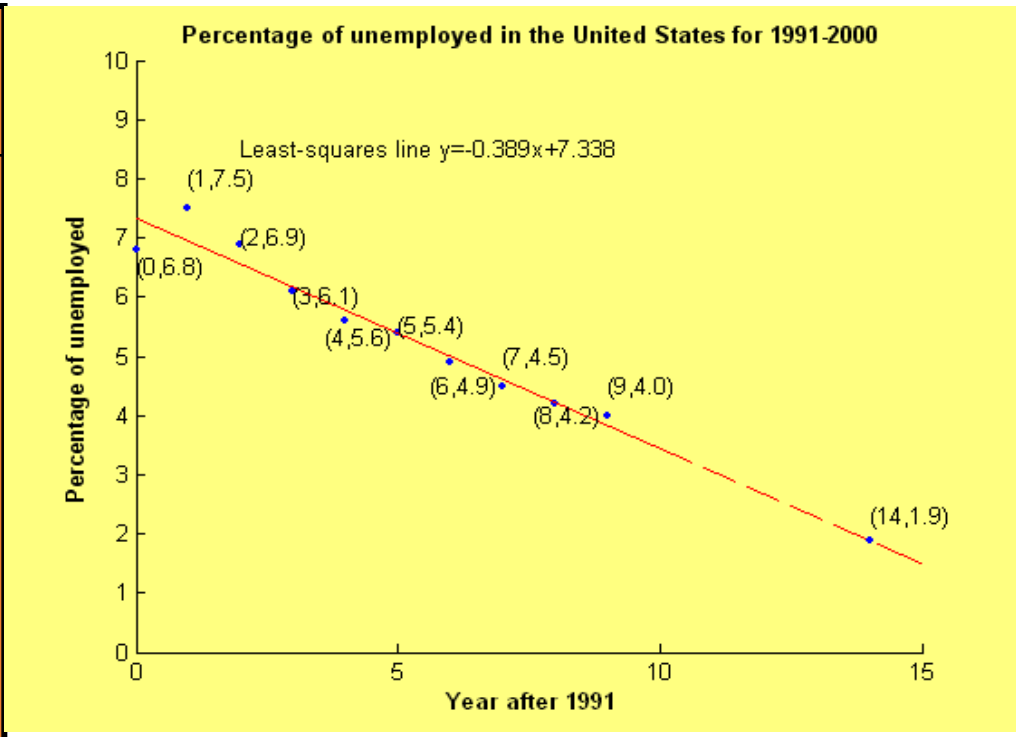
Chapter 13

Simple Linear Regression Analysis

Table 14.1 lists the percentage of the labour force that was unemployed during the decade *1991-2000*. Plot a graph with **the time (years after 1991)** on the *x* axis and **percentage of unemployment** on the *y* axis. Do the points follow a clear pattern? Based on these data, what would you expect **the percentage of unemployment** to be in the year *2005*?

Table 14.1 Percentage of Civilian Unemployment

Year	Number of Years from 1991	Percentage of Unemployed
1991	0	6.8
1992	1	7.5
1993	2	6.9
1994	3	6.1
1995	4	5.6
1996	5	5.4
1997	6	4.9
1998	7	4.5
1999	8	4.2
2000	9	4.0



The pattern does suggest that we may be able to get useful information by finding a line that "best fits" the data in some meaningful way. It produces the "best-fitting line".

$$y = -0.389x + 7.338$$

Based on this formula, we can attempt a prediction of the unemployment rate in the year *2005*:

$$y(14) = -0.389(14) + 7.338 = 1.892$$

Learning Objectives

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients b_0 and b_1
- To make inferences about the slope
- To estimate mean values and predict individual values

Managers often make decisions by studying the relationships between variables, and process improvements can often be made by understanding how changes in one or more variables affect the process output. **Regression analysis** is a statistical technique in which we use observed data to relate a variable of interest, which is called the **dependent** (or **response**) **variable,** to one or more **independent** (or **predictor**) **variables.** The objective is to build a **regression model,** or **prediction equation,** that can be used to **describe, predict,** and **control** the dependent variable on the basis of the independent variables. For example, a company might wish to improve its marketing process. After collecting data concerning the demand for a product, the product's price, and the advertising expenditures made to promote the product, the company might use regression analysis to develop an equation to predict demand on the basis of price and advertising expenditure. Predictions of demand for various price–advertising expenditure combinations can then be used to evaluate potential changes in the company's marketing strategies.

In the next two chapters we give a thorough presentation of regression analysis. We begin in this chapter by presenting **simple linear regression** analysis. Using this technique is appropriate when we are relating a dependent variable to a single independent variable and when *a straight-line model* describes the relationship between these two variables. We explain many of the methods of this chapter in the context of two new cases:

Simple Linear Regression Analysis

13.1 The Simple Linear Regression Model and the Least Square Point Estimates

13.2 Model Assumptions and the Standard Error

13.3 Testing the Significance of Slope and y-Intercept

13.4 Confidence and Prediction Intervals

13.5 Simple Coefficients of Determination and Correlation

Simple Linear Regression Analysis

13.6 Testing the Significance of the Population Correlation Coefficient (Optional)

13.7 An F Test for the Model

13.8 Residual Analysis

13.9 Some Shortcut Formulas (Optional)

13.1 The Simple Linear Regression Model and the Least Squares Point Estimates

Correlation vs. Regression

- A scatter diagram can be used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Correlation is only concerned with strength of the relationship
  - No causal effect is implied with correlation

Scatter Diagrams

- **Scatter Diagrams** are used to examine possible relationships between two numerical variables

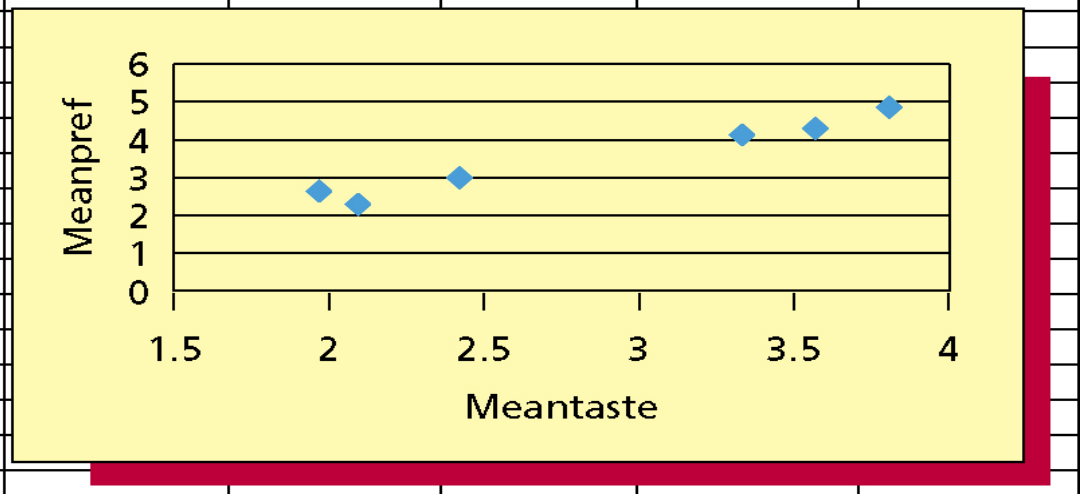
- The Scatter Diagram:
  - one variable is measured on the vertical axis and the other variable is measured on the horizontal axis

Scatter Plots

Visualize the data to see patterns, especially “trends”

Restaurant Ratings: Mean Preference vs. Mean Taste

	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee’s	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald’s	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy’s	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



Introduction to Regression Analysis

- **Regression analysis** is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable:

the variable we wish to predict or explain

Independent variable:

the variable used to explain the dependent variable

Regression Analysis

- The term **"regression"** was coined by Francis Galton in the nineteenth century to describe a biological phenomenon.
- The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).
- For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context.

Regression Analysis

- **Statistics regression analysis** includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

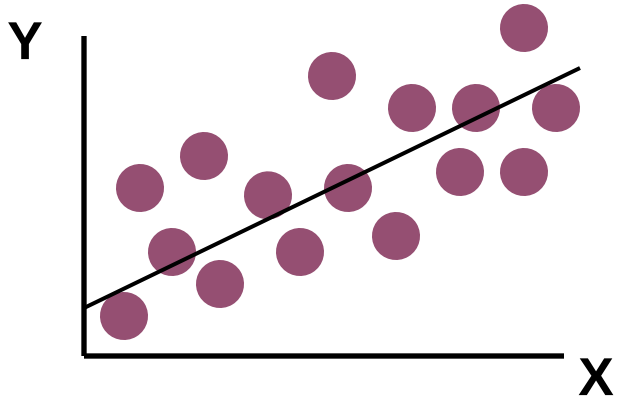
- **Regression analysis** is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

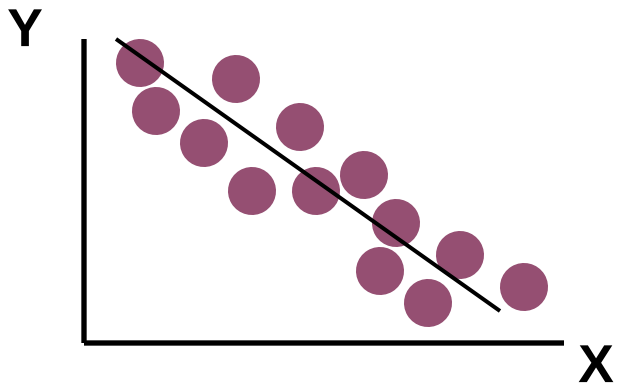
Simple Linear Regression Model

- Only **one independent variable**, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X

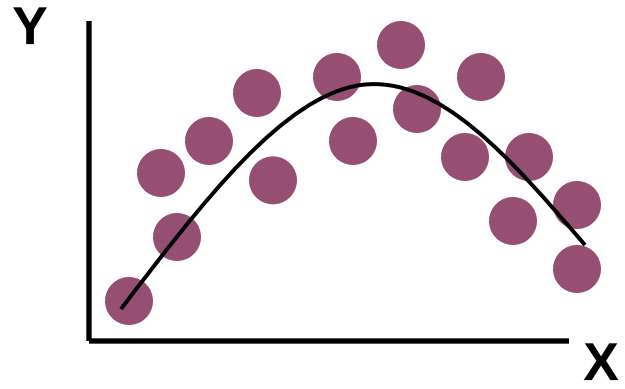
Types of Relationships

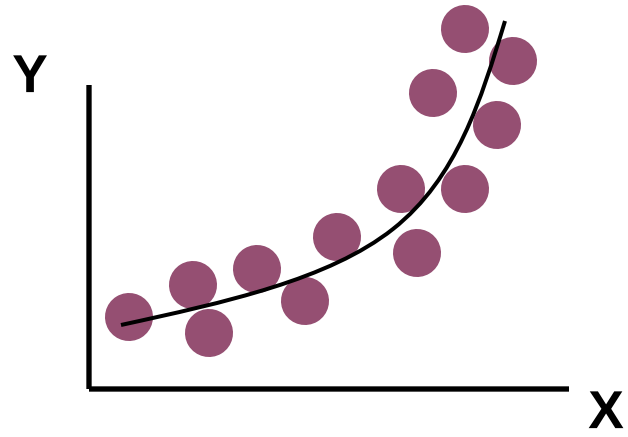
Linear relationships





Curvilinear relationships

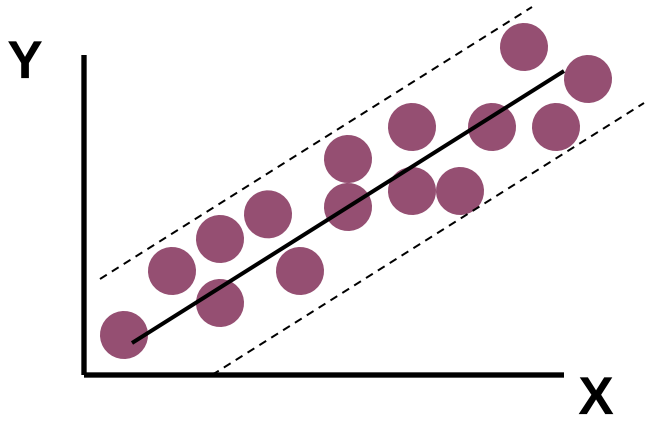


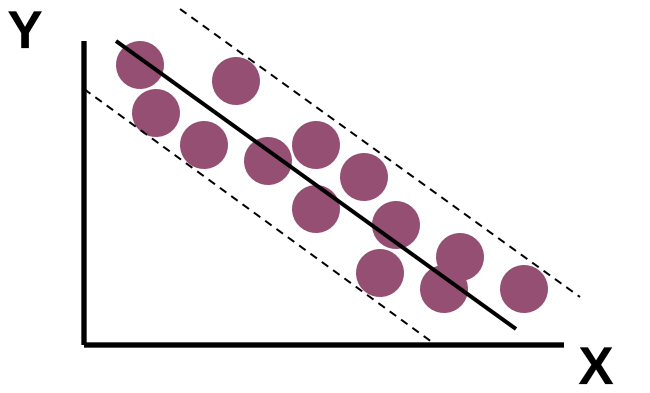


Types of Relationships

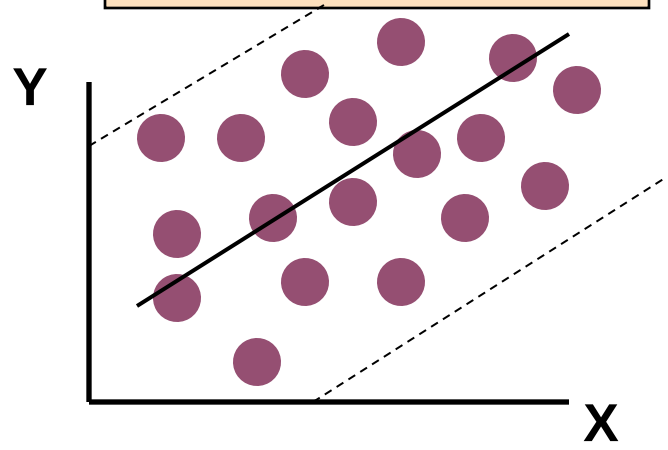
(continued)

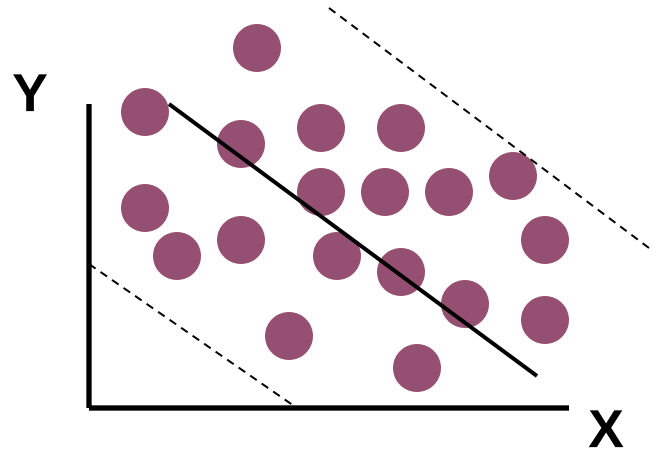
Strong relationships





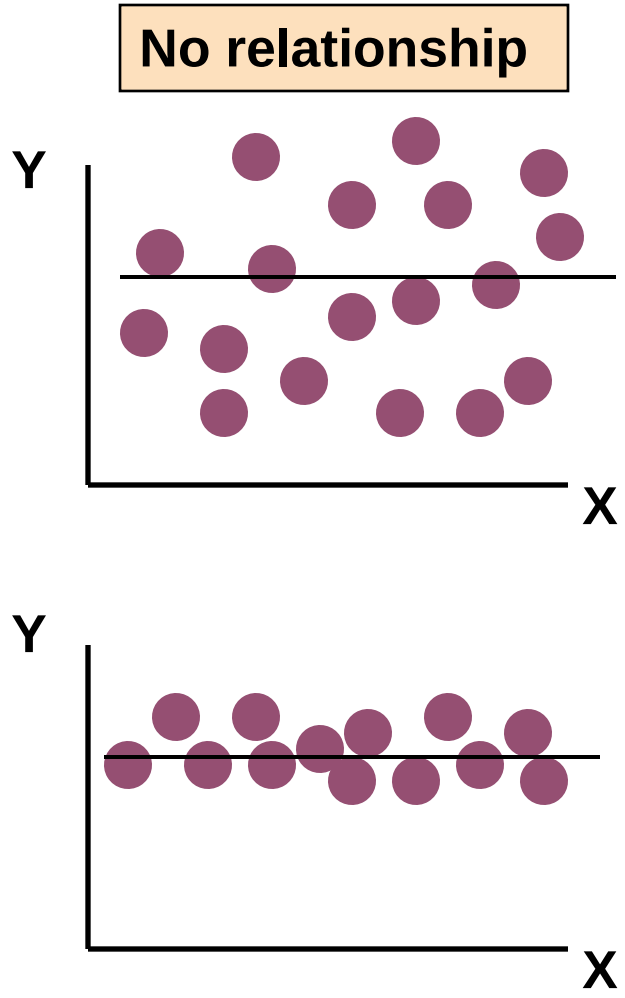
Weak relationships



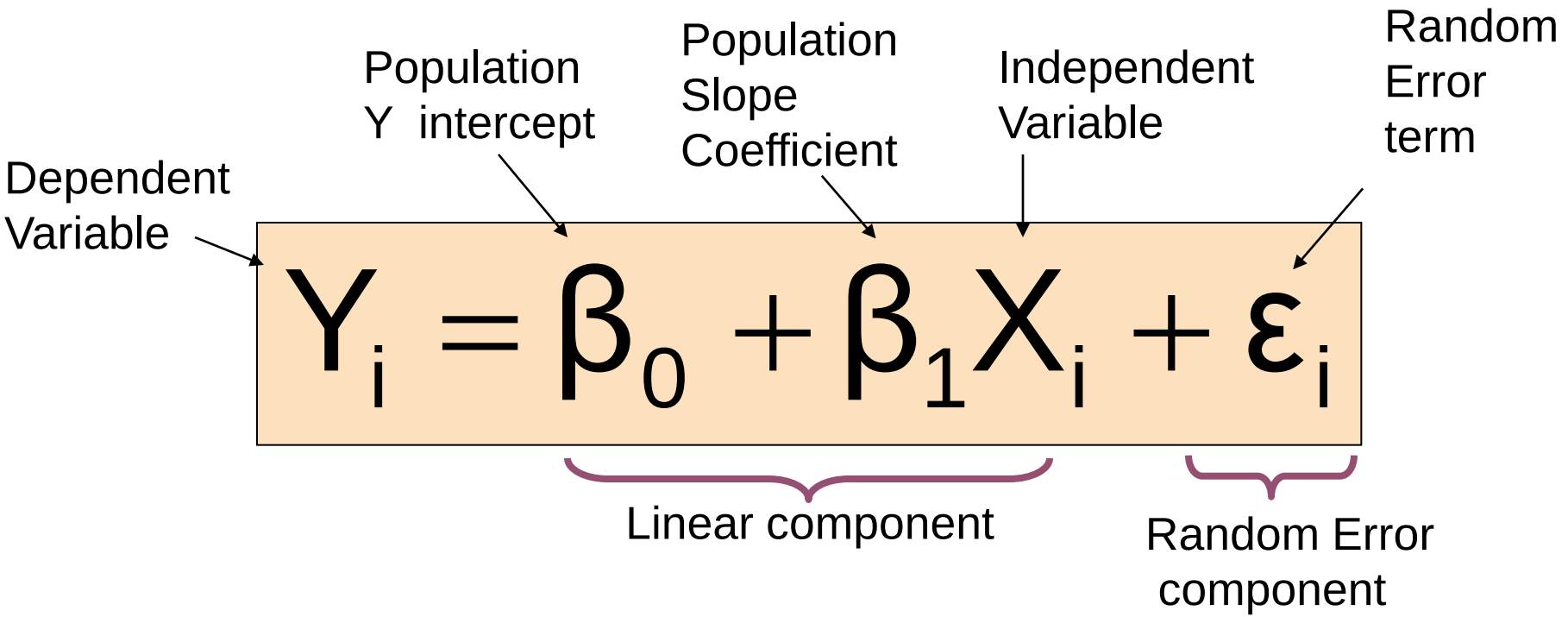


Types of Relationships

(continued)



