# Chapter 4 – Describing the Relation between Two Variables

## OUTLINE

1. Scatter Diagrams and Correlation
2. Least-Squares Regression
3. Diagnostics on the Least-Squares Regression Line
4. Contingency Tables and Association

## Putting It Together

So far, we have examined data in which a single variable was measured for each individual in the study (univariate data), such as the 5-year rate of return (the variable) for various mutual funds (the individuals). We found both graphical and numerical descriptive measures for the variable.

Now, we discuss graphical and numerical methods for describing **bivariate data**, data in which two variables are measured on an individual. For example, we might want to know whether the amount of cola consumed per week is related to one’s bone density. The individuals would be the people in the study nd the two variables would be the amount of cola consumed weekly and bone density. In this study, both variables are quantitative.

Suppose we want to know whether level of education is related to one’s employment status (employed or unemployed). Here, both variables are qualitative.

Situations may also occur in which one variable is quantitative and the other is qualitative. We have already presented a technique for describing this situation. Look back at Example 3 in Section 3.5 where we considered whether space flight affected red blood cell mass. In this case, space flight is qualitative (rat sent to space or not) and red blood cell mass is quantitative.

## Section 4.1 Scatter Diagrams and Correlation

### Objectives

1. Draw and Interpret Scatter Diagrams
2. Describe the Properties of the Linear Correlation Coefficient
3. Compute and Interpret the Linear Correlation Coefficient
4. Determine Whether a Linear Relation Exists between Two Variables
5. Explain the Difference between Correlation and Causation

#### Objective 1: Draw and Interpret Scatter Diagrams

Introduction, Page 2

Define bivariate data

In Chapters 2 and 3, we examined data

in which a single variable was measured for each individual

in the study.

We call this univariate data, such as the five-year rate

of return, the variable, for various mutual funds,

the individuals.

The graphical measures that we found

for univariate data include things such as the histogram

or box plot.

Numerical descriptive measures included

things such as the mean, standard deviation,

or interquartile range.

In this chapter, we discussed graphical and numerical methods

for describing bivariate data.

That is, data in which two variables are

measured on an individual.

For example, we might want to know

whether the amount of cola consumed per week

is related to one's bone density.

The individuals would be the people in the study

and the two variables would be the amount of cola consumed

and bone density.

In this study, both variables are quantitative.

We present methods for describing

the relation between two quantitative variables

in Sections 4.1 to 4.3.

Suppose we want to know whether the level of education

is related to one's employment status, such as employed

versus unemployed.

Here, both variables are qualitative.

We present methods for describing

the relation between two qualitative variables

in Section 4.4.

Situations may also occur in which

one variable is quantitative and the other is qualitative.

We have already presented a technique

for describing this situation.

Look back at Example 3 in Section 3.5,

where we considered whether space flight affected red blood

cell mass.

There, space flight is qualitative--

rat sent to space or not--

and red blood cell mass is quantitative.

1. In this chapter, we discuss graphical and numerical methods for describing **bivariate data**, data in which two variables are measured on an individual. For example, we might want to know whether the amount of cola consumed per week is related to a person's bone density. The individuals would be the people in the study, and the two variables would be the amount of cola consumed weekly and bone density.

Before we can represent bivariate data graphically, we must decide which variable will be used to predict the value of the other variable. For example, it seems reasonable to think that as the speed at which a golf club is swung increases, the distance a golf ball travels also increases. Therefore, we might use club-head speed to predict distance. We call distance the *response* (*dependent*) *variable* and club-head speed the *explanatory* (or *predictor* or *independent*) *variable*.

Define response variable and explanatory variable. **DEFINITION**

The **response (dependent) variable** is the variable whose value can be explained by the value of the **explanatory**(or**predictor** or **independent**)**variable**.

Objective 1, Page 1

What is a scatter diagram? How is it created? The first step in identifying the type of relation that might exist between two variables is to draw a picture. We can represent bivariate data graphically with a *scatter diagram.*

**DEFINITION**

A **scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

Objective 1, Page 2

**Example 1 *Drawing a Scatter Diagram***

A golf pro wants to investigate the relation between the club-head speed of a golf club (measured in miles per hour) and the distance (in yards) the ball will travel. He realizes that other variables besides club-head speed also determine the distance a ball will travel (such as club type, ball type, golfer, and weather conditions). To eliminate the variability due to these variables, the pro uses a single model of club and ball, one golfer, and a clear, 70-degree day with no wind. The pro records the club-head speed, measures the distance the ball travels, and collects the data in Table 1. Draw a scatter diagram of the data.

**Table 1**

| **Club-Head Speed (mph)** | **Distance (yards)** |
| --- | --- |
| 100 | 257 |
| 102 | 264 |
| 103 | 274 |
| 101 | 266 |
| 105 | 277 |
| 100 | 263 |
| 99 | 258 |
| 105 | 275 |

#### Data from Paul Stephenson, student at Joliet Junior College

#### StatCrunch

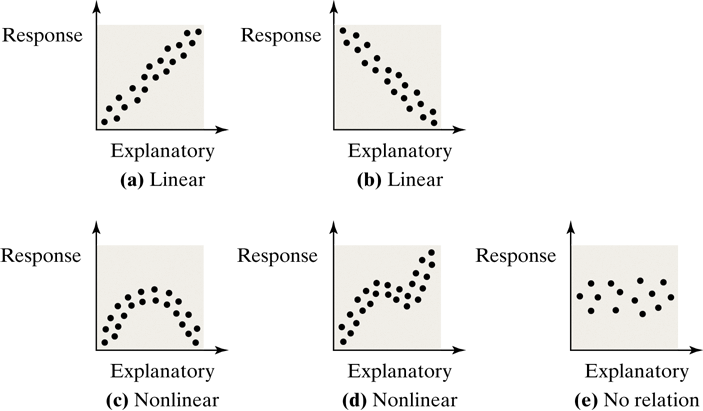
#### Draw a Scatter Diagram

1. If necessary, enter the explanatory variable in column var1 and the response variable in column var2. Name each column variable.
2. Select **Graph** and highlight **Scatter Plot.**
3. Choose the explanatory variable for the X column and the response variable for the Y column. Enter the labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.

