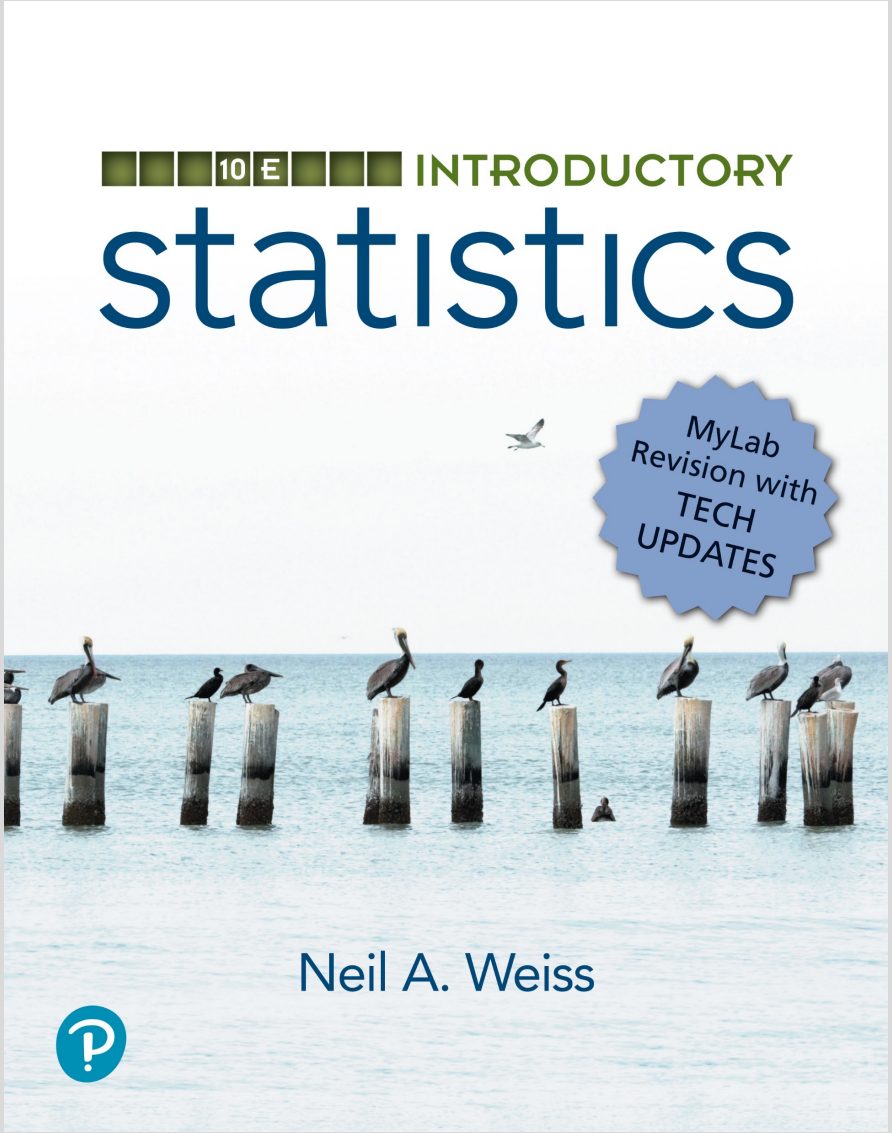
Chapter 15

Inferential Methods in Regression and Correlation



 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Section 15.1

The Regression Model; Analysis of Residuals

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Definition 15.1

Conditional Distribution, Mean, and Standard Deviation

Suppose that *x* and *y* are predictor and response variables, respectively, on a population. Let x_p denote a particular value of the predictor variable and consider the subpopulation consisting of all members of the population whose value of the predictor value is x_p.

Conditional distribution of the response variable corresponding to x_p: The distribution of all possible values of the response variable on the aforementioned subpopulation.

Conditional mean of the response variable corresponding to x_p: The mean of all possible values of the response variable on the aforementioned subpopulation.

Conditional standard deviation of the response variable corresponding to x_p: The standard deviation of all possible values of the response variable on the aforementioned subpopulation.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.1

