do not reject H_0 . Otherwise, reject H_0 .

F Test

1. An F test, based on the F probability distribution, can also be used to test for significance in regression. With only _____, the F test will provide the same conclusion as the t test.
2. But with more than one independent variable, only the F test can be used to test for an _____ relationship.
3. If the null hypothesis $H_0 : \beta_1 = 0$ is true, the mean square due to regression (_____), and is denoted _____. In general,

$$MSR = \underline{\hspace{2cm}}$$

4. The regression degrees of freedom is always equal to the _____ variables in the model. Because we consider only regression models with one independent variable in this chapter, we have _____.
5. If the null hypothesis ($H_0 : \beta_1 = 0$) is true, _____ and _____ are two independent estimates of σ^2 and the sampling distribution of _____ follows an F distribution with numerator degrees of freedom equal to one and denominator degrees of freedom equal to $n-2$. Therefore, when $\beta_1 = 0$, the value of MSR/MSE should be close to _____.
6. If the null hypothesis is false ($\beta_1 \neq 0$), MSR will _____ σ^2 and the value of MSR/MSE will be _____; thus, large values of MSR/MSE lead to the rejection of H_0 and the conclusion that the relationship between x and y is statistically significant.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}, \quad MSR = \frac{\sum(\hat{y}_i - \bar{y})^2}{1} = \frac{SSR}{1}.$$

$$E(MSE) = \underline{\hspace{2cm}}, \quad E(MSR) = \underline{\hspace{2cm}}.$$

7. If H_0 is false, MSE still provides an unbiased estimate of σ^2 and MSR overestimates σ^2 . If H_0 is true, both MSE and MSR provide unbiased estimates of σ^2 ; in this case the value of MSR/MSE should be close to 1.

8. F Test for Significance in Simple Linear Regression

(a) Hypothesis: $H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0$

(b) Test Statistic: $F = \frac{MSR}{MSE} \quad (14.21)$

(c) Rejection Rule:

i. p-value approach: Reject H_0 if $p\text{-value} \leq \alpha$

ii. Critical value approach: Reject H_0 if _____

where F_α is based on an F distribution with 1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator.

 Question (p80)

Conduct the F test for the Armand's Pizza Parlors example. ($\alpha = 0.01$)

sol:

- 10. A similar ANOVA table can be used to summarize the results of the F test for significance in regression.
- 11. (Table 14.5)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p -value
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	
Error	SSE	$n - 2$	$MSE = \frac{SSE}{n - 2}$		
Total	SST	$n - 1$			

- 12. (Table 14.6) ANOVA table with the F test computations performed for Armand's Pizza Parlors.

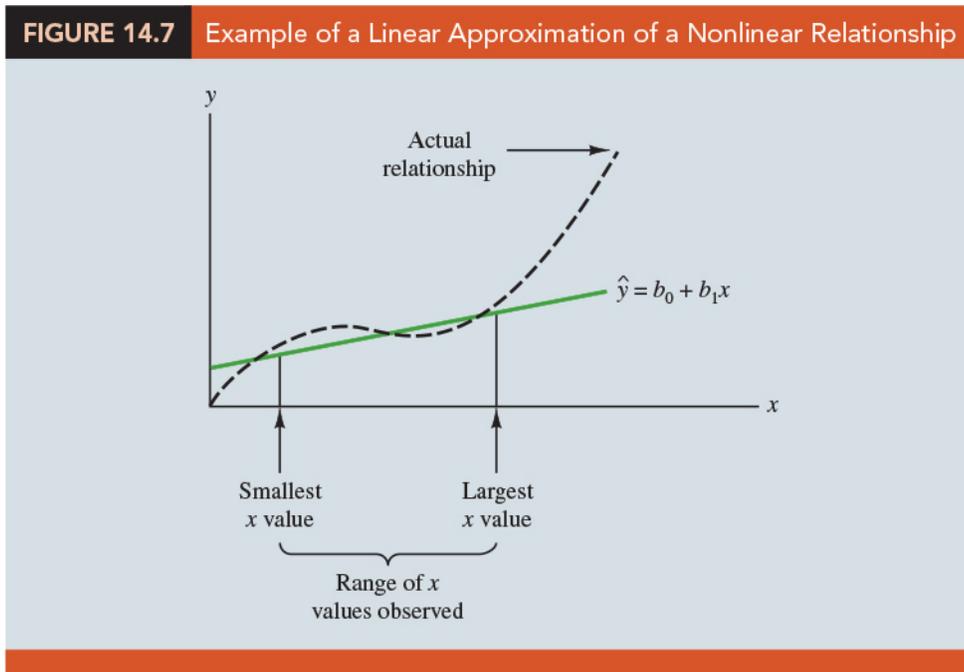
TABLE 14.6 Anova Table for the Armand's Pizza Parlors Problem

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Regression	14,200	1	$\frac{14,200}{1} = 14,200$	$\frac{14,200}{191.25} = 74.25$.000
Error	1530	8	$\frac{1530}{8} = 191.25$		
Total	15,730	9			

Some Cautions About the Interpretation of Significance Tests

1. Rejecting the null hypothesis $H_0 : \beta_1 = 0$ and concluding that the relationship between x and y is significant does not enable us to conclude that a _____ relationship is present between x and y .
2. Concluding a cause-and-effect relationship is warranted only if the analyst can provide some type of _____ that the relationship is in fact _____.
3. In the Armand's Pizza Parlors example, we can conclude that there is a significant relationship between the size of the student population x and quarterly sales y ; moreover, the estimated regression equation $\hat{y} = 60 + 5x$ provides the least squares estimate of the relationship. We cannot, however, conclude that _____ student population x _____ in quarterly sales y just because we identified a statistically significant relationship.
4. Armand's managers felt that increases in the student population were a _____ of increased quarterly sales. Thus, the result of the significance test enabled them to conclude that a cause-and-effect relationship was present.
5. We can state only that x and y are related and that a linear relationship explains a significant portion of the variability in y over the range of values for x observed in the sample.
6. (Figure 14.7) illustrates this situation. The test for significance calls for the rejection of the null hypothesis $H_0 : \beta_1 = 0$ and leads to the conclusion that x and y are

significantly related, but the figure shows that the actual relationship between x and y is not linear.



7. Although the linear approximation provided by $\hat{y} = b_0 + b_1x$ is good over the range of x values observed in the sample, it becomes poor for x values _____.
8. Given a significant relationship, we should feel confident in using the estimated regression equation for predictions corresponding to x values _____ of the x values observed in the sample.
9. For Armand's Pizza Parlors, this range corresponds to values of x _____. Unless other reasons indicate that the model is valid beyond this range, predictions outside the range of the independent variable should be made _____.

☺ **EXERCISES 14.5:** 23, 26, 27, 30