Objective 1, Page 5

#### Explain how to determine which variable is the explanatory variable and which variable is the response variable. Deciding Which Variable Is the Explanatory Variable and the Response Variable

It is not always clear which variable should be considered the response variable and which the explanatory variable. For example, does high school GPA predict a student's SAT score or can the SAT score predict GPA? The researcher must determine which variable plays the role of explanatory variable based on the questions he or she wants answered. For example, if the researcher wants to predict SAT scores based on high school GPA, then high school GPA is the explanatory variable



Objective 1, Page 6

1. Sketch a scatterplot that shows a nonlinear relation between an explanatory variable and a response variable.above
2. Sketch a scatterplot that shows no relation between an explanatory variable and a response variable.above

Objective 1, Page 7

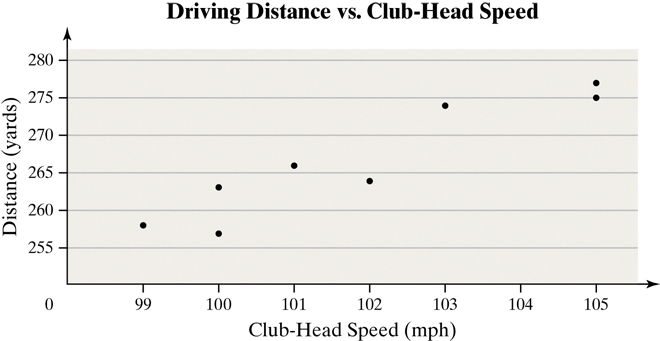
Explain how to determine if two variables are positively associated. **DEFINITION**

Two variables that are linearly related are **positively associated** when above-average values of one variable are associated with above-average values of the other variable (or below-average values of one variable are associated with below-average values of the other variable). That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases

Explain how to determine if two variables are negatively associated. Two variables that are linearly related are **negatively associated** when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

#### Objective 2: Describe the Properties of the Linear Correlation Coefficient

Objective 2, Page 1

It is dangerous to use only a scatter diagram to determine if two variables are linearly related. Just as we can manipulate the scale of graphs of univariate data, we can also manipulate the scale of graphs of bivariate data, possibly resulting in incorrect conclusions. 



Objective 2, Page 2

Define the linear correlation coefficient and state the formula for the sample linear correlation coefficient. **DEFINITION**

The **linear correlation coefficient,** or **Pearson product moment correlation coefficient,** is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formula for the sample correlation coefficient.

**Sample Linear Correlation Coefficient**

r=∑(xi−x¯¯¯/sx)(yi−y¯¯¯/sy)**/**n−1

|  |  |
| --- | --- |
| where | xi is the ith observation of the explanatory variable |
|  | x¯¯¯ is the sample mean of the explanatory variable |
|  | sx is the sample standard deviation of the explanatory variable |
|  | yi is the ith observation of the response variable |
|  | y¯¯¯ is the sample mean of the response variable |
|  | sy is the sample standard deviation of the response variable |
|  | n is the number of individuals in the sample |

Objective 2, Page 3

*Answer the following after watching the video.*

In the formula for the linear correlation coefficient, notice that the numerator is the sum of the products of *z*-scores for the explanatory (*x*) and response (*y*) variables. If the linear correlation coefficient is positive, that means that the sum of the *z*-scores for *x* and *y* must be positive. How does that occur? 00:01>> On the screen, we have a scatter diagram

that shows positive association between an explanatory variable

x and a response variable y.

What I've done on the scatter diagram

is drawn vertical lines at the mean of x and a horizontal line

at the mean for y.

What these values do is divide my scatter diagram

into four quadrants.

Let's call the points that are in this region

points in quadrant I, here's quadrant II, quadrant III,

and quadrant IV.

Now, let's focus in on the points that lie in quadrant I.

If I take any point, say this point right here,

you'll notice that that point has

a value of x that is above average,

and it has a value of y that's above average.

What does that mean in terms of the formula for correlation?

Well, because the x-value is greater than xbar,

we know the z-score for x is going to be positive.

Because the y-value is greater than ybar,

we know the z-value for y is also going to be positive.

And in general, any of the points in quadrant

I are going to have positive z-scores for x and positive

z-scores for y, and, therefore, this product

is going to be positive.

If we consider points that are in quadrant III,

say this point right here, you'll

notice that it is below average for x,

and it is below average for y.

That means the z-score for x is going to be negative,

and the z-score for y is going to be negative.

Negative times negative gives you a positive value,

so all the points in quadrant III

are going to have a positive product of the z-scores.

The point that lies in quadrant VI and the point that

lies in quadrant II are both going

to have negative products of z-scores.

And the point for quadrant VI, the x-value is above average,

but the y-value is below average,

so I'd have a positive times a negative.

And the point in quadrant II, the x-value is below average,

but the y-value is above average, so I'd have a pos--

I'm sorry-- I'd have a negative z-score for x

and a positive z-score for y giving me a negative product.

However, because an overwhelming number of points

are going to have positive z-scores,

the sum of all these z-scores is going to be positive,

and, therefore, I'm going to end up

with a positive linear correlation coefficient.

Now what if we had negative association

between the explanatory variable x and the response variable y?