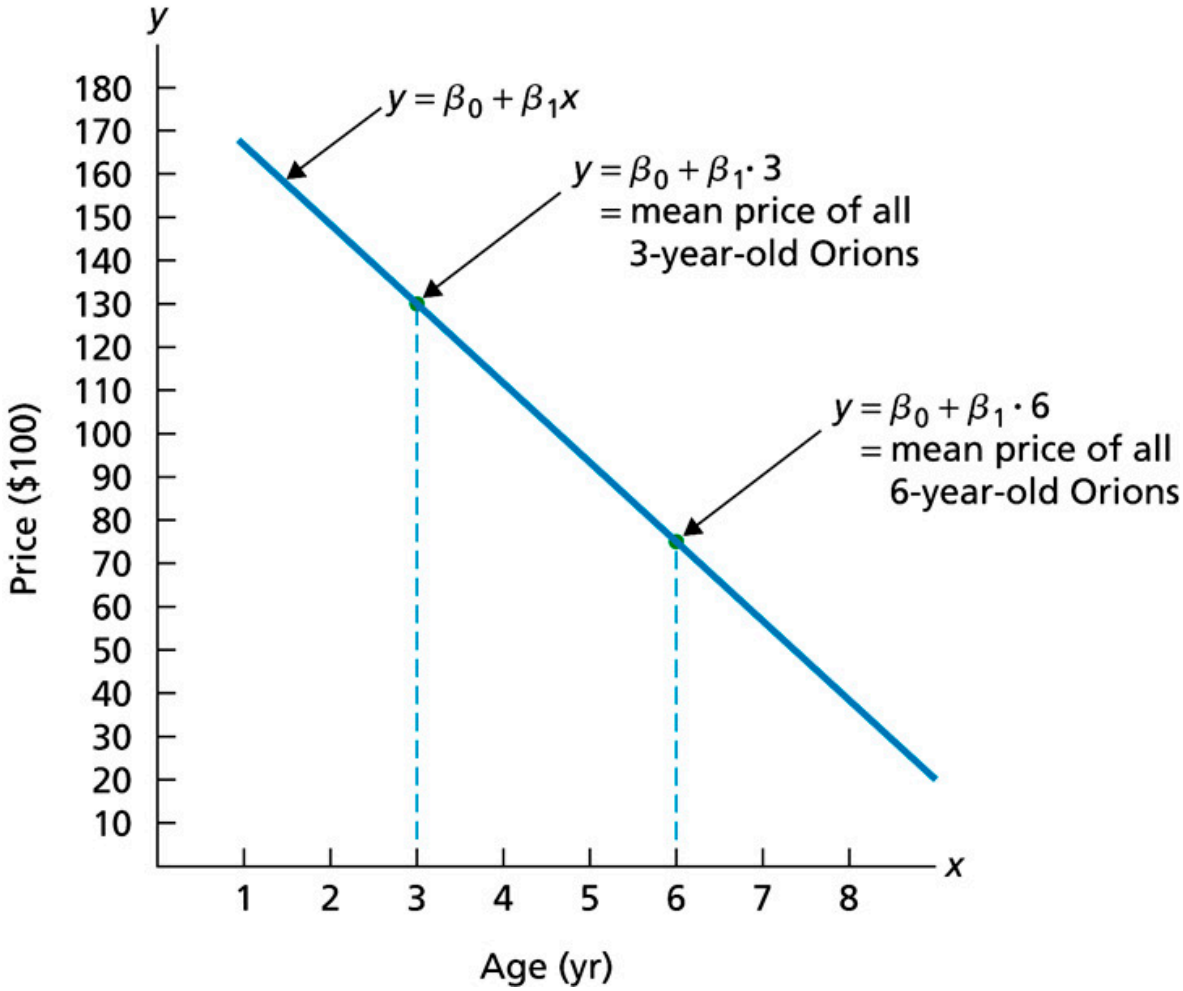
Assumptions (Conditions) for Regression Inferences

1. **Population regression line:** There are constants β_0 and β_1 such that, for each value x of the predictor variable, the conditional mean of the response variable is $\beta_0 + \beta_1 x$.
2. **Equal standard deviations:** The conditional standard deviations of the response variable are the same for all values of the predictor variable. We denote this common standard deviation σ.†
3. **Normal populations:** For each value of the predictor variable, the conditional distribution of the response variable is a normal distribution.
4. **Independent observations:** The observations of the response variable are independent of one another.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Figure 15.1