Simple Linear Regression Model

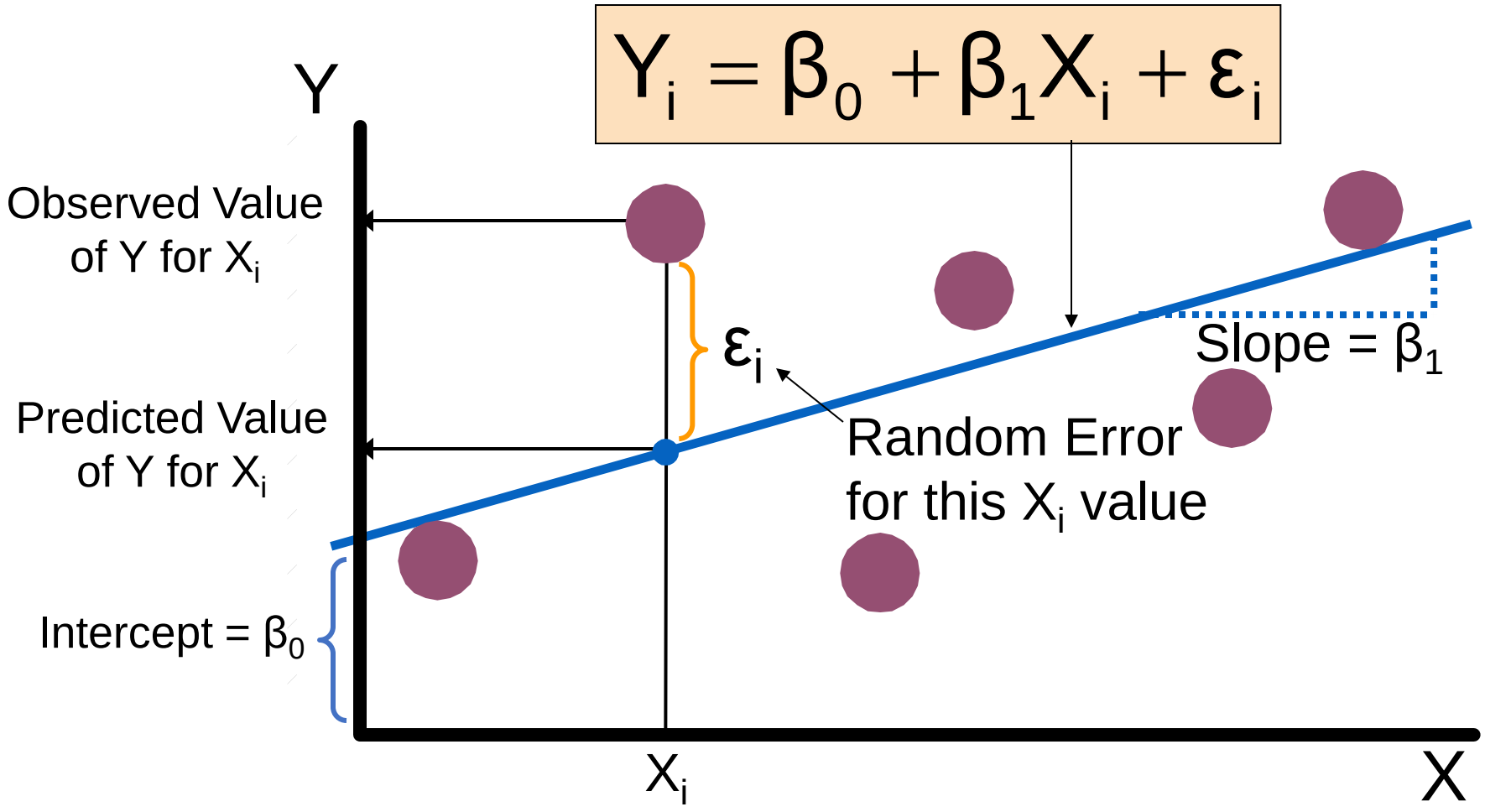


Model Assumptions

1. **Mean of Zero**: At any given value of x, the population of potential error term values has a **mean equal to zero**
2. **Constant Variance Assumption**: At any value of x, the population of potential error term values has a **variance** that does not depend on the value of x
3. **Normality Assumption**: At any given value of x, the population of potential error term values has a **normal distribution**
4. **Independence Assumption**: Any one value of the error term ε is **statistically independent** of any other value of ε

Simple Linear Regression Model

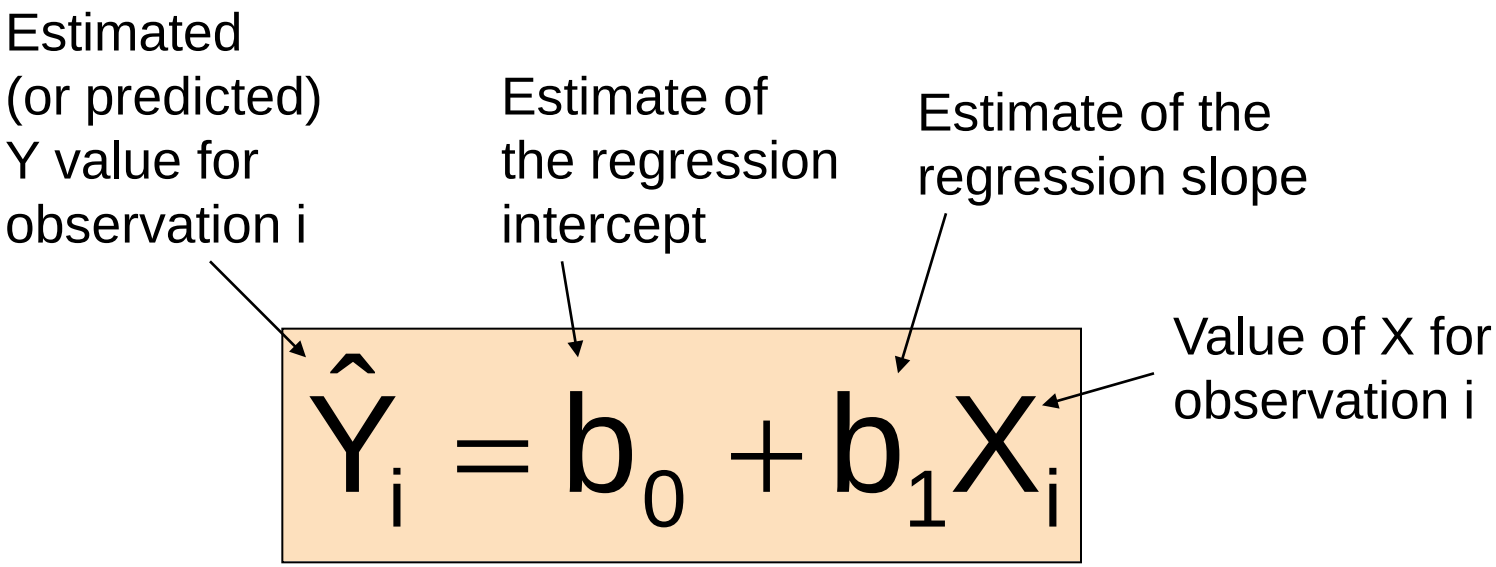
(continued)



- The **dependent** (or response) variable is the variable we wish to understand or predict
- The **independent** (or predictor) variable is the variable we will use to understand or predict the dependent variable
- **Regression analysis** is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables
- The objective is to build a regression model that can describe, predict and control the dependent variable based on the independent variable

Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an **estimate** of the population regression line



Simple Linear Regression Equation

Point Estimation and Point Prediction in Simple Linear Regression

Let b_0 and b_1 be the least squares point estimates of the y-intercept β_0 and the slope β_1 in the simple linear regression model, and suppose that x_0, a specified value of the independent variable x, is inside the experimental region. Then

$$\hat{y} = b_0 + b_1 x_0$$

1 is the **point estimate** of the **mean value of the dependent variable** when the value of the independent variable is x_0.

2 is the **point prediction** of an **individual value of the dependent variable** when the value of the independent variable is x_0. Here we predict the error term to be 0.

Regression Terms

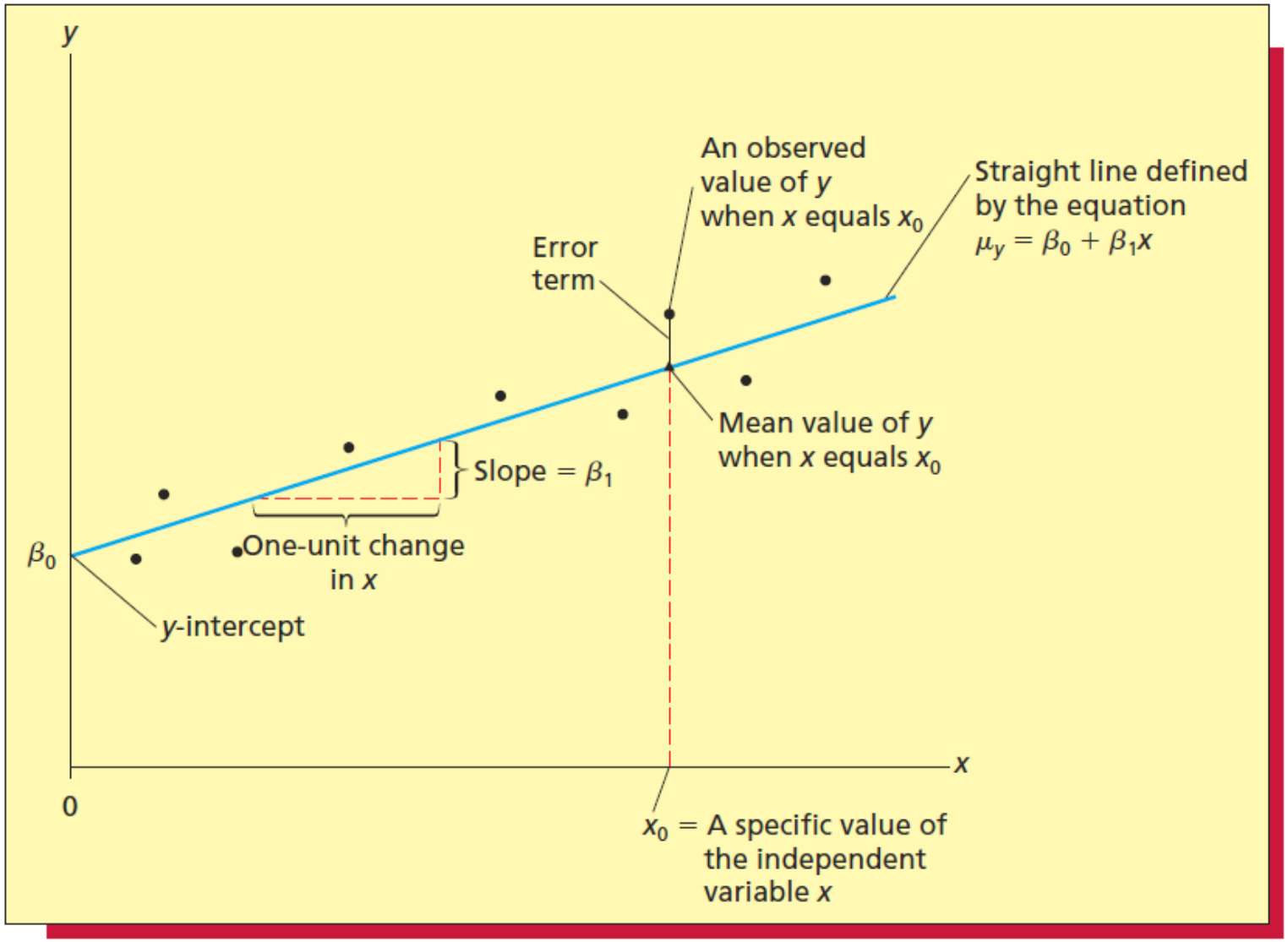
- β_0 and β_1 are called regression parameters
  - β_0 is the y-intercept
  - β_1 is the slope
- We **do not know** the true values of these parameters
- So, we must use sample data to **estimate** them
  - b_0 is the estimate of β_0
  - b_1 is the estimate of β_1

Least Squares Method
(最小二乘方法)

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that **minimize the sum of the squared differences** between Y and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

The Simple Linear Regression Model Illustrated



The Least Squares Point Estimates

Estimation/prediction equation

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

Least squares point estimate of the slope β_1

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Least squares point estimate of the y-intercept β_0

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{where} \quad \bar{y} = \frac{\sum y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum x_i}{n}$$

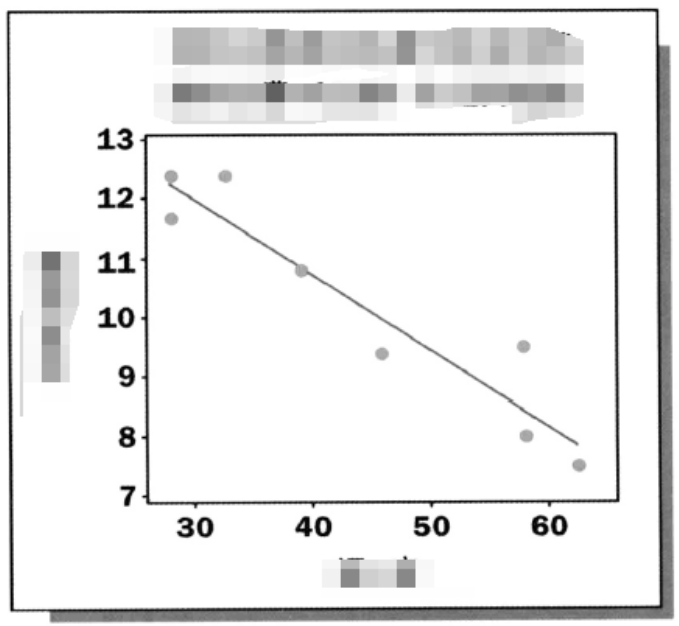
Example 13.3 The Fuel Consumption Case

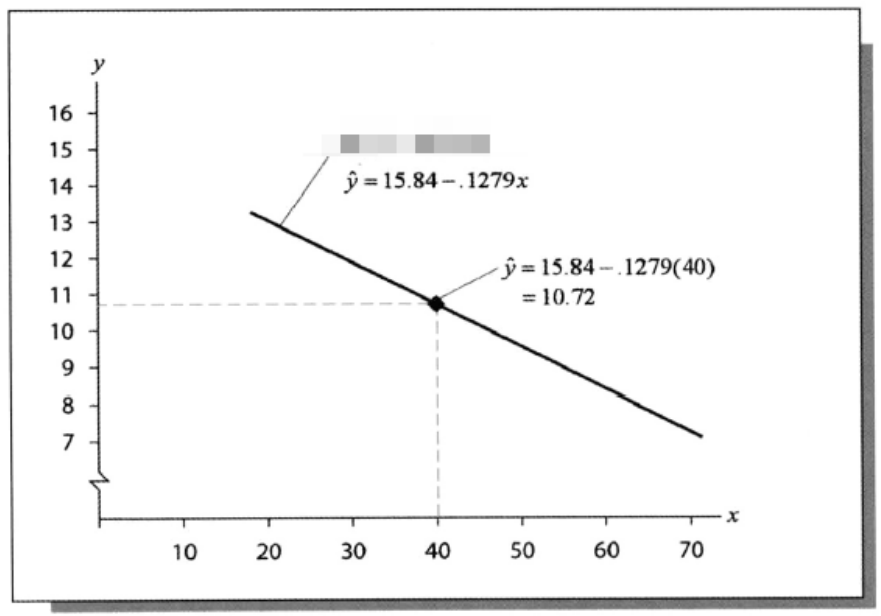
Temp	FuelCons
28.0	12.4
28.0	11.7
32.5	12.4
39.0	10.8
45.9	9.4
57.8	9.5
58.1	8.0
62.5	7.5

