

# Correlation and Linear Regression

CHAPTER

6

## CONNECTIONS: CHAPTER

In Chapter 5 we learned how to display and describe quantitative data, one variable at a time. In this chapter we learn how to display and assess the relationship between two quantitative variables. This is so important a topic that not only does it have a special name, but it will also be revisited later in the text (Chapter 14). Linear regression is the counterpart to the contingency tables in Chapter 4. You can think of the three Chapters—4, 5, and 6—as a set that cover descriptive statistics for categorical and quantitative data, with one variable or two variables.



Niar/Shutterstock

## LEARNING OBJECTIVES

- 1 Analyze a scatterplot to identify possible relationships in bivariate data
- 2 Calculate and interpret a correlation coefficient
- 3 Compute and interpret a linear regression equation
- 4 Use a linear regression equation for prediction
- 5 Calculate and interpret R-squared
- 6 Compute and interpret residuals
- 7 Distinguish between correlation and causation

## RONA

**R**ONA, Inc., Canada's largest retailer of hardware, home renovation, and gardening products, began in 1939, when hardware store operators in Quebec formed a cooperative called Les Marchands en Quincailleries Ltee. to circumvent a monopoly in the hardware supply business. It survived the rise of department store chains through the 1940s and 1950s, and in 1960 a sister company, Quincaillerie Ro-Na Inc., was established. According to popular legend, the name came from the first two letters of the first names of Rolland Dansereau and Napoleon Pottie, the company's first presidents.

Other regional hardware cooperatives sprang up in Canada, but Ro-Na remained inside Quebec. In the 1980s Ro-Na made purchases and alliances, acquiring gardening, interior decorating, and building materials stores, and brought them under one roof as a big-box retailer. Ro-Na and Home Hardware (in Western Canada) formed Alliance RONA Home Inc.

Major changes were also taking place in the U.S. hardware industry, with the rise of Home Depot. The fierce competition between Home Depot and Canadian rivals began in Ontario but only spread to Quebec in 1998.

While Ro-Na's sister company, Home Hardware, continued with small and medium-sized stores, other Canadian chains, such as

Surrey, British Columbia-based Revy Home Centres, Inc., did big-box battle with Home Depot in suburban Toronto. Ro-Na rolled out its own big-box stores (calling them “large surface” stores), but hoped to distinguish them from the competition with great customer service.

During Canada’s booming economy in the 1990s home improvement was a flourishing business. Each of the several large Canadian chains needed a strategy to keep market share. Ro-Na chose acquisition. In 1998, the company changed its name to RONA, Inc., opened new stores, and made its first foray into Ontario. By 1999, RONA had almost 500 stores under various banners in eastern Canada, and sales rose to \$2.1 billion.

RONA gained a coast-to-coast presence in the 2000s and equalled or surpassed Home Depot in market share, while making sure its stores were noticeably different from the American competitor. RONA, Inc. debuted on the Toronto Stock Exchange in October 2002. The company was now a public corporation, and no longer a cooperative. While RONA and Home Depot were the two biggest contenders in the Canadian hardware market, half the market was still shared by small independent hardware stores, along with the Home Hardware cooperative and Canadian Tire.

RONA’s strategy was to have stores in every segment of the market. It would continue to open large-surface stores as well as small neighbourhood stores. It attached RONA to the name of all of its stores. RONA wanted to “take the warehouse out of the warehouse concept.” Instead, its stores would offer an enticing shopping experience explained as “Disney meets home improvement.” RONA used ideas from other successful chains. And the company strove to please women, who made the majority of home improvement buying decisions. But recent years have also been difficult. A slow recovery following the 2008 world economic crisis, lower consumer confidence, and a slowdown in the housing market have all had a major effect on RONA’s growth. The first half of 2011 was a particularly difficult period in the Canadian renovation and construction industry. And in 2012, RONA fended off an unsolicited takeover bid by Lowe’s. Currently, RONA has over 800 corporate, franchise, and affiliate retail stores of all sizes, formats, and banners, as well as 14 distribution centres. RONA employed nearly 30 000 people and had sales of \$4.9 billion in 2012.

Competition in the sector remains fierce. With new management, store formats, smaller stores, new sales approaches, and a new corporate strategy, the leading company in home renovation is now undergoing its own corporate renovation.

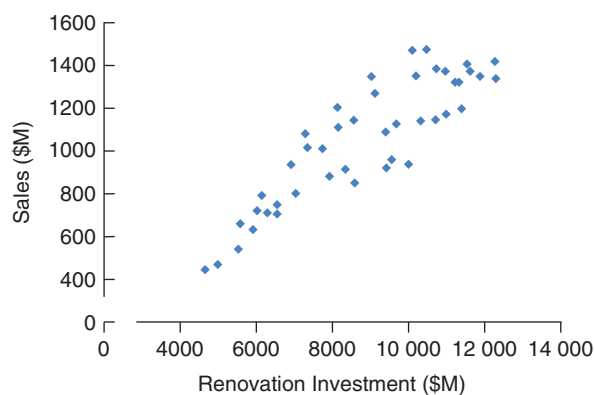


Based on information from [www.rona.ca](http://www.rona.ca)

<b>WHO</b>	Quarter years of financial data
<b>WHAT</b>	RONA's Sales and Canadian expenditures on residential Renovations
<b>UNITS</b>	Both in \$M
<b>WHEN</b>	2002–2012
<b>WHERE</b>	Canada
<b>WHY</b>	To assess RONA's sales relative to the home renovation market

RONA's quarterly sales results fluctuate widely because of the highly seasonal nature of renovation and construction activities. Over 80% of RONA's net annual earnings come from the second and third quarters. Sales in the first quarter are always lowest due to low activity in renovation and construction during the winter. Even in the summer and fall, poor weather has a major impact on sales.