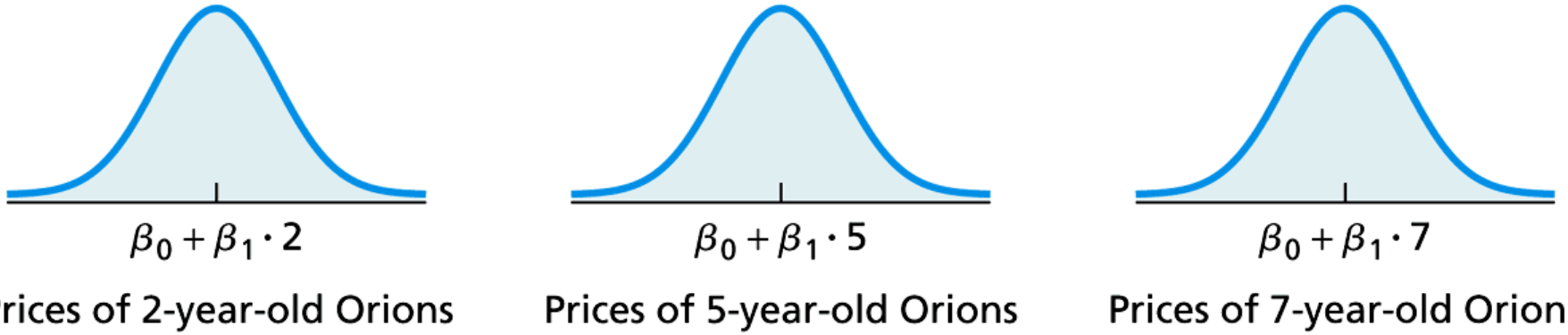
Population regression line



 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Figure 15.2

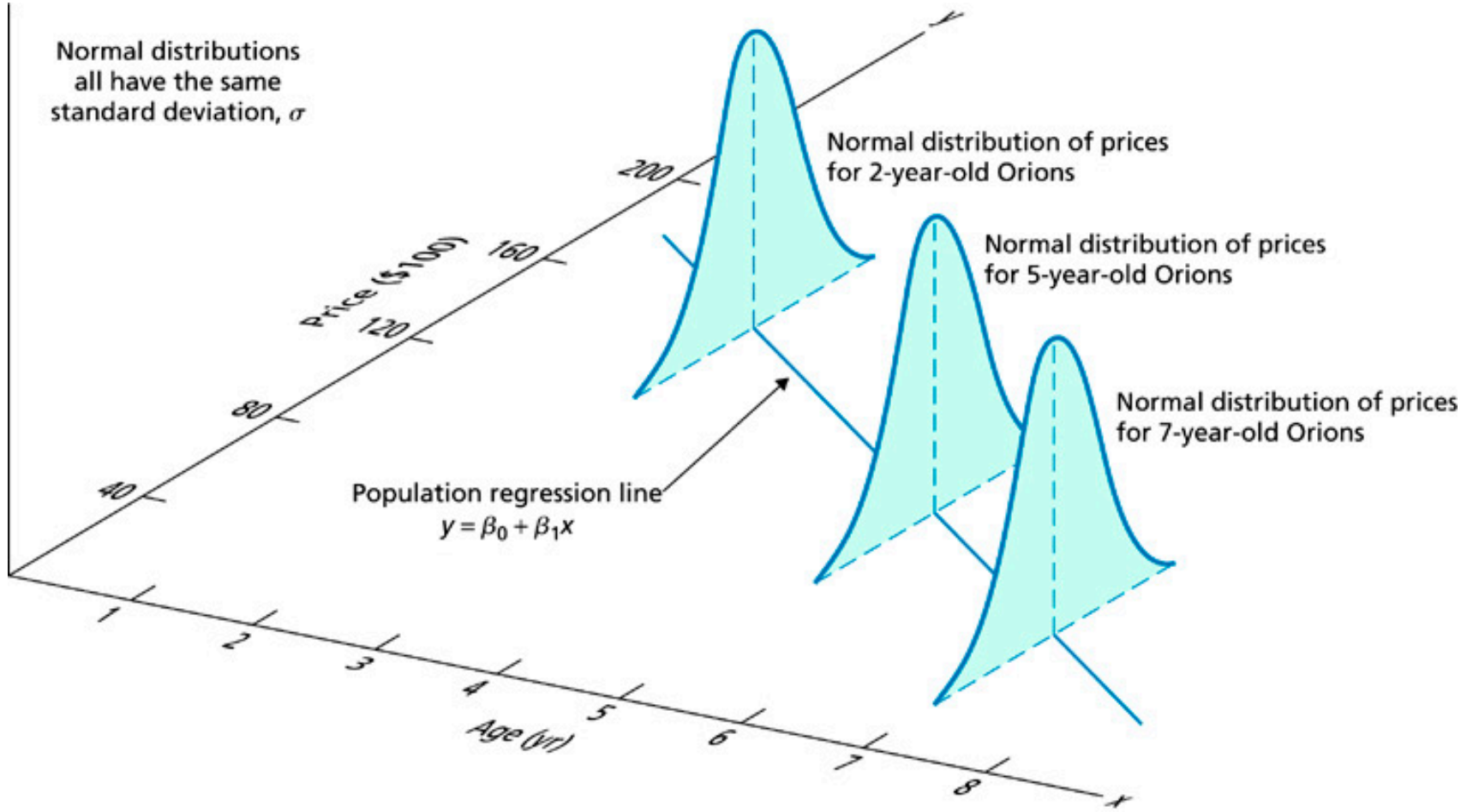
Price distributions for 2-, 5-, and 7-year-old Orions under Assumptions 2 and 3 (The means shown for the three normal distributions reflect Assumption 1)



Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Figure 15.3

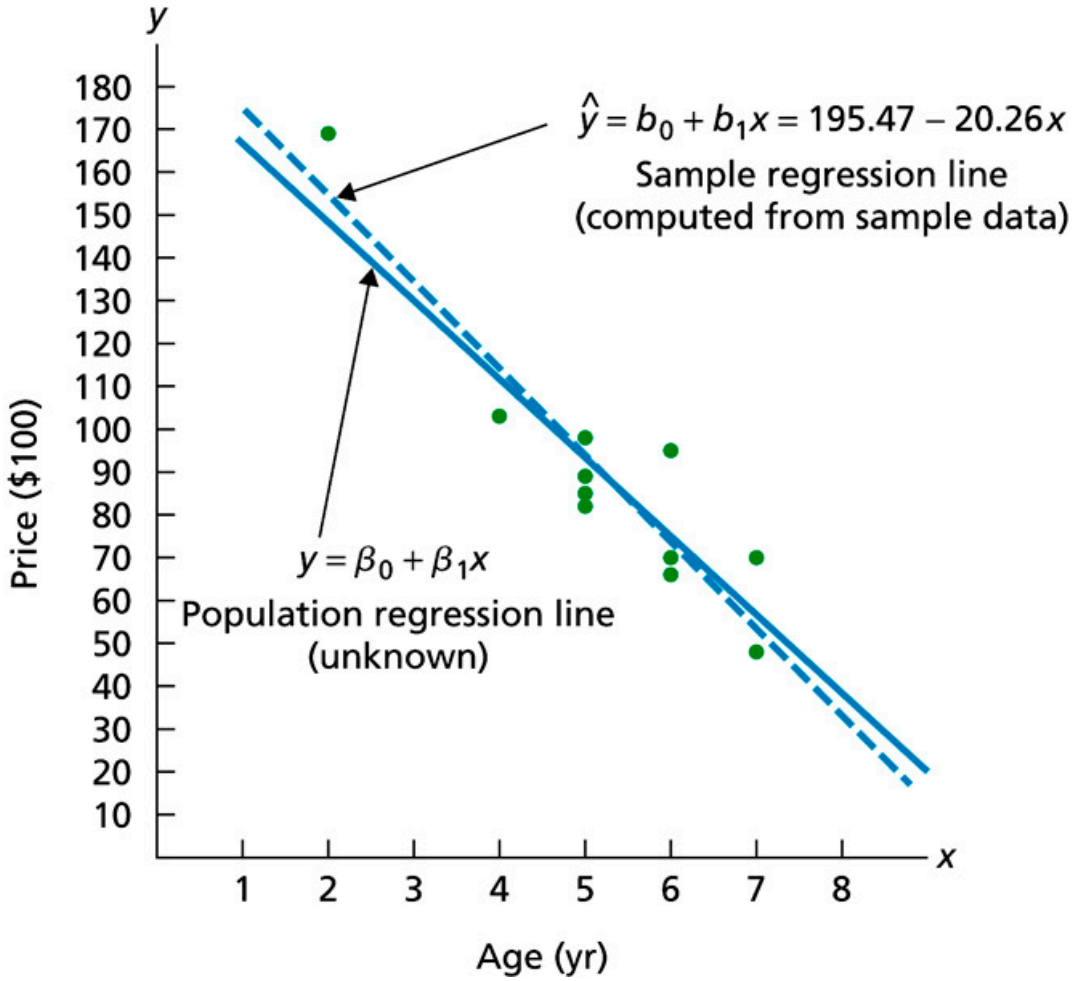
Graphical portrayal of Assumptions 1–3 for regression inferences pertaining to age and price of Orions



