統計學 (二)

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

Chapter 14: 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**

$$\_\_\_\_\_ \quad (14.1)$$

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.

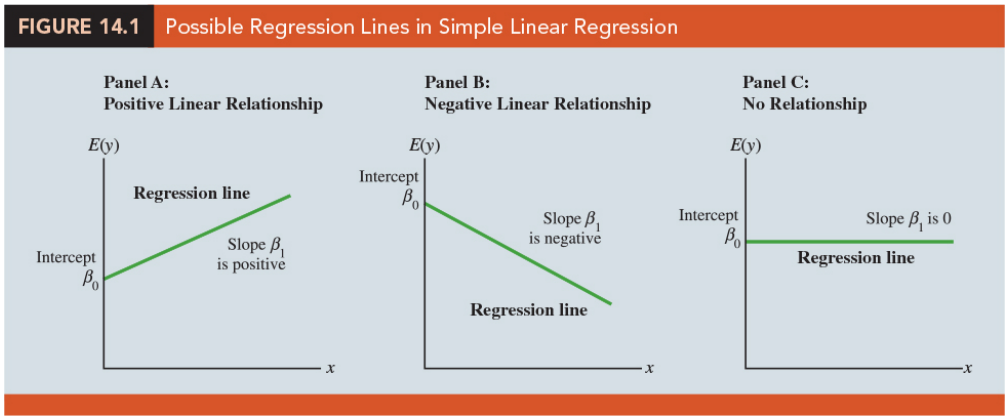
6. (**regression equation**) Each distribution of y values has its own _____ or _____. The equation that describes how the expected value of y, denoted $E(y)$, is related to x is called the _____.

7. **Simple Linear Regression Equation**

$$\_\_\_\_\_ \quad (14.2)$$

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 _____.

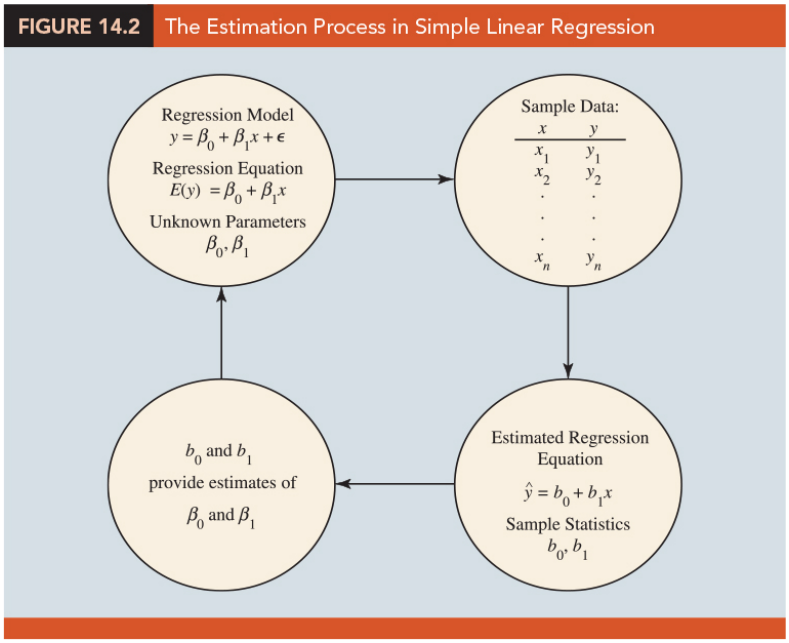
3. **Estimated Simple Linear Regression Equation**

   _____ (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

   (a) 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).

   (b) In some cases, however, Armand's may be more interested in predicting sales for one particular restaurant.

   (c) 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 _____.

   (d) 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

1. 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.

2. Any conclusions about cause and effect must be based upon the _____ of those individuals most knowledgeable about the application.

3. 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

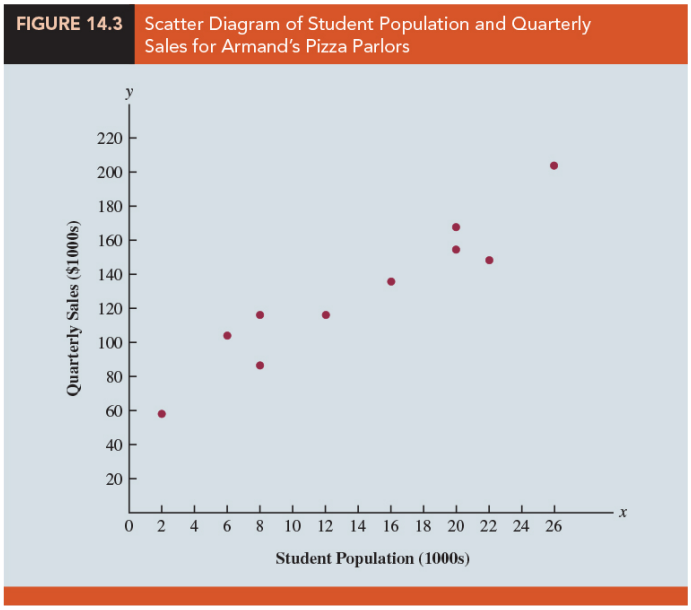
1. The _____ is a procedure for using sample data to find the estimated regression equation.

2. (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 ith 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).

TABLE 14.1 Student Population and Quarterly Sales Data for 10 Armand's Pizza Parlors

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 ith restaurant, the estimated regression equation provides

_____ (14.4)

where

- $\hat{y}_i$: predicted value of quarterly sales (\$1000s) for the ith 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 ith 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**

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

where

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

$$b_1 = \frac{\text{_____}}{\text{_____}} = \text{_____} \qquad (14.6)$$

$$b_0 = \text{_____} \qquad (14.7)$$

where

- x_i: value of the independent variable for the ith observation
- y_i: value of the dependent variable for the ith 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:

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

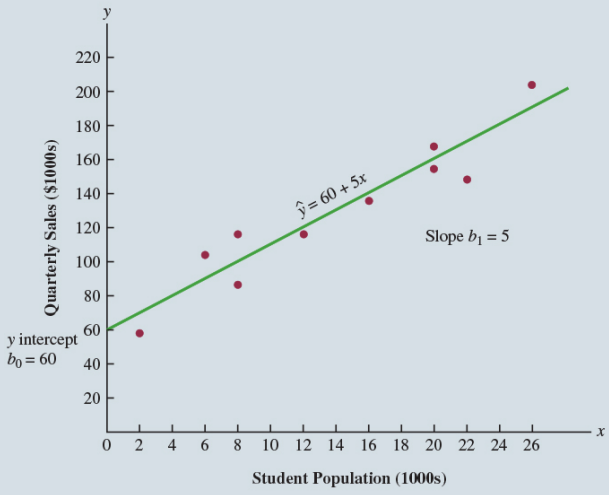
Thus, the estimated regression equation is _____.