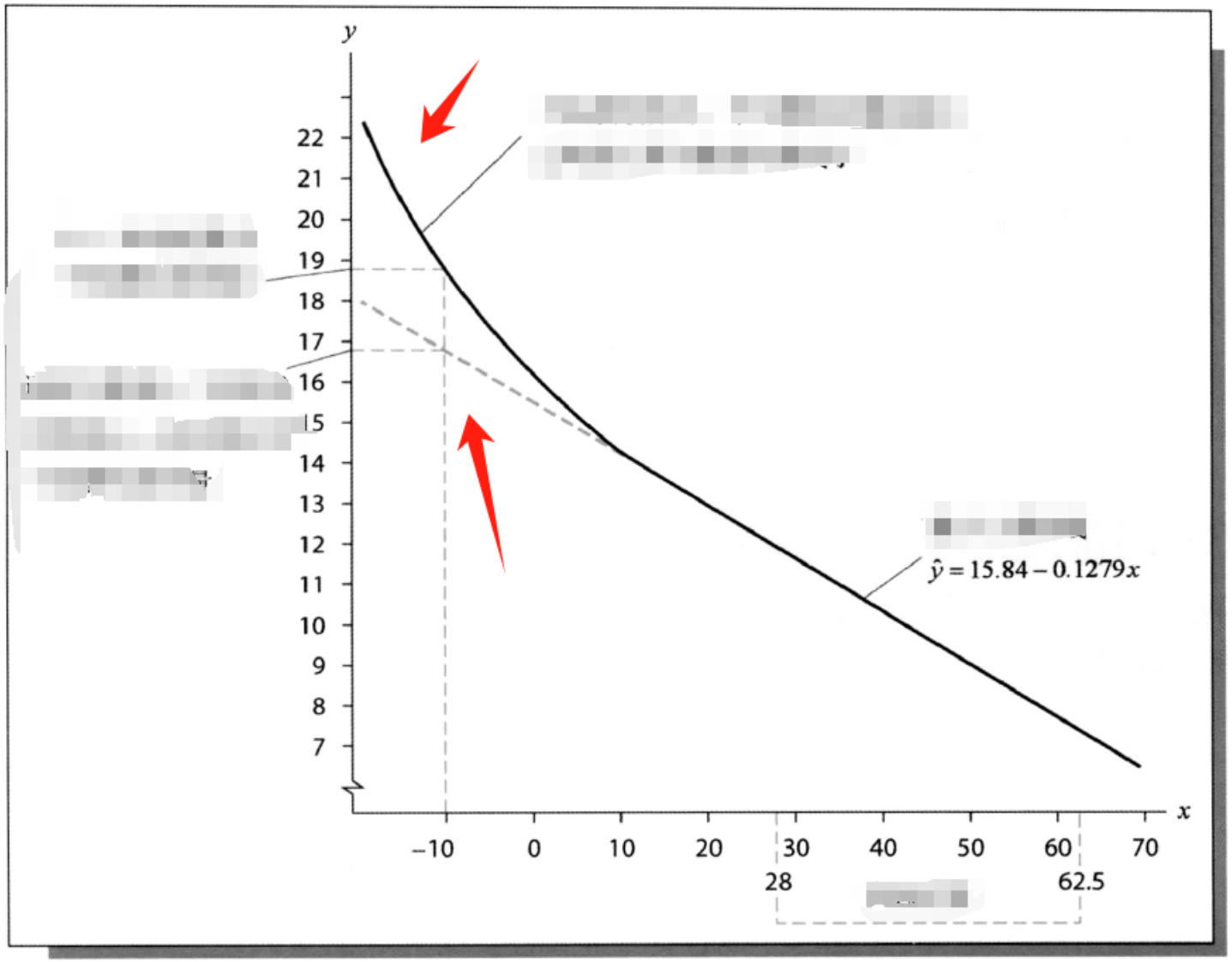
y_i	x_i	x^2_i	$x_i y_i$
12.4	28.0	$(28.0)^2 = 784$	$28.0 \times 12.4 = 347.2$
11.7	28.0	$(28.0)^2 = 784$	$28.0 \times 11.7 = 327.6$
12.4	32.5	$(32.5)^2 = 1\,056.25$	$32.5 \times 12.4 = 403$
10.8	39.0	$(39.0)^2 = 1\,521$	$39.0 \times 10.8 = 421.2$
9.4	45.9	$(45.9)^2 = 2\,106.81$	$45.9 \times 9.4 = 431.46$
9.5	57.8	$(57.8)^2 = 3\,340.84$	$57.8 \times 9.5 = 549.1$
8.0	58.1	$(58.1)^2 = 3\,375.61$	$58.1 \times 8.0 = 464.8$
7.5	62.5	$(62.5)^2 = 3\,906.25$	$62.5 \times 7.5 = 468.75$
$\sum y_i = 81.7$	$\sum x_i = 351.8$	$\sum x_i^2 = 16\,874.76$	$\sum x_i y_i = 3\,413.11$

y_i	x_i	$\hat{y}_i = 15.84 - 0.1279x_i$	$y_i - \hat{y}_i$ = resdual
12.4	28.0	$15.84 - 0.1279 \times 28.0 = 12.2588$	$12.4 - 12.2588 = 0.1412$
11.7	28.0	$15.84 - 0.1279 \times 28.0 = 12.2588$	$11.7 - 12.2588 = -0.5588$
12.4	32.5	$15.84 - 0.1279 \times 32.5 = 11.68325$	$12.4 - 11.68325 = 0.71675$
10.8	39.0	$15.84 - 0.1279 \times 39.0 = 10.8519$	$10.8 - 10.8519 = -0.0519$
9.4	45.9	$15.84 - 0.1279 \times 45.9 = 9.96939$	$9.4 - 9.96939 = -0.56939$
9.5	57.8	$15.84 - 0.1279 \times 57.8 = 8.44738$	$9.5 - 8.44738 = 1.05262$
8.0	58.1	$15.84 - 0.1279 \times 58.1 = 8.40901$	$8.0 - 8.40901 = -0.40901$
7.5	62.5	$15.84 - 0.1279 \times 62.5 = 7.84625$	$7.5 - 7.84625 = -0.34625$

$$SSE = \sum (y_i - \hat{y}_i)^2 = (0.1412)^2 + (-0.5588)^2 + \cdots + (-0.34625)^2 = 2.568$$






y
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
x
−10
0
10
20
30
40
50
60
70
28
62.5
$\hat{y} = 15.84 - 0.1279x$

Example 13.4 The QHIC Case

A	B
Value	**Upkeep**
237.00	1412.08
153.08	797.20
184.86	872.48
222.06	1003.42
160.68	852.90
99.68	288.48
229.04	1288.46
101.78	423.08
257.86	1351.74
96.28	378.04
171.00	918.08
231.02	1627.24
228.32	1204.76
205.90	857.04
185.72	775.00
168.78	869.26
247.06	1396.00
155.54	711.50

Example

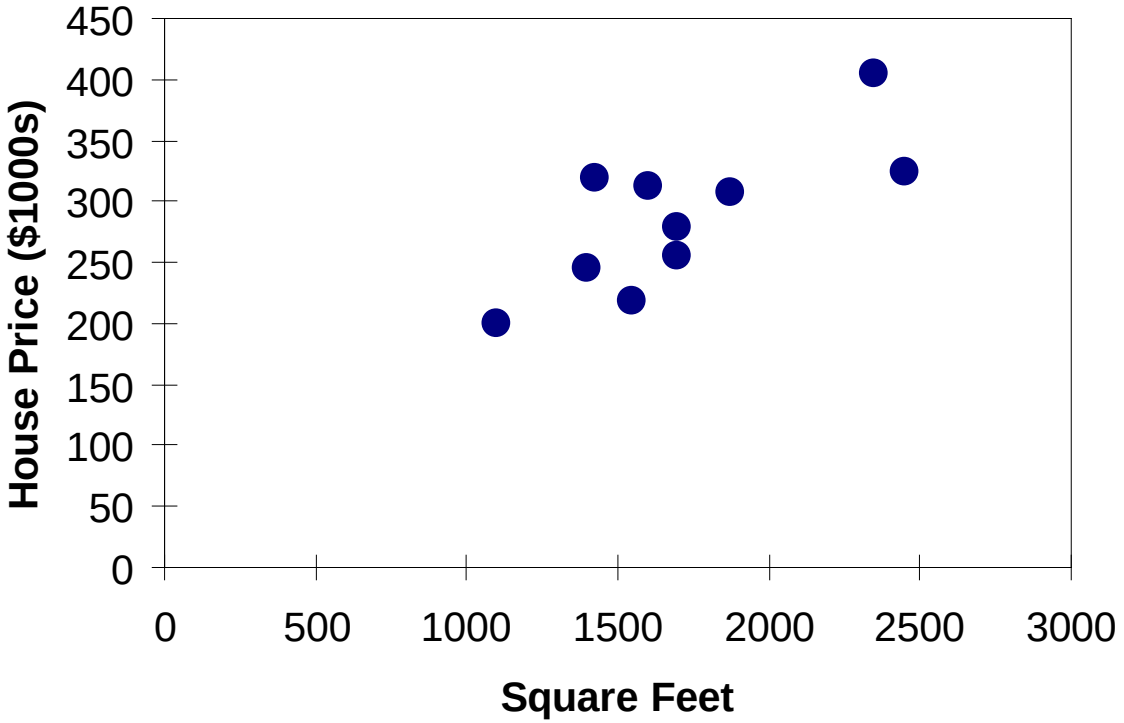
The House Price Case

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

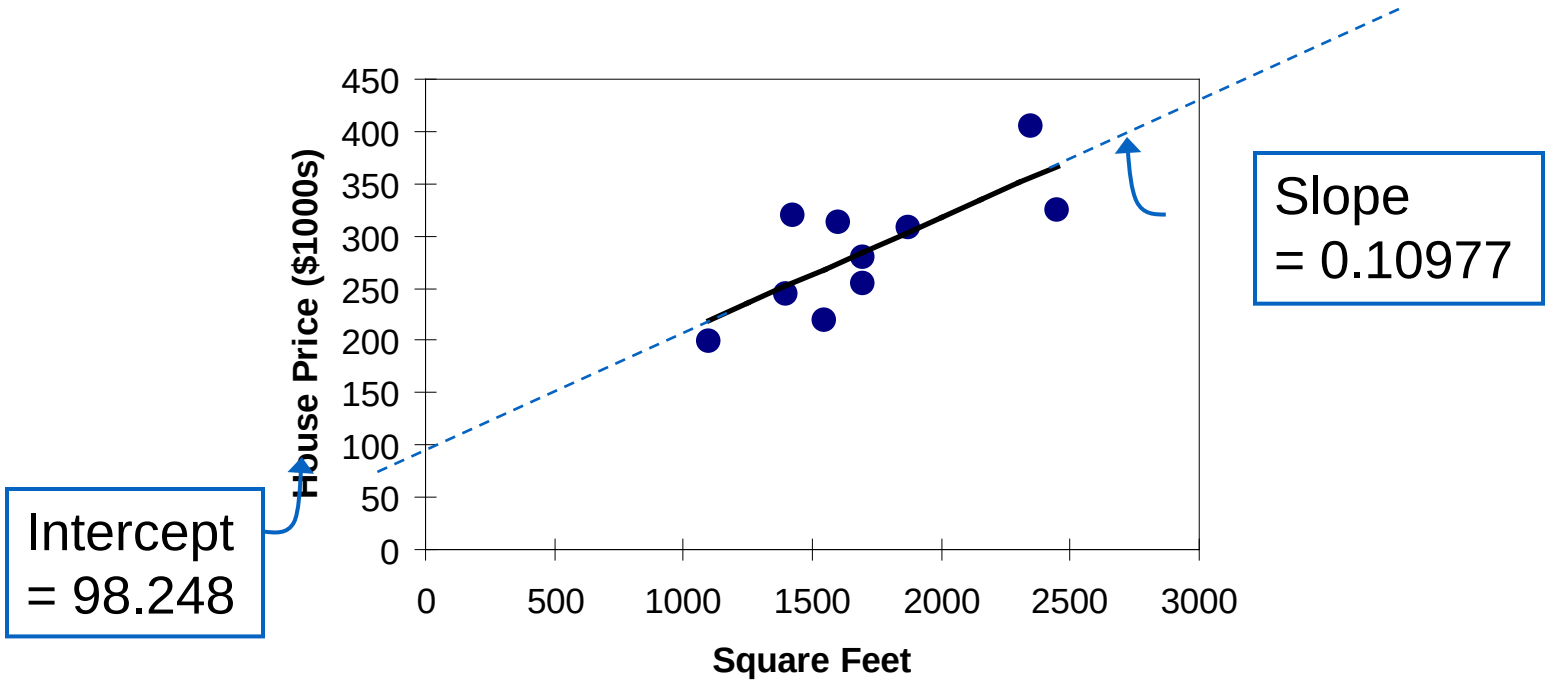
Graphical Presentation

- House price model: scatter plot



Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
  - Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $b_1 = .10977$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Predictions using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

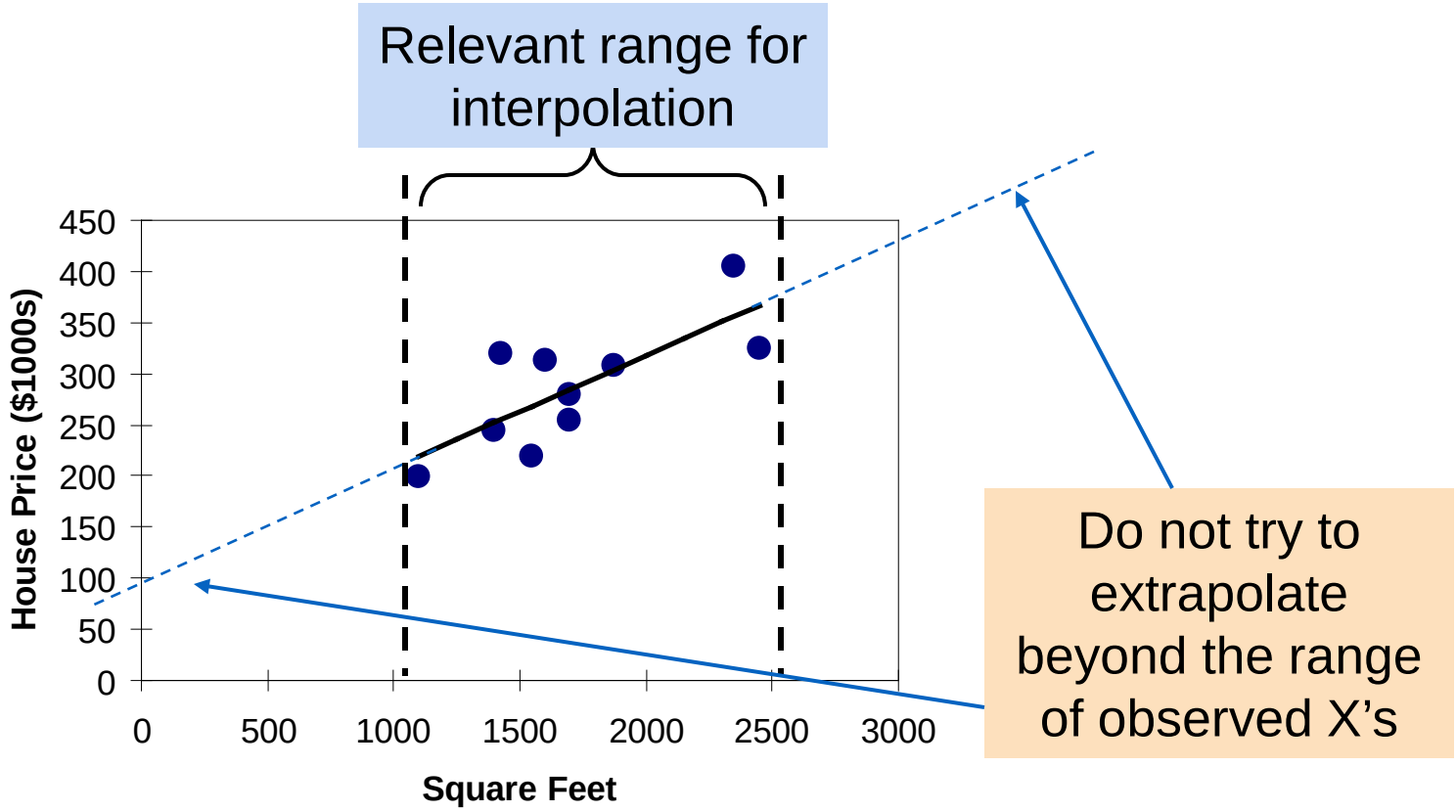
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data



Example: The Tasty Sub Shop Case

EXAMPLE The Tasty Sub Shop Case: Predicting Yearly Revenue for a Potential Restaurant Site

Part 1: Purchasing a Tasty Sub Shop franchise The Tasty Sub Shop is a restaurant chain that sells franchises to business entrepreneurs. Like Quizn does not construct a standard, recognizable building to hous entrepreneur wishing to purchase a Tasty Sub franchise finc suitable geographical location and suitable store space to r the site, an architect and a contractor are hired to remodel th the Tasty Sub Shop restaurant. Franchise regulations allow entrepreneurs understand the factors that affect restauran guidance in evaluating potential restaurant sites. However, from overpredicting profits and thus misleading potential make each individual entrepreneur responsible for predicting the profits of his or her potential restaurant sites.

In this case study we consider a business entrepreneur who has found several potential sites for a Tasty Sub Shop restaurant. Similar to most existing Tasty Sub restaurant sites, each of the entrepreneur's sites is a store rental space located in an outdoor shopping area that is close to one or more residential areas. For a Tasty Sub restaurant built on such a site, yearly revenue is known to partially depend on (1) the number of residents living near the site and (2) the amount of business and shopping near the site. Referring to the number of residents living near a site as *population size* and to the yearly revenue for a Tasty Sub restaurant built on the site as *yearly revenue,* the entrepreneur will—in this chapter—try to predict the **dependent (response) variable** yearly revenue (y) on the basis of the **independent (predictor) variable** population size (x). (In the next chapter the entrepreneur will also use the amount of business and shopping near a site to help predict yearly revenue.) To predict yearly revenue on the basis of population size, the entrepreneur randomly selects 10 existing Tasty Sub restaurants that are built on sites similar to the sites that the entrepreneur is considering. The entrepreneur then asks the owner of each existing restaurant what the restaurant's revenue y was last year and estimates—with the help of the owner and published demographic information—the number of residents, or population size x, living near the site. The values of y (measured in thousands of dollars) and x (measured in thousands of residents) that are obtained are given in Table 14.1. In Figure 14.1 we give an Excel output of a scatter plot of y versus x. This plot shows (1) a tendency for the yearly revenues to increase in a straight-line fashion as the population sizes increase and (2) a scattering of points around the straight line. A **regression model** describing the relationship between y and x must represent these two characteristics. We now develop such a model.