Similar logic applies.

Let's consider the points in quadrant II.

For this point, we have a below average value for x,

but we have an above average value for y.

So the z-scores for x for all the points in quadrant II

are going to be negative, and the z-scores

for y for the points in quadrant II are going to be positive.

So we're going to have a negative times a positive,

which gives us a negative result.

For the points in quadrant VI, say this point right here,

you'll notice that the value of x

is above average, while the value of y is below average.

And so the z-score for x for all the points in quadrant IV

is going to be positive, while the z-score for y

is going to be negative.

So we're going have a positive times a negative,

which is also negative.

And so for the points in quadrants II and VI,

we're basically summing together a bunch

of negative values, which in itself is going to be negative.

And that's why the correlation coefficient

when you have negative association

is a negative value.

Lastly, if you had no association between x and y,

in other words, all the points were equally dispersed

amongst the four quadrants, basically,

* the positive products for the z-scores
* would be offset by the negative products for the z-scores.

And you would end up with a correlation coefficient

that is close to 0.

Objective 2, Page 4

*Answer the following as you work through Activity 1: Properties of the Correlation Coefficient.*

1. When the points are aligned in a straight line with positive slope, what is the value of the linear correlation coefficient?
2. When the points are aligned in a straight line with negative slope, what is the value of the linear correlation coefficient?
3. Does a correlation coefficient close to 0 imply that there is no relation? Why or why not?
4. Is the linear correlation coefficient resistant?

Objective 2, Page 5

*Watch the video for a summary of the ideas from Activity 1.*

Objective 2, Page 6

**Note: Properties of the Linear Correlation Coefficient**

The linear correlation coefficient is always between ** and 1, inclusive. That is, 

If *r* = + 1, then a perfect positive linear relation exists between the two variables.

If  then a perfect negative linear relation exists between the two variables.

The closer *r* is to + 1, the stronger is the evidence of positive association between the two variables.

The closer *r* is to ** the stronger is the evidence of negative association between the two variables.

If *r* is close to 0, then little or no evidence exists of a linear relation between the two variables. So *r* close to 0 does not imply no relation, just no linear relation.

The linear correlation coefficient is a unitless measure of association. So the unit of measure for *x* and *y* plays no role in the interpretation of *r*. INSTRUCTOR: First and foremost, the correlation coefficient

is a number between negative 1 and 1.

It's a number between negative 1 and 1.

Now, this data set that I'm creating here,

what kind of relationship does it

look like I'm creating between x and y?

STUDENT: Negative [INAUDIBLE].

INSTRUCTOR: Negative what?

STUDENT: [INAUDIBLE]

INSTRUCTOR: Negative-- so she [INAUDIBLE],,

yeah, near-perfect.

When I ask it to show R, what it that linear correlation

coefficient down there?

Can you see it?

It says negative 0.9985.

Pretty close to negative 1.

In fact, if you add a linear correlation coefficient

of negative 1, the data would lie on a straight line,

perfectly on a straight line.

What I'm going to do is flip-flop this

and make it positive slope.

And so when I show R, the correlation is 0.999.

So when you have data that lies on a straight line

with no deviation, the correlation

is going to equal positive 1.

Now, if I asked you to describe that relation,

would you say there's a linear relation between these two

variables?

STUDENT: No.

INSTRUCTOR: Not really at all.

In other words, if I said, hey, use x to figure out y,

would you have much confidence?

STUDENT: [INAUDIBLE]

INSTRUCTOR: No.

And when I show R, I get negative 0.02,

basically meaning there is virtually-- it's

pretty darn close to 0, right?

So when correlation coefficients are close to 0,

it would imply what kind of relation?

STUDENT: No relation.

INSTRUCTOR: All right, no relation

is what we're saying at this stage of the game.

But I'm going to throw a curve ball at you, OK?

Would you say that this is no relation?

STUDENT: It's quadratic.

INSTRUCTOR: Yeah, this looks like it's pretty quadratic.

In other words, if I told you what x was, you might say,

I can use this relationship to get

a pretty decent guess for a y.

Would you agree?

Look at the correlation, 0.044.

STUDENT: [INAUDIBLE] the correlation functionality

applies to [INAUDIBLE]?

INSTRUCTOR: There you go.

The correlation coefficient only applies to data

that is linearly related.

Just because a correlation is 0, or near 0,

does not mean that there is no relation, just

no linear relation between the two variables.

What kind of correlation do you think that is?

STUDENT: [INAUDIBLE]

INSTRUCTOR: Not so much.

Let's see.

0.2, not so much.

How about that one?

What kind of slope?

STUDENT: Positive.

INSTRUCTOR: Positive slope.

Correlation of 0.8.

How about that one?

Pretty close, 0.88.

Little bit better, even.

How about this one?

Negative what?

Negative with a linear relation, so what

would you guess a good correlation might be?

Probably not negative 1, because it

would be on a straight line with a negative slope, correct?

Negative 0.8 we're going to guess.

Let's see.

Hey, that is a dang good guess.

Negative 0.8.

Now, there's one other idea that I want to show you here.

What kind of relationship appears to exist there?

Not too much, right?

I get R is negative 0.13.

So you might not feel too comfortable

using the explanatory variable to figure out the response.

But there is a slight negative association, but probably

not very, to use a word that we're going to learn later,

significant.

But watch this.

I'm just taking this one point and dragging it way over there.

Now remember, the correlation when the point was over here

was native 0.13, agree?

Now it's negative 0.83.

If I said to you, the correlation

between two variables is negative 0.83,

would you feel pretty good about using x to explain y?

Now, don't look at the scatter diagram.

I'm just telling you this correlation is negative 0.83.