RONA sells to both contractors and homeowners. Perhaps knowing how much Canadians spend on home renovation nationally can help predict RONA's sales. Here's a plot showing RONA's quarterly sales against Statistics Canada's quarterly data on spending for residential renovations.<sup>1</sup>



**Figure 6.1** RONA's Sales (\$M) and residential Renovation Investment, quarterly, 2002–2012.

If you were asked to summarize this relationship, what would you say? Clearly RONA's sales grew when home renovation spending grew. This plot is an example of a **scatterplot, which plots one quantitative variable against another**. Just by looking at a scatterplot, you can see patterns, trends, relationships, and even the occasional unusual values standing apart from the others. Scatterplots are the best way to start observing the relationship between two *quantitative* variables.

Relationships between variables are often at the heart of what we'd like to learn from data.

- ◆ *Is consumer confidence related to oil prices?*
- ◆ *What happens to customer satisfaction as sales increase?*
- ◆ *Is an increase in money spent on advertising related to sales?*
- ◆ *What is the relationship between a stock's sales volume and its price?*

Questions such as these relate two quantitative variables and ask whether there is an **association** between them. Scatterplots are the ideal way to *picture* such associations.

Why is this topic the most logical one to follow Chapters 4 and 5? Chapter 4 began with graphs and numerical summaries of categorical data, one variable at a time (the formal term is “univariate”). Then it moved on to contingency tables to examine the association between two categorical variables (called “bivariate” analysis). Similarly, Chapter 5 discussed univariate graphs and numerical summaries of quantitative data, one variable at a time, but stopped short of bivariate descriptive statistics; so there is no analogue to contingency tables. That is the role of scatterplots and correlation and that is why Chapter 6 comes next!

<sup>1</sup><http://www5.statcan.gc.ca/cansim/ Table 026-0013 Residential values, by type of investment quarterly>

**WHO** Years of financial data

**WHAT** Official *Exchange Rate* and inflation-adjusted *Price per Barrel* of oil

**UNITS** *Exchange Rate* (Canadian \$ relative to US \$); *Price per Barrel* (US \$)

**WHEN** 1980–2011

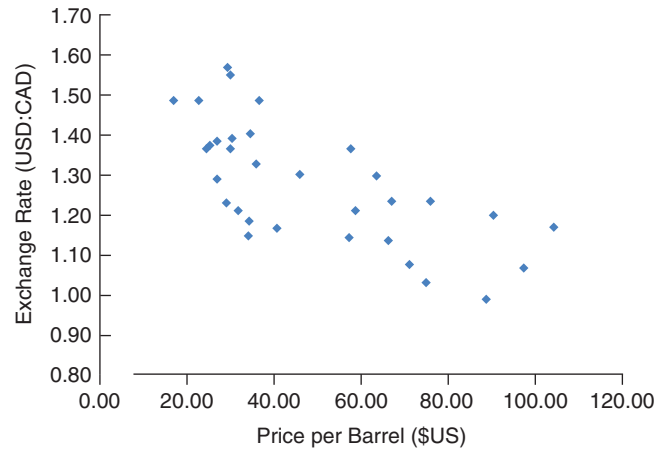
**WHERE** Canada

**WHY** To examine the relationship between exchange rate and price per barrel of oil

## 6.1 Looking at Scatterplots

The value of the Canadian dollar affects Canadians in a multitude of ways—from pricing of foreign-made products to travel costs to investment returns. For Canadian manufacturers selling products in the United States, a more valuable Canadian dollar can significantly reduce profitability. For example, if the CAD:USD exchange rate is 1.2:1, a \$5 USD sale is worth \$6 CAD. If the Canadian dollar increases in value to parity (1:1), a \$5 USD purchase is now worth only \$5 CAD to the manufacturer. This has caused many to argue that the higher Canadian dollar is damaging the Canadian manufacturing sector.

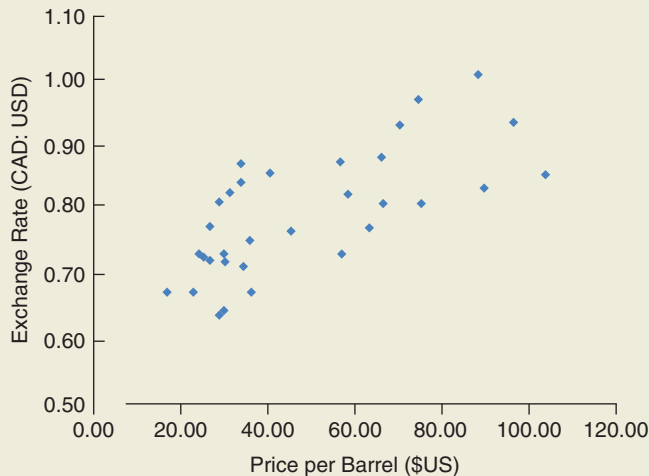
One major driver of the value of the Canadian dollar is oil. Oil from the Alberta oil sands makes up a significant portion of the Canada’s total exports. To buy Canadian oil, foreign buyers need to purchase Canadian dollars with foreign



**Figure 6.2** Official *Exchange Rate* (Canadian dollar relative to the U.S. dollar; Source: World Bank) vs. Inflation Adjusted *Price Per Barrel* of Oil (in U.S. dollars; Source: U.S. Bureau of Labor Statistics) for the period 1980–2011.

Look for *Direction*: What’s the sign—positive, negative, or neither?