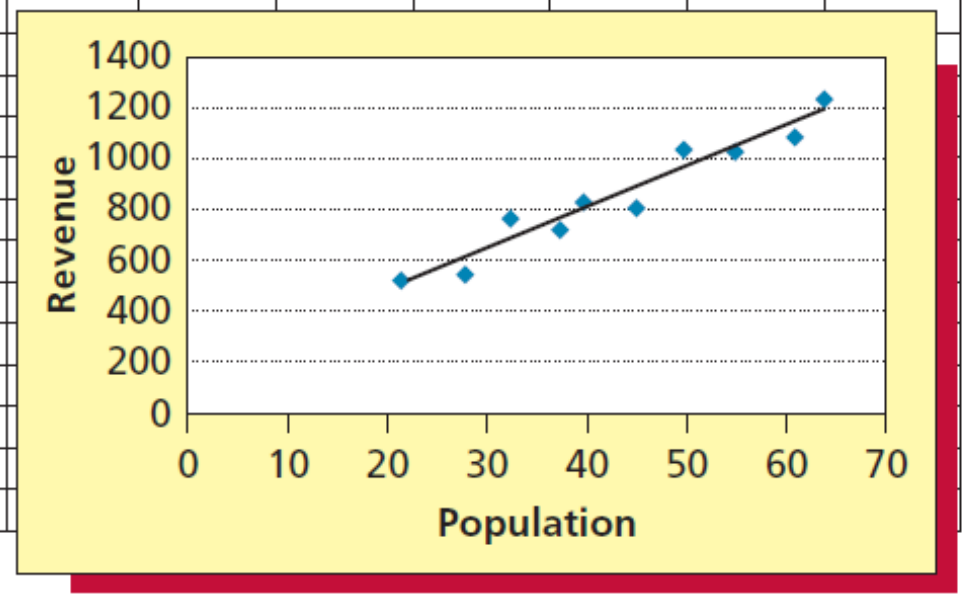
The Tasty Sub Shop Revenue Data

DS TastySub1

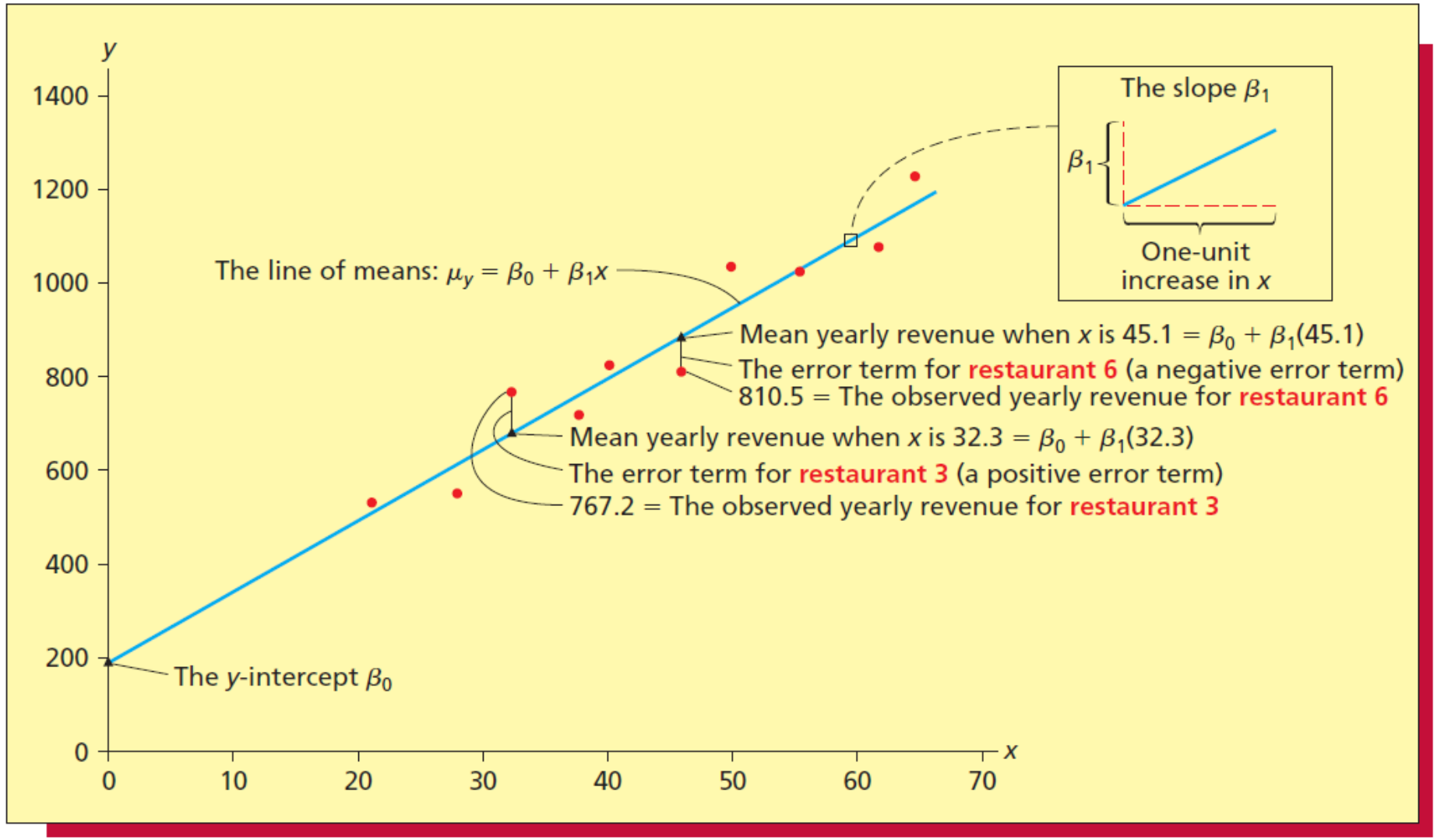
Restaurant	Population Size, x (Thousands of Residents)	Yearly Revenue, y (Thousands of Dollars)
1	20.8	527.1
2	27.5	548.7
3	32.3	767.2
4	37.2	722.9
5	39.6	826.3
6	45.1	810.5
7	49.9	1040.7
8	55.4	1033.6
9	61.7	1090.3
10	64.6	1235.8

Excel Output of a Scatter Plot of y versus x

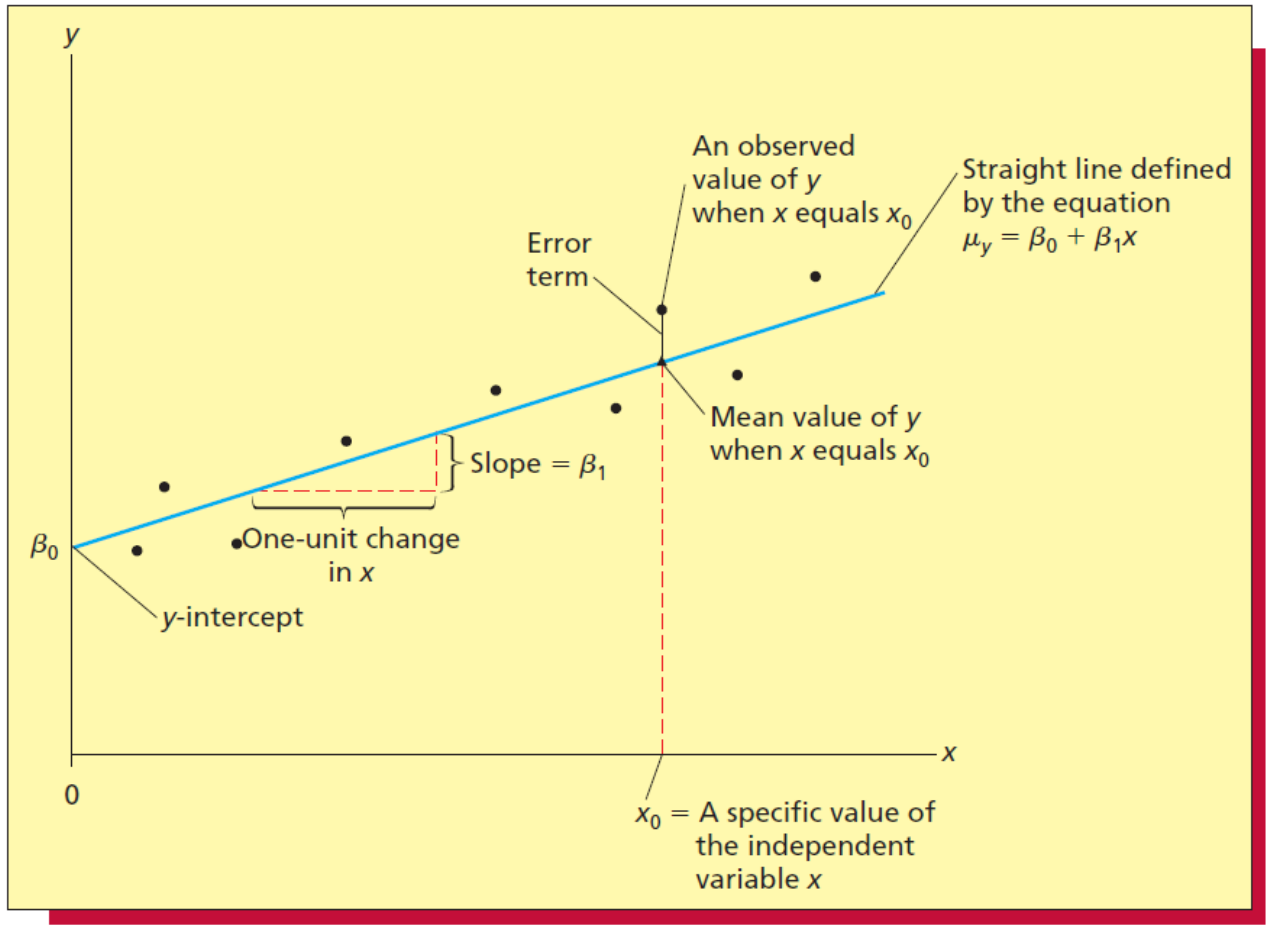
	A	B	C	D	E	F	G	H	I
1	Population	Revenue							
2	20.8	527.1							
3	27.5	548.7							
4	32.3	767.2							
5	37.2	722.9							
6	39.6	826.3							
7	45.1	810.5							
8	49.9	1040.7							
9	55.4	1033.6							
10	61.7	1090.3							
11	64.6	1235.8							
12									
13									



The Simple Linear Regression Model Relating Yearly Revenue (y) to Population (x)

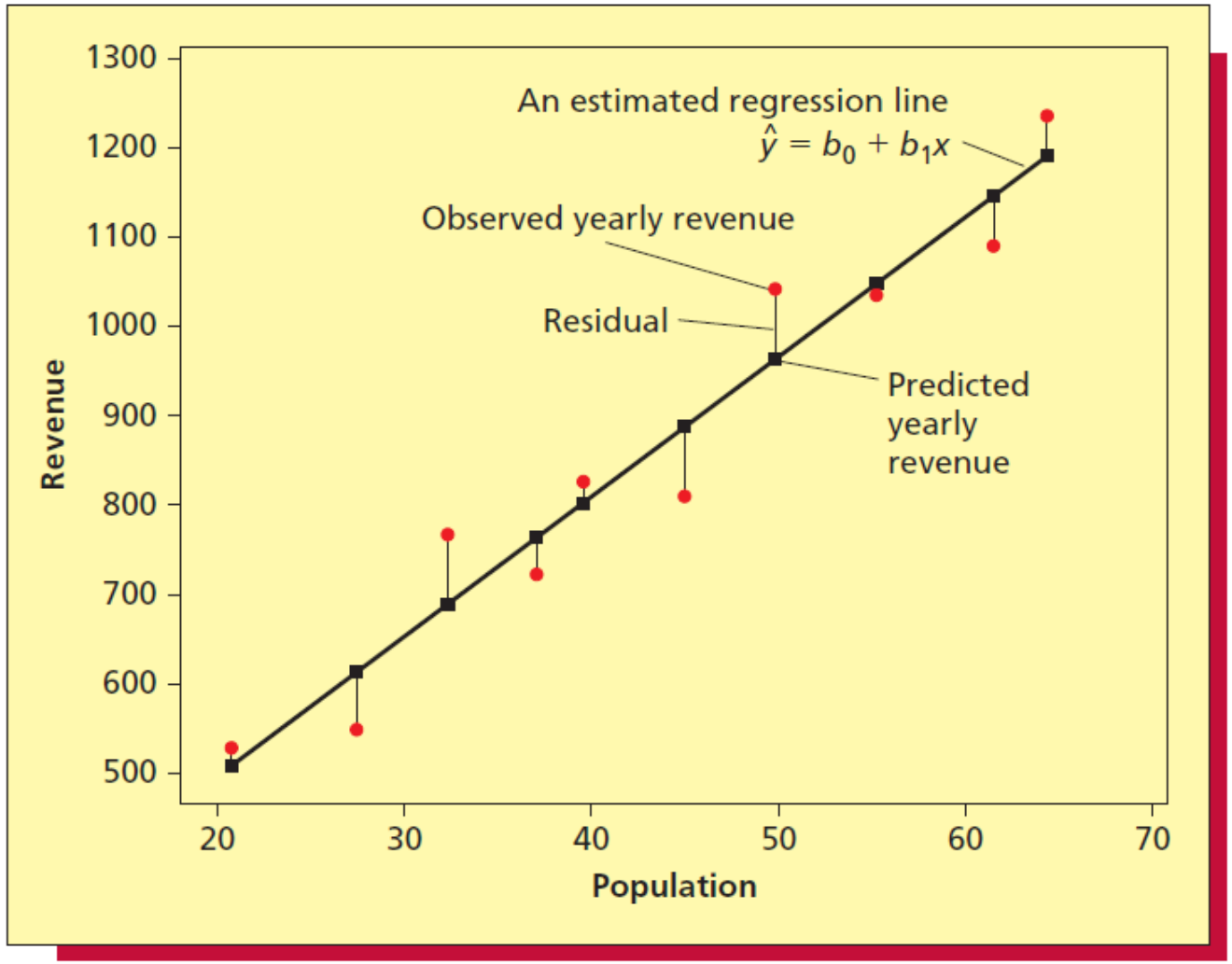


The Simple Linear Regression Model (Here the Slope β_1 Is Positive)



With the Tasty Sub Shop example as background, we are ready to define the **simple linear regression model relating the dependent variable *y* to the independent variable *x*.** We suppose that we have gathered n observations—each observation consists of an observed value of x and its corresponding value of y. Then:

FIGURE 1 An Estimated Regression Line Drawn through the Tasty Sub Shop Revenue Data



y_i	x_i	x_i^2	$x_i y_i$
527.1	20.8	$(20.8)^2 = 432.64$	$(20.8)(527.1) = 10963.68$
548.7	27.5	$(27.5)^2 = 756.25$	$(27.5)(548.7) = 15089.25$
767.2	32.3	$(32.3)^2 = 1{,}043.29$	$(32.3)(767.2) = 24780.56$
722.9	37.2	$(37.2)^2 = 1{,}383.84$	$(37.2)(722.9) = 26891.88$
826.3	39.6	$(39.6)^2 = 1{,}568.16$	$(39.6)(826.3) = 32721.48$
810.5	45.1	$(45.1)^2 = 2{,}034.01$	$(45.1)(810.5) = 36553.55$
1040.7	49.9	$(49.9)^2 = 2{,}490.01$	$(49.9)(1040.7) = 51930.93$
1033.6	55.4	$(55.4)^2 = 3{,}069.16$	$(55.4)(1033.6) = 57261.44$
1090.3	61.7	$(61.7)^2 = 3{,}806.89$	$(61.7)(1090.3) = 67271.51$
1235.8	64.6	$(64.6)^2 = 4{,}173.16$	$(64.6)(1235.8) = 79832.68$
$\sum y_i = 8603.1$	$\sum x_i = 434.1$	$\sum x_i^2 = 20{,}757.41$	$\sum x_i y_i = 403{,}296.96$

- From last slide,
  - $\Sigma y_i = 8{,}603.1$
  - $\Sigma x_i = 434.1$
  - $\Sigma x^2_i = 20{,}757.41$
  - $\Sigma x_i y_i = 403{,}296.96$
- Once we have these values, we no longer need the raw data
- Calculation of b_0 and b_1 uses these totals

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$= 403{,}296.96 - \frac{(434.1)(8{,}603.1)}{10} = 29{,}836.389$$

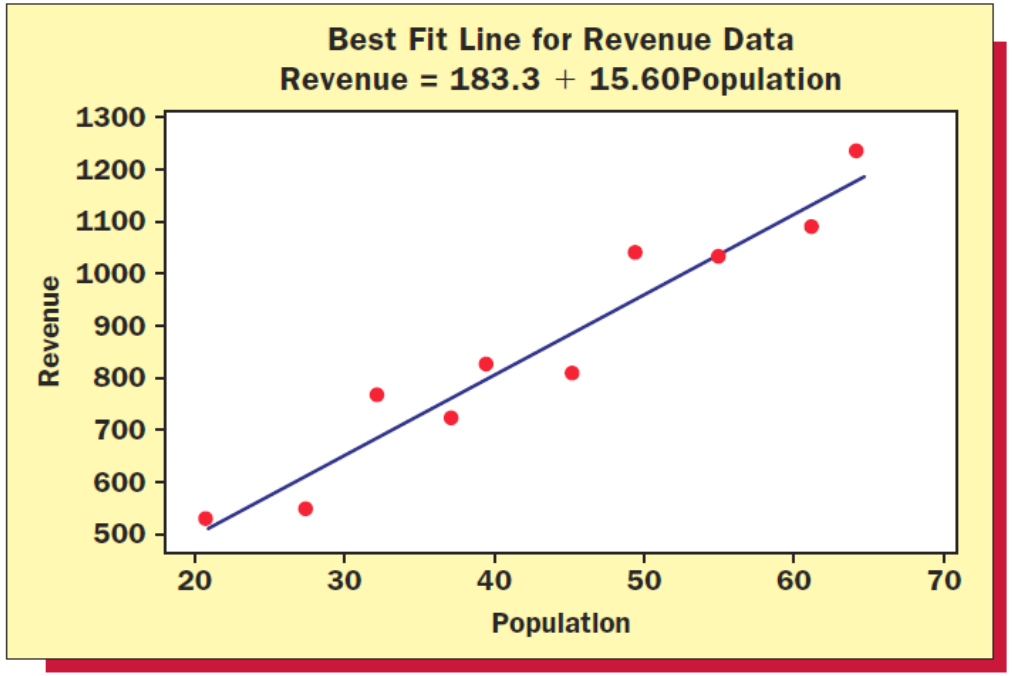
$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$= 120{,}757.41 - \frac{(434.1)^2}{10} = 1{,}913.129$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{29{,}836.389}{1{,}913.129} = 15.596$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{8{,}603.1}{10} = 860.31$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{434.1}{10} = 43.41$$