STUDENT: [INAUDIBLE]

INSTRUCTOR: Most people would probably

say, wow, that's pretty strong linear relation.

But then I show you this picture.

Now what do you say?

No way.

It's out-- yeah, well, we know what outlier means now, right?

Yeah.

So when one observation has a significant effect

on the value of a statistic, what do we call that?

STUDENT: Not resistant.

INSTRUCTOR: Not resistant.

So is the linear correlation coefficient

a resistant measure of linear association?

STUDENT: [INAUDIBLE]

INSTRUCTOR: No, it's not.

No, it's not.

So we learn quite a bit about correlation

just from playing around with this little lab,

but did we not?

The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

#### Objective 3: Compute and Interpret the Linear Correlation Coefficient

Objective 3, Page 1

Linear correlation coefficients are typically found using technology. Example 2 shows how it would be computed by hand to help show how it measures the strength of a linear relation.

Objective 3, Page 2

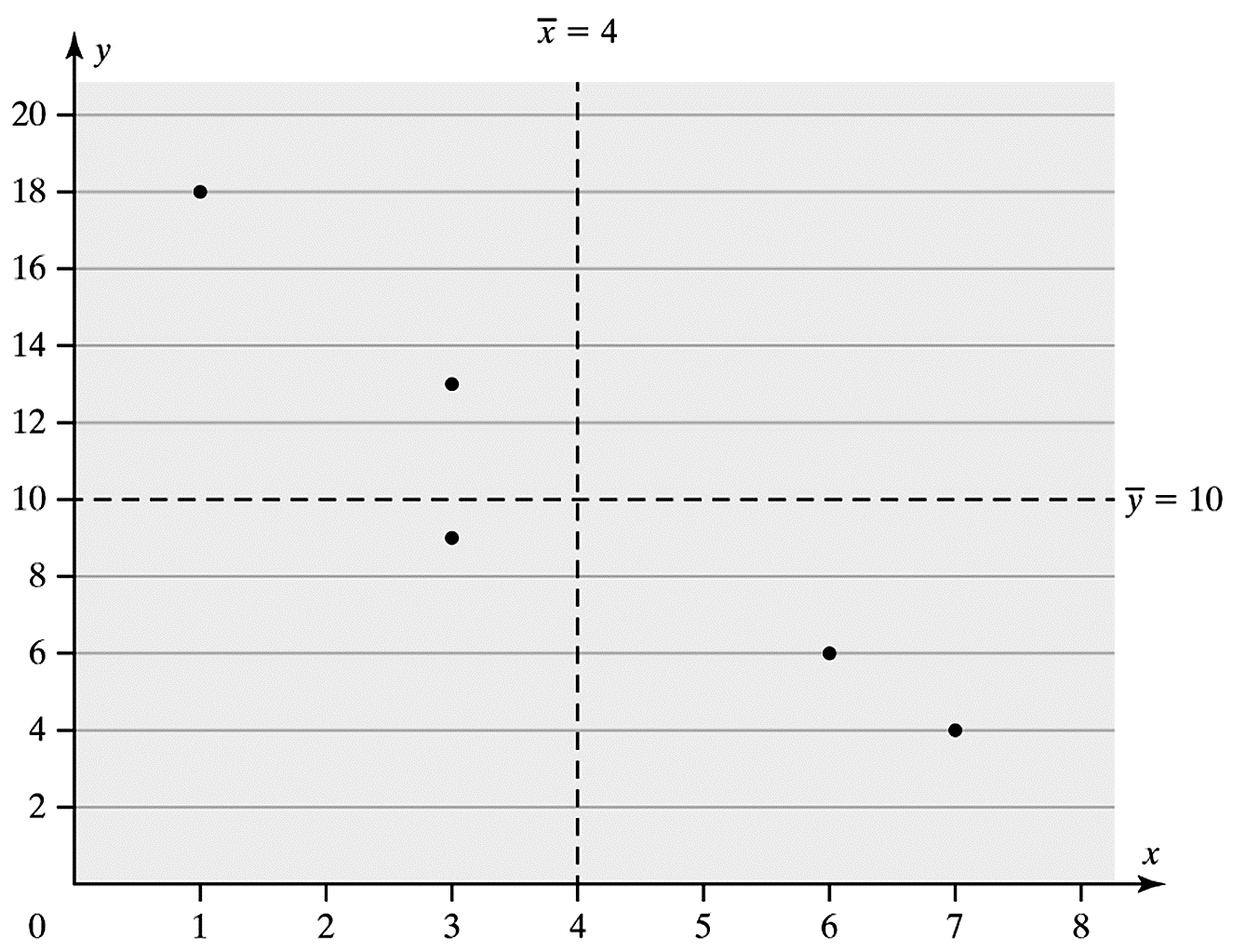
**Example 2 *Computing the Correlation Coefficient by Hand***

For the data shown in Table 2, compute the linear correlation coefficient. A scatter diagram of the data is shown in Figure 5. The dashed lines on the scatter diagram represent the mean of x and y.

**Table 2**

| ***x*** | ***y*** |
| --- | --- |
| 1 | 8 |
| 3 | 13 |
| 3 | 9 |
| 6 | 6 |
| 7 | 4 |

**Figure 5**



Objective 3, Page 3

In the scatter diagram, notice that below-average values of *x* are associated with above-average values of *y* and above-average values of *x* are associated with below-average values of *y*. This helps to explain why the linear correlation coefficient is negative.

Objective 3, Page 5

**Example 3 *Determining the Linear Correlation Coefficient Using Technology***

Use a statistical spreadsheet or a graphing calculator with advanced statistical features to determine the linear correlation coefficient between club-head speed and distance from the data in Table 1. Interpret the linear correlation coefficient.