Of course, we could have computed the exchange rate as the U.S. dollar relative to the Canadian dollar. That would mean computing the reciprocal of the exchange rate data presented in Figure 6.2, and then the scatterplot would look like this. It’s the same pattern but with positive association. As the price of oil increases, the value of the Canadian dollar increases.



currencies. An increase in the world price of oil increases total exports, subsequently increasing demand for Canadian dollars, driving up the value. To measure the effect of oil prices, we have gathered historical financial data on both oil prices and the value of the Canadian dollar. Figure 6.2 shows a scatterplot of the Inflation Adjusted *Price per Barrel* of Oil (in U.S. dollars) vs. the official *Exchange Rate* (Canadian dollar relative to the U.S. dollar).

Everyone looks at scatterplots. But, if asked, many people would find it hard to say what to look for in a scatterplot. What do *you* see? Try to describe the scatterplot of *Price per Barrel* against *Exchange Rate*.

First, you might say that the **direction** of the association is important. As the price of a barrel of oil goes up, the exchange rate goes down. A pattern that runs from the

upper left to the lower right  is said to be **negative**.


A pattern running the other way  is called **positive**.


The second thing to look for in a scatterplot is its **form**. If there is a straight line relationship, it will appear

as a cloud or swarm of points stretched out in a generally consistent, straight form. For example, the scatterplot of oil prices has an underlying **linear** form, although some points stray away from it.


Scatterplots can reveal many different kinds of patterns. Often they will not be straight, but straight line patterns are both the most common and the most useful for statistics.


If the relationship isn't straight, but curves gently, while still increasing or

decreasing steadily,  we can often find ways to straighten it out. But

if it curves sharply—up and then down, for example, —then you'll need more advanced methods.

The third feature to look for in a scatterplot is the **strength** of the relationship. At one extreme, do the points appear tightly clustered in a single stream

 (whether straight, curved, or bending all over the place)? Or, at the other extreme, do the points seem to be so variable and spread out that we can barely

discern any trend or pattern?  The oil prices plot shows moderate scatter around a generally straight form. That indicates that there's a moderately strong linear relationship between price and exchange rate.

Finally, always look for the unexpected. Often the most interesting discovery in a scatterplot is something you never thought to look for. One example of such a surprise is **an unusual observation, or outlier, standing away from the overall pattern of the scatterplot**. Such a point is almost always interesting and deserves special attention. You may see entire clusters or subgroups that stand away or show a trend in a different direction than the rest of the plot. That should raise questions about why they are different. They may be a clue that you should split the data into subgroups instead of looking at them all together.

Look for **Form**: Straight, curved, something exotic, or no pattern?

Look for **Strength**: How much scatter?

Look for **Unusual Features**: Are there unusual observations or subgroups?

## 6.2 Assigning Roles to Variables in Scatterplots

Scatterplots were among the first modern mathematical displays. The idea of using two axes at right angles to define a field on which to display values can be traced back to René Descartes (1596–1650), and the playing field he defined in this way is formally called a *Cartesian plane*, in his honour.

The two axes Descartes specified characterize the scatterplot. The axis that runs up and down is, by convention, called the *y*-axis, and the one that runs from side to side is called the *x*-axis. These terms are standard.<sup>2</sup>

To make a scatterplot of two quantitative variables, assign one to the *y*-axis and the other to the *x*-axis. As with any graph, be sure to label the axes clearly, and indicate the scales of the axes with numbers. Scatterplots display *quantitative* variables.

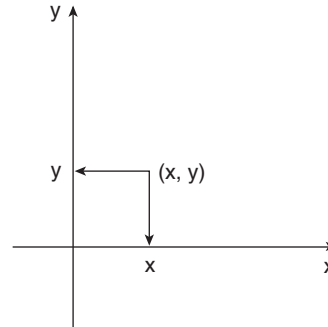


Library of congress  
Descartes was a philosopher, famous for his statement *cogito, ergo sum*: I think, therefore I am.

<sup>2</sup> The axes are also called the “ordinate” and the “abscissa”—but we can never remember which is which because statisticians don't generally use these terms. In Statistics (and in all statistics computer programs) the axes are generally called “*x*” (abscissa) and “*y*” (ordinate) and are usually labelled with the names of the corresponding variables.



Each variable has units, and these should appear with the display—usually near each axis. Each point is placed on a scatterplot at a position that corresponds to values of these two variables. Its horizontal location is specified by its  $x$ -value, and its vertical location is specified by its  $y$ -value variable. Together, these are known as *coordinates* and written  $(x, y)$ .



Scatterplots made by computer programs often do not—and usually should not—show the *origin*, the point at  $x = 0$ ,  $y = 0$  where the axes meet. If both variables have values near or on both sides of zero, then the origin will be part of the display. If the values are far from zero, though, there’s no reason to include the origin. In fact, it’s far better to focus on the part of the Cartesian plane that contains the data. In our example about oil prices, the exchange rate was, of course, nowhere near zero, so the scatterplot in Figure 6.2 has axes that don’t quite meet.

Which variable should go on the  $x$ -axis and which on the  $y$ -axis? What we want to know about the relationship can tell us how to make the plot. We often have questions such as:

- ◆ *Is RONA’s employee satisfaction related to productivity?*
- ◆ *Are increased sales at RONA’s reflected in the share price?*
- ◆ *What other factors besides residential renovations are related to RONA’s sales?*

### NOTATION ALERT:

So  $x$  and  $y$  are reserved letters as well, but not just for labeling the axes of a scatterplot. In Statistics, the assignment of variables to the  $x$ - and  $y$ -axes (and choice of notation for them in formulas) often conveys information about their roles as predictor or response.

