Chapter 13

Simple Linear Regression Analysis

Review and Outlook

- 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.

Review and Outlook

- 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.

Review and Outlook

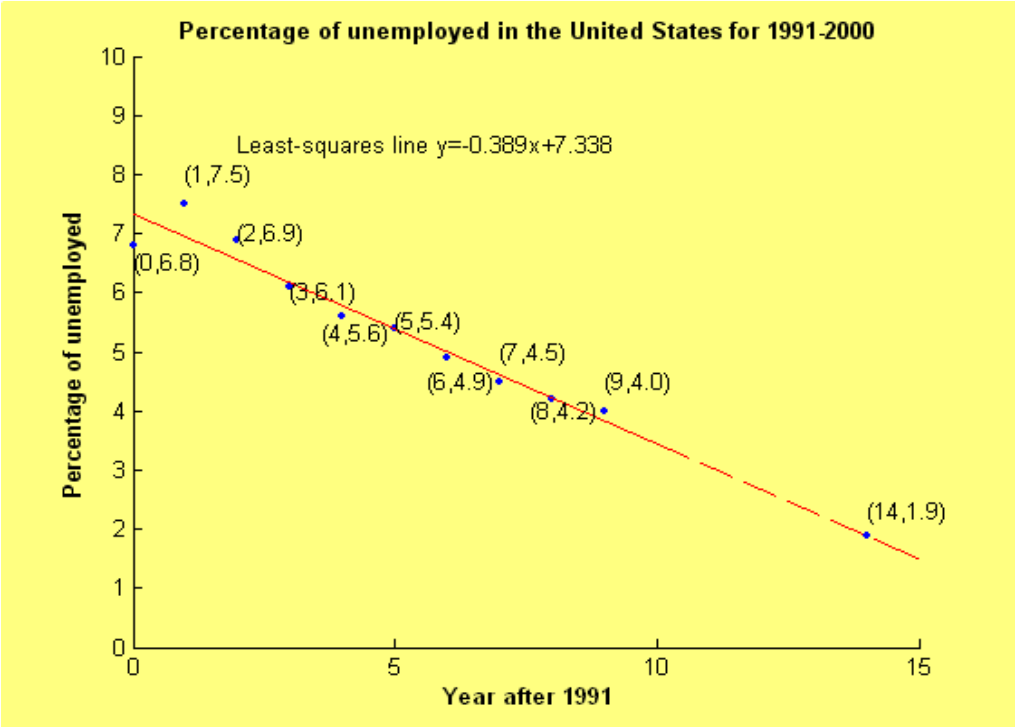
- We begin in this chapter by presenting simple linear regression analysis. Using this technique is appropriate when we are relating a dependent variable to a single independent variable and when a straight-line model describes the relationship between these two variables.

Example

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



Example

The pattern does suggest that we can get some useful information to predict future trends 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

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 (Optional)

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 (Optional)

13.8 Residual Analysis (Optional)

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

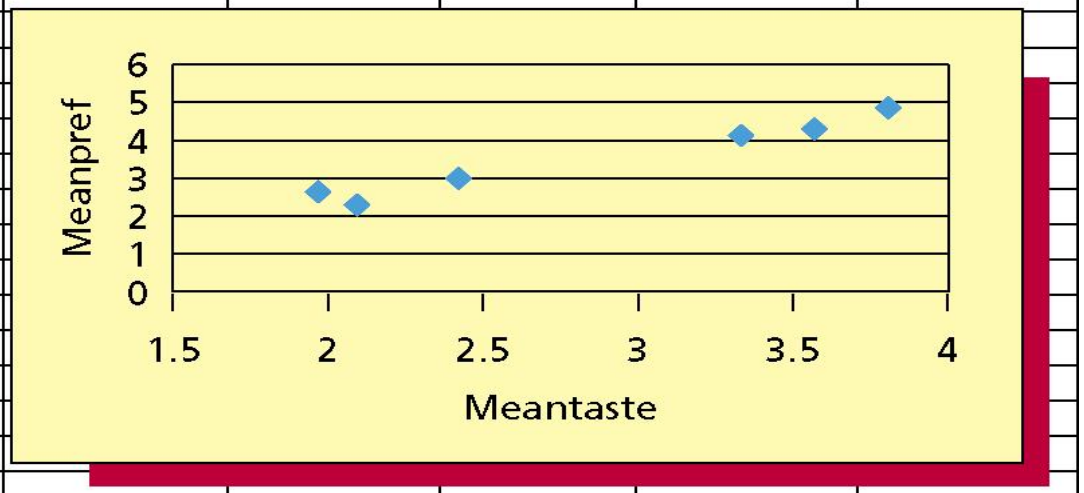
- The Scatter Diagram:
  - one variable is measured on the vertical axis and the other variable is measured on the horizontal axis

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.

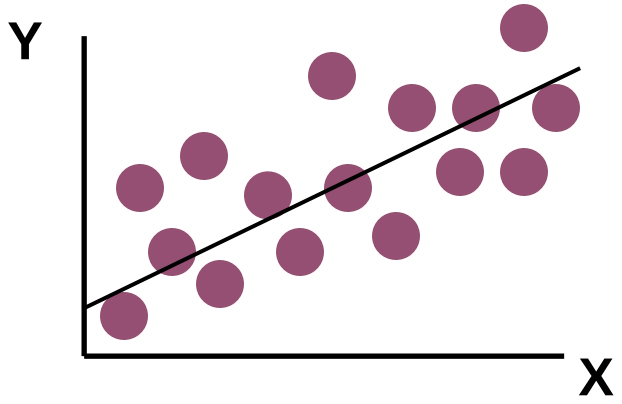
- **Regression analysis** is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

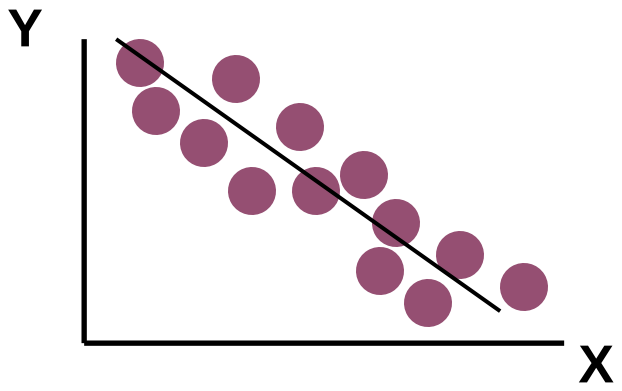
Simple Linear Regression Model

- 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

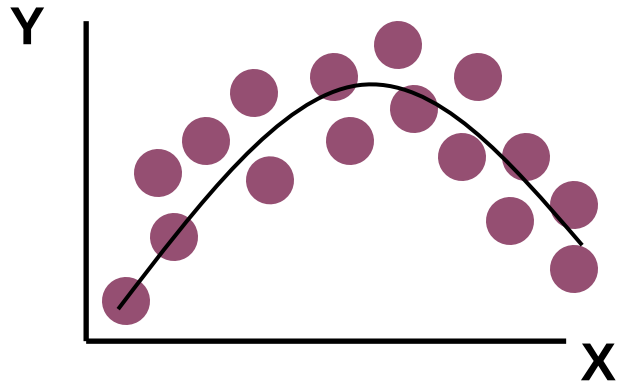
Types of Relationships

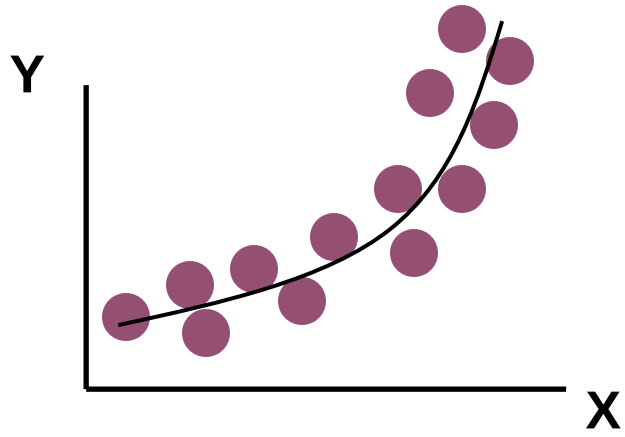
Linear relationships





Curvilinear relationships

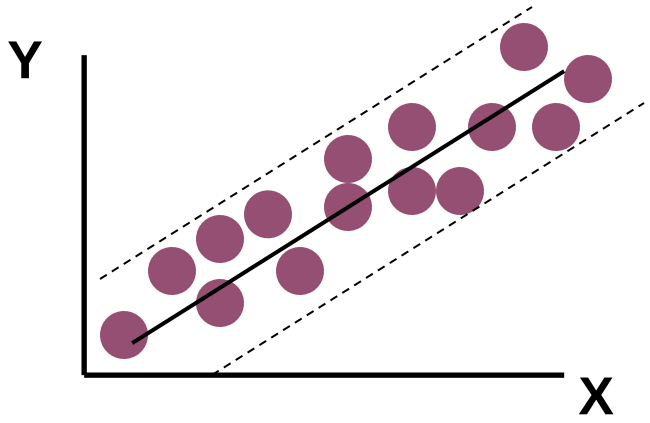


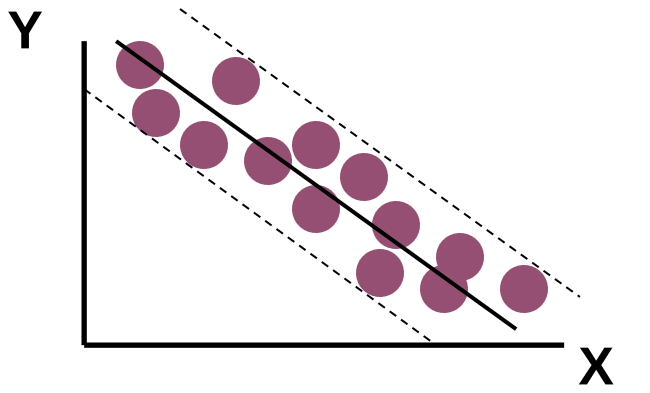


Types of Relationships

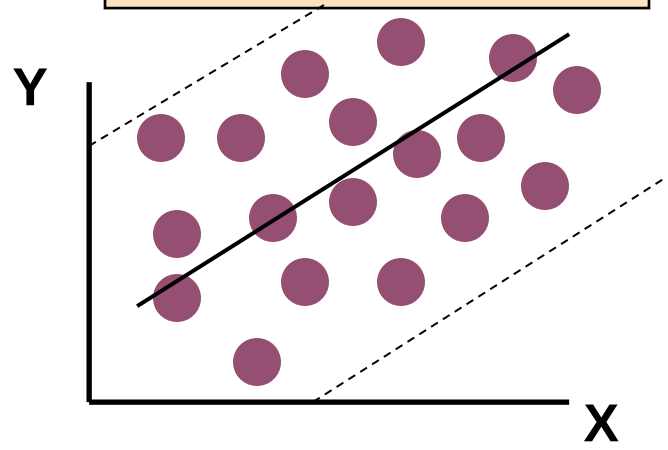
(continued)