**Table 1**

| **Club-Head Speed (mph)** | **Distance (yards)** |
| --- | --- |
| 100 | 257 |
| 102 | 264 |
| 103 | 274 |
| 101 | 266 |
| 105 | 277 |
| 100 | 263 |
| 99 | 258 |
| 105 | 275 |

Data from Paul Stephenson, student at Joliet Junior College

#### Objective 4: Determine Whether a Linear Relation Exists between Two Variables

Objective 4, Page 1

1. List the three steps for testing for a linear relation.

Objective 4, Page 2

**Example 4 *Does a Linear Relation Exist?***

We have been analyzing the association between club-head speed and distance the ball travels. A scatter diagram suggests a positive association between the variables. The linear correlation coefficient of 0.939 also suggests a positive association between the variables. Use these results to determine whether a linear relation exists between club-head speed and distance.

#### Objective 5: Explain the Difference between Correlation and Causation

Objective 5, Page 1

1. If data are obtained from an experiment, can we claim a causal relationship between the explanatory and response variables? How about if the data are obtained from an observational study?
2. Is there another way two variables can be correlated without a causal relationship existing?

Objective 5, Page 2

**Example 5 *Lurking Variables in a Bone Mineral Density Study***

Because cola tends to replace healthier beverages and cola contains caffeine and phosphoric acid, researchers Katherine L. Tucker and associates wanted to know if cola consumption is associated with lower bone mineral density in women. Table 4 lists the typical number of cans of cola consumed in a week and the bone mineral density for a sample of 15 women. The data were collected through a prospective cohort study.

Figure 7 shows the scatter diagram of the data. The correlation between number of colas per week and bone mineral density is  The critical value for correlation with *n* = 15 from Table II is 0.514. Because  we conclude that a negative linear relation exists between number of colas consumed and bone mineral density. Can the authors conclude that an increase in the number of colas consumed causes a decrease in bone mineral density? Identify some lurking variables in the study.

## Section 4.2 Least-Squares Regression

### Objectives

1. Find the Least-Squares Regression Line and Use the Line to Make Predictions
2. Interpret the Slope and the *y*-Intercept of the Least-Squares Regression Line
3. Compute the Sum of Squared Residuals

Introduction, Page 2

**Example 1 *Finding an Equation That Describes Linearly Related Data***

The data in Table 1 represent the club-head speed and the distance a golf ball travels for eight swings of the club.

**Table 1**

| **Club-Head Speed (mph)** | **Distance (yards)** | **(*x*, *y*)** |
| --- | --- | --- |
| 100 | 257 | (100, 257) |
| 102 | 264 | (102, 264) |
| 103 | 274 | (103, 274) |
| 101 | 266 | (101, 266) |
| 105 | 277 | (105, 277) |
| 100 | 263 | (100, 263) |
| 99 | 258 | (99, 258) |
| 105 | 275 | (105, 275) |

Data from Paul Stephenson, student at Joliet Junior College

1. Find a linear equation that relates club-head speed x (the explanatory variable) and distance y (the response variable) by selecting two points and finding the equation of the line containing the points.
2. Graph the line on the scatter diagram.
3. Use the equation to predict the distance a golf ball will travel if the club-head speed is 104 miles per hour.

#### Objective 1: Find the Least-Squares Regression Line and Use the Line to Make Predictions

Objective 1, Page 1

1. What is the residual for an observation?

Objective 1, Page 3

1. What is the least-squares regression line?

Objective 1, Page 4

*Answer the following as you work through Activity 1: What is Least Squares?*

1. As your line gets closer and closer to the regression line, what happens to the SSE?

Objective 1, Page 5

1. List the formulas associated with the least-squares regression line.

Objective 1, Page 6

**Key Ideas about the Least-Squares Regression Line**

* The least-squares regression line, **, always contains the point .
* Because  and  must both be positive, the sign of the linear correlation coefficient, *r*, and the sign of the slope of the least-squares regression line, , are the same.
* The predicted value of *y*, , has an interesting interpretation. It is an estimate of the mean value of the response variable for any value of the explanatory variable.