 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Figure 15.4

Population regression line and sample regression line for age and price of Orions



 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Definition 15.2

Standard Error of the Estimate

The **standard error of the estimate, s_e,** is defined by

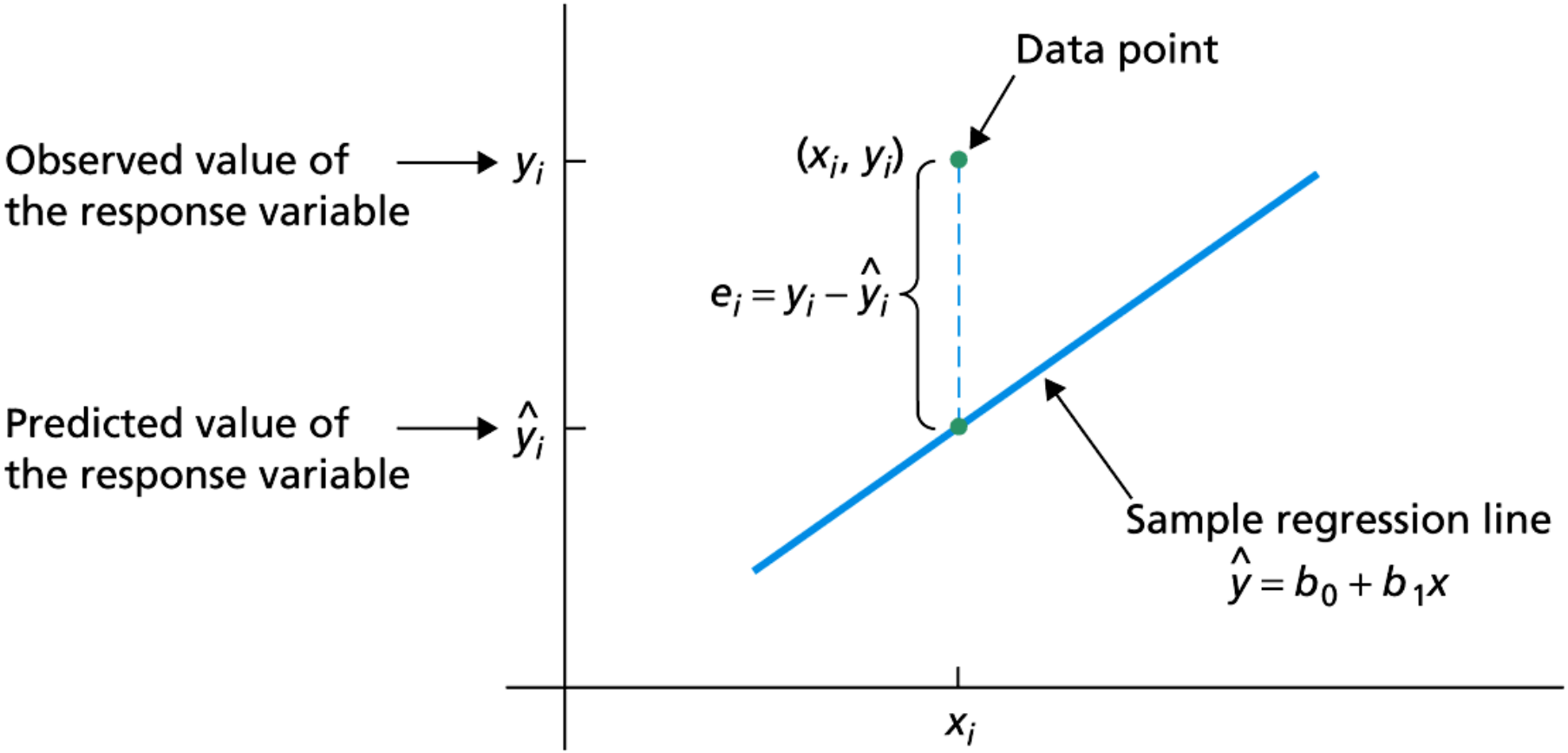
$$s_e = \sqrt{\frac{SSE}{n-2}},$$

where *SSE* is the error sum of squares.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Figure 15.5