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

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

FIGURE 14.4 Graph of the Estimated Regression Equation for Armand's Pizza Parlors: $\hat{y} = 60 + 5x$



(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

$$\hat{y} = \_\_\_\_\_$$

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 ith observation, the difference between the observed value of the dependent variable, _____, and the predicted value of the dependent variable, _____, is called the ith residual. The ith residual represents the error in using $\hat{y}_i$ to estimate y_i. Thus, for the ith 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 _____:

$$SSE = \_\_\_\_\_ \qquad (14.8)$$

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 ith 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 _____.

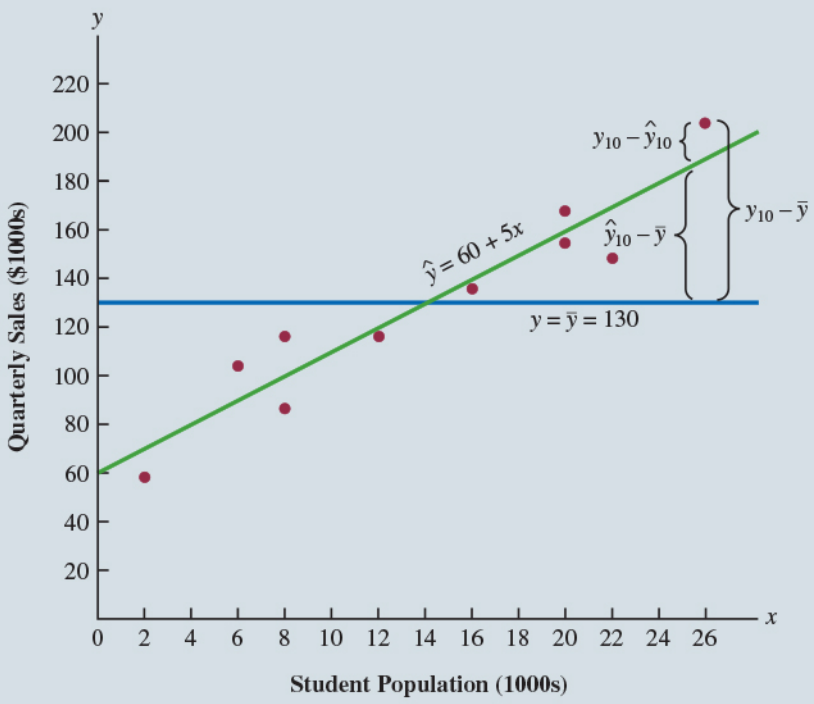
$$SST = \_\_\_\_\_ \qquad (14.9)$$

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

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 $\bar{y}$:

$$SSR = \_\_\_\_\_ \qquad (14.10)$$

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

$$\_\_\_\_\_ \qquad (14.11)$$

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

$$SSR = \_\_\_\_\_ = 15,730 - 1530 = 14,200$$

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: _____).

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

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 quar

(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

1. 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 _____.

2. 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:

$$r_{xy} = \underline{\qquad\qquad\qquad\qquad} = \underline{\qquad\qquad} \tag{14.13}$$

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

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

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 $+1$.

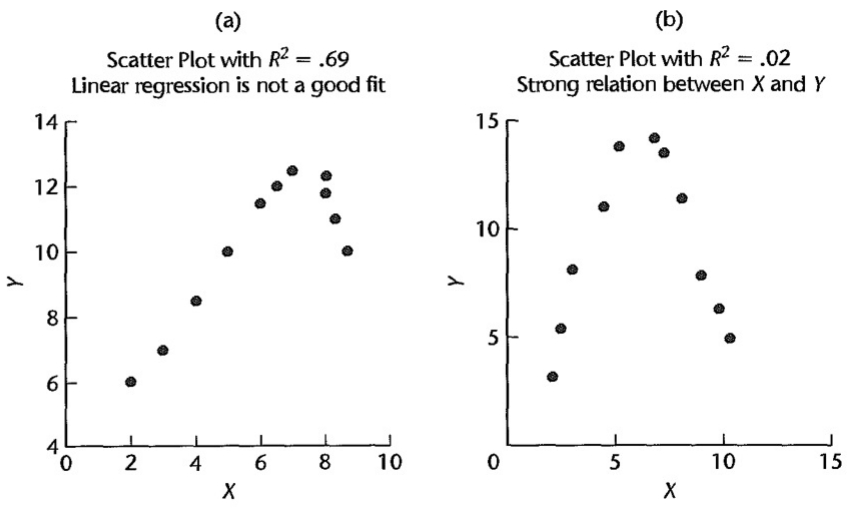
6. Although the sample correlation coefficient is restricted to a linear relationship between two variables, the coefficient of determination can be used for _____ relationships and for relationships that have _____. Thus, the coefficient of determination provides a wider range of applicability.

(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. **Misanderstanding 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 Misunderstandings 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

   _____ (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 _____.

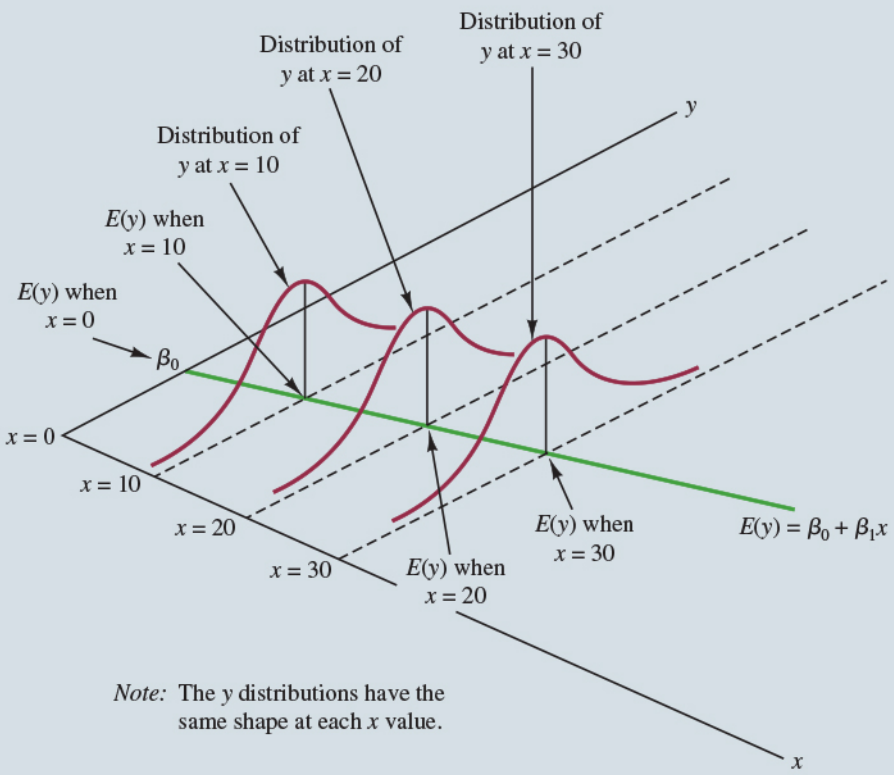
   Implication: The value of ϵ for a particular value of x is not related to the value of ϵ for any other value of x; thus, the value of y for a particular value of x is not related to the value of y for any other value of x.