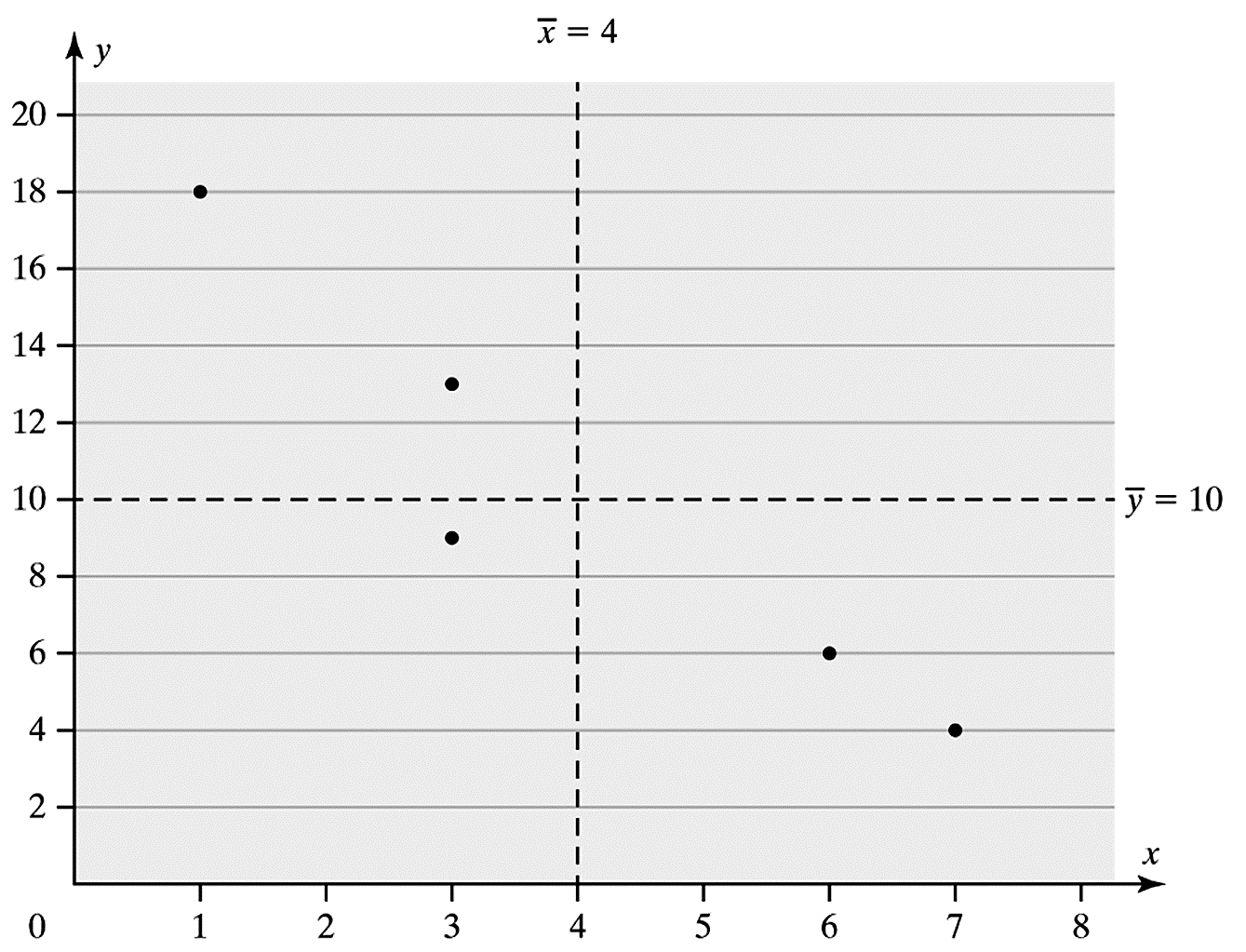
Objective 1, Page 8

**Example 2 *Finding the Least-Squares Regression Line by Hand***

Find the least-squares regression line for the data in Table 2 from Section 4.1. For convenience, we also provide a scatter diagram of the data.

**Table 2**

| ***x*** | ***y*** |
| --- | --- |
| 1 | 18 |
| 3 | 13 |
| 3 | 9 |
| 6 | 6 |
| 7 | 4 |



Objective 1, Page 9

Throughout the course, we agree to round the slope and *y*-intercept to four decimal places.

In Example 2 we found a least-squares regression line to see the role that the correlation coefficient, *r*, the standard deviation of *y*, and the standard deviation of *x* play in finding the slope.

In practice, the least-squares regression line is found using technology.

Objective 1, Page 10

**Example 3 *Finding the Least-Squares Regression Line Using Technology***

Use the golf data in Table 1.

**Table 1**

| **Club-Head Speed (mph)** | **Distance (yards)** |
| --- | --- |
| 100 | 257 |
| 102 | 264 |
| 103 | 274 |
| 101 | 266 |
| 105 | 277 |
| 100 | 263 |
| 99 | 258 |
| 105 | 275 |

Data from Paul Stephenson, student at Joliet Junior College

1. Find the equation of the least-squares regression line using technology.
2. Draw the least-squares regression line on the scatter diagram of the data.
3. Predict the distance a golf ball will travel when hit with a club-head speed of 103 miles per hour (mph).
4. Determine the residual for the predicted value found in part (c). Is the distance above or below average among all balls hit with a swing speed of 103 mph?

#### Objective 2: Interpret the Slope and the y-Intercept of the Least-Squares Regression Line

Objective 2, Page 1

1. The slope of the least-squares regression line from Example 3 is 3.1661 yards per mph. List two interpretations of the slope that are acceptable.

Objective 2, Page 2

1. Before interpreting the *y*-intercept, what two questions must be asked?
2. What does it mean to say that we should not use the regression model to make predictions outside the scope of the model?

Objective 2, Page 3

1. When the correlation coefficient indicates no linear relation between the explanatory and response variables and the scatter diagram indicates no relation between the variables, how do we find a predicted value for the response variable?

#### Objective 3: Compute the Sum of Squared Residuals

Objective 3, Page 1

1. What does the least-squares regression line minimize?

Objective 3, Page 2

**Example 4 *Computing the Sum of Squared Residuals***

Compare the sum of squared residuals for the linear models in Examples 1 and 3.

Objective 3, Page 3

The total area of the red squares represent the sum of the squared residuals for the least-squares regression line. Any other line will produce a larger sum of the squared residuals, such as the sum of the areas of the green squares.

## Section 4.3 Diagnostics on the Least-Squares Regression Line

### Objectives

1. Compute and Interpret the Coefficient of Determination
2. Perform Residual Analysis on a Regression Model
3. Identify Influential Observations

#### Objective 1: Compute and Interpret the Coefficient of Determination

Objective 1, Page 1

1. Define the coefficient of determination, .
2. The coefficient of determination is a number between 0 and 1, inclusive. That is,  What does it mean if ? What does it mean if ?

Objective 1, Page 2

The differences between the observed value (*y*), the mean value  of the response variable, and the predicted value  are called deviations.

Objective 1, Page 4

1. Explain the meaning of total deviation, explained deviation, and unexplained deviation.
2. The closer the observed *y*’s are to the regression line (the predicted *y*’s), how is  affected?