Residual of a data point



Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.2

Residual Analysis for the Regression Model

If the assumptions for regression inferences are met, the following two conditions should hold:

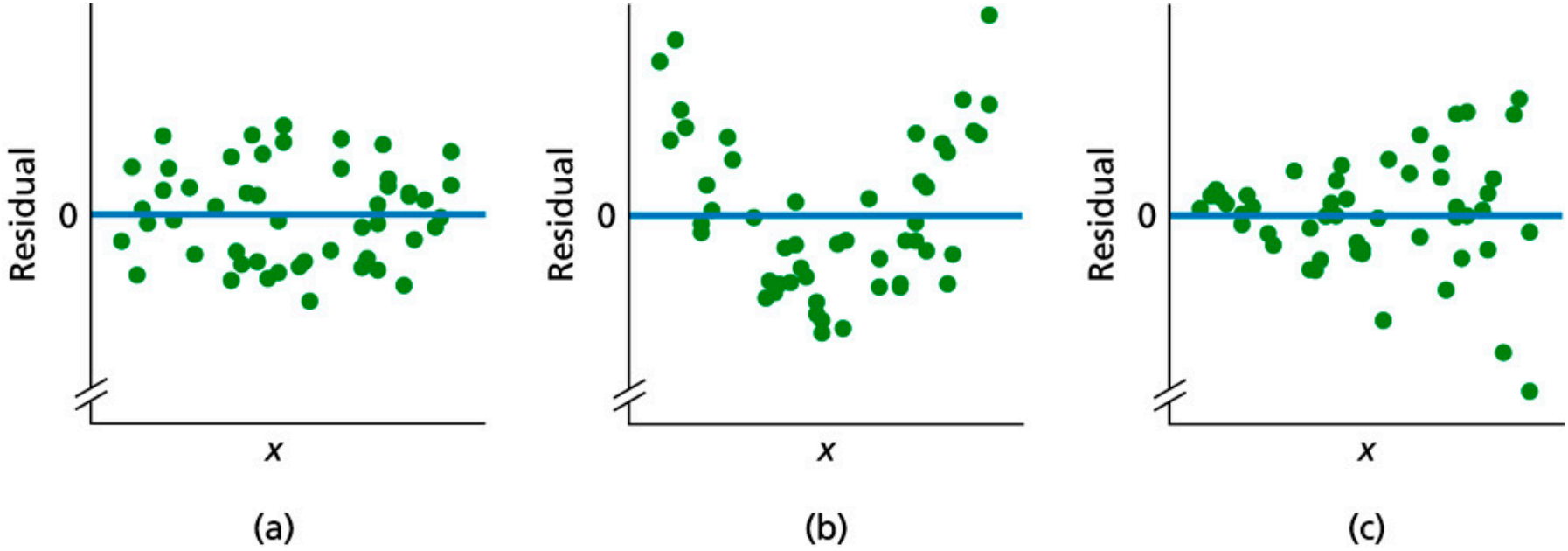
- A plot of the residuals against the observed values of the predictor variable should fall roughly in a horizontal band centered and symmetric about the *x*-axis.
- A normal probably plot of the residuals should be roughly linear.

Failure of either of these two conditions casts doubt on the validity of one or more of the assumptions for regression inferences for the variables under consideration.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Figure 15.6

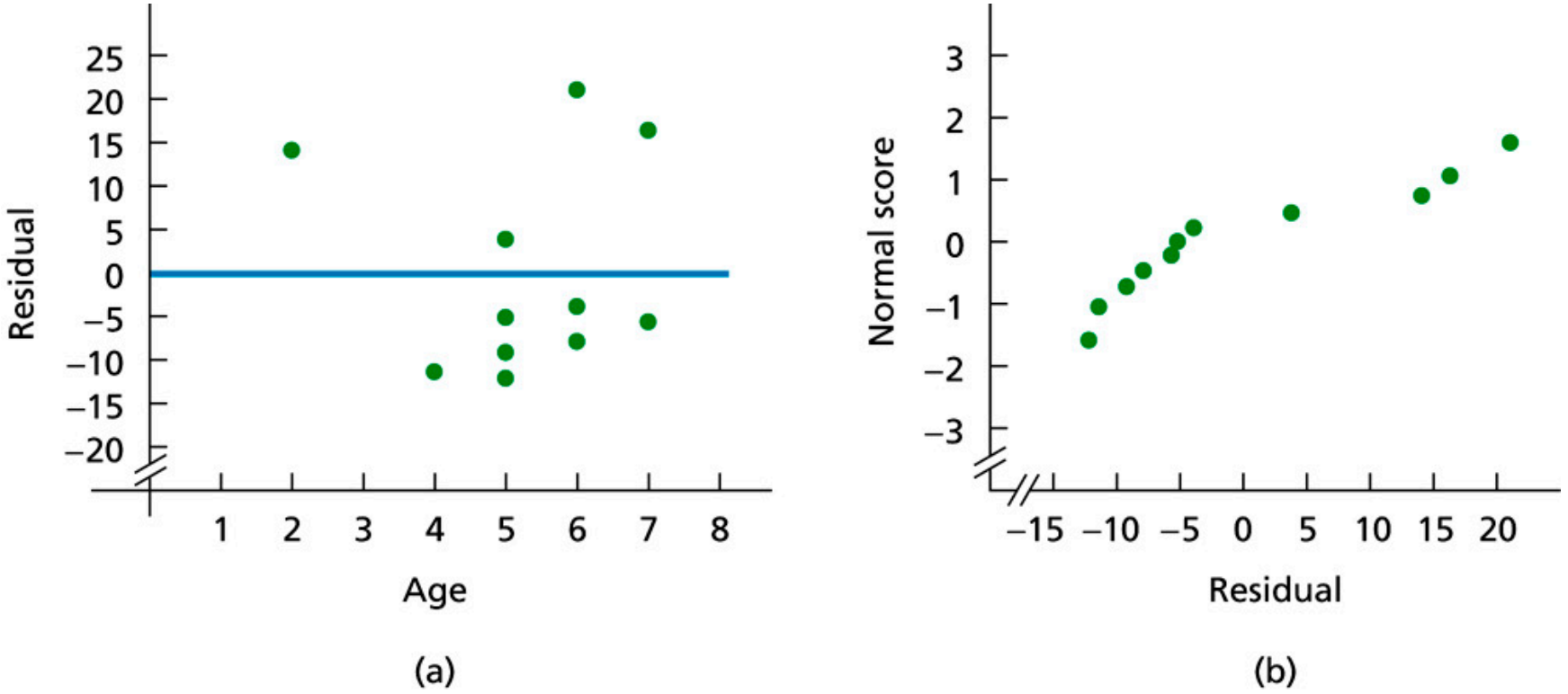
Residual plots suggesting (a) no violation of linearity or constant standard deviation, (b) violation of linearity, and (c) violation of constant standard deviation