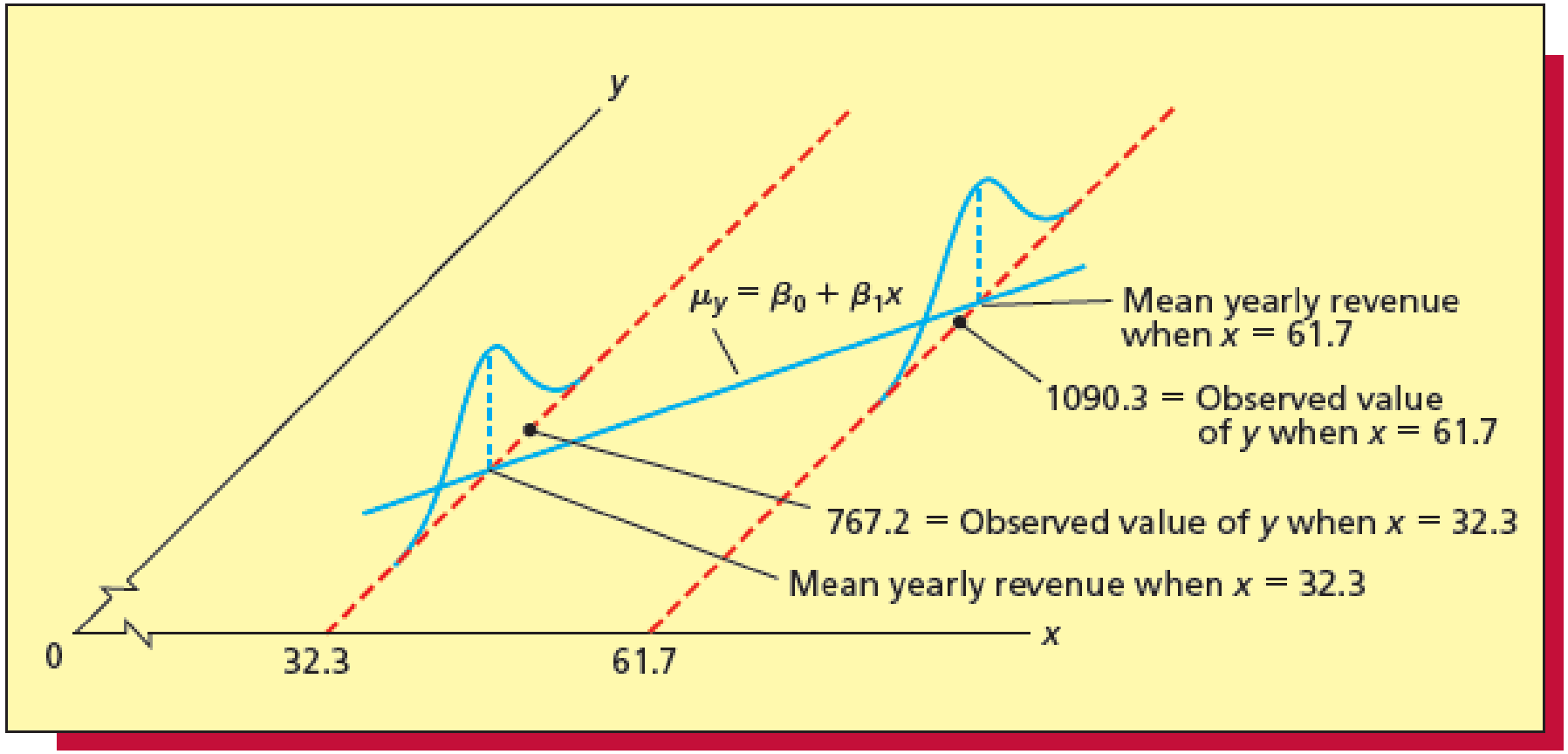


$$b_0 = \bar{y} - b_1\bar{x}$$
$$= 860.31 - (15.596)(43.41)$$
$$= 183.31$$

- Prediction (x = 20.8)
- $\hat{y} = b_0 + b_1x = 183.31 + (15.59)(20.8)$
- $\hat{y} = 507.69$
- Residual is 527.1 – 507.69 = 19.41

13.2 Model Assumptions and the Standard Error

1. **Mean of Zero**: At any given value of x, the population of potential error term values has a mean equal to zero
2. **Constant Variance Assumption**: At any value of x, the population of potential error term values has a variance that does not depend on the value of x
3. **Normality Assumption**: At any given value of x, the population of potential error term values has a normal distribution
4. **Independence Assumption**: Any one value of the error term ε is statistically independent of any other value of ε


y
$\mu_y = \beta_0 + \beta_1 x$
Mean yearly revenue when $x = 61.7$
1090.3 = Observed value of y when $x = 61.7$
767.2 = Observed value of y when $x = 32.3$
Mean yearly revenue when $x = 32.3$
0
32.3
61.7
x

The Mean Square Error and the Standard Error

Sum of squared errors

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Mean square error : point estimate of the residual variance σ^2

$$s^2 = MSE = \frac{SSE}{n\text{-}2}$$

Standard error : point estimate of residual standard deviation σ

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n\text{-}2}}$$

Example 13.5 The Fuel Consumption Case

- Point estimate of σ^2

- Point estimate of σ

13.3 Testing the Significance of the Slope and y-Intercept

- A regression model is not likely to be useful unless there is a significant relationship between x and y
- To test significance, we use the null hypothesis:

  $H_0: \beta_1 = 0$

- Versus the alternative hypothesis:

  $H_a: \beta_1 \neq 0$

Testing the Significance of the Slope

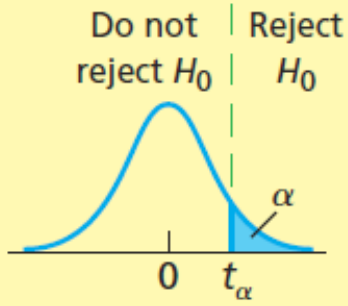
Null Hypothesis $H_0: \beta_1 = 0$

Test Statistic $t = \dfrac{b_1}{s_{b_1}}$ where $s_{b_1} = \dfrac{s}{\sqrt{SS_{xx}}}$

Assumptions The regression assumptions

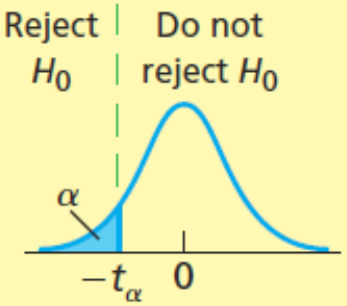
Critical Value Rule

$H_a: \beta_1 > 0$



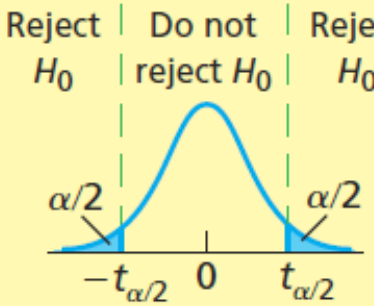
Reject H_0 if $t > t_\alpha$

$H_a: \beta_1 < 0$



Reject H_0 if $t < -t_\alpha$

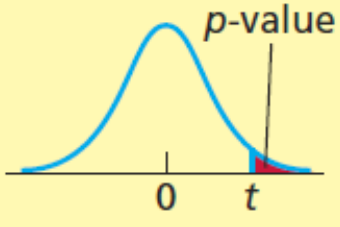
$H_a: \beta_1 \neq 0$



Reject H_0 if $|t| > t_{\alpha/2}$—that is, $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

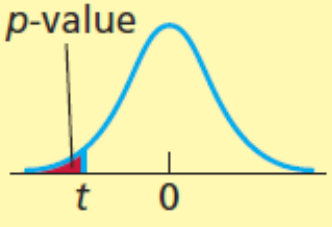
p-Value (Reject H_0 if p-Value $< \alpha$)

$H_a: \beta_1 > 0$



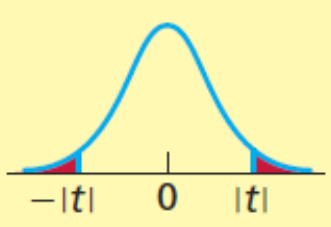
p-value = area to the right of t

$H_a: \beta_1 < 0$



p-value = area to the left of t

$H_a: \beta_1 \neq 0$



p-value = twice the area to the right of $|t|$

Here $t_{\alpha/2}$, t_α, and all p-values are based on $n - 2$ degrees of freedom. **If we can reject $H_0: \beta_1 = 0$ at a given value of α, then we conclude that the slope (or, equivalently, the regression relationship) is significant at the α level.**

We usually use the two-sided alternative H_a: $\beta_1 \neq 0$ for this test of significance. However, sometimes a one-sided alternative is appropriate. For example, in the Tasty Sub Shop problem we can say that if the slope β_1 is not 0, then it must be positive. A positive β_1 would say that mean yearly revenue increases as the population size x increases. Because of this, it would be appropriate to decide that x is significantly related to y if we can reject H_0: $\beta_1 = 0$ in favor of the one-sided alternative H_a: $\beta_1 > 0$. Although this test would be slightly more effective than the usual two tailed test, there is little practical difference between using the one tailed or two tailed test. Furthermore, computer packages (such as Excel and MINITAB) present results for the two tailed test. For these reasons we will emphasize the two tailed test in future discussions.

It should also be noted that

1. **If we can decide that the slope is significant at the .05 significance level,** then we have concluded that x is significantly related to y by using a test that allows only a .05 probability of concluding that x is significantly related to y when it is not. **This is usually regarded as strong evidence that the regression relationship is significant.**
2. **If we can decide that the slope is significant at the .01 significance level, this is usually regarded as very strong evidence that the regression relationship is significant.**
3. **The smaller the significance level α at which H_0 can be rejected, the stronger is the evidence that the regression relationship is significant.**

EXAMPLE The Tasty Sub Shop Case: Testing the Significance of the Slope C

Again consider the Tasty Sub Shop revenue model. For this model $SS_{xx} = 1913.129$, $b_1 = 15.596$, and $s = 61.7052$ [see Examples 14.2 (page 493) and 14.3 (page 502)]. Therefore,

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}} = \frac{61.7052}{\sqrt{1913.129}} = 1.411$$

and

$$t = \frac{b_1}{s_{b_1}} = \frac{15.596}{1.411} = 11.05$$

A Confidence Interval for the Slope

If the regression assumptions hold, a **100(1 − α) percent confidence interval for the true slope β_1** is $[b_1 \pm t_{\alpha/2} s_{b_1}]$. Here $t_{\alpha/2}$ is based on $n - 2$ degrees of freedom.

EXAMPLE The Tasty Sub Shop Case: A Confidence Interval for the Slope

The Excel and MINITAB outputs in Figure 14.8 tell us that $b_1 = 15.596$ and $s_{b_1} = 1.411$. Thus, for instance, because $t_{.025}$ based on $n - 2 = 10 - 2 = 8$ degrees of freedom equals 2.306, a 95 percent confidence interval for β_1 is

$$[b_1 \pm t_{.025} s_{b_1}] = [15.596 \pm 2.306(1.411)]$$
$$= [12.342, 18.849]$$

(where we have used more decimal place accuracy than shown to obtain the final result). This interval says we are 95 percent confident that, if the population size increases by one thousand residents, then mean yearly revenue will increase by at least \$12,342 and by at most \$18,849. Also, because the 95 percent confidence interval for β_1 does not contain 0, we can reject H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ at level of significance .05. Note that the 95 percent confidence interval for β_1 is given on the Excel output but not on the MINITAB output (see Figure 14.8).

13.4 Confidence and Prediction Intervals

- The point on the regression line corresponding to a particular value of x_0 of the independent variable x is $\hat{y} = b_0 + b_1x_0$
- It is unlikely that this value will equal the mean value of y when x equals x_0
- Therefore, we need to place bounds on how far the predicted value might be from the actual value
- We can do this by calculating a confidence interval mean for the value of y and a prediction interval for an individual value of y

A Confidence Interval and a Prediction Interval

If the regression assumptions hold,

1 A **100(1 − α) percent confidence interval for the mean value of *y*** when x equals x_0 is

$$\left[\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}\right]$$

2 A **100(1 − α) percent prediction interval for an individual value of *y*** when x equals x_0 is

$$\left[\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}\right]$$

Here, $t_{\alpha/2}$ is based on $(n - 2)$ degrees of freedom.

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

$$s_{(y-\hat{y})} = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

Distance Value

- Both the confidence interval for the mean value of y and the prediction interval for an individual value of y employ a quantity called the distance value

$$\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}$$

- The distance value is a measure of the distance between the value x_0 of x and $\bar{x}$
- Notice that the further x_0 is from $\bar{x}$, the larger the distance value

A Confidence Interval and Prediction Interval

- Assume that the regression assumption holds
- The formula for a 100(1-α) confidence interval for the mean value of y is as follows:

$$[\hat{y} \pm t_{\alpha/2} s\sqrt{\text{Distance value}}]$$

- The formula for a 100(1-α) prediction interval for an individual value of y is as follows:

$$[\hat{y} \pm t_{\alpha/2} s\sqrt{1+\text{Distance value}}]$$

- This is based on n-2 degrees of freedom

Which to Use?

- The prediction interval is useful if it is important to predict an individual value of the dependent variable
- A confidence interval is useful if it is important to estimate the mean value
- The prediction interval will always be wider than the confidence interval

EXAMPLE The Tasty Sub Shop Case: Predicting Revenue and Profit C

In the Tasty Sub Shop problem, recall that one of the business entrepreneur's potential sites is near a population of 47,300 residents. Also, recall that

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_0 \\ &= 183.31 + 15.596(47.3) \\ &= 921.0 \text{ (that is, \$921,000)}\end{aligned}$$

is the point estimate of the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents and is the point prediction of the yearly revenue for a single Tasty Sub restaurant that is built near a population of 47,300 residents. Using the information in Example 14.2 (page 493), we compute

$$\begin{aligned}\text{distance value} &= \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \\ &= \frac{1}{10} + \frac{(47.3 - 43.41)^2}{1913.129} \\ &= .1079\end{aligned}$$

Example: The Tasty Sub Shop Case

Because $s = 61.7052$ (see Example 14.3 on page 502) and because $t_{\alpha/2} = t_{.025}$ based on $n - 2 = 10 - 2 = 8$ degrees of freedom equals 2.306, it follows that a 95 percent confidence interval for the mean yearly revenue when $x = 47.3$ is

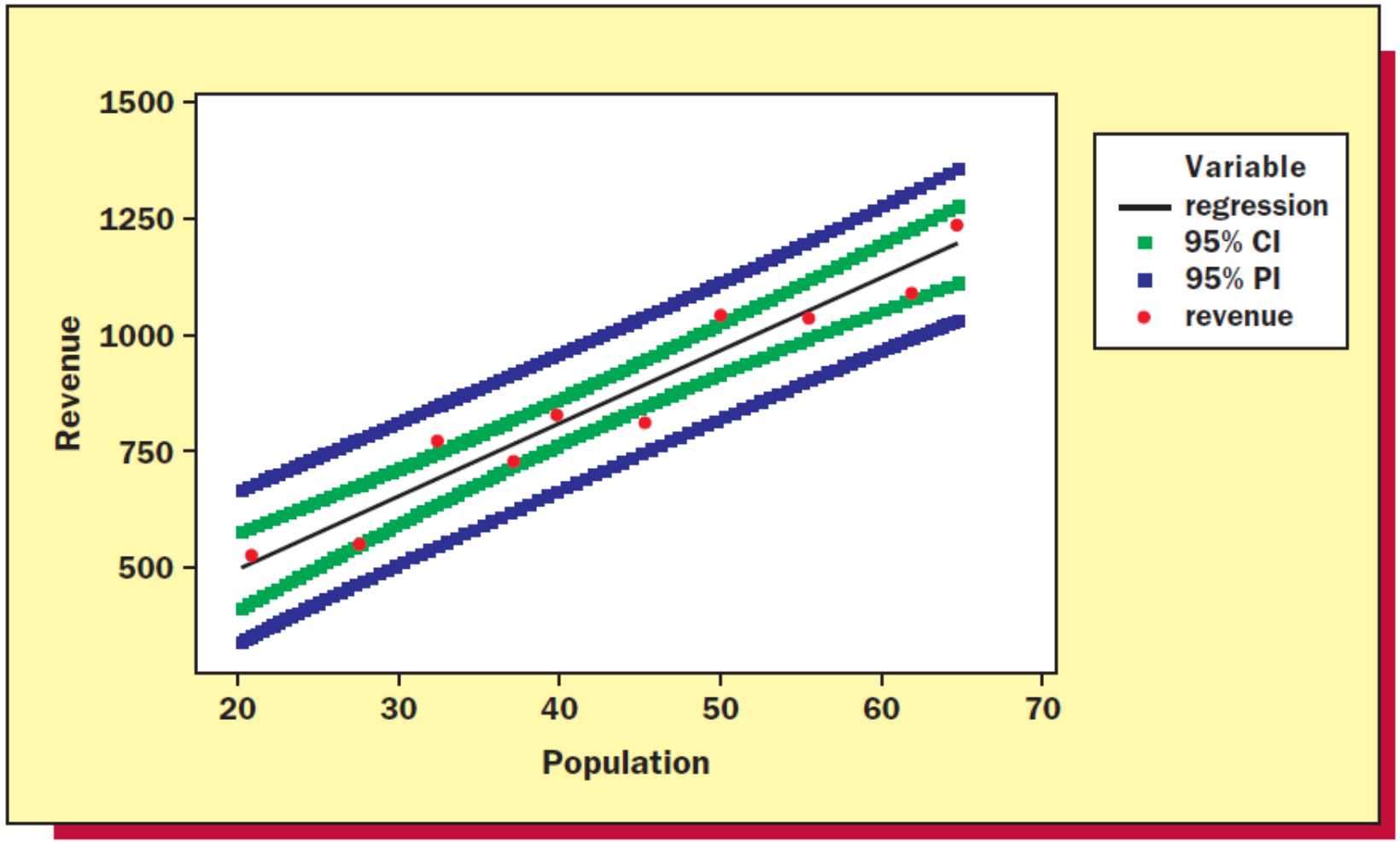
$$\begin{aligned}
&[\hat{y} \pm t_{\alpha/2} s \sqrt{\text{distance value}}] \\
&= [921.0 \pm 2.306(61.7052)\sqrt{.1079}] \\
&= [921.0 \pm 46.74] \\
&= [874.3, 967.7]
\end{aligned}$$

This interval says we are 95 percent confident that the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents is between \$874,300 and \$967,700.