   (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 _____.

FIGURE 14.6 Assumptions for the Regression Model



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: $E(y) = \beta_0 + \beta_1 x$. 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 _____.

2. To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of _____.

3. Two tests are commonly used. Both require an estimate of _____, the variance of ϵ in the regression model.

Estimate of σ^2

1. 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.

2. Thus, _____, the sum of squared residuals, is a measure of the variability of the actual observations about the estimated regression line.

$$SSE = \_\_\_\_\_ = \_\_\_\_\_$$

3. Statisticians have shown that SSE has _____ degrees of freedom because two parameters (β_0 and β_1) must be estimated to compute SSE.

4. The _____ provides the estimate of σ^2; it is SSE divided by its degrees of freedom.

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

$$s^2 = MSE = \frac{\quad}{\quad} \qquad (14.15)$$

6. **Standard Error of the Estimate**

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

7. 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**

   $$s_{b_1} = \frac{\qquad\qquad}{\qquad\qquad} \quad (14.18)$$

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:

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

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

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

   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,

   $$MSR = \_\_\_\_\_$$

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.

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

$$E(MSE) = \_\_\_\_\_, \qquad E(MSR) = \_\_\_\_\_.$$

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

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

   (c) Rejection Rule:

      i. p-value approach: Reject H_0 if p-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)

TABLE 14.5 General Form of the Anova Table for Simple Linear Regression

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

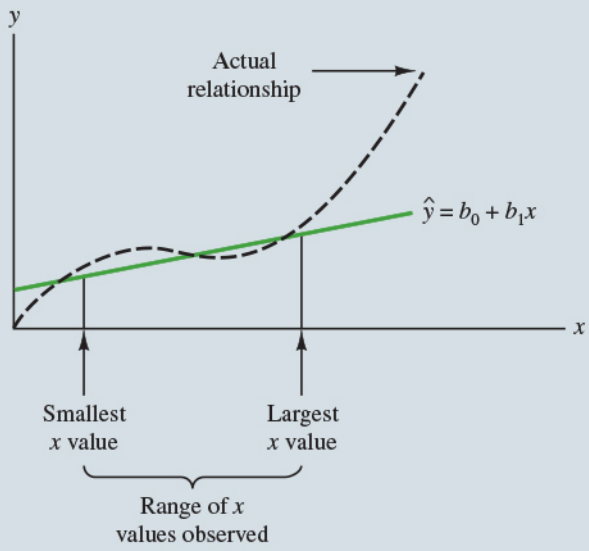
4. Armand's managers felt that increases in the student population were a _____ of increased quarterly sales. Thus, the result of the significance test enabled them to conclude that a cause-and-effect relationship was present.

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.

FIGURE 14.7 Example of a Linear Approximation of a Nonlinear Relationship



7. Although the linear approximation provided by $\hat{y} = b_0 + b_1 x$ 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

14.6 Using the Estimated Regression Equation for Estimation and Prediction

1. When using the simple linear regression model, we are making an _____ about the relationship between x and y. We then use the _____ method to obtain the estimated simple linear regression equation.

2. If a _____ relationship exists between x and y and the _____ _____ shows that the fit is good, the estimated regression equation should be useful for estimation and prediction.

3. Example Armand's Pizza Parlors example

   (a) The estimated regression equation is $\hat{y} = 60 + 5x$. $\hat{y}$ can be used as a point estimator of _____, the mean or expected value of y for a given value of x, and as a predictor of an individual value of _____.

   (b) For example, a point estimate of the mean quarterly sales for all restaurant locations near campuses with $x = 10$ (10,000 students) students is

   $$\hat{y} = \_\_\_\_\_(\$110,000).$$

   In this case we are using $\hat{y}$ as the _____ of the mean value of y when $x = 10$.

   (c) For example, to predict quarterly sales for a new restaurant Armand's is considering building near Talbot College, a campus with 10,000 students, we would compute

   $$\hat{y} = \_\_\_\_\_.$$

   Hence, we would predict quarterly sales of \$110,000 for such a new restaurant. In this case, we are using $\hat{y}$ as the _____ of y for a new observation when $x = 10$.

4. Notations:

   (a) _____ = the given value of the independent variable x

   (b) _____ = the random variable denoting the possible values of the dependent variable y when $x = x^*$

(c) _____ = the mean or expected value of the dependent variable y when $x = x^*$

(d) _____ = the point estimator of $E(y^*)$ and the predictor of an individual value of y^* when $x = x^*$

5. Example Armand's Pizza Parlors example

   (a) To illustrate the use of this notation, suppose we want to estimate the mean value of quarterly sales for all Armand's restaurants located near a campus with 10,000 students.

   (b) For this case, _____ and _____ denotes the unknown mean value of quarterly sales for all restaurants where $x^* = 10$.

   (c) Thus, the point estimate of $E(y^*)$ is provided by _____, or \$110,000.

   (d) But, using this notation, $\hat{y}^* = 110$ is also the _____ of quarterly sales for the new restaurant located near Talbot College, a school with 10,000 students.

Interval Estimation

1. Point estimators and predictors do not provide any information about the _____ associated with the estimate and/or prediction. For that we must develop _____ intervals and _____ intervals.

   (a) A confidence interval is an interval estimate of the _____ for a given value of x.

   (b) A prediction interval is used whenever we want to predict an _____ for a new observation corresponding to a given value of x.

2. Although the predictor of y for a given value of x is the same as the point estimator of the mean value of y for a given value of x, the _____ we obtain for the two cases are different.

3. The margin of error is _____ for a prediction interval.

4. Prediction intervals resemble confidence intervals. However, they differ conceptually. A confidence interval represents an _____ and is an interval that is intended to cover the value of the parameter. A prediction interval is a statement about the value to be taken by a _____, the new observation y^*_{new}.

Confidence Interval for the Mean Value of y

1. In general, we cannot expect $\hat{y}^*$ to equal $E(y^*)$ exactly. If we want to make an inference about how close $\hat{y}^*$ is to the true mean value $E(y^*)$, we will have to estimate the variance of $\hat{y}^*$.

2. The formula for estimating the variance of $\hat{y}^*$, denoted by _____, is

$$s^2_{\hat{y}^*} = \underline{\qquad\qquad\qquad\qquad} \qquad (14.22)$$

where $s^2 =$ _____.