Since the  $y$ -axis variable will be the outcome, that is, what happened, and the  $x$ -axis variable will be the predictor or explanation, here’s a suggestion on how to remember which is which: “ $x$ ” “explains” “why” “ $y$ ” happened. It’s a bit corny, but it works!

In all of these examples, one variable plays the role of the **explanatory** variable or predictor variable, while the other takes on the role of the **response variable**. We place the explanatory variable on the  $x$ -axis and the response variable on the  $y$ -axis. When you make a scatterplot, you can assume that those who view it will think this way, so choose which variables to assign to which axes carefully.

The roles that we choose for variables have more to do with how we *think* about them than with the variables themselves. Just placing a variable on the  $x$ -axis doesn’t necessarily mean that it explains or predicts *anything*, and the variable on the  $y$ -axis may not respond to it in any way. We plotted *Exchange Rate* against *Price per Barrel* thinking that as the price per barrel increases, the exchange rate would decrease. But maybe changing the exchange rate would increase the price of oil. If we were examining that option, we might choose to plot *Exchange Rate* as the explanatory variable and *Price per Barrel* as the response variable.

Perhaps an easier example to understand is the relationship between *Height* and *Weight* of young people. As a person grows taller, he/she gains weight. It makes more sense to think of *Height* explaining *Weight* than the other way around. In that case, we would be thinking that gaining weight might increase one’s height. That has been a failed experiment in North America, hence the problems with obesity!

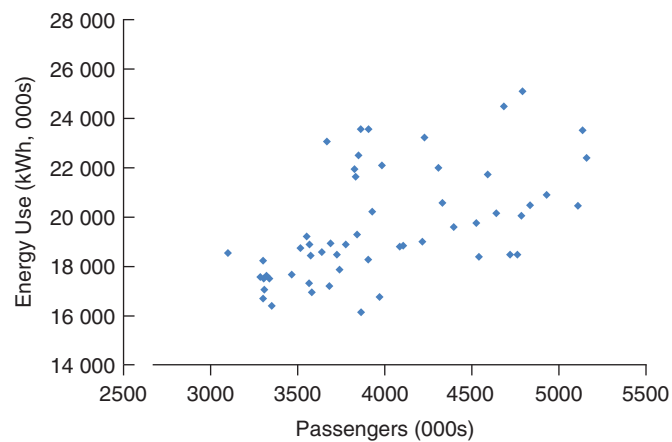
The  $x$ - and  $y$ -variables are sometimes referred to as the **independent and dependent** variables, respectively. The idea is that the  $y$ -variable *depends* on the

$x$ -variable and the  $x$ -variable acts *independently* to make  $y$  respond. These names, however, conflict with other uses of the same terms in Statistics. Instead, we'll sometimes use the terms “explanatory” or “predictor variable” and “response variable” when we're discussing roles, but we'll often just say  $x$ -variable and  $y$ -variable.

## 6.3 Understanding Correlation

The Vancouver International Airport Authority (YVR) recently undertook a study to examine how energy usage was related to various factors such as outside temperature, total area of the airport (since airports are always expanding!), and the number of passengers categorized as domestic, transborder (U.S.), and international. Data were collected on a monthly basis and summarized into quarterly totals. Of

<b>WHO</b>	Quarter years of YVR data
<b>WHAT</b>	Energy Use and total Passengers
<b>UNITS</b>	Energy Use (thousands of kWh) and total Passengers (thousands)
<b>WHEN</b>	1997–2010
<b>WHERE</b>	YVR (Vancouver International Airport Authority)
<b>WHY</b>	To examine the relationship between energy use and number of passengers in order to forecast and budget future energy costs



**Figure 6.3** Energy Use at YVR (kWh, 000s) and number of Passengers (000s), 1997 to 2010.

particular interest is how Energy Use and Total Passengers are related to each other. Figure 6.3 shows the scatterplot.

As you might expect, energy use and passenger count tend to rise and fall together. There is a clear positive association and, the scatterplot looks linear. But how strong is the association? If you had to put a number (say, between 0 and 1) on the strength of the association, what would it be? Your measure shouldn't depend on the choice of units for the variables. After all, if the data had been recorded in hundreds of kilowatt hours, or millions of passengers, the scatterplot would look the same. The direction, form, and strength won't change, so neither should our measure of the association's strength.

We saw a way to remove the units in the previous chapter. We can standardize each of the variables finding  $z_x = \left( \frac{x - \bar{x}}{s_x} \right)$  and  $z_y = \left( \frac{y - \bar{y}}{s_y} \right)$ . With these, we can compute a measure of strength that you've probably heard of: the **correlation coefficient**:

$$r = \frac{\sum z_x z_y}{n - 1}$$

Keep in mind that the  $x$ 's and  $y$ 's are paired. For each quarter, we have a measure of energy use and a passenger count. To find the correlation we multiply each

**NOTATION ALERT:**

The letter  $r$  is always used for correlation, so you can't use it for anything else in Statistics. Whenever you see an “ $r$ ,” it's safe to assume it's a correlation.

standardized value by the standardized value it is paired with and add up those *cross-products*. Then we divide the total by the number of pairs minus one,  $n - 1$ .<sup>3</sup>

For *Energy Use* and *Passengers*, the correlation coefficient is 0.54.

There are alternative formulas for the correlation in terms of the variables  $x$  and  $y$ . Here are two of the more common:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}.$$

These formulas can be more convenient for calculating correlation by hand, but the form given using  $z$ -scores is best for understanding what correlation means.