Objective 1, Page 6

1. Explain how to find the coefficient of determination, , for the least-squares regression model *.*

Objective 1, Page 7

**Example 1 *Determining and Interpreting the Coefficient of Determination***

Determine and interpret the coefficient of determination, R2, for the club-head speed versus distance data shown in Table 1.

**Table 1**

| **Club-Head Speed (mph)** | **Distance (yards)** |
| --- | --- |
| 100 | 257 |
| 102 | 264 |
| 103 | 274 |
| 101 | 266 |
| 105 | 277 |
| 100 | 263 |
| 99 | 258 |
| 105 | 275 |

Data from Paul Stephenson, student at Joliet Junior College

Objective 1, Page 9

 *Fill in the table as you work through Activity 1: Understanding the Coefficient of Determination.*

| **Data Set** | **Coefficient of Determsination,** | **Interpretation** |
| --- | --- | --- |
| A | \_\_\_\_\_\_\_ | \_\_\_\_ of the variability in y is explained by the least-squares regression line. |
| B | \_\_\_\_\_\_\_ | \_\_\_\_ of the variability in y is explained by the least-squares regression line. |
| C | \_\_\_\_\_\_\_ | \_\_\_\_ of the variability in y is explained by the least-squares regression line. |

#### Objective 2: Perform Residual Analysis on a Regression Model

Objective 2, Page 1

1. List the three purposes for which we analyze residuals.

Objective 2, Page 2

 *Answer the following after watching the video.*

1. What is a residual plot?
2. If a plot of the residuals against the explanatory variable shows a discernable pattern, what does this say about the explanatory and response variables?

Objective 2, Page 3

 *Answer the following after watching the video.*

1. Why is it important for the residuals to have constant error variance?

Objective 2, Page 4

 *Answer the following after watching the third video.*

1. List two ways that we can use residuals to identify outliers.

Objective 2, Page 5

**Example 2 *Graphical Residual Analysis***

The data in Table 6 represent the club-head speed, the distance the golf ball traveled, and the residuals (Column 4) for eight swings of a golf club. Construct a residual plot and boxplot of the residuals and comment on the appropriateness of the least-squares regression model.

**Table 6**

| **Club-Head Speed (mph)** | **Example 3 Distance(yd)** |  | **Residual** |
| --- | --- | --- | --- |
| 100 | 257 | 260.8 |  |
| 102 | 264 | 267.1 |  |
| 103 | 274 | 270.3 | 3.7 |
| 101 | 266 | 264.0 | 2.0 |
| 105 | 277 | 276.6 | 0.4 |
| 100 | 263 | 260.8 | 2.2 |
| 99 | 258 | 257.6 | 0.4 |
| 105 | 275 | 276.6 |  |

#### Objective 3: Identify Influential Observations

Objective 3, Page 1

1. What is an influential observation?

Objective 3, Page 2

 *Answer the following after working through Activity 2: Influential Observations.*

1. For the point that had the greatest effect on the slope and *y*-intercept, how did its residual compare to the other points you added? How did its *x*-value compare to the other points you added?

Objective 3, Page 5

1. What do we call the relative vertical position of an observation? What do we call the relative horizontal position of an observation?
2. List a combination of leverage and residual that indicates that an observation may be influential.

Objective 3, Page 6

**Example 3 *Identifying Influential Observations***

Suppose that our golf ball experiment calls for nine trials, but the player we are using hurts his wrist. Luckily, Bubba Watson (a professional golfer) is practicing on the same range and graciously agrees to participate in our experiment. His club-head speed is 120 miles per hour, and he hits the golf ball 305 yards. Is Bubba's shot an influential observation? For convenience, we present the data from Table 1.

**Table 1**

| **Club-Head Speed (mph)** | **Distance (yards)** |
| --- | --- |
| 100 | 257 |
| 102 | 264 |
| 103 | 274 |
| 101 | 266 |
| 105 | 277 |
| 100 | 263 |
| 99 | 258 |
| 105 | 275 |

Data from Paul Stephenson, student at Joliet Junior College

Objective 3, Page 7

As with outliers, influential observations should be removed only if there is justification to do so. When an influential observation occurs in a data set and its removal is not warranted, two possible courses of action are to (1) collect more data so that additional points near the influential observation are obtained or (2) use techniques that reduce the influence of the influential observation. (These techniques are beyond the scope of this text.)

## Section 4.4 Contingency Tables and Association

### Objectives

1. Compute the Marginal Distribution of a Variable
2. Use the Conditional Distribution to Identify Association Among Categorical Data
3. Explain Simpson's Paradox

Introduction, Page 1

1. What is a contingency table?

#### Objective 1: Compute the Marginal Distribution of a Variable

Objective 1, Page 1

1. What is a marginal distribution of a variable?

Objective 1, Page 2

**Example 1 *Determining Frequency Marginal Distributions***

The data (measured in thousands) in Table 8 represent the employment status and level of education of all civilian, noninstitutional residents (excludes inmates and individuals in the armed services) who are 25 years and older in May 2017. Find the frequency marginal distributions for employment status and level of education.

**Table 8**

| **Employment Status** | **Level of Education Did Not Finish High School** | **High School Graduate** | **Some College** | **Bachelor's Degree or Higher** |
| --- | --- | --- | --- | --- |
| Employed | 9671 | 34,211 | 35,941 | 53,760 |
| Unemployed | 628 | 1697 | 1492 | 1278 |
| Not in the Labor Force | 12,564 | 26,406 | 19,346 | 19,524 |

*Data from Bureau of Labor Statistics*

Objective 1, Page 4

**Example 2 *Determining Relative Frequency Marginal Distributions***

Determine and interpret the relative frequency marginal distribution for level of education and employment status from the data in Table 9.