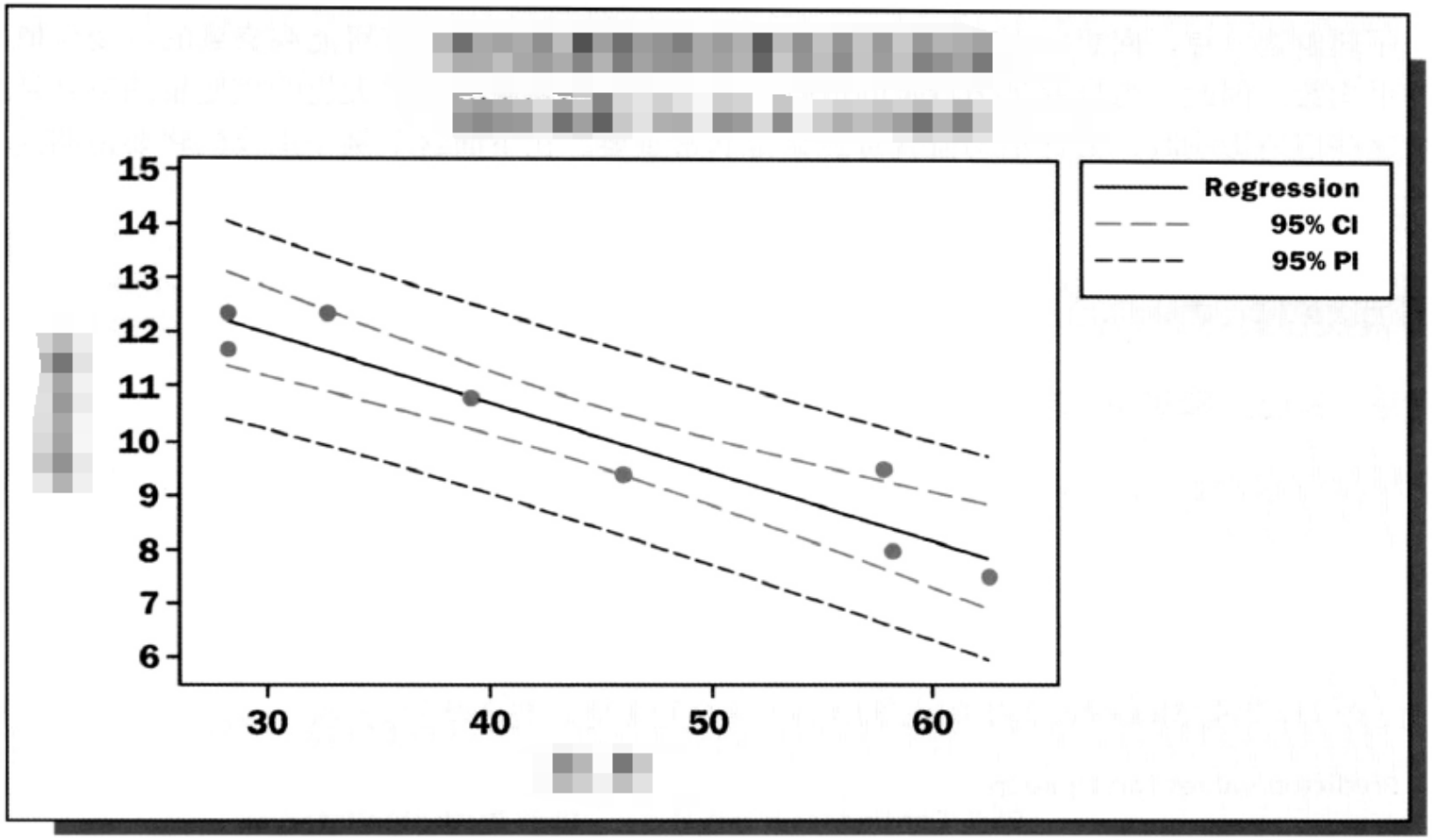
Because the entrepreneur would be operating a single Tasty Sub restaurant that is built near a population of 47,300 residents, the entrepreneur is interested in obtaining a prediction interval for the yearly revenue of such a restaurant. A 95 percent prediction interval for this revenue is

$$\begin{aligned}
&[\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \text{distance value}}] \\
&= [921.0 \pm 2.306(61.7052)\sqrt{1.1079}] \\
&= [921.0 \pm 149.77] \\
&= [771.2, 1070.8]
\end{aligned}$$

FIGURE MINITAB Output of 95% Confidence and Prediction Intervals for the Tasty Sub Shop Case



Example 13.9 The Fuel Consumption Case



13.5 Simple Coefficient of Determination and Correlation

- How useful is a particular regression model?
- One measure of usefulness is the simple coefficient of determination
- It is represented by the symbol r^2

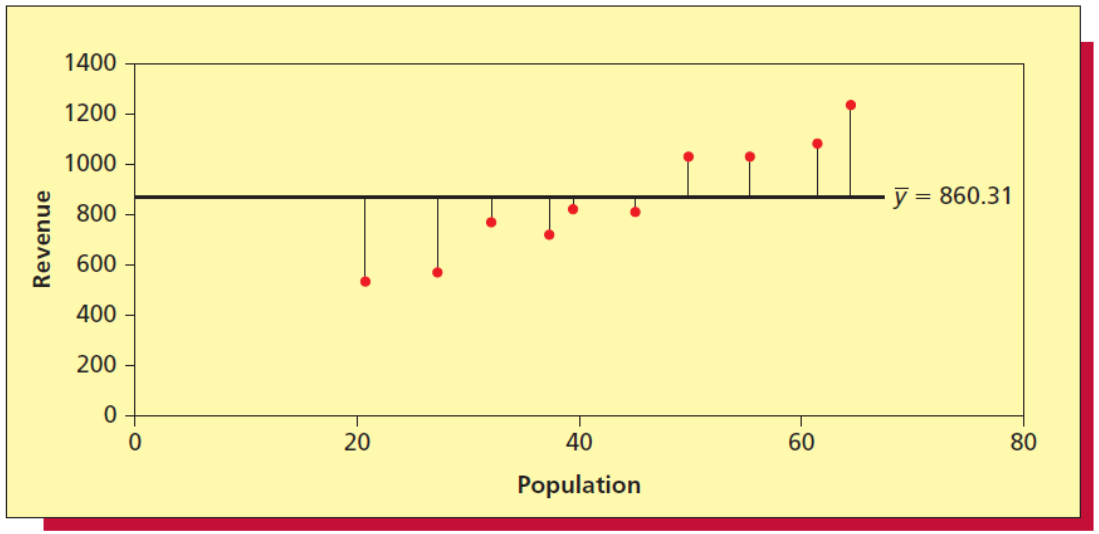
Calculating The Simple Coefficient of Determination

1. **Total variation** is $\sum(y_i - \bar{y})^2$
2. **Explained variation** is $\sum(\hat{y}_i - \bar{y})^2$
3. **Unexplained variation** is $\sum(y_i - \hat{y}_i)^2$
4. **Total variation is the sum of explained and unexplained variation**
5. **r^2 is the ratio of explained variation to total variation**
6. **r^2 is the proportion of explained variation**

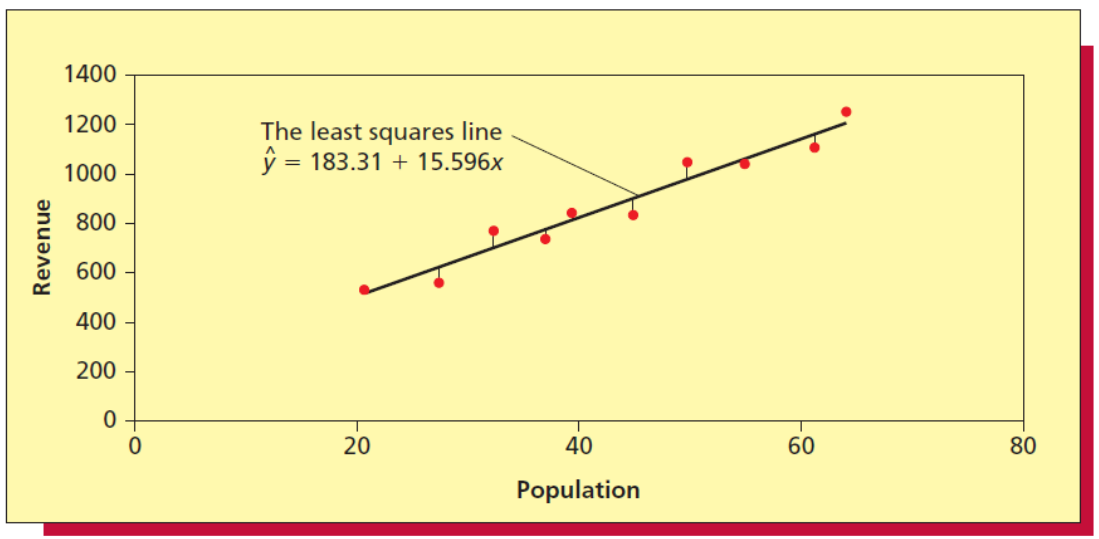
$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

FIGURE The Reduction in the Prediction Errors Accomplished by Employing the Predictor Variable x

(a) Prediction errors for the Tasty Sub Shop case when we do not use the information contributed by x



(b) Prediction errors for the Tasty Sub Shop case when we use the information contributed by x by using the least squares line



The Simple Coefficient of Determination, r^2

For the simple linear regression model

1 **Total variation** $= \sum (y_i - \bar{y})^2$

2 **Explained variation** $= \sum (\hat{y}_i - \bar{y})^2$

3 **Unexplained variation** $= \sum (y_i - \hat{y}_i)^2$

4 **Total variation = Explained variation + Unexplained variation**

5 **The simple coefficient of determination is**

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

6 r^2 is the proportion of the total variation in the n observed values of the dependent variable that is explained by the simple linear regression model.

The Simple Correlation Coefficient

The simple correlation coefficient measures the strength of the linear relationship between y and x and is denoted by r

$$r > 0 \iff b_1 > 0$$

$$r < 0 \iff b_1 < 0$$

Where, b_1 is the slope of the least squares line

r can be calculated using the formula

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

The Coefficient of Correlation (r).

It is a measure of the strength of the relationship (linear) between two variables

$$r = \frac{ss_{xy}}{\sqrt{ss_{xx}ss_{yy}}}$$

$$ss_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$ss_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$ss_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

It can range from -1.00 to 1.00.

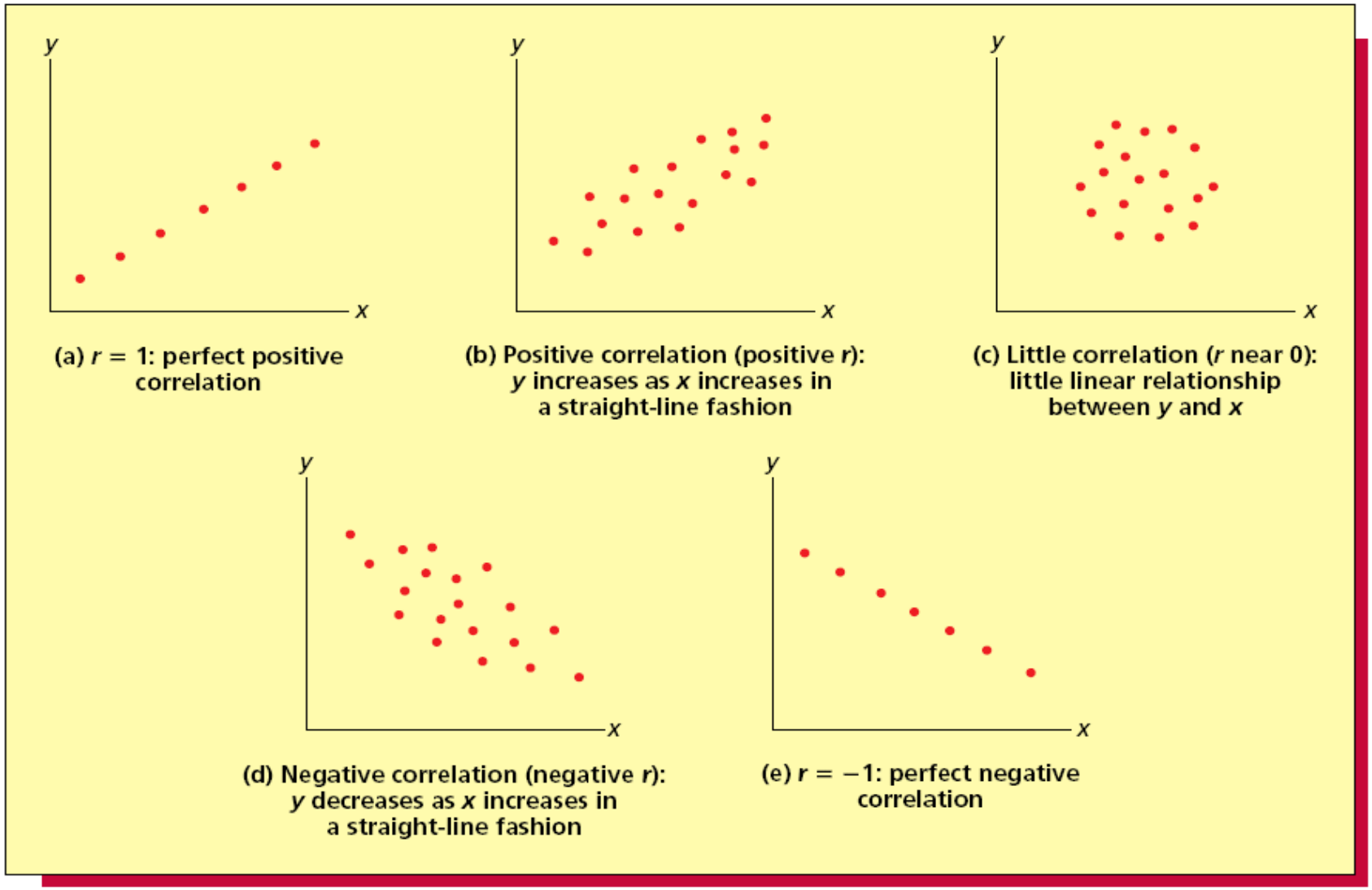
The Coefficient of Correlation (*r*).

Values of -1.00 or 1.00 indicate perfect and strong correlation.

Negative values indicate an inverse relationship and positive values indicate a direct relationship.

Values close to 0.0 indicate weak correlation.

Different Values of the Correlation Coefficient



The Simple Correlation Coefficient

- The simple correlation coefficient measures the strength of the linear relationship between y and x and is denoted by r

$$r = +\sqrt{r^2} \text{ if } b_1 \text{ is positive, and}$$

$$r = -\sqrt{r^2} \text{ if } b_1 \text{ is negative}$$

- Where b_1 is the slope of the least squares line

EXAMPLE The Tasty Sub Shop Case: Calculating and Interpreting r^2 C

For the Tasty Sub data we have seen that $\bar{y} = 860.31$ (see Example 14.2 on page 493). It follows that the total variation is

$$\sum (y_i - \bar{y})^2 = (527.1 - 860.31)^2 + (548.7 - 860.31)^2 + \cdots + (1235.8 - 860.31)^2$$

$$= 495{,}776.51$$

Furthermore, we found in Table 14.2 (page 494) that the unexplained variation is $SSE = 30{,}460.21$. Therefore, we can compute the explained variation and r^2 as follows:

$$\text{Explained variation} = \text{Total variation} - \text{Unexplained variation}$$

$$= 495{,}776.51 - 30{,}460.21 = 465{,}316.30$$

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{465{,}316.30}{495{,}776.51} = .939$$

This value of r^2 says that the regression model explains 93.9 percent of the total variation in the 10 observed yearly revenues.

- Practice

(1) The Fuel Consumption Case

(2) The QHIC Case

13.6 Testing the Significance of the Population Correlation Coefficient (Optional)

- The simple correlation coefficient (r) measures the linear relationship between the observed values of x and y from the sample
- The population correlation coefficient (ρ) measures the linear relationship between all possible combinations of observed values of x and y
- r is an estimate of ρ

Testing ρ

- We can test to see if the correlation is significant using the hypotheses

  H_0: ρ = 0
  H_a: ρ ≠ 0

- The statistic is $$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

- This test will give the same results as the test for significance on the slope coefficient b_1

13.7 An *F*-Test for Model

- For simple regression, this is another way to test the null hypothesis

  $H_0: \beta_1 = 0$
  $H_1: \beta_1 \neq 0$

- This is the only test we will use for multiple regression
- The F test tests the significance of the overall regression relationship between x and y

An *F* Test for the Simple Linear Regression Model

Suppose that the regression assumptions hold, and define the **overall *F* statistic** to be

$$F(\text{model}) = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n-2)}$$

Also define the *p*-value related to *F*(model) to be the area under the curve of the *F* distribution (having 1 numerator and $n-2$ denominator degrees of freedom) to the right of *F*(model)—see Figure 14.20(b).

We can reject H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ at level of significance α if either of the following equivalent conditions holds:

1. $F(\text{model}) > F_\alpha$
2. $p\text{-value} < \alpha$

Here the point F_α is based on 1 numerator and $n-2$ denominator degrees of freedom.

Mechanics of the F Test

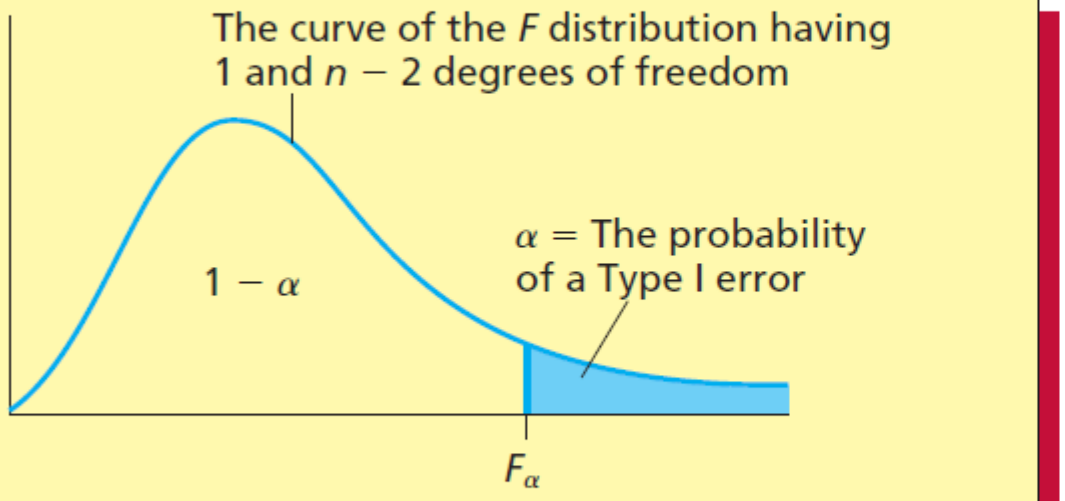
- To test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ at the α level of significance

$$\text{Test statistic}$$

$$F = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n-2)}$$

- Reject H_0 if F(model) > F_α or p-value < α
- F_α is based on 1 numerator and n-2 denominator degrees of freedom

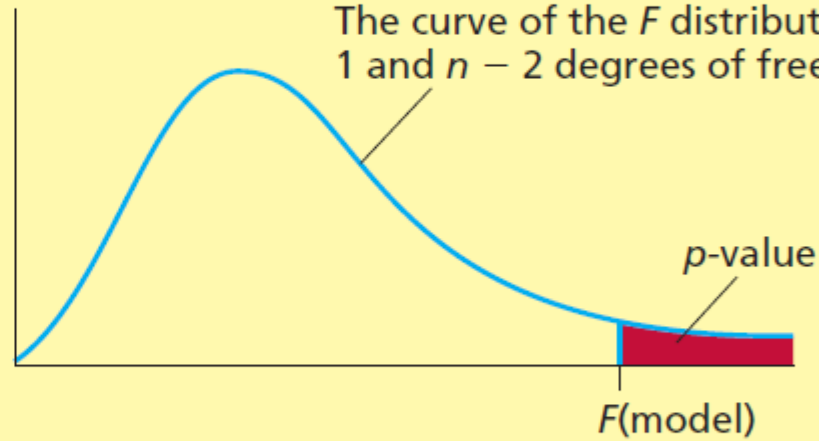
The *F*-Test Critical Value



If $F(\text{model}) \leq F_\alpha$, do not reject H_0 in favor of H_a

If $F(\text{model}) > F_\alpha$, reject H_0 in favor of H_a

The *F*-Test *p*-Value



If the *p*-value is smaller than α, then $F(\text{model}) > F_\alpha$ and we reject H_0.