3. The estimate of the standard deviation of $\hat{y}^*$ is given by the square root of equation (14.22).

$$s_{\hat{y}^*} = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]} \qquad (14.23)$$

4. NOTE:

$$\begin{aligned}
y_i &= \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i, \beta_0^* = \beta_0 + \beta_1\bar{x} \quad \text{(alternative model)} \\
\hat{y}^* &= b_0^* + b_1(x^* - \bar{x}), b_0^* = b_0 + b_1\bar{x} = \bar{y} \\
\hat{y}^* &= \bar{y} + b_1(x^* - \bar{x}) \\
E(\hat{y}^*) &= E(y^*) \\
\sigma^2_{\hat{y}^*} = Var(\hat{y}^*) &= Var(\bar{y} + b_1(x^* - \bar{x})) \\
&= Var(\bar{y}) + Var(b_1(x^* - \bar{x})) \\
&= \frac{\sigma^2}{n} + (x^* - \bar{x})^2\frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\
&= \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right].
\end{aligned}$$

5. Example Armand's Pizza Parlors

   $s = 13.829$. With $x^* = 10$, $\bar{x} = 14$, and $\sum(x_i - \bar{x})^2 = 568$, we can use equation (14.23) to obtain

   $$s_{\hat{y}^*} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

6. Theorem:

   $$\frac{\hat{y}^* - E(y^*)}{s_{\hat{y}^*}} \sim t_{(n-2)}$$

7. **Confidence Interval for** $E(y^*)$

   $$\underline{\qquad\qquad\qquad} \qquad (14.24)$$

   where the confidence coefficient is $1-\alpha$ and $t_{\alpha/2}$ is based on the t distribution with (n–2) degrees of freedom.

8. Example Armand's Pizza Parlors

   (a) Develop a 95% confidence interval of the mean quarterly sales for all Armand's restaurants located near campuses with 10,000 students.

   (b) We have _____. Thus, with _____ and a margin of error of _____, the 95% confidence interval estimate is 110 ± 11.415.

   (c) In dollars, the 95% confidence interval for the mean quarterly sales of all restaurants near campuses with 10,000 students is \$110,000 ± \$11,415. Therefore, the 95% confidence interval for the _____ when the student population is 10,000 is _____.

9. Note that the estimated standard deviation of $\hat{y}^*$ given by equation (14.23) is smallest when _____.

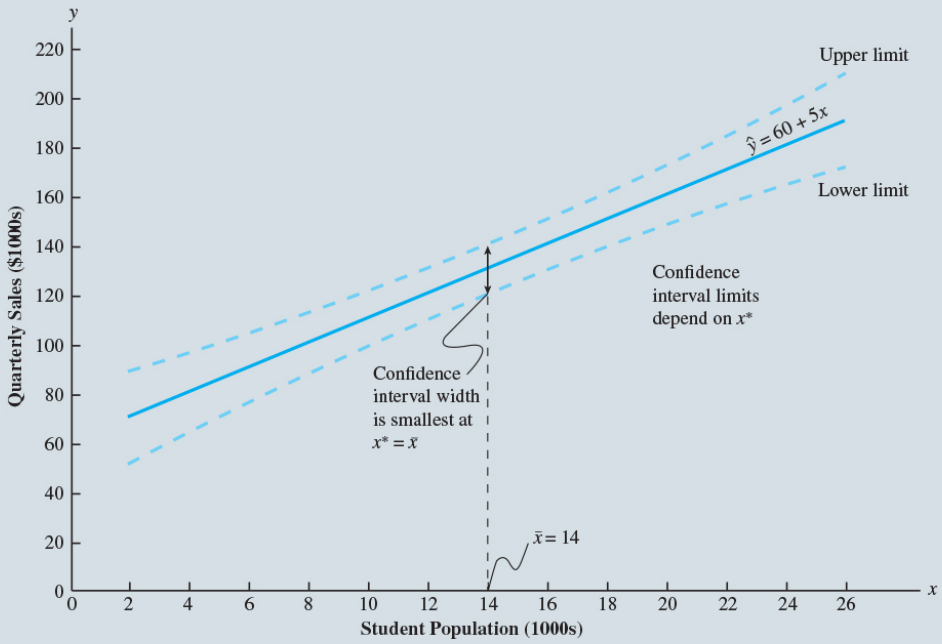
10. In this case the estimated standard deviation of $\hat{y}^*$ becomes

    $$s_{\hat{y}^*} = s\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]} = \underline{\qquad\qquad}$$

    This result implies that we can make the best or most precise estimate of the mean value of y whenever $x^* = \bar{x}$.

11. (Figure 14.8) the further x^* is from $\bar{x}$, the larger $x^*-\bar{x}$ becomes. As a result, the confidence interval for the mean value of y will become wider as x^* deviates more from $\bar{x}$.

FIGURE 14.8 Confidence Intervals for the Mean Sales y at Given Values of Student Population x



Prediction Interval for an Individual Value of y

1. The predictor of y^*, the value of y corresponding to the given x^*, is $\hat{y}^* = \beta_0 + \beta_1 x^*$.

2. For the new restaurant located near Talbot College, $x^* = 10$ and the prediction of quarterly sales is $\hat{y}^* = 60 + 5(10) = 110$, or \$110,000. Note that the prediction of quarterly sales for the new Armand's restaurant near Talbot College is the _____ as the point estimate of the mean sales for all Armand's restaurants located near campuses with 10,000 students.

3. Determine the variance associated with using $\hat{y}^*$ as a predictor of y when $x = x^*$. This variance is made up of the sum of the following two components.

   (a) The (estimated) variance of the y^* values about the mean $E(y^*)$: _____.

   (b) The (estimated) variance associated with using $\hat{y}^*$ to estimate $E(y^*)$: _____.

4. The formula for estimating the variance corresponding to the prediction of the value of y when $x = x^*$, denoted s^2_{pred}, is

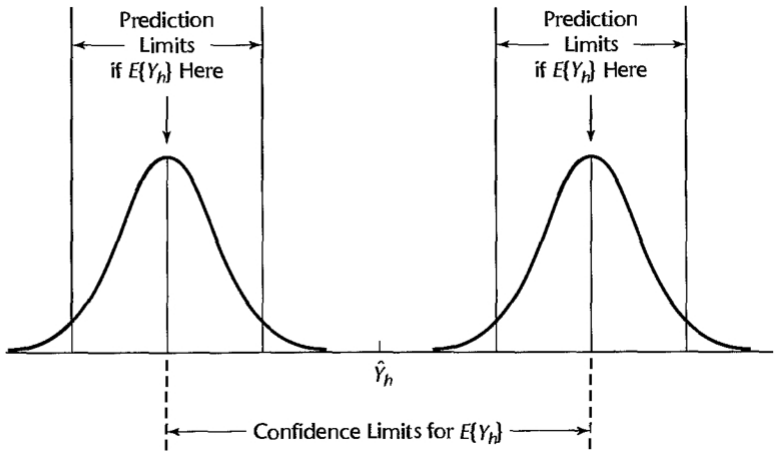
$$\begin{aligned} s^2_{pred} &= \underline{\qquad\qquad} \\ &= \underline{\qquad\qquad\qquad\qquad} \\ &= \underline{\qquad\qquad\qquad\qquad} \end{aligned} \tag{14.25}$$

5. Theorem:

$$\frac{\hat{y}^* - y^*_{new}}{s_{pred}} \sim t_{(n-2)}$$

$$\begin{aligned} \sigma^2_{pred} &= Var(\hat{y}^* - y^*_{new}) \\ &= Var(\hat{y}^*) + Var(y^*_{new}) \\ &= Var(\hat{y}^*) + \sigma^2 \end{aligned}$$

FIGURE 2.5 Prediction of $Y_{h(new)}$ when Parameters Unknown.