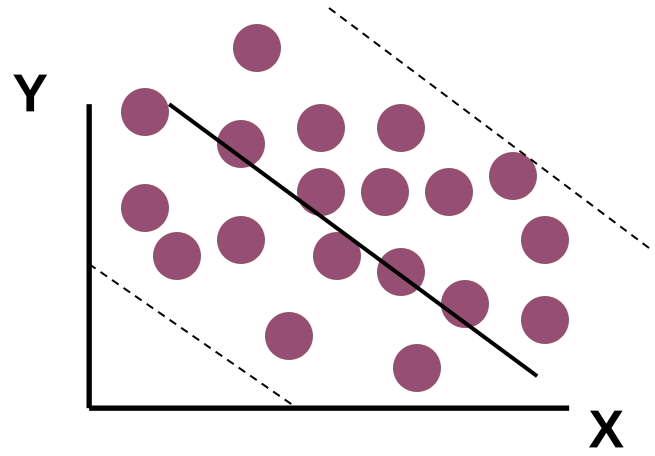
Strong relationships





Weak relationships

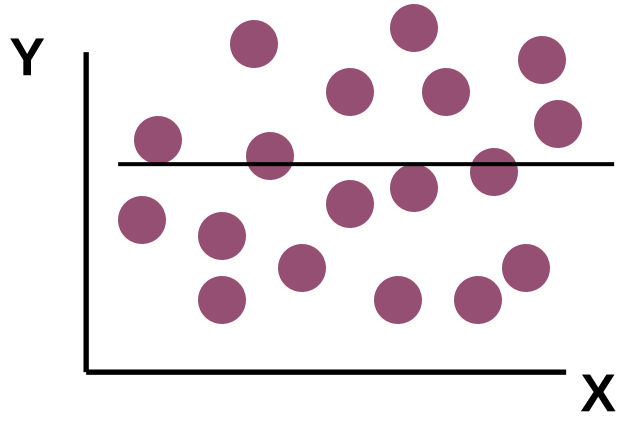




Types of Relationships

(continued)

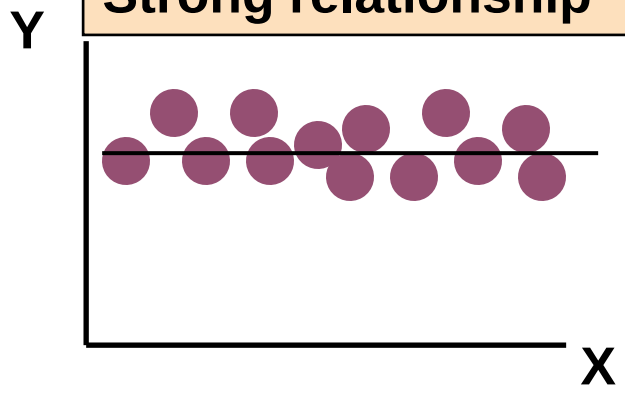
No relationship



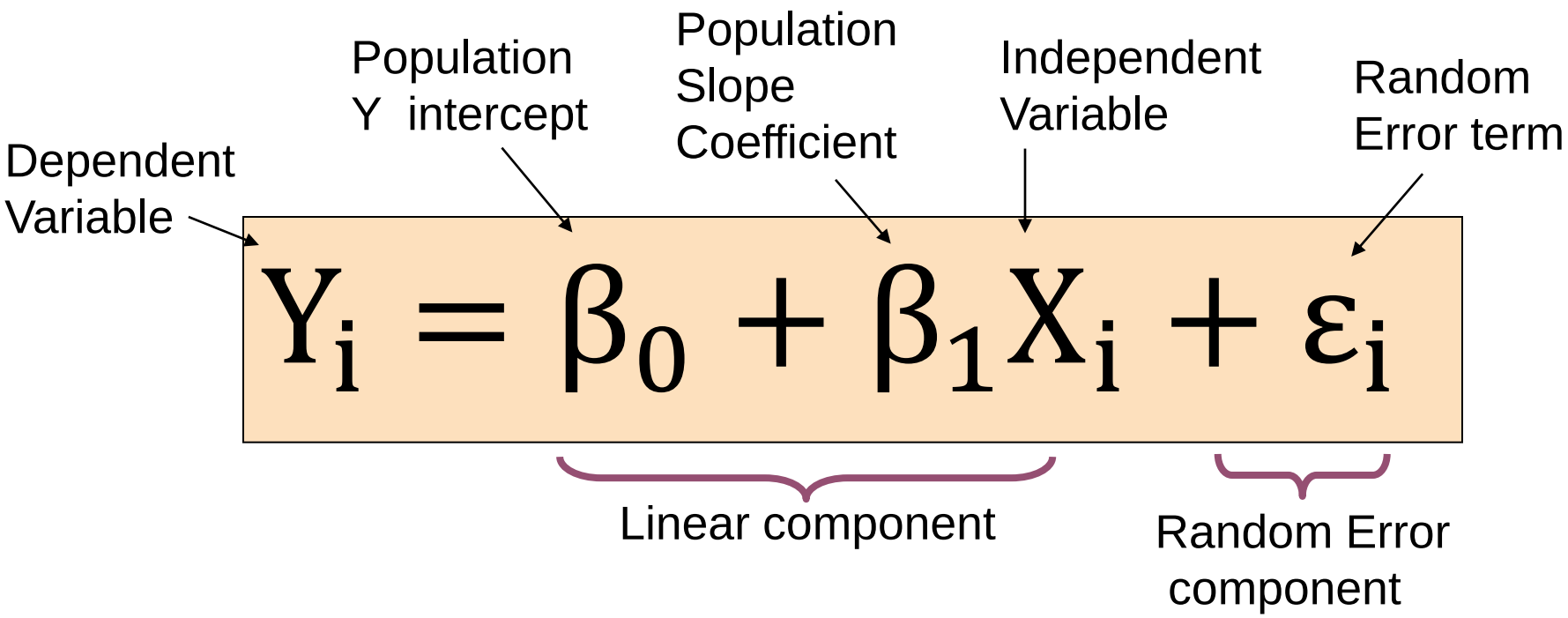
Randomly distributed

Variance is too large to find any relation or trend

Strong relationship

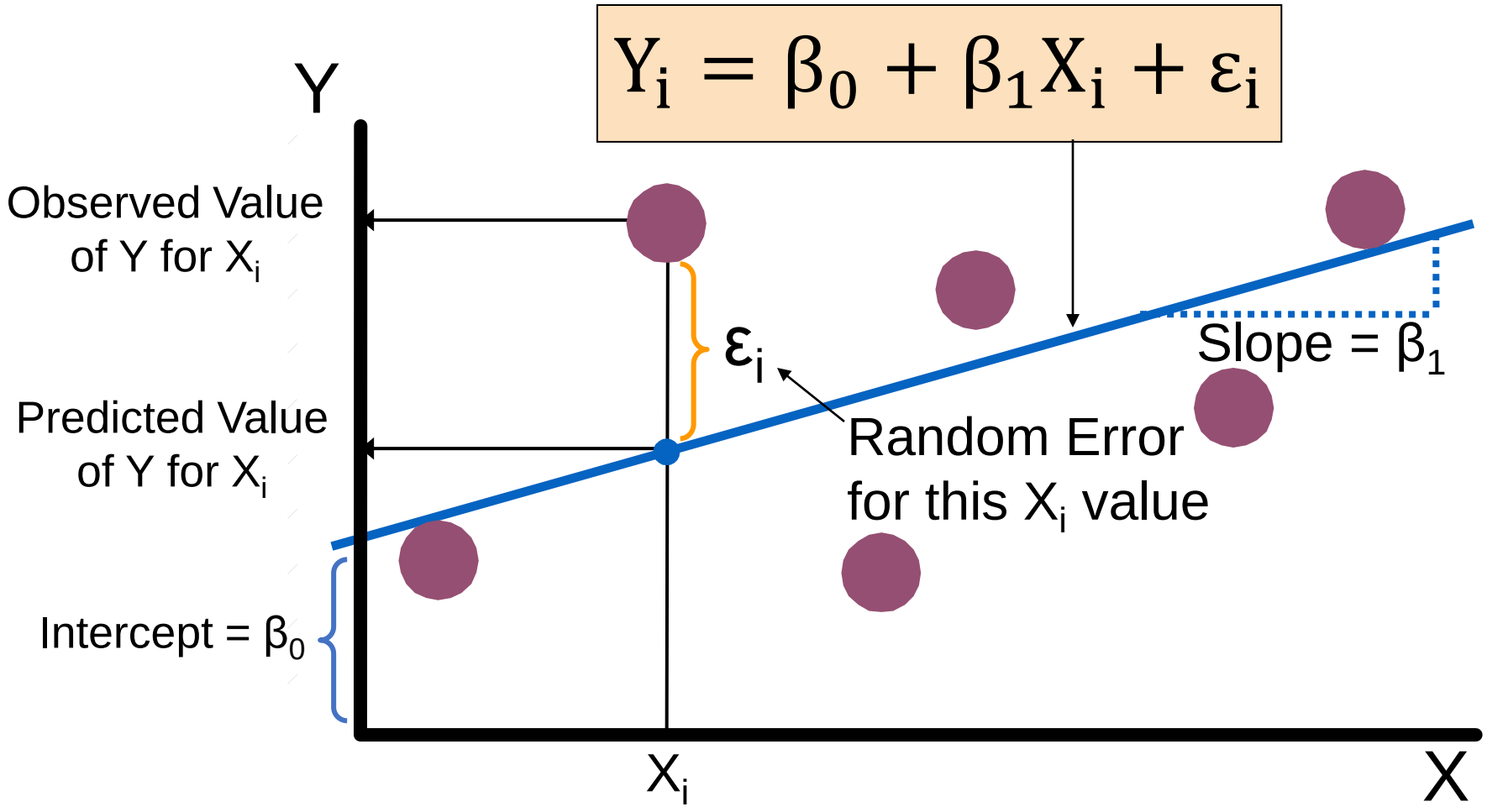


Simple Linear Regression Model



Simple Linear Regression Model

(continued)