## Correlation Conditions

### Finding the correlation coefficient by hand

To find the correlation coefficient by hand, we'll use a formula in original units, rather than  $z$ -scores. This will save us the work of having to standardize each individual data value first. Start with the summary statistics for both variables:  $\bar{x}$ ,  $\bar{y}$ ,  $s_x$ , and  $s_y$ . Then find the deviations as we did for the standard deviation, but now in both  $x$  and  $y$ :  $(x - \bar{x})$  and  $(y - \bar{y})$ . For each data pair, multiply these deviations together:  $(x - \bar{x}) \times (y - \bar{y})$ . Add the products up for all data pairs. Finally, divide the sum by the product of  $(n - 1) \times s_x \times s_y$  to get the correlation coefficient.

Here we go.

Suppose the data pairs are:

$x$	6	10	14	19	21
$y$	5	3	7	8	12

Then  $\bar{x} = 14$ ,  $\bar{y} = 7$ ,  $s_x = 6.20$ , and  $s_y = 3.39$

Deviations in $x$	Deviations in $y$	Product
$6 - 14 = -8$	$5 - 7 = -2$	$-8 \times -2 = 16$
$10 - 14 = -4$	$3 - 7 = -4$	16
$14 - 14 = 0$	$7 - 7 = 0$	0
$19 - 14 = 5$	$8 - 7 = 1$	5
$21 - 14 = 7$	$12 - 7 = 5$	35

Add up the products:  $16 + 16 + 0 + 5 + 35 = 72$

Finally, we divide by  $(n - 1) \times s_x \times s_y = (5 - 1) \times 6.20 \times 3.39 = 84.07$

The ratio is the correlation coefficient:

$$r = 72/84.07 = 0.856$$

Correlation measures the strength of the *linear* association between two *quantitative* variables. Before you use correlation, you must check three *conditions*:

- ◆ **Quantitative Variables Condition:** Correlation applies only to quantitative variables. Don't apply correlation to categorical data masquerading as quantitative. Check that you know the variables' units and what they measure.
- ◆ **Linearity Condition:** Sure, you can *calculate* a correlation coefficient for any pair of variables. But correlation measures the strength only of the *linear* association and will be misleading if the relationship is not straight enough. What is “straight enough”? This question may sound too informal for a statistical condition, but that's really the point. We can't verify whether a relationship is linear or not. Very few relationships between variables are perfectly linear, even in theory, and scatterplots of real data are never perfectly straight. How nonlinear looking would the scatterplot have to be to fail the condition? This is a judgment call that you just have to think about. Do you think that the underlying relationship is curved? If so, then summarizing its strength with a correlation would be misleading.
- ◆ **Outlier Condition:** Unusual observations can distort the correlation and can make an otherwise small correlation look big or, on the other hand, hide a large correlation. It can even give an otherwise positive association a negative correlation coefficient (and vice versa). When you see an outlier, it's often a good idea to report the correlation both with and without the point.

Each of these conditions is easy to check with a scatterplot. Many correlations are reported without supporting data or plots. You should still think about the conditions. You should be cautious in interpreting (or accepting others' interpretations of) the correlation when you can't check the conditions for yourself.

<sup>3</sup> The same  $n - 1$  we used for calculating the standard deviation.





## JUST CHECKING

For the years 1992 to 2002, the quarterly stock price of the semiconductor companies Cypress and Intel have a correlation of 0.86.

- 1 Before drawing any conclusions from the correlation, what would you like to see? Why?
- 2 If your co-worker tracks the same prices in euros, how will this change the correlation? Will you need to know the exchange rate between euros and U.S. dollars to draw conclusions?
- 3 If you standardize both prices, how will this affect the correlation?
- 4 In general, if on a given day the price of Intel is relatively low, is the price of Cypress likely to be relatively low as well?
- 5 If on a given day the price of Intel shares is high, is the price of Cypress shares definitely high as well?

## GUIDED EXAMPLE

## Customer Spending

A major credit card company sends an incentive to its best customers in hope that the customers will use the card more. They wonder how often they can offer the incentive. Will repeated offerings of the incentive result in repeated increased credit card use?

To examine this question, an analyst took a random sample of 184 customers from their highest use segment and investigated the charges in the two months in which the customers had received the incentive.

### PLAN

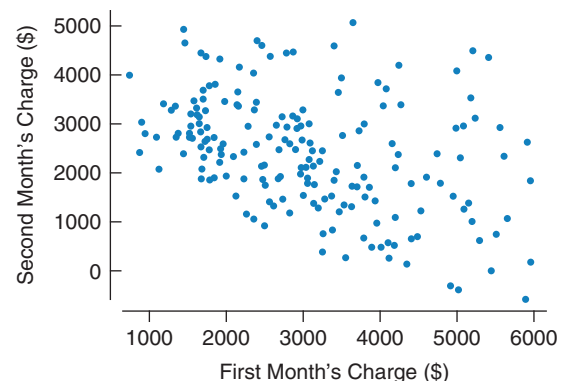
**Setup** State the objective. Identify the quantitative variables to examine. Report the time frame over which the data have been collected and define each variable. (State the W's.)



*Make* the scatterplot and clearly label the axes to identify the scale and units.

Our objective is to investigate the association between the amount that a customer charges in the two months in which they received an incentive. The customers have been randomly selected from among the highest use segment of customers. The variables measured are the total credit card charges (in \$) in the two months of interest.

✓ **Quantitative Variable Condition.** Both variables are quantitative. Both charges are measured in dollars.

Because we have two quantitative variables measured on the same cases, we can make a scatterplot.