6. (Armand's Pizza Parlors) the estimated standard deviation corresponding to the prediction of quarterly sales for a new restaurant located near Talbot College, a campus with 10,000 students, is computed as follows.

$$s_{pred} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$$

7. **Prediction Interval For y^***

$$\underline{\qquad\qquad\qquad} \tag{14.27}$$

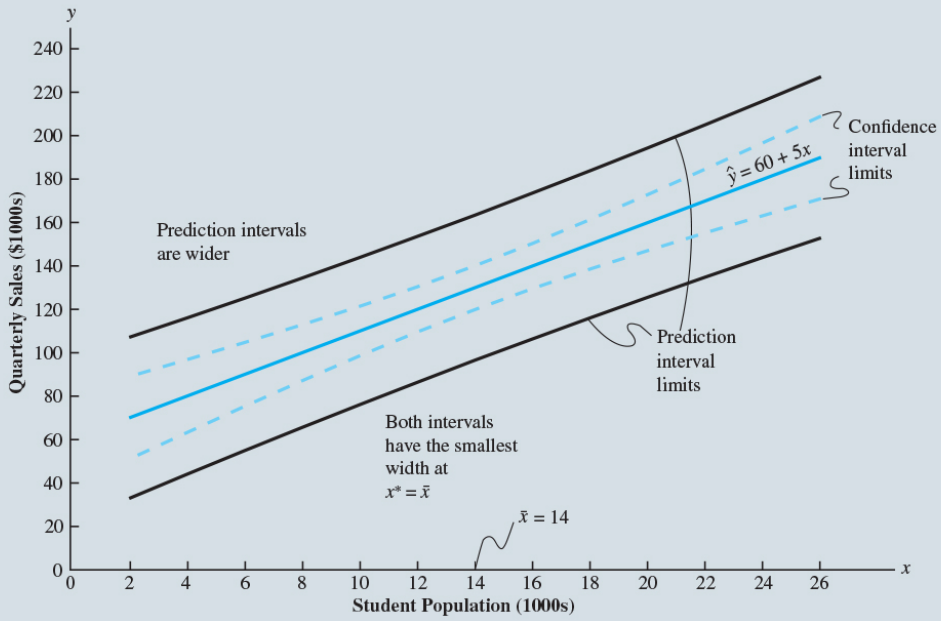
where the confidence coefficient is $1-\alpha$ and $t_{\alpha/2}$ is based on the t distribution with n–2 degrees of freedom.

8. (Armand's Pizza Parlors) The 95% prediction interval for quarterly sales for the new Armand's restaurant located near Talbot College, with $\hat{y}^* = 110$ and a margin of error of _____, the 95% prediction interval is 110 ± 33.875 ($76,125 to $143,875).

9. Note that the prediction interval for the new restaurant located near Talbot College, a campus with 10,000 students, is wider than the confidence interval for the mean quarterly sales of all restaurants located near campuses with 10,000 students. The difference reflects the fact that we are able to estimate the mean value of y _____ than we can predict an individual value of y.

10. (Figure 14.9) Confidence intervals and prediction intervals are both more precise when the value of the independent variable x^* is closer to $\bar{x}$.

FIGURE 14.9 Confidence and Prediction Intervals for Sales y at Given Values of Student Population x



😍 **EXERCISES 14.6**: 32, 36, 37

14.7 Computer Solution

FIGURE 14.10 Output for the Armand's Pizza Parlors Problem

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	14200.0	14200.0	74.25	.000
Error	8	1530.0	191.2		
Total	9	15730.0			

← ANOVA table

Model Summary

S	R-sq	R-sq(adj)
13.8293	90.27%	89.06%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	60.00	9.23	6.50	.000
Population	5.000	.580	8.62	.000

Regression Equation

Sales = 60.00 + 5.000 Population ← Estimated Regression Equation

Prediction for Sales

Variable	Setting
Population	10

Fit	SE Fit	95% CI		95% PI	
110	4.95099	(98.5830,	121.417)	(76.1275,	143.873)

← Interval Estimates

😍 **EXERCISES 14.7**: 40, 41

14.8 Residual Analysis: Validating Model Assumptions

1. Residual for observation i: the difference between the observed value of the dependent variable (y_i) and the predicted value of the dependent variable ($\hat{y}_i$), _____.

2. An analysis of the corresponding residuals will help determine whether the assumptions made about the regression model are appropriate.

3. (Table 14.7)

TABLE 14.7 Residuals for Armand's Pizza Parlors

Student Population x_i	Sales y_i	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

4. Example Armand's Pizza Parlors

   (a) A simple linear regression model was assumed.

   $$y = \beta_0 + \beta_1 x + \epsilon \qquad (14.29)$$

   This model indicates that we assumed quarterly sales (y) to be a linear function of the size of the student population (x) plus an error term ϵ. In Section 14.4 we made the following assumptions about the error term ϵ.

   1. _____.
   2. The variance of ϵ is the same for all values of x. _____.
   3. The values of ϵ are _____.
   4. The error term ϵ has a _____.

   (b) These assumptions provide the theoretical basis for the _____ and the _____ used to determine whether the relationship between x and y is significant, and for the _____ estimates presented in Section 14.6.

5. If the assumptions about the error term ϵ appear _____, the hypothesis tests about the significance of the regression relationship and the interval estimation results _____.

6. Much of residual analysis is based on an examination of graphical plots:

   (a) A plot of the _____ against values of the independent variable _____.

   (b) A plot of _____ against the _____ of the dependent variable y

   (c) A _____ plot.

   (d) A _____ plot.