Model Assumptions

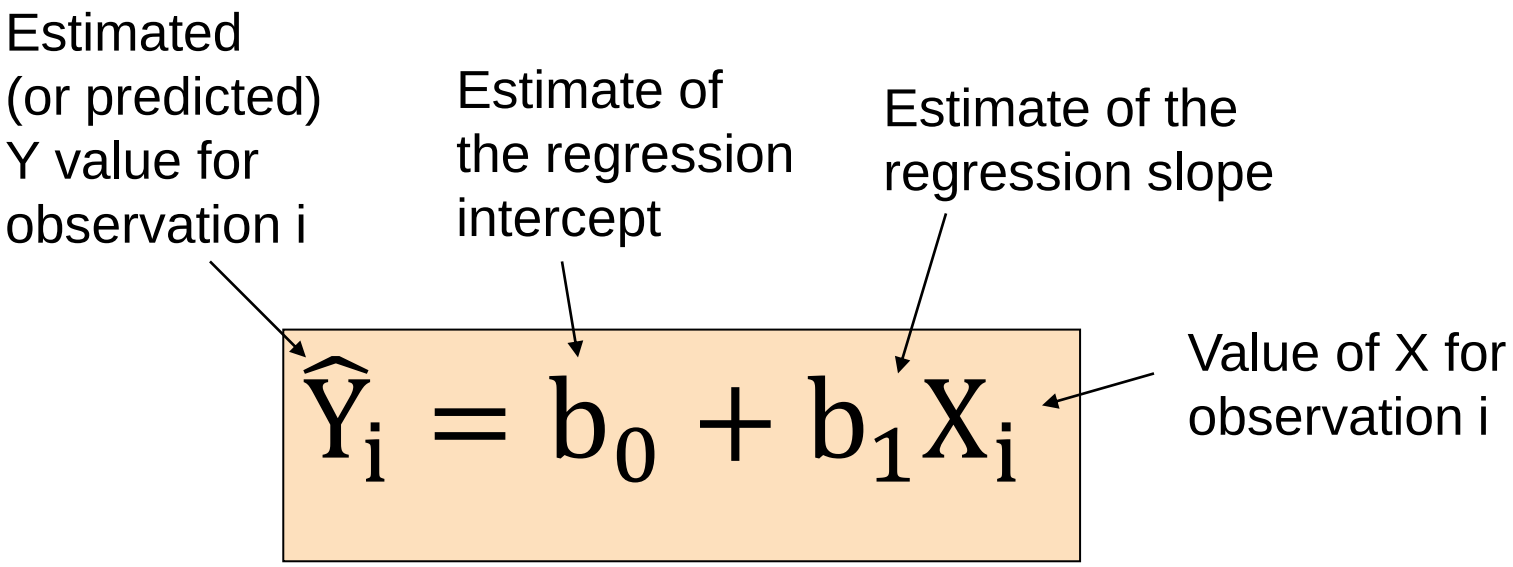
1. **Mean of Zero**: At any given value of x, the population of potential error term values has a **mean equal to zero**
2. **Constant Variance Assumption**: At any value of x, the population of potential error term values has a **variance** that does not depend on the value of x
3. **Normality Assumption**: 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

- 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

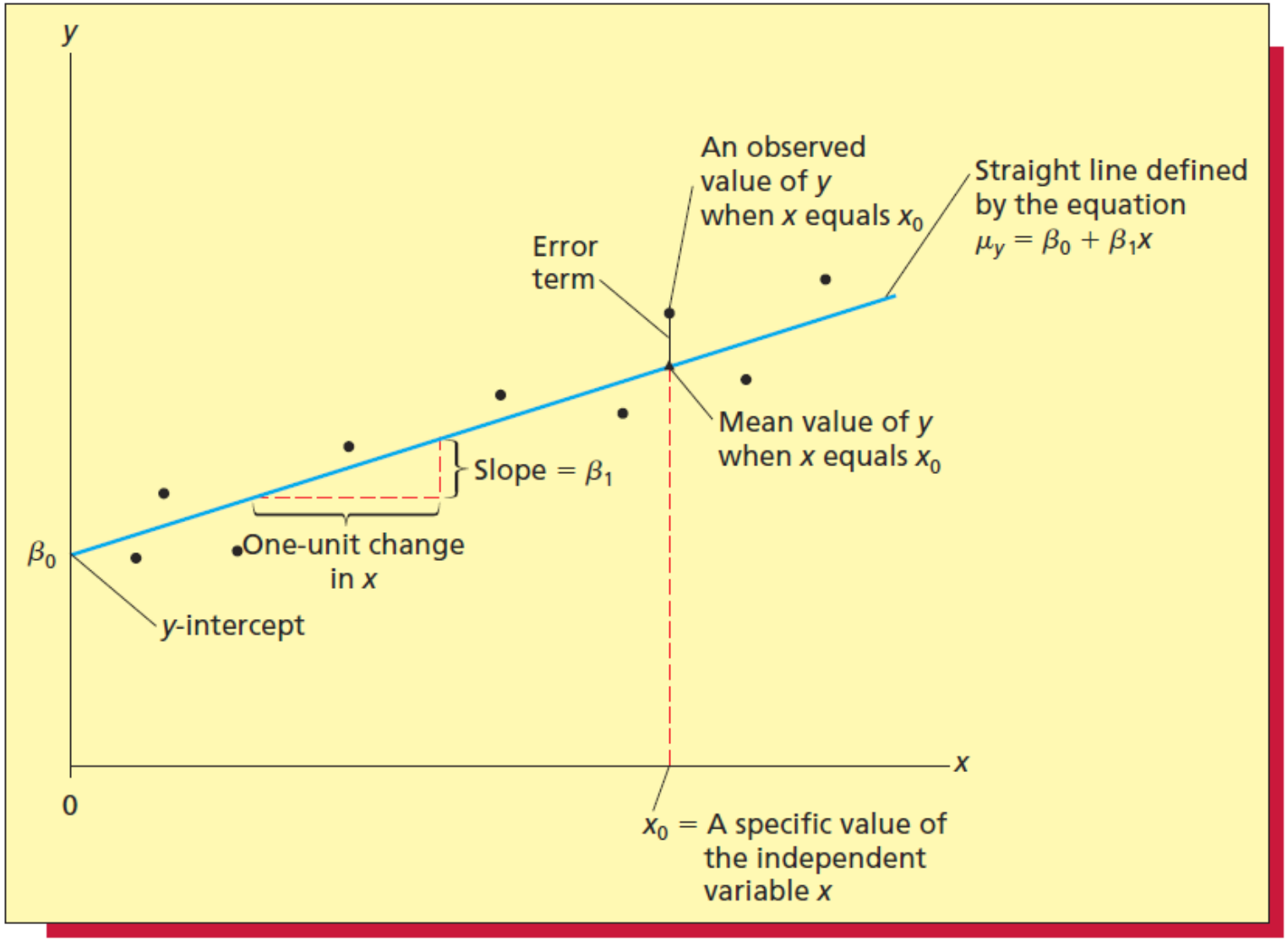
- β_0 and β_1 are called regression parameters
  - β_0 is the y-intercept
  - β_1 is the slope
- We **do not know** the true values of these parameters
- So, we must use sample data to **estimate** them
  - b_0 is the estimate of β_0
  - b_1 is the estimate of β_1

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

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

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

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

Least squares point estimate of the y−intercept β_0

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

where $\bar{y} = \frac{\sum y_i}{n}$ and $\bar{x} = \frac{\sum x_i}{n}$

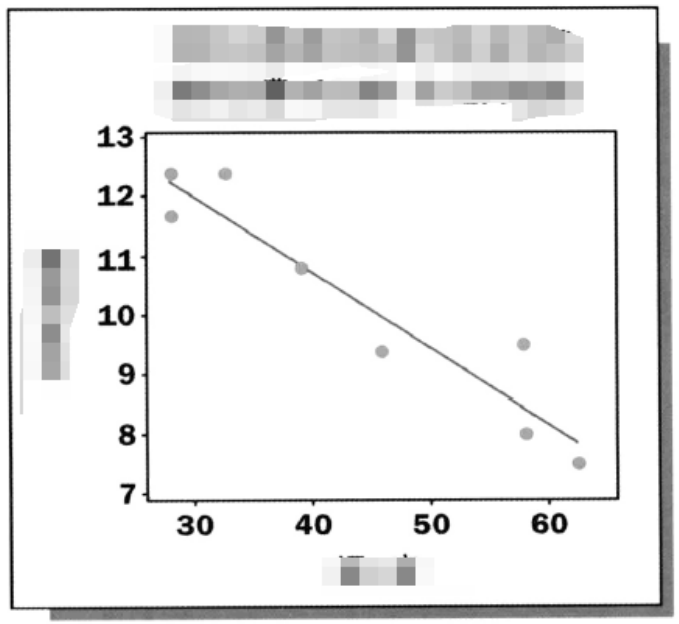
Example 13.3 The Fuel Consumption Case

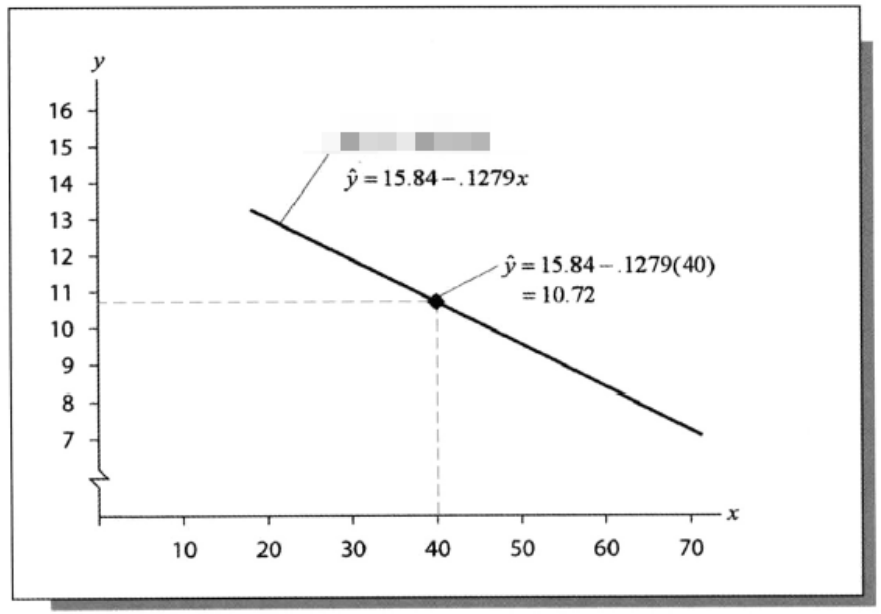
Temp	FuelCons
28.0	12.4
28.0	11.7
32.5	12.4
39.0	10.8
45.9	9.4
57.8	9.5
58.1	8.0
62.5	7.5