	<p><b>Check</b> the conditions.</p>	<p>✓ <b>Linearity Condition.</b> The scatterplot is straight enough.</p> <p>✓ <b>Outlier Condition.</b> There are no obvious outliers.</p>
	<p><b>Mechanics</b> Once the conditions are satisfied, calculate the correlation with technology.</p>	<p>The correlation is <math>-0.391</math>.</p> <p>The negative correlation coefficient confirms the impression from the scatterplot.</p>
	<p><b>Conclusion</b> Describe the direction, form, and the strength of the plot, along with any unusual points or features. Be sure to state your interpretation in the proper context.</p>	<p><b>MEMO:</b></p> <p><b>Re: Credit Card Spending</b></p> <p>We have examined some of the data from the incentive program. In particular, we looked at the charges made in the first two months of the program. We noted that there was a negative association between charges in the second month and charges in the first month. The correlation was <math>-0.391</math>, which is only moderately strong, and indicates substantial variation.</p> <p>We've concluded that although the observed pattern is negative, these data do not allow us to find the causes of this behaviour. It is likely that some customers were encouraged by the offer to increase their spending in the first month, but then returned to former spending patterns. It is possible that others didn't change their behaviour until the second month of the program, increasing their spending at that time. Without data on the customers' pre-incentive spending patterns it would be hard to say more.</p> <p>We suggest further research, and we suggest that the next trial extend for a longer period of time to help determine whether the patterns seen here persist.</p>

### Correlation Properties

Because correlation is so widely used as a measure of association, it's a good idea to remember some of its basic properties. Here's a useful list of facts about the correlation coefficient:

- ◆ **The sign of a correlation coefficient gives the direction of the association.**
- ◆ **Correlation is always between  $-1$  and  $+1$ .** Correlation *can* be exactly equal to  $-1.0$  or  $+1.0$ , but watch out. These values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.
- ◆ **Correlation treats  $x$  and  $y$  symmetrically.** The correlation of  $x$  with  $y$  is the same as the correlation of  $y$  with  $x$ .

- ◆ **Correlation has no units.** This fact can be especially important when the data's units are somewhat vague to begin with (customer satisfaction, worker efficiency, productivity, and so on).
- ◆ **Correlation is not affected by changes in the centre or scale of either variable.** Changing the units or baseline of either variable has no effect on the correlation coefficient because the correlation depends only on the  $z$ -scores.
- ◆ **Correlation measures the strength of the *linear* association between the two variables.** Variables can be strongly associated but still have a small correlation if the association is not linear.
- ◆ **Correlation is sensitive to unusual observations.** A single outlier can make a small correlation large or make a large one small.

**How strong is strong?** Be careful when using the terms “weak,” “moderate,” or “strong,” because there's no agreement on exactly what those terms mean. The same numerical correlation might be strong in one context and weak in another. You might be thrilled to discover a correlation of 0.7 between an economic index and stock market prices, but finding “only” a correlation of 0.7 between a drug treatment and blood pressure might be viewed as a failure by a pharmaceutical company. Using general terms like “weak,” “moderate,” or “strong” to describe a linear association can be useful, but be sure to report the correlation and show a scatterplot so others can judge for themselves.

It is very difficult to estimate the numerical correlation by eye. See Exercises 7 and 8.

## Correlation Tables

Sometimes you'll see the correlations between each pair of variables in a data set arranged in a table. The rows and columns of the table name the variables, and the cells hold the correlations.

Correlation tables are compact and give a lot of summary information at a glance. They can be an efficient way to start to look at a large data set. The diagonal cells of a correlation table always show correlations of exactly 1.000, and the upper half of the table is symmetrically the same as the lower half (can you see why?), so by convention, only the lower half is shown. A table like this can be convenient, but be sure to check for linearity and unusual observations or the correlations in the table may be misleading or meaningless. Can you be sure, looking at Table 6.1, that the variables are linearly associated? Correlation tables are often produced by statistical software packages. Fortunately, these same packages often offer simple ways to make all the scatterplots you need to look at.<sup>4</sup>

You can also call a correlation table a correlation matrix if you want a more impressive-sounding term.

	Volume	Close	Interest Rate	Unemployment Rate
Volume	1.000			
Close	0.187	1.000		
Interest Rate	0.337	0.750	1.000	
Unemployment Rate	-0.381	-0.504	-0.924	1.000

**Table 6.1** A correlation table for variables measured monthly during the period 2006 through 2012. Volume = number of shares of RONA traded, Close = closing price of RONA stock, Interest Rate = prevailing Bank of Canada prime interest rate, Unemployment Rate = in Canada, as a percent.

<sup>4</sup> A table of scatterplots arranged just like a correlation table is sometimes called a *scatterplot matrix*, or SPLOM, and is easily created using a statistics package.

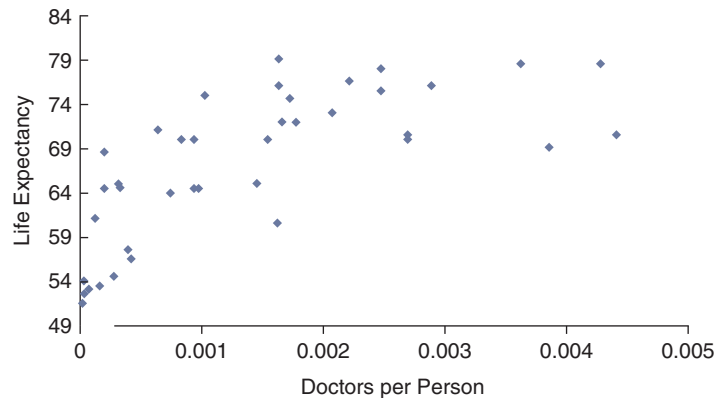


Crystal Kirk/Shutterstock

## 6.4 Lurking Variables and Causation

An educational researcher finds a strong association between height and reading ability among elementary school students in a nationwide survey. Taller children tend to have higher reading scores. Does that mean that students' height *causes* their reading scores to go up? No matter how strong the correlation is between two variables, there's no simple way to show from observational data that one variable causes the other. A high correlation just increases the temptation to think and to say that the  $x$ -variable *causes* the  $y$ -variable. Just to make sure, let's repeat the point again.