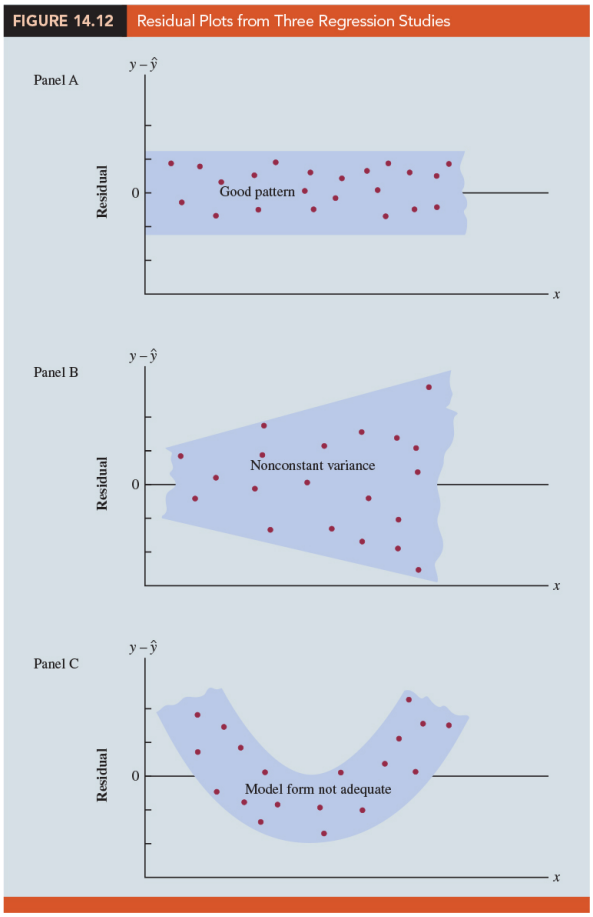
Residual Plot Against x

1. (Figure 14.12)

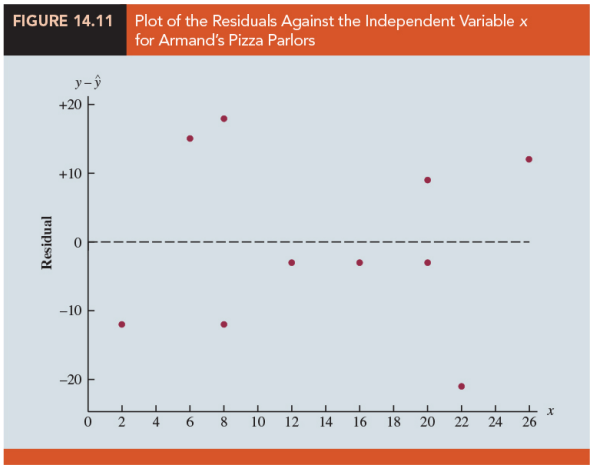
   (a) Panel A: If the assumption that the _____ is the same for all values of x and the assumed regression model is an adequate representation of the relationship between the variables, the residual plot should give an overall impression of a _____.

   (b) Panel B: if the _____ is not the same for all values of x—for example, if variability about the regression line is greater for larger values of x.

   (c) Panel C: we would conclude that the assumed regression model is not an adequate representation of the relationship between the variables. A _____ regression model or _____ regression model should be considered.



2. (Figure 14.11) Example Armand's Pizza Parlors:



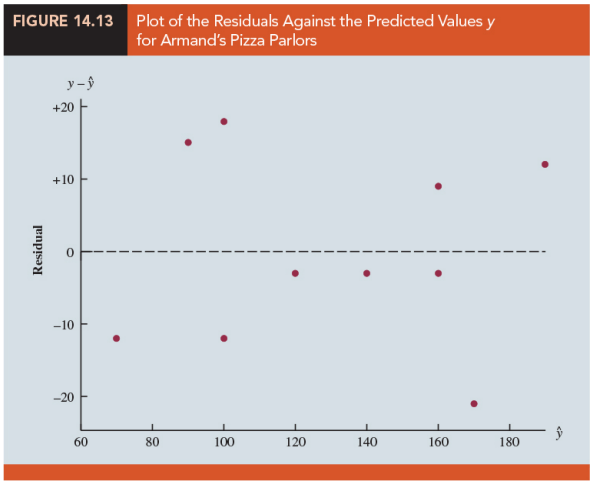
The residual plot does not provide evidence that the assumptions made for Armand's regression model should be challenged. At this point, we are confident in the conclusion that Armand's simple linear regression model is _____.

3. Experience and good judgment are always factors in the effective interpretation of residual plots.

Residual Plot Against $\hat{y}$

1. Another residual plot represents the predicted value of the dependent variable $\hat{y}$ on the horizontal axis and the residual values on the vertical axis.

2. (Figure 14.13) With the Armand's data from Table 14.7,



   Note that the pattern of this residual plot is the same as the pattern of the residual plot against the independent variable x.

3. For _____ analysis, the residual plot against $\hat{y}$ is more widely used because of the presence of more than one independent variable.

Standardized Residuals

1. A random variable is standardized by subtracting its mean and dividing the result by its standard deviation.

2. With the least squares method, the mean of the residuals is _____. Thus, simply dividing each residual by its _____ provides the standardized residual.

3. **Standard Deviation of the ith Residual**

$$s_{y_i-\hat{y}_i} = \underline{\qquad\qquad} \tag{14.30}$$

$s_{y_i-\hat{y}_i}$ = the standard deviation of residual i

s = the standard error of the estimate

h_i = $\underline{\qquad\qquad}$ (14.31)

4. **Standardized Residual for Observation i**

$$\underline{\qquad\qquad} \tag{14.32}$$

5. (Table 14.8) the standardized residuals for Armand's Pizza Parlors.

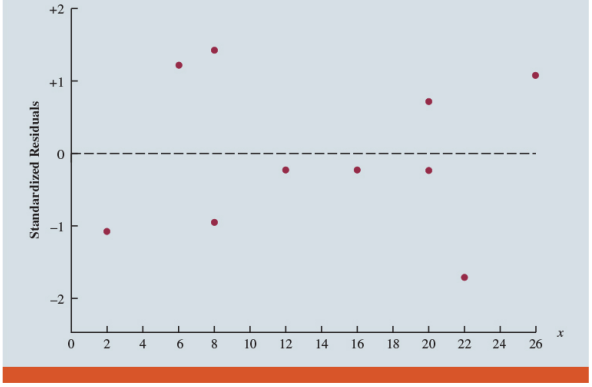
TABLE 14.8 Computation of Standardized Residuals for Armand's Pizza Parlors

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	−12	144	.2535	.3535	11.1193	−12	−1.0792
2	6	−8	64	.1127	.2127	12.2709	15	1.2224
3	8	−6	36	.0634	.1634	12.6493	−12	−.9487
4	8	−6	36	.0634	.1634	12.6493	18	1.4230
5	12	−2	4	.0070	.1070	13.0682	−3	−.2296
6	16	2	4	.0070	.1070	13.0682	−3	−.2296
7	20	6	36	.0634	.1634	12.6493	−3	−.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	−21	−1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
		Total	568					

Note: The values of the residuals were computed in Table 14.7.

6. (Figure 14.14)

FIGURE 14.14 Plot of the Standardized Residuals Against the Independent Variable x for Armand's Pizza Parlors



7. The standardized residual plot can provide insight about the assumption that the error term ϵ has a _____. If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a _____ probability distribution.

8. Thus, when looking at a standardized residual plot, we should expect to see approximately _____ of the standardized residuals between _____.

9. We see in Figure 14.14 that for the Armand's example all standardized residuals are between -2 and $+2$. Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that ϵ has a normal distribution.