Example 13.15 The Fuel Consumption Case

Example: The Tasty Sub Shop Case

Analysis of Variance					
Source	DF	SS	MS	F	P-value
Regression	1	465316	465316	122.21	0.000
Residual Error	8	30460	3808		
Total	9	495777			

Looking at this output, we see that the explained variation is 465,316 and the unexplained variation is 30,460. It follows that

$$F(\text{model}) = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n-2)}$$

$$= \frac{465{,}316}{30{,}460/(10-2)} = \frac{465{,}316}{3808}$$

$$= 122.21$$

Note that this overall F statistic is given on the MINITAB output and is also given on the following partial Excel output:

ANOVA	df	SS	MS	F	Significance F
Regression	1	465316.3004	465316.3004	122.2096	0.0000
Residual	8	30460.2086	3807.5261		
Total	9	495776.5090			

13.8 Residual Analysis (Optional)

- Checks of regression assumptions are performed by analyzing the regression residuals
- **Residuals** (e) are defined as the difference between the observed value of y and the predicted value of y, $e = y - \hat{y}$
  - Note that e is the point estimate of ε
- If regression assumptions valid, the population of potential error terms will be normally distributed with mean zero and variance σ^2
- Different error terms will be statistically independent

For any particular observed value of y, the corresponding **residual** is

$$e = y - \hat{y} = (\text{observed value of } y - \text{predicted value of } y)$$

where the predicted value of y is calculated using the **least squares prediction equation**

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

Residual Analysis #2

- Residuals should as if they are randomly and independently selected from normal populations with mean zero and variance σ^2
- With any real data, assumptions will not hold exactly
- Mild departures do not affect our ability to make statistical inferences
- In checking assumptions, we are looking for pronounced departures from the assumptions
- So, only require residuals to approximately fit the description above

Example 13.11 The QHIC Case: Constructing Residual Plots

- Quality Home Improvement Center (QHIC) operates five stores
- Wish to study relationship between home value and yearly expenditure on home upkeep
- Random sample of 40 homeowners
  - Intercept = –348.3921
  - Slope 7.2583

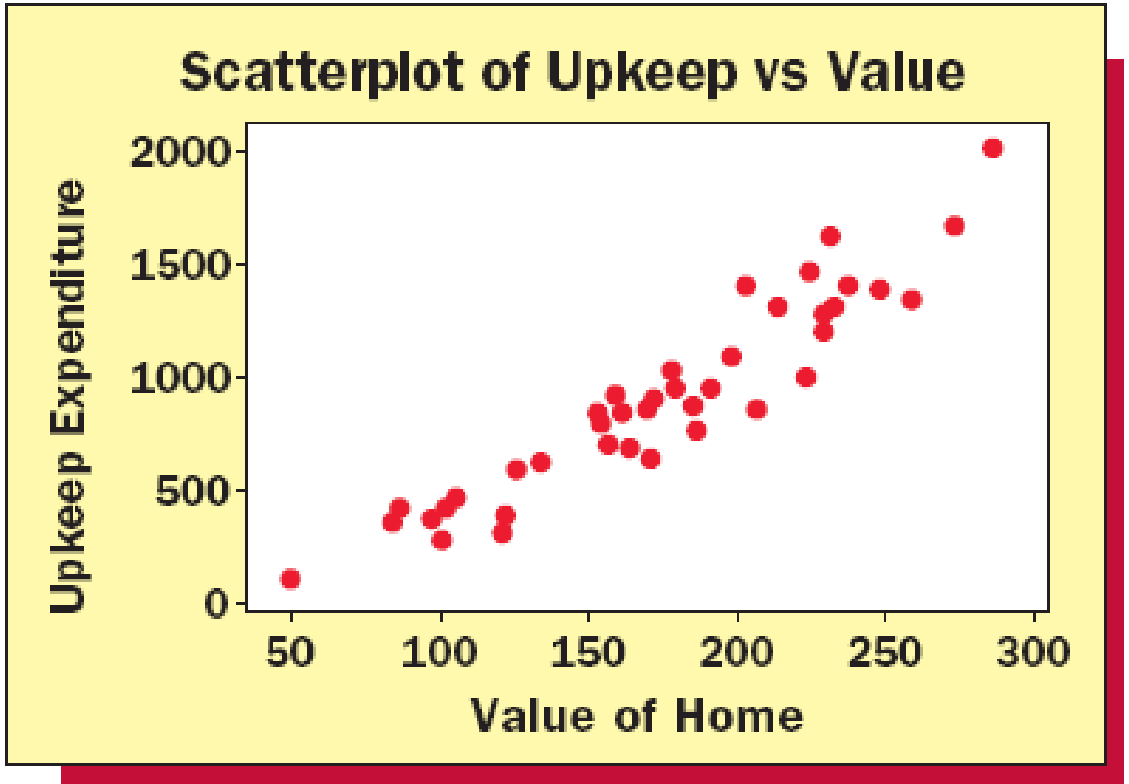
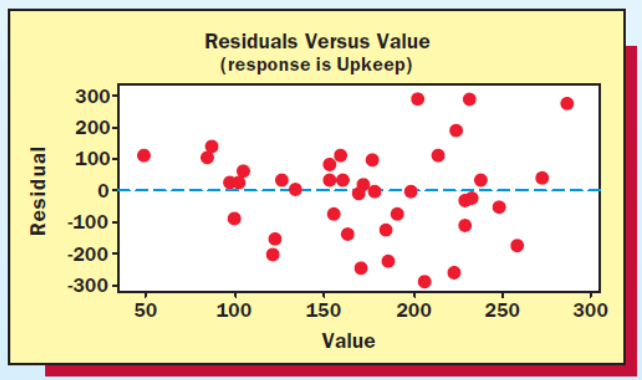


FIGURE [illegible] Residuals and Residual Plots for the QHIC Simple Linear Regression Model

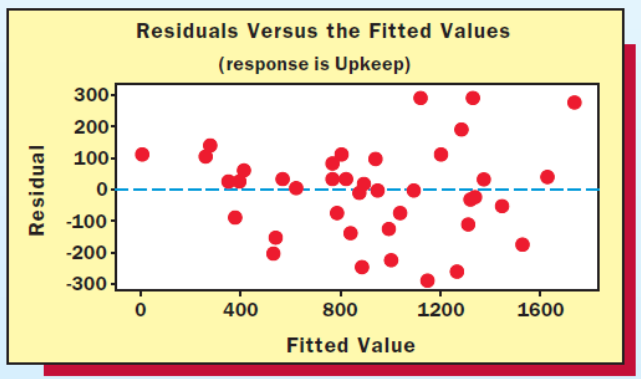
(a) Excel add-in (MegaStat) output of the residuals

Observation	Upkeep	Predicted	Residual	Observation	Upkeep	Predicted	Residual
1	1,412.080	1,371.816	40.264	21	849.140	762.413	86.727
2	797.200	762.703	34.497	22	1,313.840	1,336.832	−22.992
3	872.480	993.371	−120.891	23	602.060	562.085	39.975
4	1,003.420	1,263.378	−259.958	24	642.140	884.206	−242.066
5	852.900	817.866	35.034	25	1,038.800	938.353	100.447
6	288.480	375.112	−86.632	26	697.000	833.398	−136.398
7	1,288.460	1,314.041	−25.581	27	324.340	525.793	−201.453
8	423.080	390.354	32.726	28	965.100	1,038.662	−73.562
9	1,351.740	1,523.224	−171.484	29	920.140	804.075	116.065
10	378.040	350.434	27.606	30	950.900	947.208	3.692
11	918.080	892.771	25.309	31	1,670.320	1,627.307	43.013
12	1,627.240	1,328.412	298.828	32	125.400	6.537	118.863
13	1,204.760	1,308.815	−104.055	33	479.780	410.532	69.248
14	857.040	1,146.084	−289.044	34	2,010.640	1,728.778	281.862
15	775.000	999.613	−224.613	35	368.360	259.270	109.090
16	869.260	876.658	−7.398	36	425.600	277.270	148.330
17	1,396.000	1,444.835	−48.835	37	626.900	621.167	5.733
18	711.500	780.558	−69.058	38	1,316.940	1,196.602	120.338
19	1,475.180	1,278.911	196.269	39	390.160	537.261	−147.101
20	1,413.320	1,118.068	295.252	40	1,090.840	1,088.889	1.951

(b) MINITAB output of a residual plot versus x



(c) MINITAB output of a residual plot versus $\hat{y}$



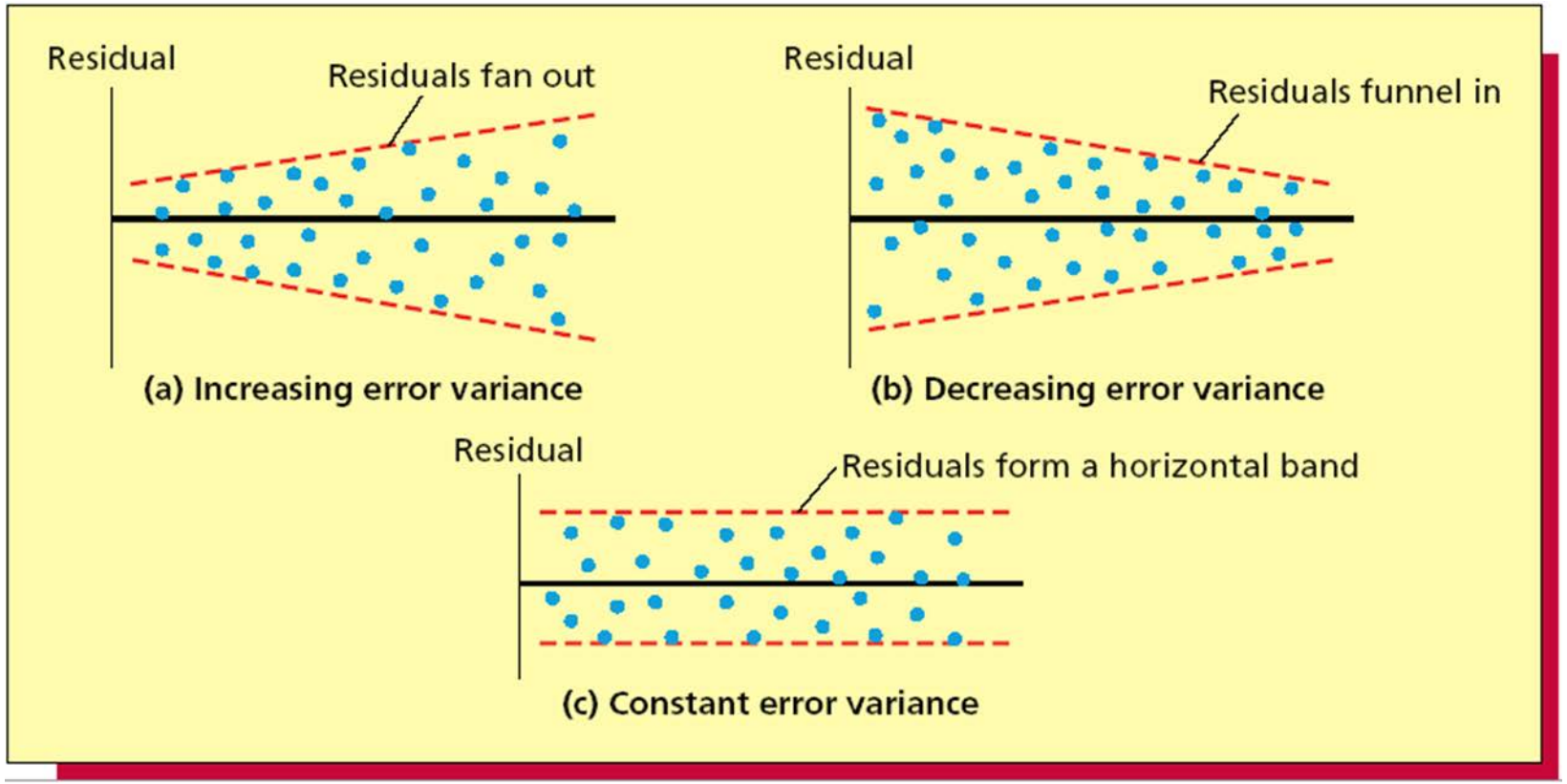
Residual Plots

1. Residuals versus independent variable
2. Residuals versus predicted y’s
3. Residuals in time order (if the response is a time series)

Constant Variance Assumptions

- To check the validity of the constant variance assumption, examine residual plots against
  - The x values
  - The predicted y values
  - Time (when data is time series)
- A pattern that fans out says the variance is increasing rather than staying constant
- A pattern that funnels in says the variance is decreasing rather than staying constant
- A pattern that is evenly spread within a band says the assumption has been met

Constant Variance Visually



Assumption of Correct Functional Form

- If the relationship between x and y is something other than a linear one, the residual plot will often suggest a form more appropriate for the model
- For example, if there is a curved relationship between x and y, a plot of residuals will often show a curved relationship

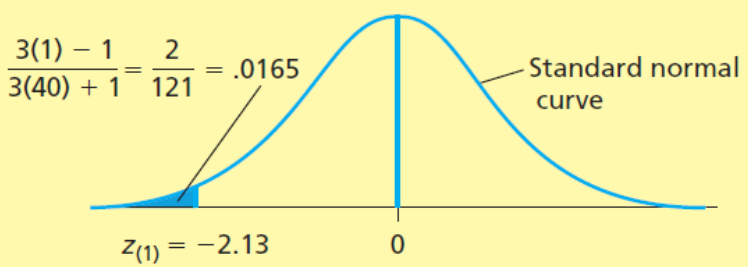
Normality Assumption

- If the normality assumption holds, a histogram or stem-and-leaf display of residuals should look bell-shaped and symmetric
- Another way to check is a normal plot of residuals
  1. Order residuals from smallest to largest
  2. Plot $e_{(i)}$ on vertical axis against $z_{(i)}$
     - $Z_{(i)}$ is the point on the horizontal axis under the z curve so the area under this curve to the left is (3i-1)/(3n+1)
  3. If the normality assumption holds, the plot should have a straight-line appearance

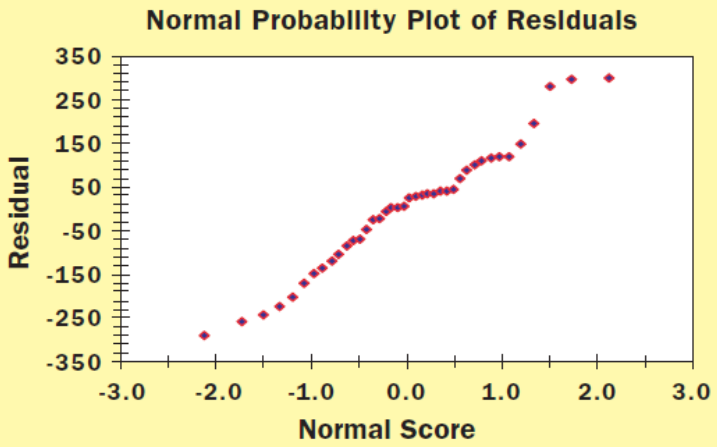
FIGURE Stem-and-Leaf Display and Normal Plots of the Residuals from the Simple Linear Regression Model Describing the QHIC Data

```
Stem-and-leaf of RESI1 N = 40
Leaf Unit = 10
 2     -2  85
 5     -2  420
 6     -1  7
 10    -1  4320
 13    -0  876
 17    -0  4220
(11)    0  00022333344
 12     0  68
 10     1  001124
 4      1  9
 3      2
 3      2  899
```

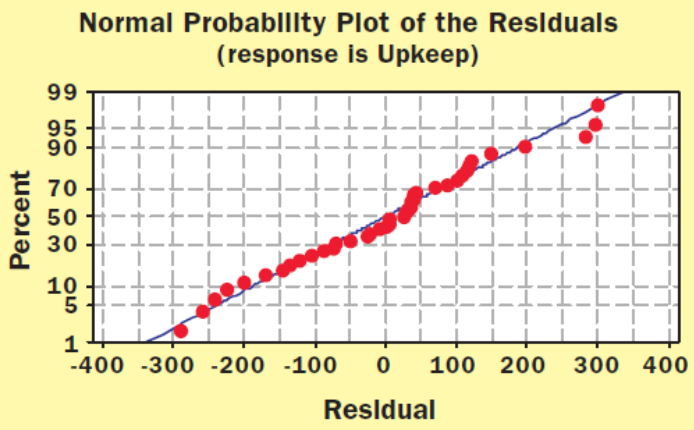
(a) MINITAB output of the stem-and-leaf display



(b) Calculating $z_{(1)}$ for a normal plot



(c) Excel add-in (MegaStat) normal plot

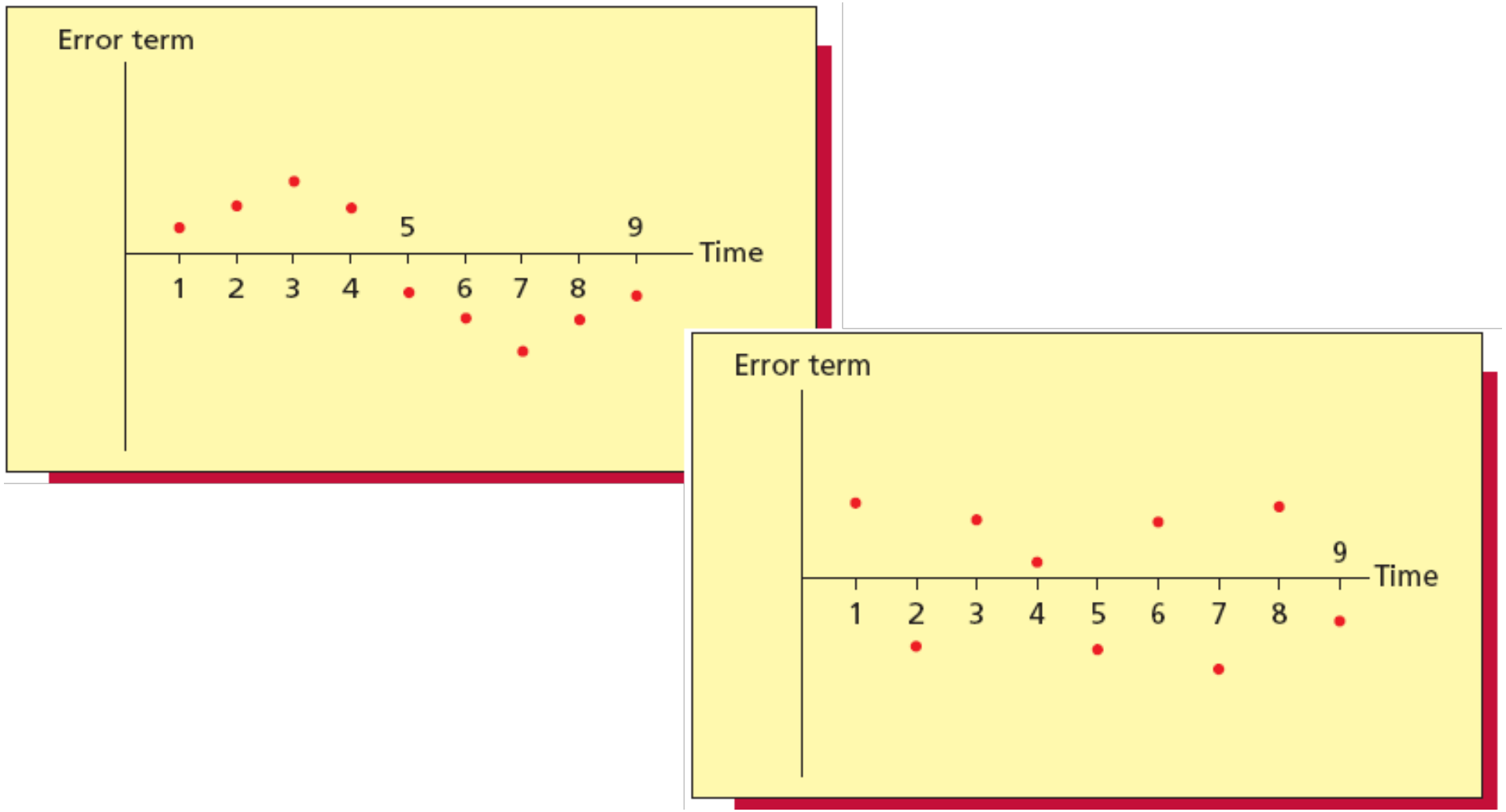


(d) MINITAB normal plot

Independence Assumption

- Independence assumption most likely violated by time-series data
  - If the data is not time series, it can be reordered without affecting it
- For time-series data, the time-ordered error terms can be autocorrelated
  - Positive autocorrelation is when a positive error term in time period i tends to be followed by another positive value in i+k
  - Negative autocorrelation is when a positive error term tends to be followed by a negative value
- Either one will cause a cyclical error term over time