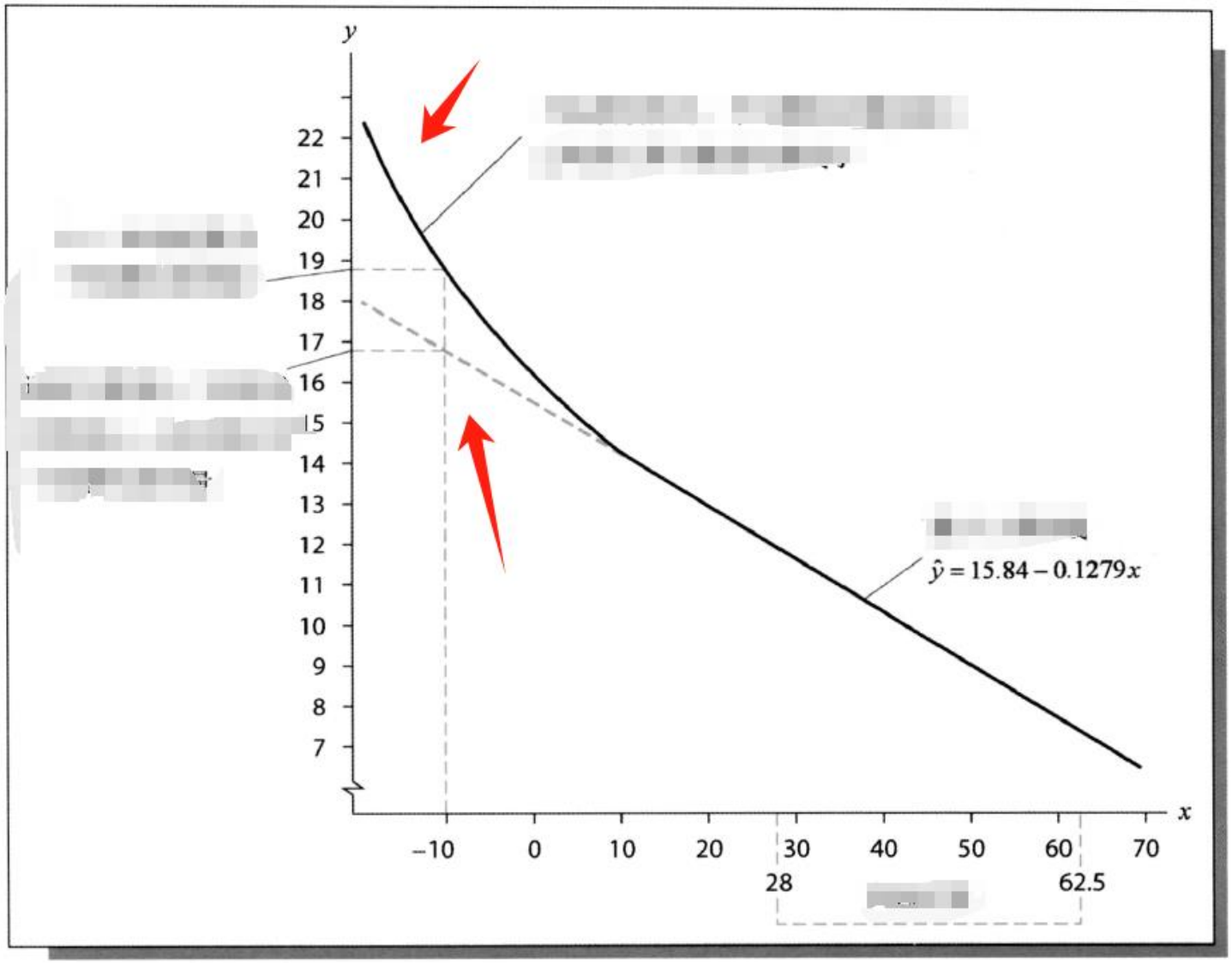
y_i	x_i	x^2_i	$x_i y_i$
12.4	28.0	$(28.0)^2 = 784$	$28.0 \times 12.4 = 347.2$
11.7	28.0	$(28.0)^2 = 784$	$28.0 \times 11.7 = 327.6$
12.4	32.5	$(32.5)^2 = 1\,056.25$	$32.5 \times 12.4 = 403$
10.8	39.0	$(39.0)^2 = 1\,521$	$39.0 \times 10.8 = 421.2$
9.4	45.9	$(45.9)^2 = 2\,106.81$	$45.9 \times 9.4 = 431.46$
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	$15.84 - 0.1279 \times 62.5 = 7.84625$	$7.5 - 7.84625 = -0.34625$

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

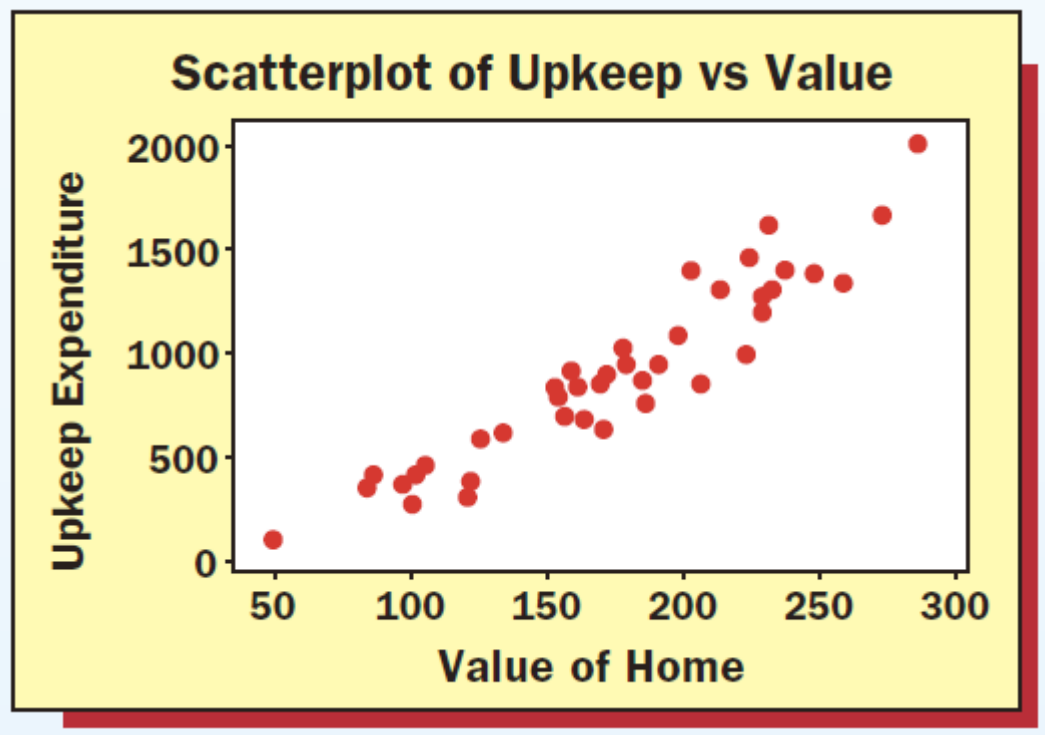





y
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
x
−10
0
10
20
30
40
50
60
70
28
62.5
$\hat{y} = 15.84 - 0.1279x$