**Table 9**

| **Employment Status** | **Did Not Finish High School** | **High School Graduate** | **Some College** | **Bachelor's Degree or Higher** | **Frequency Marginal Distribution for Employment Status** |
| --- | --- | --- | --- | --- | --- |
| **Employed** | 9671 | 34,211 | 35,941 | 53,760 | 133,583 |
| **Unemployed** | 628 | 1697 | 1492 | 1278 | 5095 |
| **Not in the Labor Force** | 12,564 | 26,406 | 19,346 | 19,524 | 77,840 |
| **Frequency Marginal Distribution for Level of Education** | 22,863 | 62,314 | 56,779 | 74,562 | 216,518 |

#### Objective 2: Use the Conditional Distribution to Identify Association Among Categorical Data

Objective 2, Page 1

Marginal distributions allow us to see the distribution of either the row variable or the column variable, but we do not get a sense of association between the two variables from these tables.

To learn about any association that may exist, we need a different table.

Objective 2, Page 2

**Example 3 *Comparing Two Categories of a Variable***

Use the data in Table 9 to answer parts (A) through (D).

* 1. What proportion of those who did not finish high school is employed?
  2. What proportion of those who are high school graduates is employed?
  3. What proportion of those who finished some college is employed?
  4. What proportion of those who have at least a Bachelor's degree is employed?

**Table 9**

| **Employment Status** | **Did Not Finish High School** | **High School Graduate** | **Some College** | **Bachelor's Degree or Higher** | **Frequency Marginal Distribution for Employment Status** |
| --- | --- | --- | --- | --- | --- |
| **Employed** | 9671 | 34,211 | 35,941 | 53,760 | 133,583 |
| **Unemployed** | 628 | 1697 | 1492 | 1278 | 5095 |
| **Not in the Labor Force** | 12,564 | 26,406 | 19,346 | 19,524 | 77,840 |
| **Frequency Marginal Distribution for Level of Education** | 22,863 | 62,314 | 56,779 | 74,562 | 216,518 |

Objective 2, Page 3

1. What is a conditional distribution?

The variable we condition upon represents the explanatory variable in a conditional distribution, and the remaining variable becomes the response variable.

Objective 2, Page 4

**Example 4 *Constructing a Conditional Distribution***

Find the conditional distribution of the response variable employment status by level of education, the explanatory variable, for the data in in Table 9. What is the association between level of education and employment status?

**Table 9**

| **Employment Status** | **Did Not Finish High School** | **High School Graduate** | **Some College** | **Bachelor's Degree or Higher** | **Frequency Marginal Distribution for Employment Status** |
| --- | --- | --- | --- | --- | --- |
| **Employed** | 9671 | 34,211 | 35,941 | 53,760 | 133,583 |
| **Unemployed** | 628 | 1697 | 1492 | 1278 | 5095 |
| **Not in the Labor Force** | 12,564 | 26,406 | 19,346 | 19,524 | 77,840 |
| **Frequency Marginal Distribution for Level of Education** | 22,863 | 62,314 | 56,779 | 74,562 | 216,518 |

Objective 2, Page 7

**Example 5 *Drawing a Bar Graph of a Conditional Distribution***

Using the results of Example 4, draw a bar graph that represents the conditional distribution of employment status by level of education.

Objective 2, Page 9

The methods presented for identifying the association between two categorical variables are different from the methods used for measuring association between two quantitative variables. The measure of association is based on whether there are differences in the relative frequencies of the response variable (employment status) for the different categories of the explanatory variable (level of education). If differences exist, we might attribute these differences to the explanatory variable.

#### Objective 3: Explain Simpson's Paradox

Objective 3, Page 1

A lurking variable can cause two quantitative variables to be correlated even though they are unrelated. The same phenomenon exists when we explore the relation between two qualitative variables.

Objective 3, Page 2

**Example 6 *Gender Bias at the University of California, Berkeley***

The data in Table 12 show the admission status and gender of students who applied to the University of California, Berkeley. From the data in Table 12, the proportion of accepted applications is . The proportion of accepted men is , and the proportion of accepted women is . On the basis of these proportions, a gender bias suit was brought against the university. The university was shocked and claimed that program of study is a lurking variable that created the apparent association between admission status and gender. The university supplied Table 13 in its defense. Develop a conditional distribution by program of study and use it to defend the university's admission policies.

(Data from P. J. Bickel, E. A. Hammel, and J. W. O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley." Science 187(4175): 398-404, 1975.)

**Table 12**

|  | **Accepted (A)** | **Not Accepted (NA)** | **Total** |
| --- | --- | --- | --- |
| **Men** | 1191 | 1399 | 2590 |
| **Women** | 557 | 1278 | 1835 |
| **Total** | 1748 | 2677 | 4425 |

**Table 13**

**ADMISSION STATUS (ACCEPTED, A, OR NOT ACCEPTED, NA), FOR SIX PROGRAMS OF STUDY (A, B, C, D, E, F) BY GENDER**

|  | **A** | **A** | **B** | **B** | **C** | **C** | **D** | **D** | **E** | **E** | **F** | **F** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Men** | A | NA | A | NA | A | NA | A | NA | A | NA | A | NA |
| **Men** | 511 | 314 | 353 | 207 | 120 | 205 | 138 | 279 | 53 | 138 | 16 | 256 |
| **Women** | A | NA | A | NA | A | NA | A | NA | A | NA | A | NA |
| **Women** | 89 | 19 | 17 | 8 | 202 | 391 | 131 | 244 | 94 | 299 | 24 | 317 |

Objective 3, Page 3

Simpson’s Paradox describes a situation in which an association between two variables inverts or goes away when a third variable is introduced to the analysis.