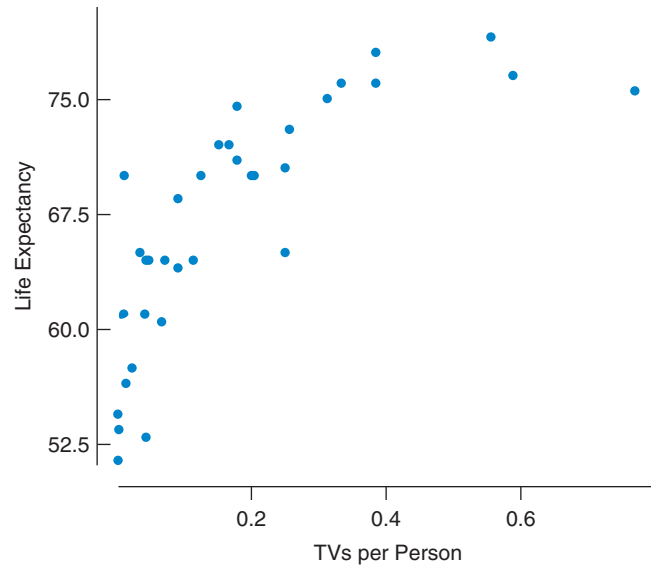
No matter how strong the association, no matter how large the  $r$  value, no matter how straight the form, there is no way to conclude from a high correlation *alone* that one variable causes the other. **There's always the possibility that some third variable—a lurking variable—is affecting both of the variables you have observed.** In the reading score example, you may have already guessed that the lurking variable is the age of the child. Older children tend to be taller and have stronger reading skills. But even when the lurking variable isn't as obvious, resist the temptation to think that a high correlation implies causation. Here's another example.



**Figure 6.4** Life Expectancy and numbers of Doctors per Person in 40 countries shows a fairly strong, positive linear relationship with a correlation of 0.705.

The scatterplot in Figure 6.4 shows the *Life Expectancy* (average of men and women, in years) for each of 40 countries of the world, plotted against the number of *Doctors per Person* in each country. The strong positive association ( $r = 0.705$ ) seems to confirm our expectation that more *Doctors per Person* improves health care, leading to longer lifetimes and a higher *Life Expectancy*. Perhaps we should send more doctors to developing countries to increase life expectancy.

If we increase the number of doctors, will the life expectancy increase? That is, would adding more doctors *cause* greater life expectancy? Could there be another explanation of the association? Figure 6.5 shows another scatterplot. *Life Expectancy* is still the response, but this time the predictor variable is not the number of doctors, but the number of *Televisions per Person* in each country. The positive association in this scatterplot looks even *stronger* than the association in the previous plot. If we wanted to calculate a correlation, we should straighten the plot first, but even from this plot, it's clear that higher life expectancies are associated with more televisions per person. Should we conclude that increasing the number of televisions extends lifetimes? If so, we should send televisions instead of doctors to developing countries. Not only is the association with life expectancy stronger, but televisions are cheaper than doctors.

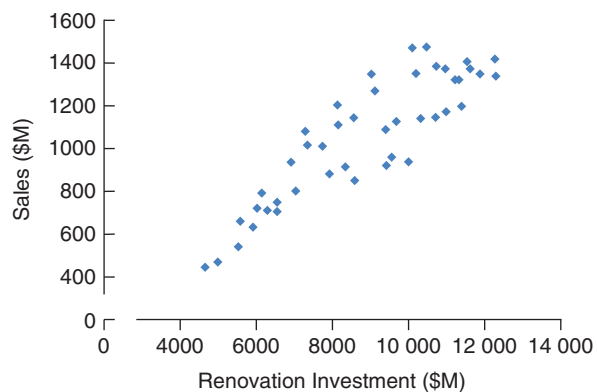


**Figure 6.5** Life Expectancy and number of Televisions per Person shows a strong, positive (although clearly not linear) relationship.

What's wrong with this reasoning? Maybe we were a bit hasty earlier when we concluded that doctors *cause* greater life expectancy. Maybe there's a lurking variable here. Countries with higher standards of living have both longer life expectancies *and* more doctors. Could higher living standards cause changes in the other variables? If so, then improving living standards might be expected to prolong lives, increase the number of doctors, and increase the number of televisions. From this example, you can see how easy it is to fall into the trap of mistakenly inferring causality from a correlation. For all we know, doctors (or televisions) *do* increase life expectancy. But we can't tell that from data like these no matter how much we'd like to. Resist the temptation to conclude that  $x$  causes  $y$  from a correlation, no matter how obvious that conclusion seems to you.

## 6.5 The Linear Model

Let's return to the relationship between RONA's sales and home renovation expenditures between 2002 and 2012. In Figure 6.1 (repeated here) we saw a strong, positive, linear relationship, so we can summarize its strength with a correlation. For this relationship, the correlation is 0.885.





“Statisticians, like artists, have the bad habit of falling in love with their models.”

—George Box, famous statistician

That’s quite strong, but the strength of the relationship is only part of the picture. RONA’s management might want to predict sales based on Statistics Canada’s estimate of residential renovation expenditures for the next four quarters. That’s a reasonable business question, but to answer it we’ll need a model for the trend. The correlation says that there seems to be a strong linear association between the variables, but it doesn’t tell us what that association is.