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

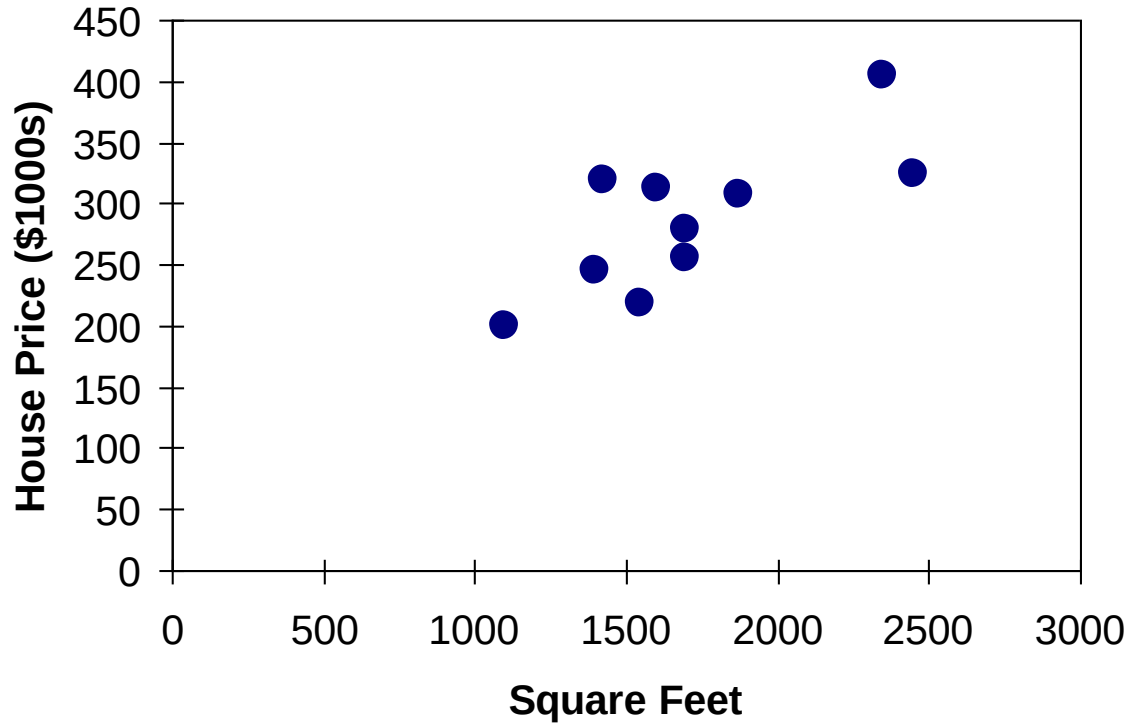
The House Price Case

- 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

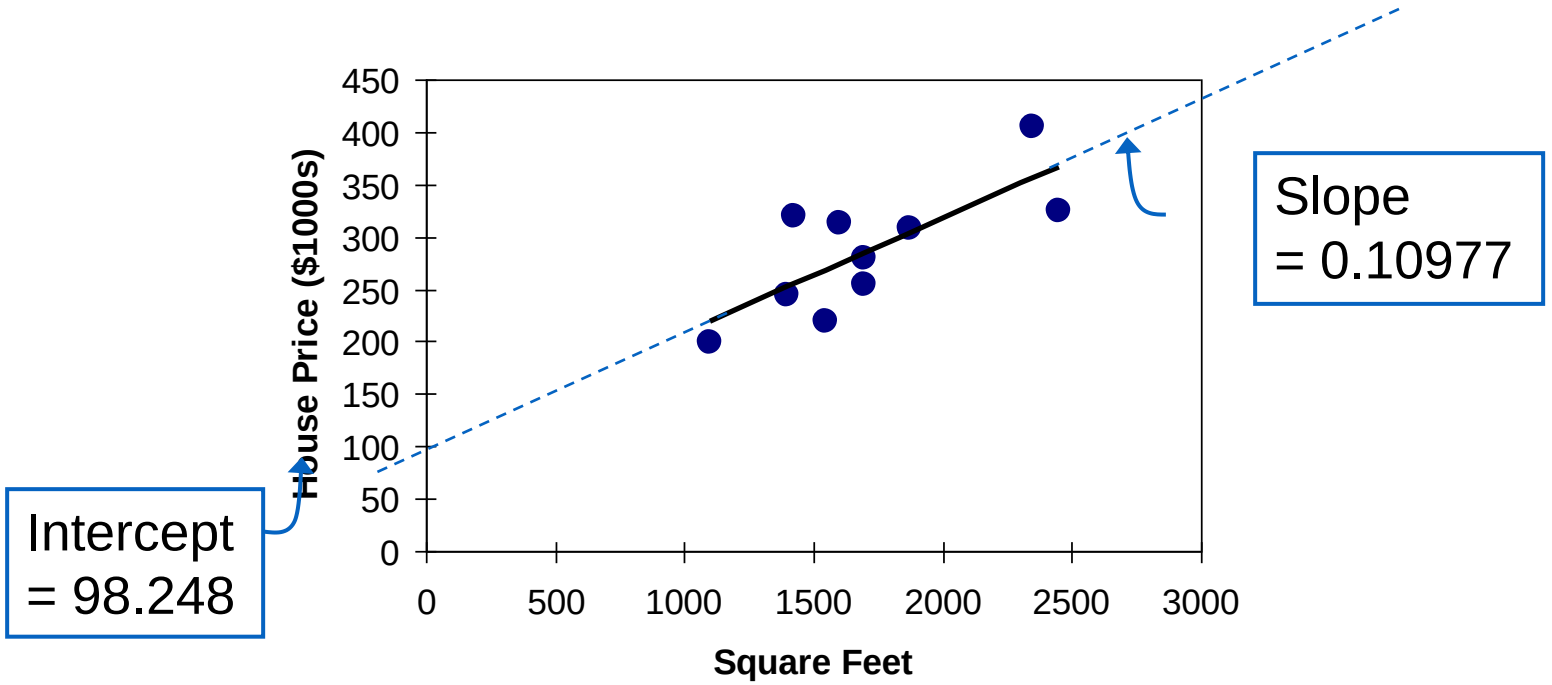
Graphical Presentation

- 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 x(1000) = $109.77, on average, for each additional one square foot of size

Predictions using Regression Analysis

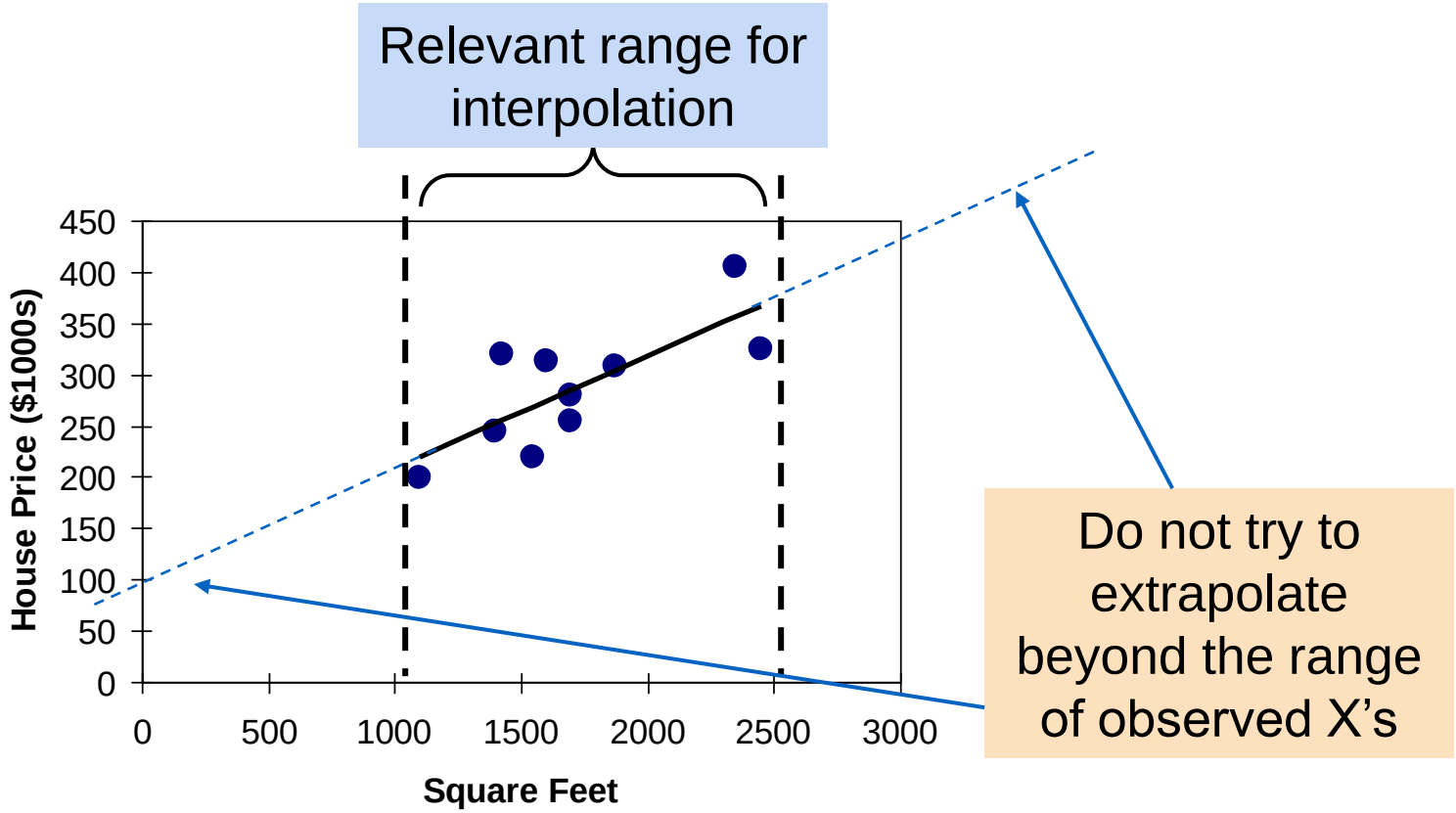
Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$
$$= 98.25 + 0.1098(2000)$$
$$= 317.85$$

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. The entrepreneur wishing to purchase a Tasty Sub franchise finds a suitable site, which consists of a suitable geographical location and suitable store space to rent. 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. 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). 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.