Normal Probability Plot

1. Another approach for determining the validity of the assumption that the error term has a normal distribution is the normal probability plot.

2. To show how a normal probability plot is developed, we introduce the concept of _____.

   (a) Suppose 10 values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of 10 _____.

   (b) The random variable representing the smallest value obtained in repeated sampling is called the _____.

   (c) Statisticians show that for samples of size 10 from a standard normal probability distribution, the expected value of the first-order statistic is -1.55. This expected value is called a _____.

   NOTE: Compute the expected values of order statistics for a random sample from a standard normal distribution: `evNormOrdStats {EnvStats}`

   https://search.r-project.org/CRAN/refmans/EnvStats/html/evNormOrdStats.html

   (d) (Table 14.9) For the case with a sample of size $n = 10$, there are 10 order statistics and 10 normal.

TABLE 14.9

Normal Scores For $n = 10$

Order Statistic	Normal Score
1	−1.55
2	−1.00
3	−.65
4	−.37
5	−.12
6	.12
7	.37
8	.65
9	1.00
10	1.55

TABLE 14.10

Normal Scores and Ordered Standardized Residuals for Armand's Pizza Parlors

Normal Scores	Ordered Standardized Residuals
−1.55	−1.7114
−1.00	−1.0792
−.65	−.9487
−.37	−.2372
−.12	−.2296
.12	−.2296
.37	.7115
.65	1.0792
1.00	1.2224
1.55	1.4230

```
> data.frame(p, qnorm(p))
             p   qnorm.p.
1  0.00000000       -Inf
2  0.09090909 -1.3351777
3  0.18181818 -0.9084579
4  0.27272727 -0.6045853
5  0.36363636 -0.3487557
6  0.45454545 -0.1141853
7  0.54545455  0.1141853
8  0.63636364  0.3487557
9  0.72727273  0.6045853
10 0.81818182  0.9084579
11 0.90909091  1.3351777
12 1.00000000        Inf
```

(e) Let us now show how the 10 normal scores can be used to determine whether the standardized residuals for Armand's Pizza Parlors appear to come from a standard normal probability distribution.

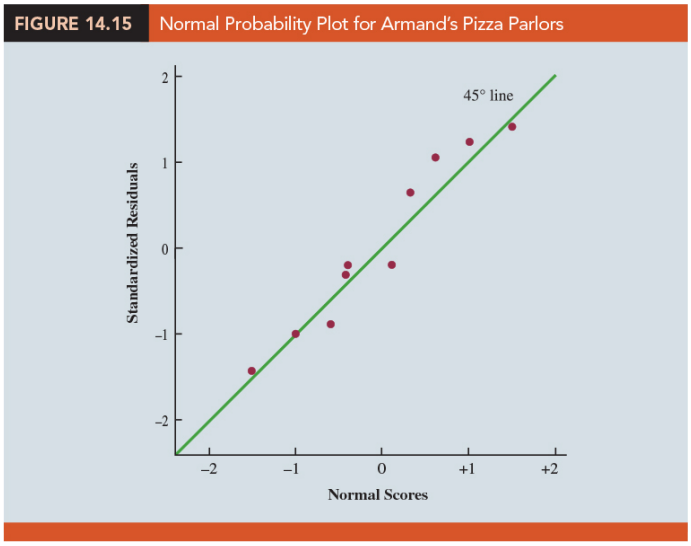
(f) (Table 14.10) The 10 normal scores and the ordered standardized residuals are shown together in Table 14.10. If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest

normal score, and so on.

(g) A normal probability plot: a plot with the _____ on the horizontal axis and the corresponding _____ on the vertical axis.

(h) If the standardized residuals are approximately normally distributed, the plotted points should cluster closely around a _____ passing through the _____.

3. (Figure14.15) the normal probability plot for the Armand's Pizza Parlors example: conclude that the assumption of the error term having a normal probability distribution is reasonable.



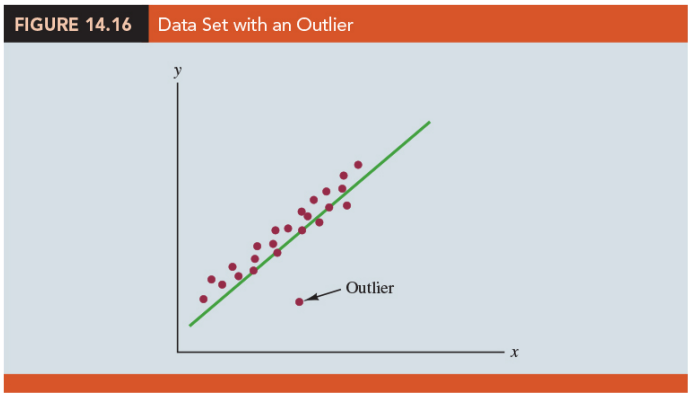
4. Any substantial curvature in the normal probability plot is evidence that the residuals have not come from a normal distribution.

😍 **EXERCISES 14.8**: 45, 47

14.9 Residual Analysis: Outliers and Influential Observations

Detecting Outliers

1. (Figure 14.16) is a scatter diagram for a data set that contains an _____, a data point (observation) that does not fit the trend shown by the remaining data.



2. Outliers represent observations that are suspect and warrant careful examination. They may represent _____ data; if so, the data should be _____.

3. They may signal a violation of model assumptions; if so, _____ should be considered.

4. Finally, they may simply be _____ values that occurred by chance. In this case, they should be retained.

5. (Table 14.11) The process of detecting outliers: Except for observation 4 ($x_4 = 3$, $y_4 = 75$), a pattern suggesting a negative linear relationship is apparent. Indeed, given the pattern of the rest of the data, we would expect y_4 to be much smaller and hence would identify the corresponding observation as an outlier.

6. For the case of simple linear regression, one can often detect outliers by simply examining the _____.

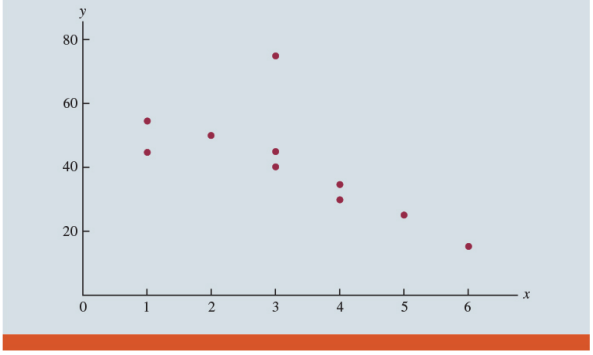
7. The _____ can also be used to identify outliers. If an observation deviates greatly from the pattern of the rest of the data, the corresponding standardized residual will be large in absolute value.

TABLE 14.11

Data Set Illustrating the Effect of an Outlier

x_i	y_i
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

FIGURE 14.17 Scatter Diagram for Outlier Data Set



8. (Figure 14.18) the output from a regression analysis. The highlighted portion of the output shows that the standardized residual for observation 4 is 2.67. With normally distributed errors, standardized residuals should be outside the range of −2 to +2 approximately 5% of the time.