 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Figure 15.7

(a) Residual plot; (b) normal probability plot for residuals



 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Section 15.2

Inferences for the Slope of the Population Regression Line

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.3

The Sampling Distribution of the Slope of the Regression Line

Suppose that the variables x and y satisfy the four assumptions for regression inferences. Then, for samples of size n, each with the same values $x_1, x_2, \ldots, x_n$ for the predictor variable, the following properties hold for the slope, b_1, of the sample regression line:

- The mean of b_1 equals the slope of the population regression line; that is, we have $\mu_{b_1} = \beta_1$ (i.e., the slope of the sample regression line is an unbiased estimator of the slope of the population regression line).
- The standard deviation of b_1 is $\sigma_{b_1} = \sigma / \sqrt{S_{xx}}$.
- The variable b_1 is normally distributed.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.4

t-Distribution for Inferences for β_1

Suppose that the variables x and y satisfy the four assumptions for regression inferences. Then, for samples of size n, each with the same values $x_1, x_2, \ldots, x_n$ for the predictor variable, the variable

$$t = \frac{b_1 - \beta_1}{s_e / \sqrt{S_{xx}}}$$

has the t-distribution with $\text{df} = n - 2$.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.1

Pooled *t*-Interval Procedure

Purpose To find a confidence interval for the difference between two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1 **For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n_1 + n_2 - 2$.**

Step 2 **The endpoints of the confidence interval for $\mu_1 - \mu_2$ are**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot s_p \sqrt{(1/n_1) + (1/n_2)},$$

where s_p is the pooled sample standard deviation.

Step 3 **Interpret the confidence interval.**

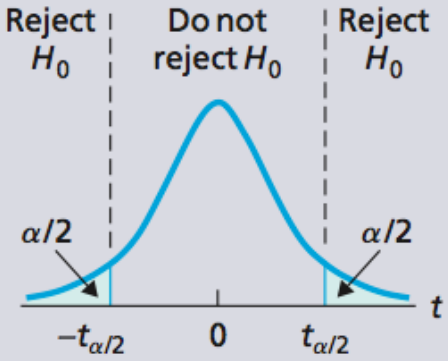
Note: The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.1 (cont.)

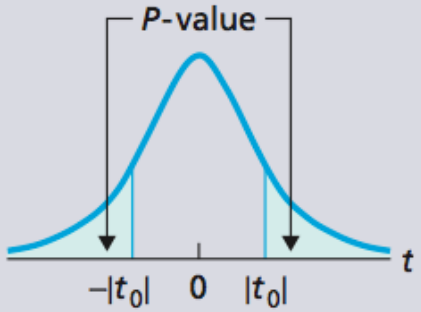
CRITICAL-VALUE APPROACH	OR	P-VALUE APPROACH

Step 4 The critical values are $\pm t_{\alpha/2}$ with $\text{df} = n - 2$. Use Table IV to find the critical values.



Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

Step 4 The t-statistic has $\text{df} = n - 2$. Use Table IV to estimate the P-value, or obtain it exactly by using technology.



Step 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

Step 6 Interpret the results of the hypothesis test.

Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.2

Regression *t*-Interval Procedure

Purpose To find a confidence interval for the slope, β_1, of the population regression line

Assumptions The four assumptions for regression inferences

Step 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n - 2$.

Step 2 The endpoints of the confidence interval for β_1 are

$$b_1 \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{S_{xx}}}.$$

Step 3 Interpret the confidence interval.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Section 15.3

Estimation and Prediction

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.5

Distribution of the Predicted Value of a Response Variable

Suppose that the variables x and y satisfy the four assumptions for regression inferences. Let x_p denote a particular value of the predictor variable, and let $\hat{y}_p$ be the corresponding value predicted for the response variable by the sample regression equation; that is, $\hat{y}_p = b_0 + b_1 x_p$. Then, for samples of size n, each with the same values $x_1, x_2, \ldots, x_n$ for the predictor variable, the following properties hold for $\hat{y}_p$.