Example: The Tasty Sub Shop Case

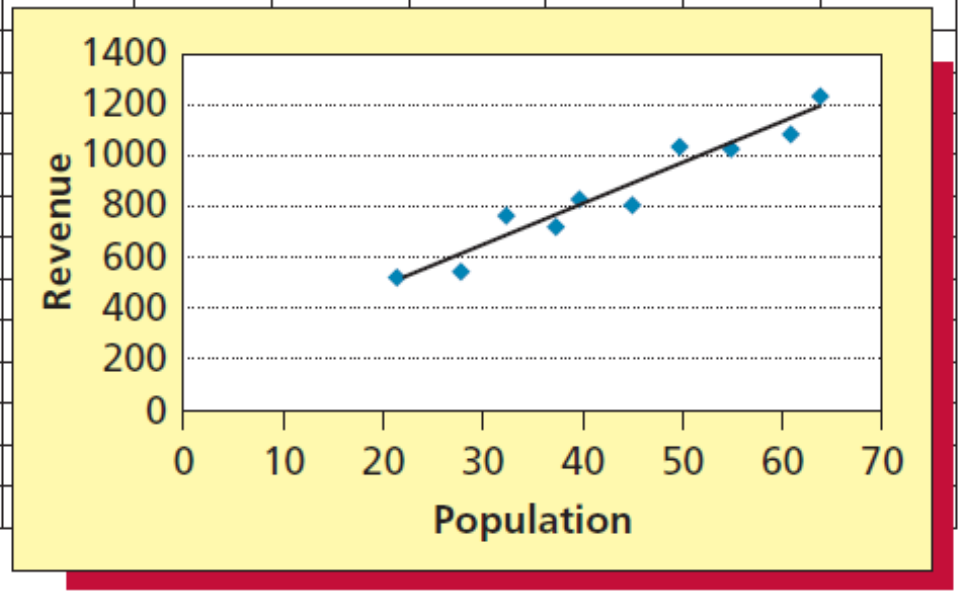
The Tasty Sub Shop Revenue Data

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

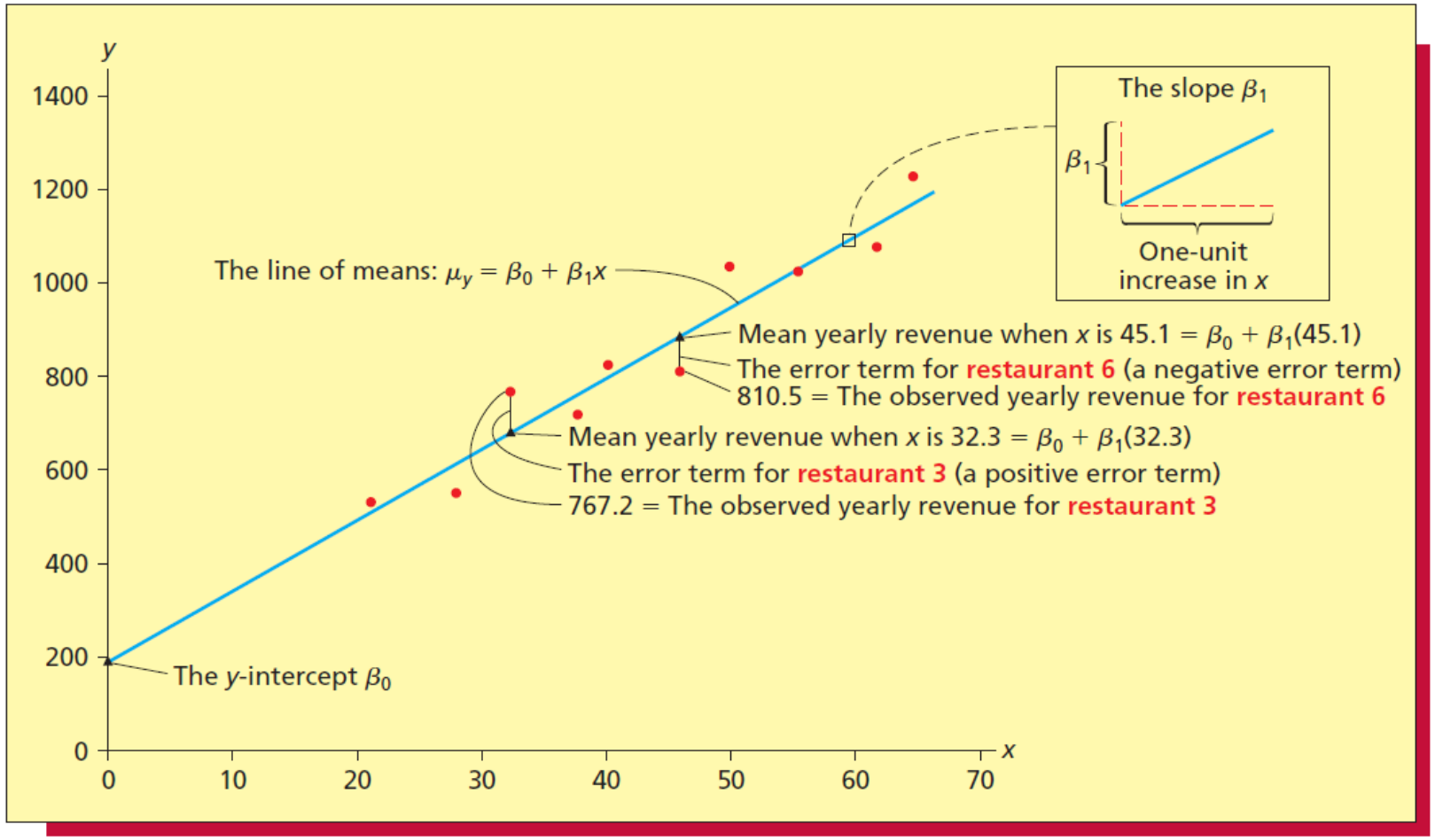
Excel Output of a Scatter Plot of y versus x

	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									



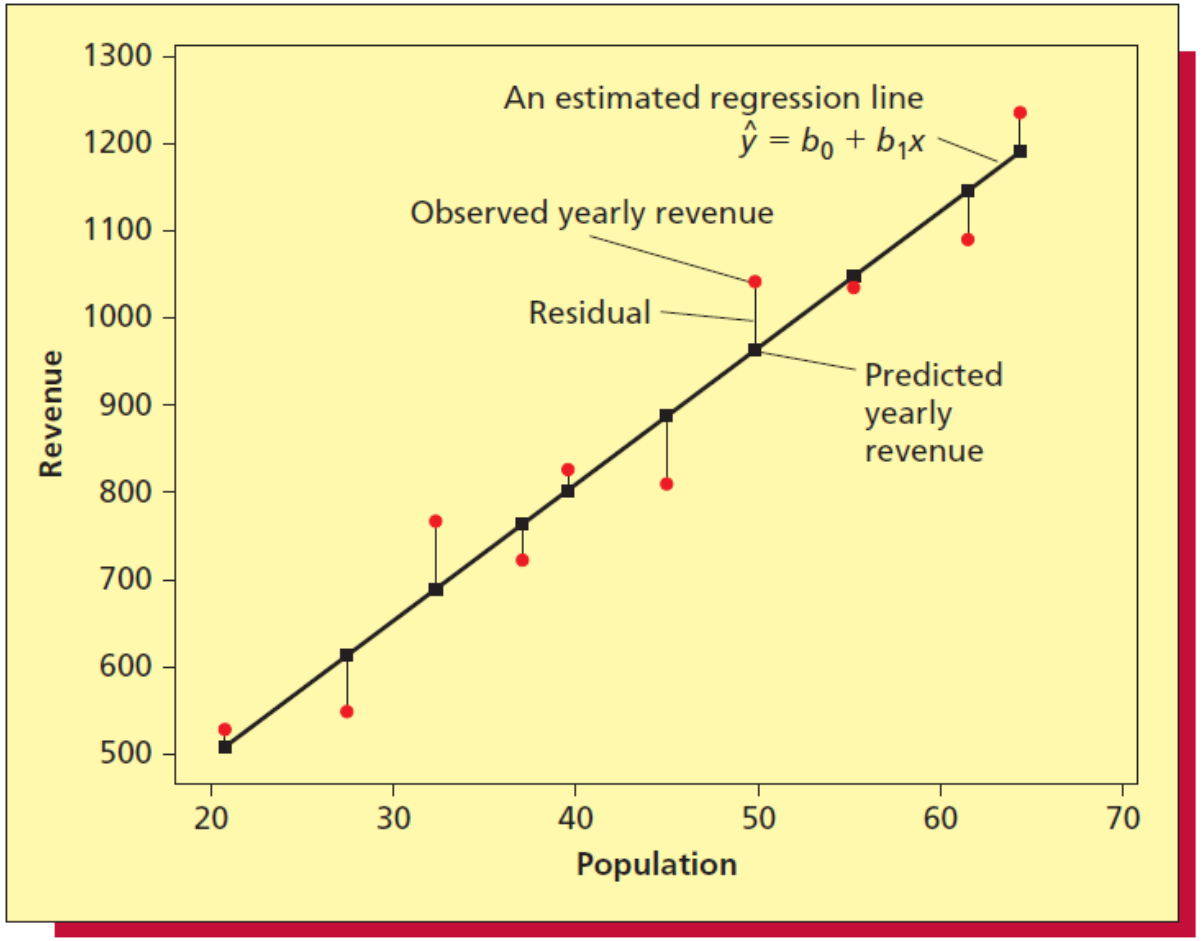
Example: The Tasty Sub Shop Case

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



Example: The Tasty Sub Shop Case

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



Example: The Tasty Sub Shop Case

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	$(64.6)^2 = 4{,}173.16$	$(64.6)(1235.8) = 79832.68$
$\sum y_i = 8603.1$	$\sum x_i = 434.1$	$\sum x_i^2 = 20{,}757.41$	$\sum x_i y_i = 403{,}296.96$

Example: The Tasty Sub Shop Case

- 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

Example: The Tasty Sub Shop Case

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

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

$$= 29{,}836.389$$

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

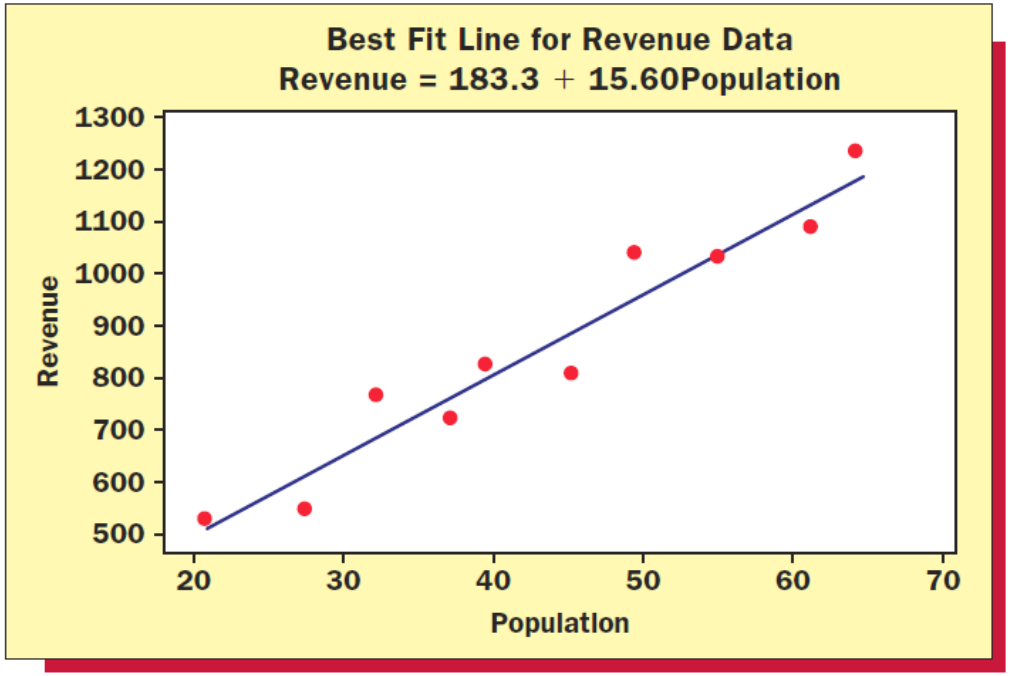
$$= 120{,}757.41 - \frac{(434.1)^2}{10} = 1{,}913.129$$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{29{,}836.389}{1{,}913.129} = 15.596$$