FIGURE 14.18 Output for Regression Analysis of the Outlier Data Set

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1268.2	1268.2	7.90	.023
Error	8	1284.3	160.5		
Total	9	2552.5			

Model Summary

S	R-sq	R-sq(adj)
12.6704	49.68%	43.39%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	64.96	9.26	7.02	.000
x	−7.33	2.6	−2.81	.023

Regression Equation

y = 64.96 - 7.33 x

Observation	Predicted y	Residuals	Standard Residuals
1	57.6271	−12.6271	−1.0570
2	57.6271	−2.6271	−.2199
3	50.2966	−.2966	−.0248
4	42.9661	32.0339	2.6816
5	42.9661	−2.9661	−.2483
6	42.9661	2.0339	.1703
7	35.6356	−5.6356	−.4718
8	35.6356	−.6356	−.0532
9	28.3051	−3.3051	−.2767
10	20.9746	−5.9746	−.5001

9. In deciding how to handle an outlier, we should first check to see whether it is a _____. Perhaps an _____ was made in initially recording the data or in entering the data into the computer file.

10. (Figure 14.19) For example, suppose that in checking the data for the outlier in Table 14.11, we find an error; the correct value for observation 4 is $x_4 = 3, y_4 = 30$. Figure 14.19 is a portion of the output obtained after correction of the value of y_4. We see that using the incorrect data value substantially affected the goodness of fit. With the correct data, the value of _____ increased from 49.68% to 83.8% and the value of _____ decreased from 64.96 to 59.24. The _____ of the line changed from −7.33 to −6.95.

FIGURE 14.19 Output for the Revised Outlier Data Set

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1139.66	1139.66	41.38	.000
Error	8	220.34	27.54		
Total	9	1360.00			

Model Summary

S	R-sq	R-sq(adj)
5.24808	83.80%	81.77%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	59.24	3.83	15.45	.000
x	−6.95	1.08	−6.43	.000

Regression Equation

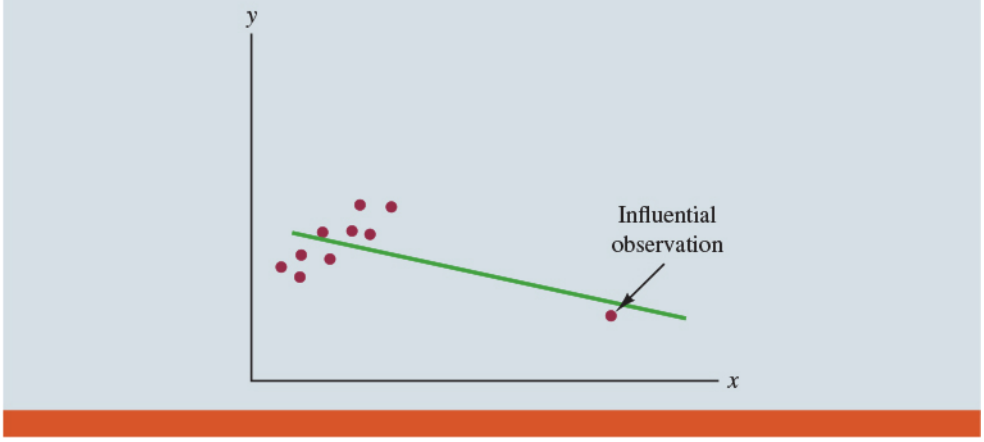
y = 59.24 − 6.95 x

11. The identification of the outlier enabled us to correct the data error and improve the regression results.

Detecting Influential Observations

1. (Figure 14.20) shows an example of an influential observation in simple linear regression.

FIGURE 14.20 Data Set with an Influential Observation



The estimated regression line has a negative slope. However, if the influential observation were dropped from the data set, the slope of the estimated regression line would change from negative to positive and the y-intercept would be smaller. Clearly, this one observation is much more influential in determining the estimated regression line than any of the others.

2. Influential observations can be identified from a _____ when only one independent variable is present.

3. An influential observation may be an _____ (an observation with a y value that deviates substantially from the trend), it may correspond to an x value far away from its mean (e.g., see Figure 14.20), or it may be caused by a combination of the two (a somewhat off-trend y value and a somewhat extreme x value).

4. The presence of the influential observation in Figure 14.20, if valid, would suggest trying to obtain data on intermediate values of x to understand better the relationship between x and y.

5. Observations with _____ for the independent variables are called high _____. The influential observation in Figure 14.20 is a point with high leverage.

6. The leverage of an observation is determined by how far the values of the independent variables are from their _____.

7. For the single-independent-variable case, the leverage of the ith observation, denoted h_i, can be computed by using equation (14.33).

$$h_i = \underline{\qquad\qquad\qquad\qquad}$$

Definition and properties of leverages:
`https://online.stat.psu.edu/stat501/lesson/11/11.2`

8. From the formula, it is clear that the farther x_i is from its mean $\bar{x}$, the higher the leverage of observation i.

9. (Figure 14.21) a scatter diagram for the data set in Table 14.12, it is clear that observation 7 ($x = 70, y = 100$) is an observation with an extreme value of x. Hence, we would expect it to be identified as a point with high leverage:

$$h_7 = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2621.43} = .094$$

10. For the case of simple linear regression, observations have high leverage if $h_i > 6/n$ or 0.99, whichever is smaller.

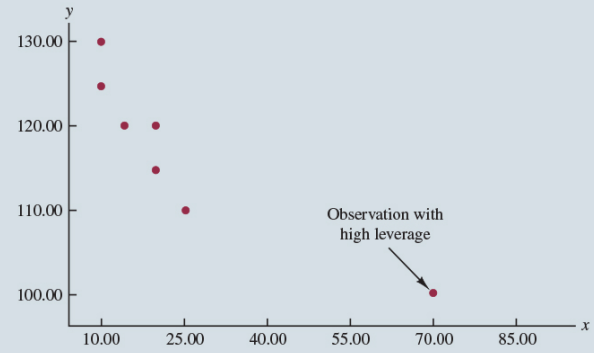
11. For the data set in Table 14.12, $6/n = 6/7 = 0.86$. Because $h_7 = 0.94 > 0.86$, we will identify observation 7 as an observation whose x value gives it large influence.

12. Influential observations that are caused by an interaction of large residuals and high leverage can be difficult to detect. Diagnostic procedures are available that take both into account in determining when an observation is influential. One such measure, called _____, will be discussed in Chapter 15.

TABLE 14.12
Data Set with a High Leverage Observation

x_i	y_i
10	125
10	130
15	120
20	115
20	120
25	110
70	100

FIGURE 14.21 Scatter Diagram for the Data Set with a High Leverage Observation



😍 **EXERCISES 14.9**: 50, 52

😍 **SUPPLEMENTARY EXERCISES**: 59, 67