- The mean of $\hat{y}_p$ equals the conditional mean of the response variable corresponding to the value x_p of the predictor variable: $\mu_{\hat{y}_p} = \beta_0 + \beta_1 x_p$.
- The standard deviation of $\hat{y}_p$ is

$$\sigma_{\hat{y}_p} = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}.$$

- The variable $\hat{y}_p$ is normally distributed.

In particular, for fixed values of the predictor variable, the possible predicted values of the response variable corresponding to x_p have a normal distribution with mean $\beta_0 + \beta_1 x_p$.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.6

t-Distribution for Confidence Intervals for Conditional Means in Regression

Suppose that the variables x and y satisfy the four assumptions for regression inferences. Then, for samples of size n, each with the same values $x_1, x_2, \ldots, x_n$ for the predictor variable, the variable

$$t = \frac{\hat{y}_p - (\beta_0 + \beta_1 x_p)}{s_e \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}}$$

has the t-distribution with $\text{df} = n - 2$.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.3

Conditional Mean *t*-Interval Procedure

Purpose To find a confidence interval for the conditional mean of the response variable corresponding to a particular value of the predictor variable, x_p

Assumptions The four assumptions for regression inferences

Step 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with df $= n - 2$.

Step 2 Compute the point estimate, $\hat{y}_p = b_0 + b_1 x_p$.

Step 3 The endpoints of the confidence interval for the conditional mean of the response variable are

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}.$$

Step 4 Interpret the confidence interval.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.7

Distribution of the Difference between the Observed and Predicted Values of the Response Variable

Suppose that the variables x and y satisfy the four assumptions for regression inferences. Let x_p denote a particular value of the predictor variable, and let $\hat{y}_p$ be the corresponding value predicted for the response variable by the sample regression equation. Furthermore, let y_p be an independently observed value of the response variable corresponding to the value x_p of the predictor variable. Then, for samples of size n, each with the same values $x_1, x_2, \ldots, x_n$ for the predictor variable, the following properties hold for $y_p - \hat{y}_p$, the difference between the observed and predicted values.

- The mean of $y_p - \hat{y}_p$ equals zero: $\mu_{y_p - \hat{y}_p} = 0$.
- The standard deviation of $y_p - \hat{y}_p$ is

$$\sigma_{y_p - \hat{y}_p} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}.$$

- The variable $y_p - \hat{y}_p$ is normally distributed.

In particular, for fixed values of the predictor variable, the possible differences between the observed and predicted values of the response variable corresponding to x_p have a normal distribution with a mean of 0.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.8

t-Distribution for Prediction Intervals in Regression

Suppose that the variables x and y satisfy the four assumptions for regression inferences. Then, for samples of size n, each with the same values $x_1, x_2, \ldots, x_n$ for the predictor variable, the variable

$$t = \frac{y_p - \hat{y}_p}{s_e\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}}$$

has the t-distribution with $\text{df} = n - 2$.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.4

Predicted Value *t*-Interval Procedure

Purpose To find a prediction interval for the value of the response variable corresponding to a particular value of the predictor variable, x_p

Assumptions The four assumptions for regression inferences

Step 1 For a prediction level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $\text{df} = n - 2$.

Step 2 Compute the predicted value, $\hat{y}_p = b_0 + b_1 x_p$.

Step 3 The endpoints of the prediction interval for the value of the response variable are

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}.$$

Step 4 Interpret the prediction interval.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Section 15.4

Inferences in Correlation

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Key Fact 15.9

t-Distribution for a Correlation Test

Suppose that the variables x and y satisfy the four assumptions for regression inferences and that $\rho = 0$. Then, for samples of size n, the variable

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

has the t-distribution with $\text{df} = n - 2$.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.5

Correlation *t*-Test

Purpose To perform a hypothesis test for a population linear correlation coefficient, ρ

Assumptions The four assumptions for regression inferences

Step 1 The null hypothesis is H_0: $\rho = 0$, and the alternative hypothesis is

$$H_a: \rho \neq 0 \text{ (Two tailed)} \quad \text{or} \quad H_a: \rho < 0 \text{ (Left tailed)} \quad \text{or} \quad H_a: \rho > 0 \text{ (Right tailed)}.$$

Step 2 Decide on the significance level, α.

Step 3 Compute the value of the test statistic

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

and denote that value t_0.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

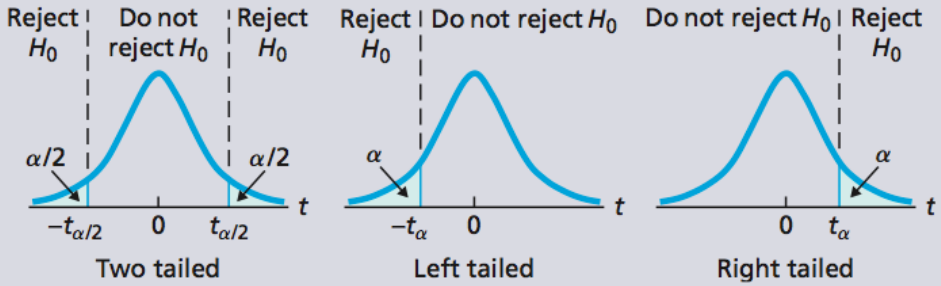
Procedure 15.5 (cont.)