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

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

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



- 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

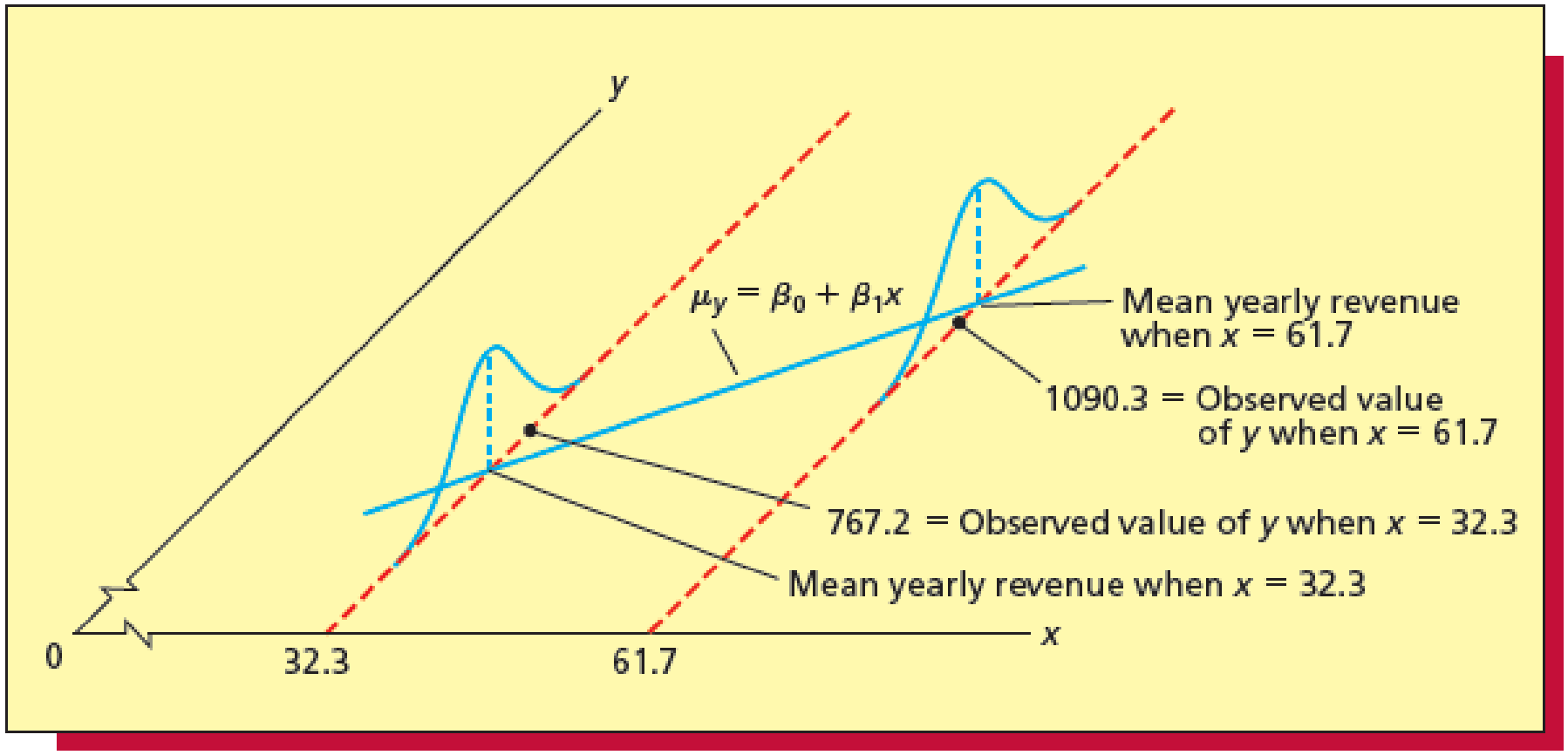
Simple Linear Regression Analysis

13.2 Model Assumptions and the Standard Error

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

Normality Assumption



- At any given value of x, the population of potential error term values has a normal distribution

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

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

Standard error: point estimate of residual standard deviation σ

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

Simple Linear Regression Analysis

13.4 Confidence and Prediction Intervals

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

Confidence interval for the slope

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).

Confidence intervals and prediction intervals

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

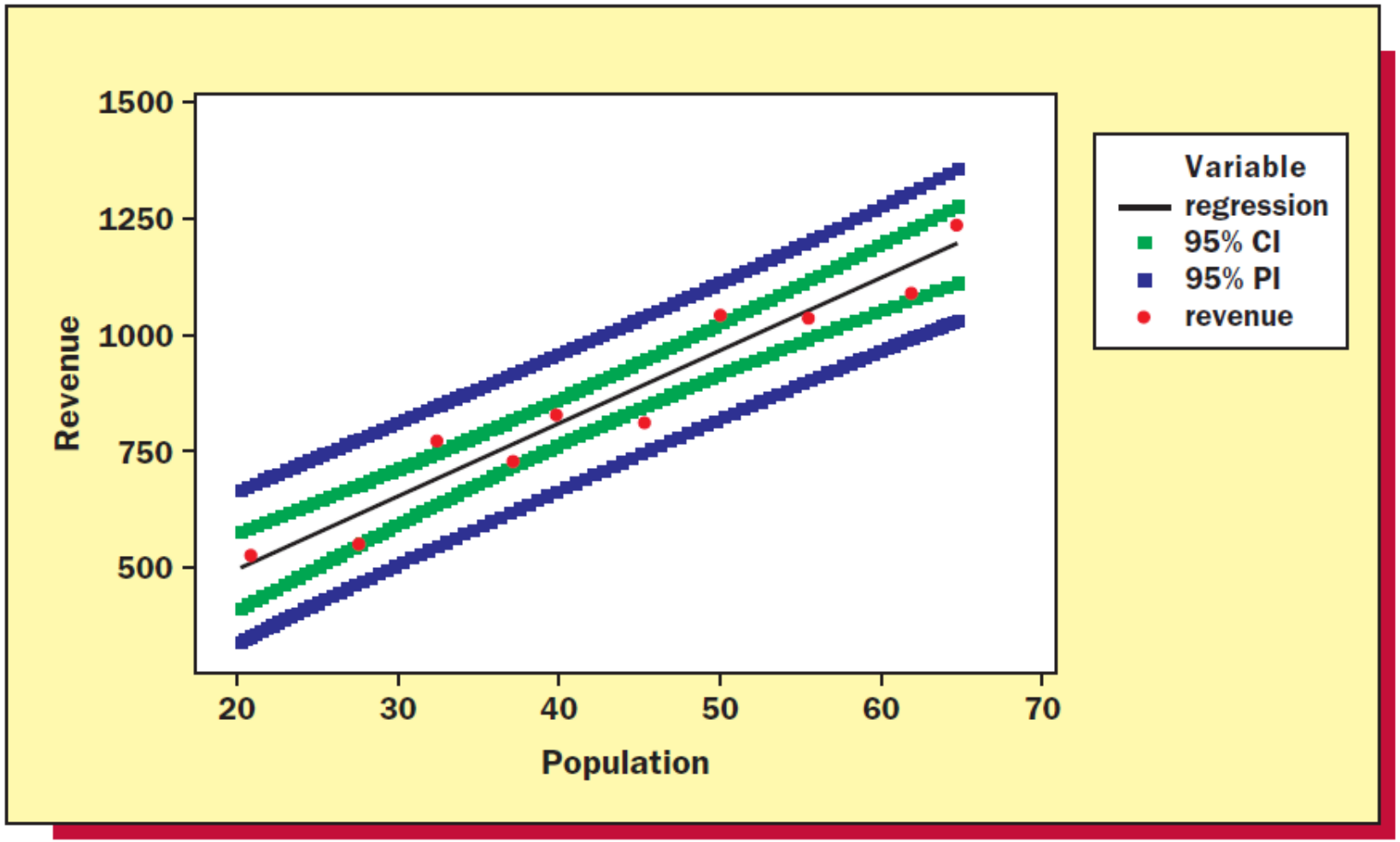
$$\begin{aligned}
&[\hat{y} \pm t_{\alpha/2} s \sqrt{\text{distance value}}] \\
&= [921.0 \pm 2.306(61.7052)\sqrt{.1079}] \\
&= [921.0 \pm 46.74] \\
&= [874.3, 967.7]
\end{aligned}$$

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.

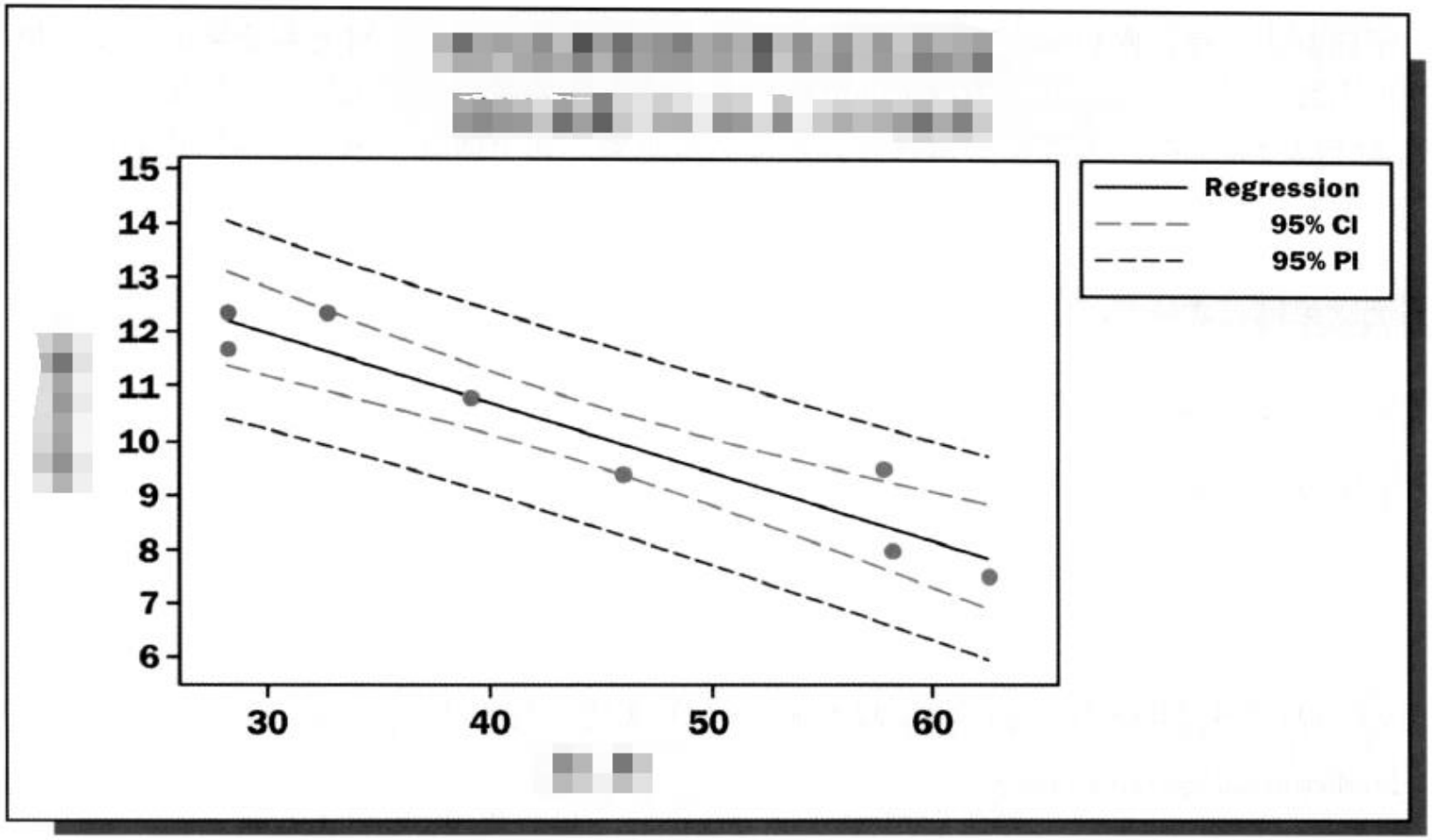
Because the entrepreneur would be operating a single Tasty Sub restaurant that is built near a population of 47,300 residents, the entrepreneur is interested in obtaining a prediction interval for the yearly revenue of such a restaurant. A 95 percent prediction interval for this revenue is