Independence Assumption Visually



Example 13.17

EXAMPLE Pages Bookstore: Positive Autocorrelation

Figure 14.26(a) presents data concerning weekly sales at Pages Bookstore (Sales), Pages weekly advertising expenditure (Adver), and the weekly advertising expenditure of Pages main competitor (Compadv). Here the sales values are expressed in thousands of dollars, and the advertising expenditure values are expressed in hundreds of dollars. Figure 14.26(a) also gives the MINITAB output of the residuals that are obtained when a simple linear regression analysis is performed relating Pages sales to Pages advertising expenditure. These residuals are plotted versus time in Figure 14.26(b). We see that the residual plot has a cyclical pattern. This tells us that the error terms for the model are positively autocorrelated and the independence assumption is violated. Furthermore, there tend to be positive residuals when the competitor's advertising expenditure is lower (in weeks 1 through 8 and weeks 14, 15, and 16) and negative residuals when the competitor's advertising expenditure is higher (in weeks 9 through 13). Therefore, the competitor's advertising expenditure seems to be causing the positive autocorrelation.

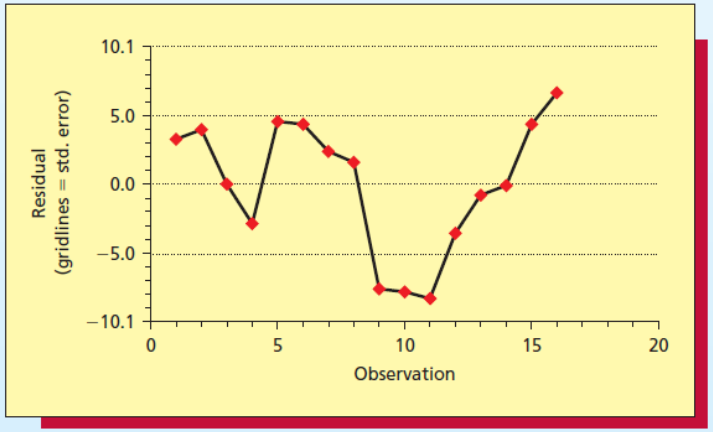
FIGURE Pages Bookstore Sales and Advertising Data, and Residual Analysis

(a) The data and the MINITAB output of the residuals from a simple linear regression relating Pages sales to Pages advertising expenditure DS BookSales

Observation	Adver	Compadv	Sales	Predicted	Residual
1	18	10	22	18.7	3.3
2	20	10	27	23.0	4.0
3	20	15	23	23.0	−0.0
4	25	15	31	33.9	−2.9
5	28	15	45	40.4	4.6
6	29	20	47	42.6	4.4
7	29	20	45	42.6	2.4
8	28	25	42	40.4	1.6
9	30	35	37	44.7	−7.7
10	31	35	39	46.9	−7.9
11	34	35	45	53.4	−8.4
12	35	30	52	55.6	−3.6
13	36	30	57	57.8	−0.8
14	38	25	62	62.1	−0.1
15	41	20	73	68.6	4.4
16	45	20	84	77.3	6.7

Durbin-Watson = 0.65

(b) A plot of the residuals in Figure 14.26(a) versus time

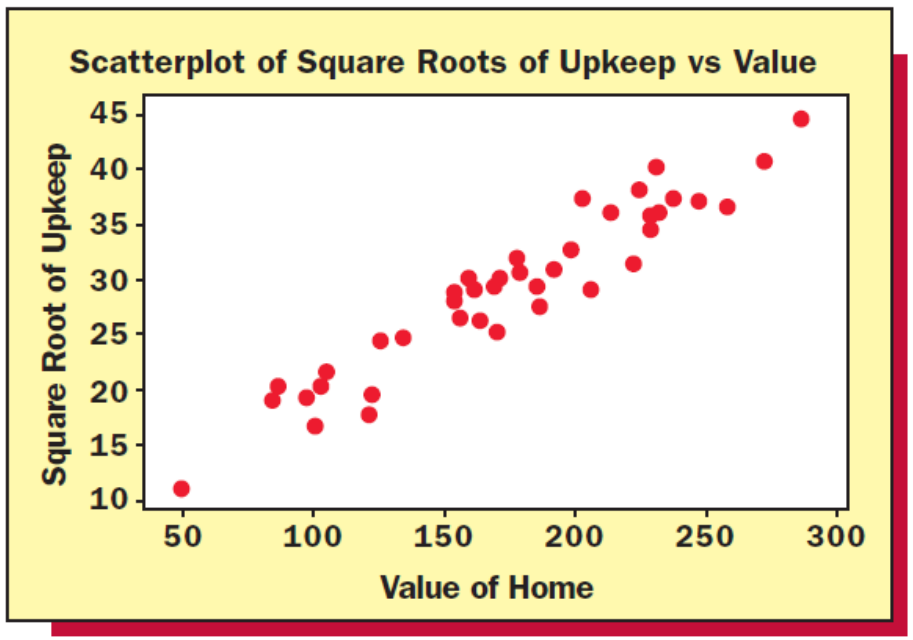


EXAMPLE The QHIC Case: Using a Data Transformation

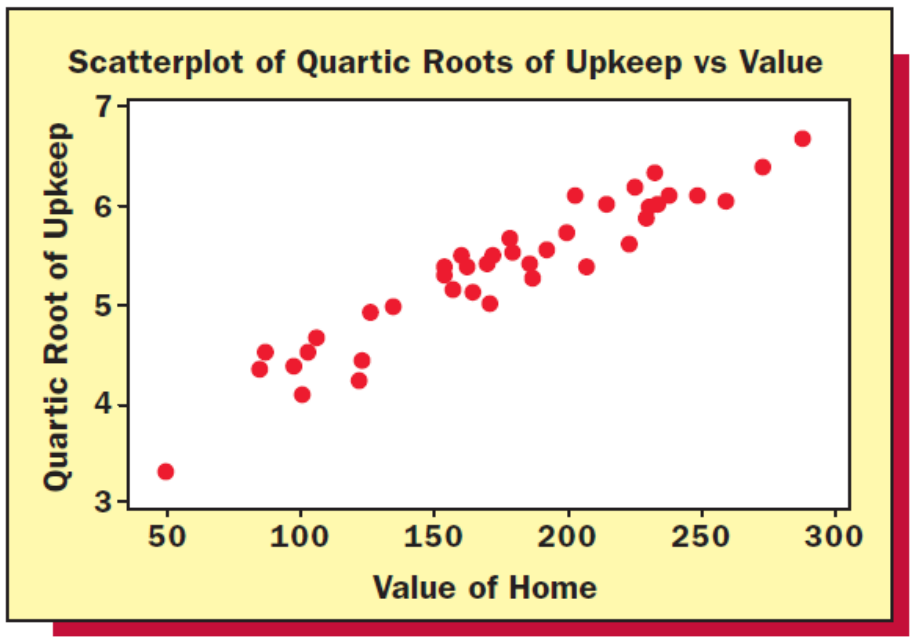
C

Consider the QHIC upkeep expenditures in Figure 14.21. In Figures 14.27, 14.28, and 14.29 we show the plots that result when we take the square root, quartic root, and natural logarithmic transformations of the upkeep expenditures and plot the transformed values versus the home values. The square root transformation seems to best equalize the error variance and straighten out the curved data plot in Figure 14.21. Note that the natural logarithm transformation seems to

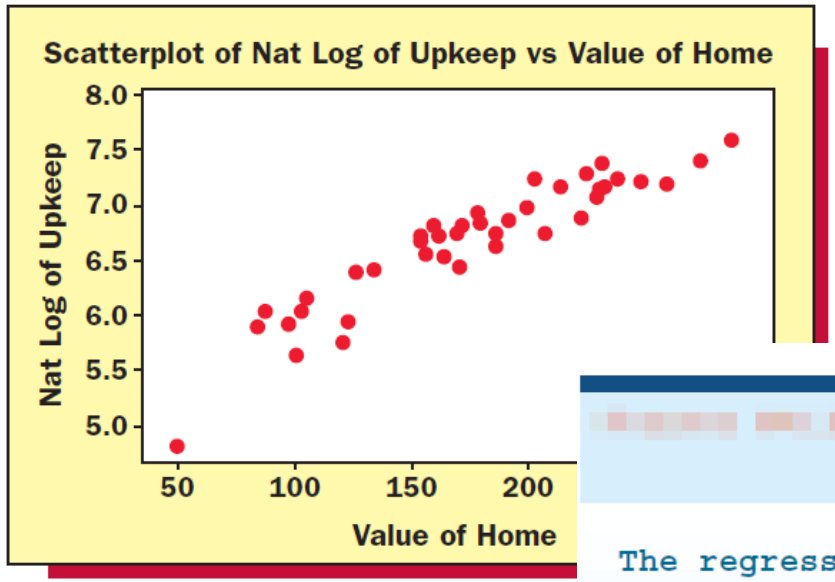
MINITAB Plot of the Square Roots of the Upkeep Expenditures versus the Home Values



MINITAB Plot of the Quartic Roots of the Upkeep Expenditures versus the Home Values



MINITAB Plot of the Natural Logarithms of the Upkeep Expenditures versus the Home Values



MINITAB Output of a Regression Analysis of the Upkeep Expenditure Data by Using the Model $y^* = \beta_0 + \beta_1 x + \varepsilon$ where $y^* = y^{.5}$, and a Residual Plot versus x

```
The regression equation is
SqRtUpkeep = 7.20 + 0.127 Value

Predictor       Coef    SE Coef       T       P
Constant       7.201      1.205    5.98   0.000
Value       0.127047   0.006577   19.32   0.000

S = 2.32479   R-Sq = 90.8%   R-Sq(adj) = 90.5%

Analysis of Variance
Source           DF      SS      MS       F       P
Regression        1  2016.8  2016.8  373.17   0.000
Residual Error   38   205.4     5.4
Total            39  2222.2

Values of Predictors for New Obs
New Obs  Value
      1    220

Predicted Values for New Observations
New Obs      Fit  SE Fit         95% CI             95% PI
      1   35.151   0.474  (34.191, 36.111)  (30.348, 39.954)
```

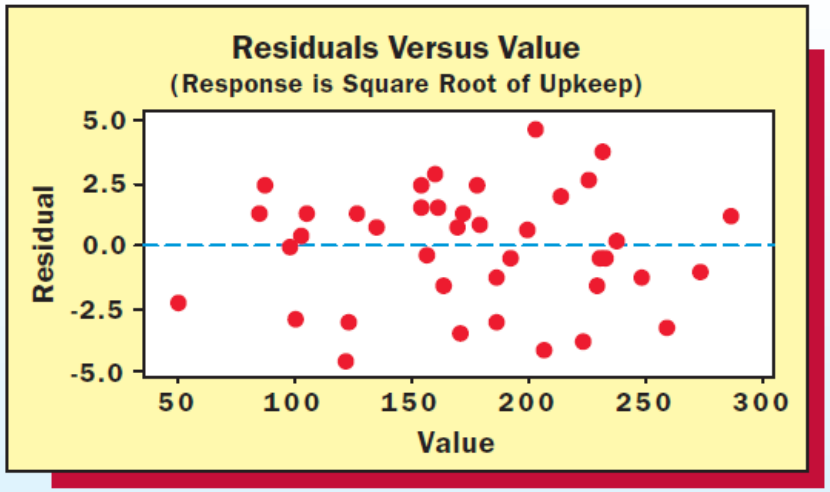
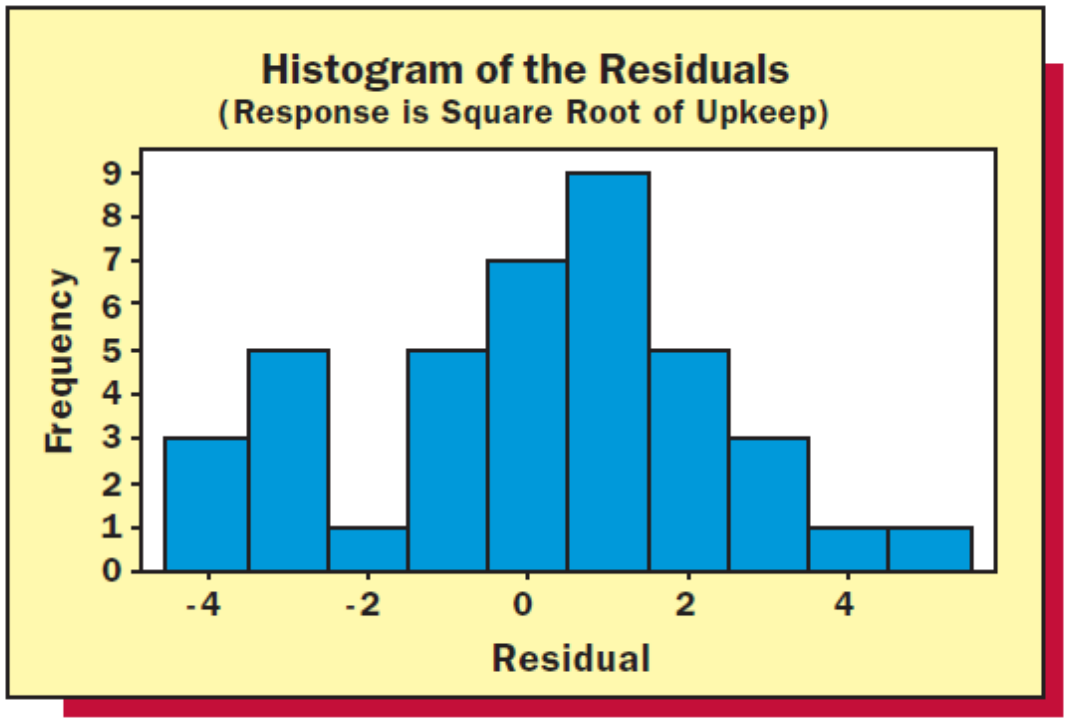


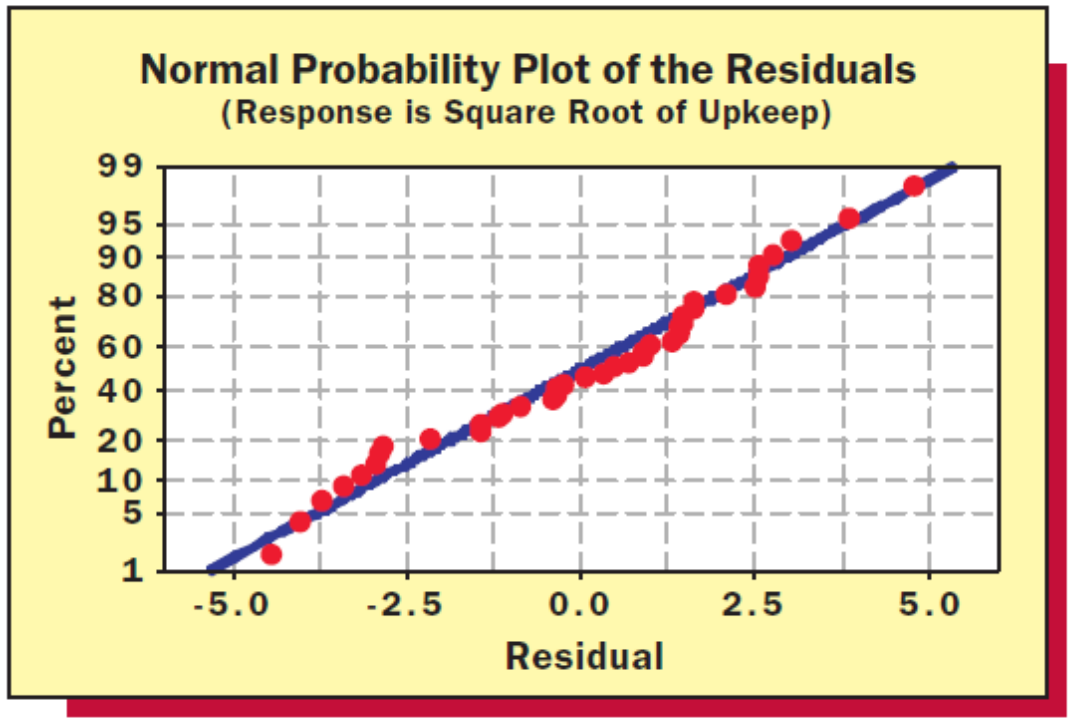
FIGURE MINITAB Output of Normality Assumption Analysis for the Upkeep Expenditure Model

$y^* = \beta_0 + \beta_1 x + \varepsilon$ where $y^* = y^{.5}$

(a) Histogram of the residuals



(b) Normal plot of the residuals



Chapter Summary

- Introduced types of regression models
- Reviewed assumptions of regression and correlation
- Discussed determining the simple linear regression equation
- Described inference about the slope
- Discussed correlation -- measuring the strength of the association
- Addressed estimation of mean values and prediction of individual values

Thank you!