CRITICAL-VALUE APPROACH

Step 4 The critical value(s) are

$\pm t_{\alpha/2}$ (Two tailed) or $-t_{\alpha}$ (Left tailed) or t_{α} (Right tailed)

with $df = n - 2$. Use Table IV to find the critical value(s).

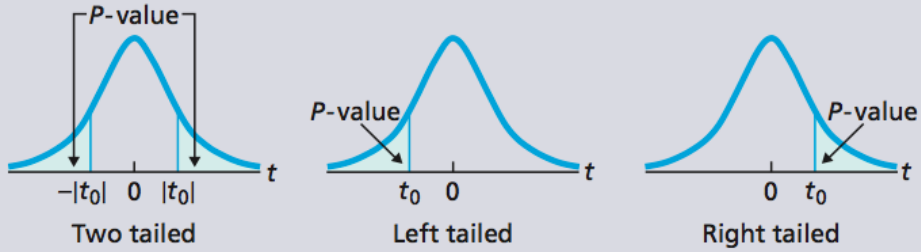


Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

OR

P-VALUE APPROACH

Step 4 The t-statistic has $df = n - 2$. Use Table IV to estimate the P-value, or obtain it exactly by using technology.



Step 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

Step 6 Interpret the results of the hypothesis test.

Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Section 15.5

Testing for Normality

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.6

Correlation Test for Normality

Purpose To perform a hypothesis test to decide whether a variable is not normally distributed

Assumption Simple random sample

Step 1 **The null and alternative hypotheses are, respectively,**

H_0: **The variable is normally distributed**

H_a: **The variable is not normally distributed.**

Step 2 **Decide on the significance level, α.**

Step 3 **Compute the value of the test statistic**

$$R_p = \frac{\Sigma x_i w_i}{\sqrt{S_{xx} \Sigma w_i^2}},$$

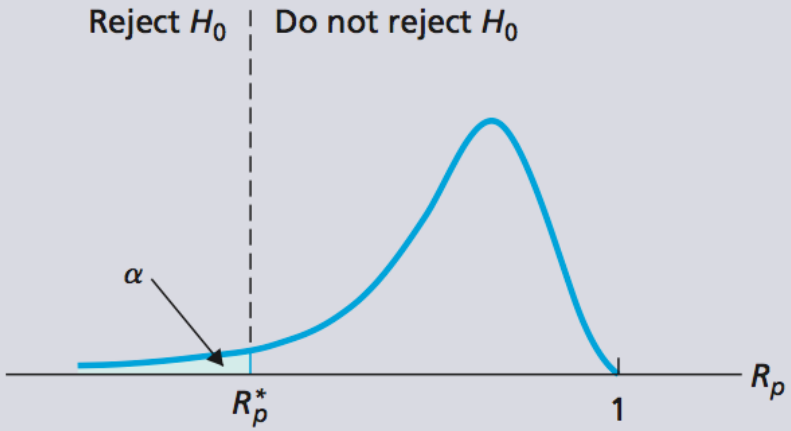
where x and w denote observations of the variable and the corresponding normal scores, respectively. Denote the value of the test statistic R_p^0.

 Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Procedure 15.6 (cont.)

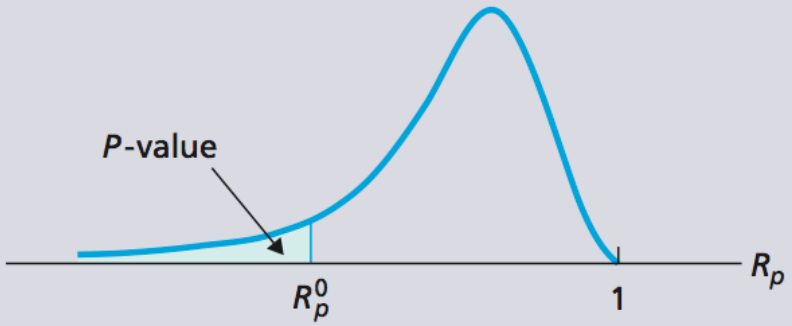
CRITICAL-VALUE APPROACH	OR	P-VALUE APPROACH

Step 4 The critical value is R_p^*. Use Table XII to find the critical value.



Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

Step 4 Use Table XII to estimate the P-value, or obtain it exactly by using technology.



Step 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

Step 6 Interpret the results of the hypothesis test.

Copyright © 2020, 2016, 2012 Pearson Education, Inc.

Table 15.7

Table for computing R_p

Adjusted gross income x	Normal score w	xw	x^2	w^2
7.8	−1.64	−12.792	60.84	2.6896
9.7	−1.11	−10.767	94.09	1.2321
10.6	−0.79	−8.374	112.36	0.6241
12.7	−0.53	−6.731	161.29	0.2809
12.8	−0.31	−3.968	163.84	0.0961
18.1	−0.10	−1.810	327.61	0.0100
21.2	0.10	2.120	449.44	0.0100
33.0	0.31	10.230	1,089.00	0.0961
43.5	0.53	23.055	1,892.25	0.2809
51.1	0.79	40.369	2,611.21	0.6241
81.4	1.11	90.354	6,625.96	1.2321
93.1	1.64	152.684	8,667.61	2.6896
395.0	0.00	274.370	22,255.50	9.8656

Copyright © 2020, 2016, 2012 Pearson Education, Inc.