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

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



Example 13.9 The Fuel Consumption Case



Simple Linear Regression Analysis

13.5 Simple Coefficients of Determination and Correlation

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

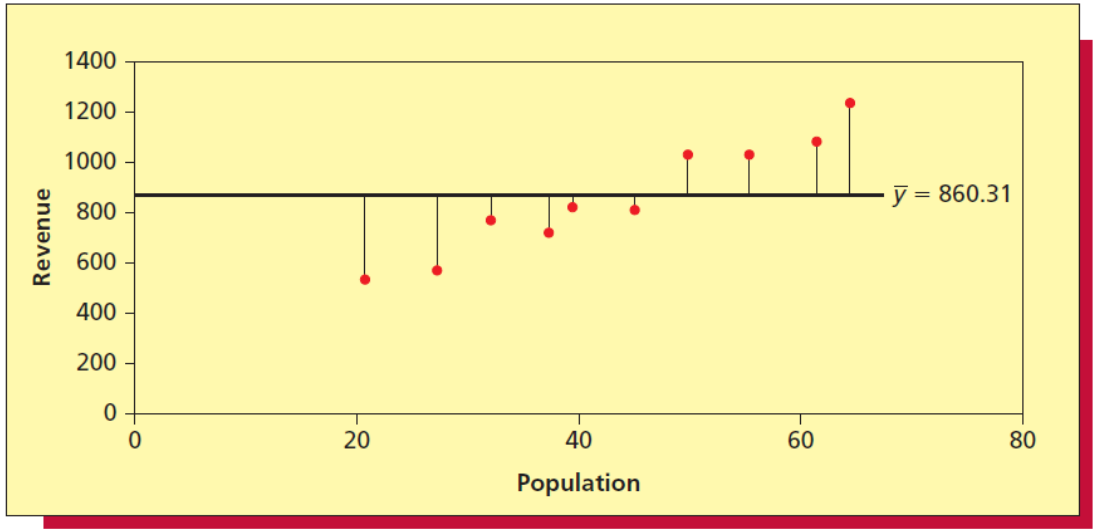
Calculating The Simple Coefficient of Determination

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**

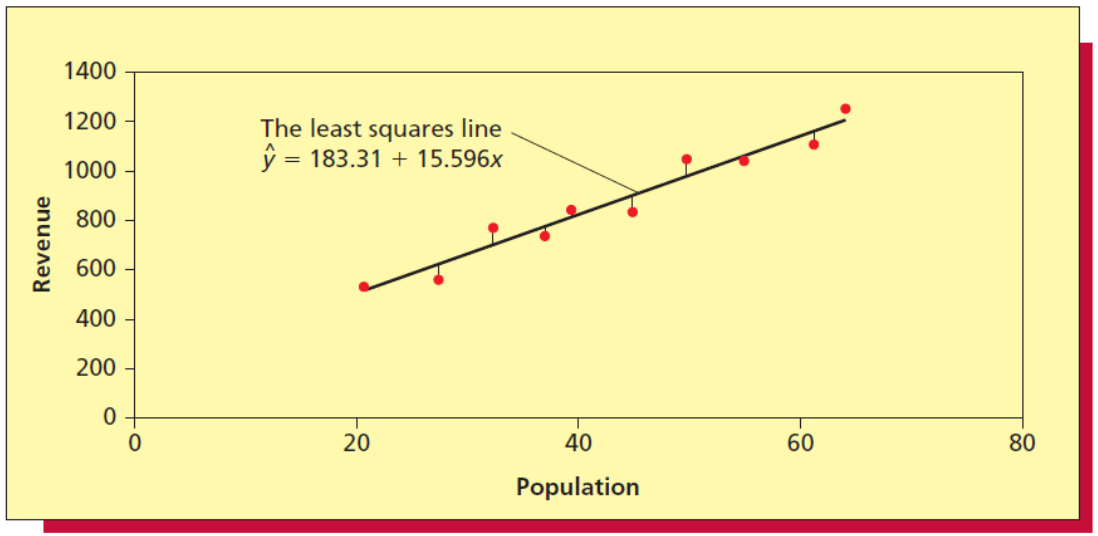
$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

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.

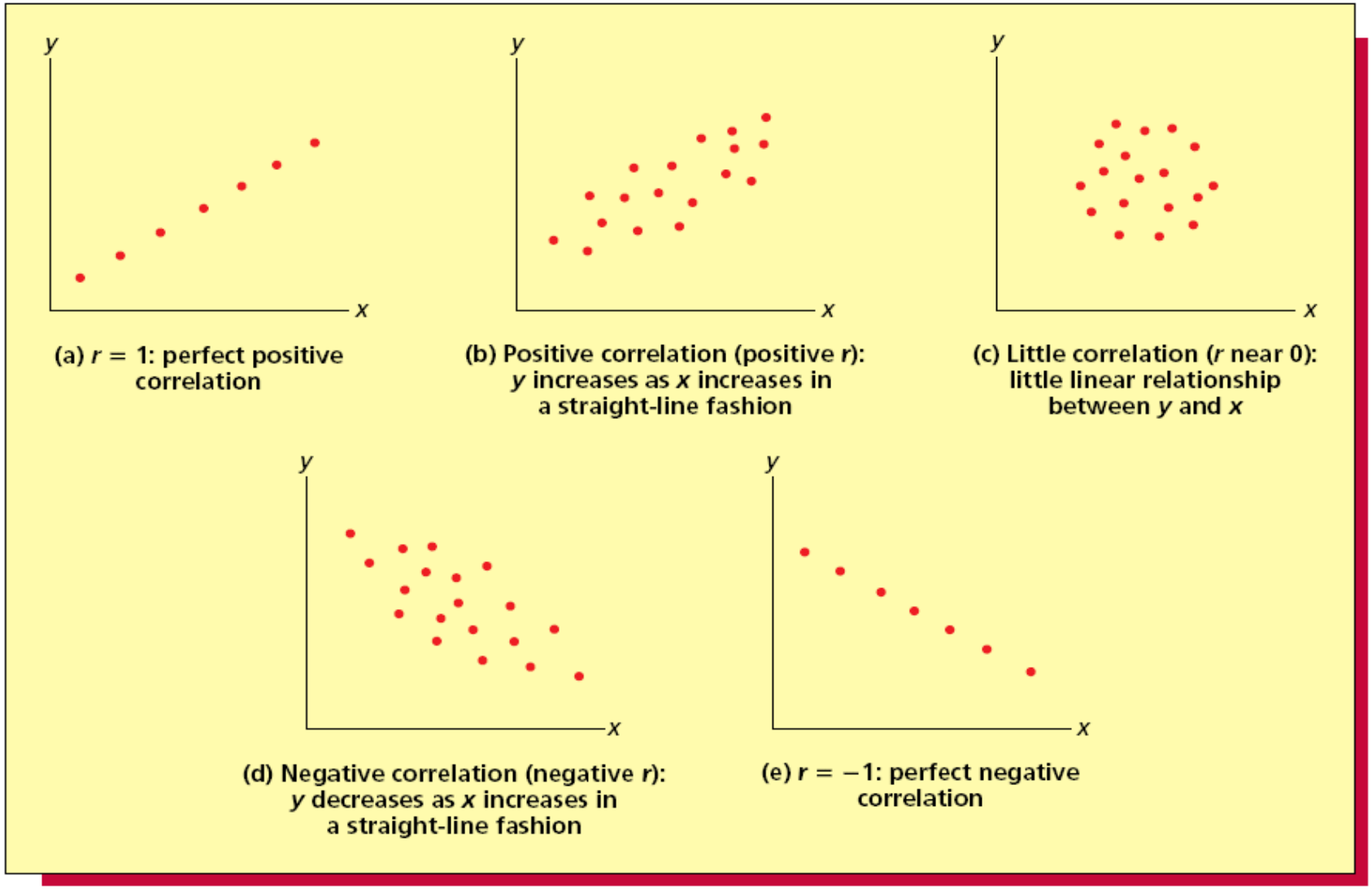
The Coefficient of Correlation (*r*).

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

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

- 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

Chapter Summary

- Discussed simple linear regression analysis, which relates a **dependent variable** to a single **independent** (predictor) **variable.**
- Simple linear regression model, which employs two parameters: the slope and y intercept
- **Least squares point estimates** of these parameters and how to use these estimates to calculate a **point estimate of the mean value of the dependent variable** and a **point prediction of an individual value** of the dependent variable.
- **Testing the significance of the regression relationship (slope),** calculating a **confidence interval** for the mean value of the dependent variable, and calculating a **prediction interval** for an individual value of the dependent variable.
- **Simple coefficient of determination**

Thank you!