Of course, we can say more. We can model the relationship with a line and give the equation. For RONA, we can find a linear model to describe the relationship we saw in Figure 6.1 between RONA’s *Sales* and residential *Renovations*. **A linear model is just an equation of a straight line through the data.** The points in the scatterplot don’t all line up, but a straight line can summarize the general pattern with only a few parameters. This model can help us understand how the variables are associated.

### ! NOTATION ALERT:

“Putting a hat on it” is standard Statistics notation to indicate that something has been predicted by a model. Whenever you see a hat over a variable name or symbol, you can assume it is the predicted version of that variable or symbol.

## Residuals

We know the model won’t be perfect. No matter what line we draw, it won’t go through many of the points. The best line might not even hit any of the points. Then how can it be the “best” line? We want to find the line that somehow comes *closer* to all the points than any other line. Some of the points will be above the line and some below. **A linear model can be written as  $\hat{y} = b_0 + b_1x$ , where  $b_0$  and  $b_1$  are numbers estimated from the data and  $\hat{y}$  (pronounced *y-hat*) is the predicted value. We use the *hat* to distinguish the predicted value from the observed value  $y$ . The difference between these two is called the residual:**

$$e = y - \hat{y}.$$

The residual value tells us how far the model’s prediction is from the observed value at that point. To find the residuals, we always subtract the predicted values from the observed ones.

Our question now is how to find the right line.

## The Line of “Best Fit”

When we draw a line through a scatterplot, some residuals are positive, and some are negative. We can’t assess how well the line fits by adding up all the residuals—the positive and negative ones would just cancel each other out. We need to find the line that’s closest to all the points, and to do that, we need to make all the distances positive. We faced the same issue when we calculated a standard deviation to measure spread. And we deal with it the same way here: by squaring the residuals to make them positive. The sum of all the squared residuals tells us how well the line we drew fits the data—the smaller the sum, the better the fit. A different line will produce a different sum, maybe bigger, maybe smaller. **The line of best fit is the line for which the sum of the squared residuals is smallest—often called the least squares line.**

This line has the special property that the variation of the data around the model, as seen in the residuals, is the smallest it can be for any straight line model for these data. No other line has this property. Speaking mathematically, we say that this line minimizes the sum of the squared residuals. You might think that finding this “least squares line” would be difficult. Surprisingly,

A *negative* residual means the predicted value is too big—an overestimate. A *positive* residual shows the model makes an underestimate. These may seem backwards at first.

### Who Was First?

French mathematician Adrien-Marie Legendre was the first to publish the “least squares” solution to the problem of fitting a line to data when the points don’t all fall exactly on the line. The main challenge was how to distribute the errors “fairly.” After considerable thought, he decided to minimize the sum of the squares of what we now call the residuals. After Legendre published his paper in 1805, Carl Friedrich Gauss, the German mathematician and astronomer, claimed he had been using the method since 1795 and, in fact, had used it to calculate the orbit of the asteroid Ceres in 1801. Gauss later referred to the “least squares” solution as “*our method*” (principium *nostrum*), which certainly didn’t help his relationship with Legendre.

it's not, although it was an exciting mathematical discovery when Legendre published it in 1805.

Other criteria for “best fit” are theoretically possible. We have chosen to minimize squared residuals, that is, squared vertical distances from data points to the line. But, mathematically, the shortest distance from a point to a line is the perpendicular. There is also the horizontal distance from a data point to the line. Putting the vertical and horizontal distances together makes a triangle, so a creative criterion would be the minimum area of all the formed triangles. Of course, none of these other criteria work in the context of the linear model we have developed here.

## 6.6 Correlation and the Line

Any straight line can be written as:

$$y = b_0 + b_1x.$$

If we were to plot all the  $(x, y)$  pairs that satisfy this equation, they'd fall exactly on a straight line. We'll use this form for our linear model. Of course, with real data, the points won't all fall on the line. So, we write our model as  $\hat{y} = b_0 + b_1x$ , using  $\hat{y}$  for the predicted values, because it's the predicted values (not the data values) that fall on the line. If the model is a good one, the data values will scatter closely around it.

For the RONA sales data, the line is:

$$\widehat{Sales} = 12.13 + 0.117 \text{ Renovations}$$

What does this mean? The **slope** 0.117 says that we can expect a year in which residential renovation spending is 1 million dollars higher to be one in which RONA sales will be about 0.117 \$M (\$117,000) higher. Slopes are always expressed in  $y$ -units per  $x$ -units. They tell you how the response variable changes for a one unit step in the predictor variable. So we'd say that the slope is 0.117 million dollars of *Sales* per million dollars of *Renovations*.

The **intercept**, 12.13, is the value of the line when the  $x$ -variable is zero. What does it mean here? The intercept often serves just as a starting value for our predictions. We don't interpret it unless a 0 value for the predictor variable would really mean something under the circumstances. The RONA model is based on quarters in which spending on residential renovation is between 4 and 14 billion dollars. It's unlikely to be appropriate if there were no such spending at all. In this case, we wouldn't interpret the intercept.

How do we find the slope and intercept of the least squares line? The formulas are simple. The model is built from the summary statistics we've used before. We'll need the correlation (to tell us the strength of the linear association), the standard deviations (to give us the units), and the means (to tell us where to locate the line).

The slope of the line is computed as:

$$b_1 = r \frac{s_y}{s_x}.$$

We've already seen that the correlation tells us the sign and the strength of the relationship, so it should be no surprise to see that the slope inherits this sign as well. If the correlation is positive, the scatterplot runs from lower left to upper right, and the slope of the line is positive.

Correlations don't have units, but slopes do. How  $x$  and  $y$  are measured—what units they have—doesn't affect their correlation, but does change the slope. The slope gets its units from the ratio of the two standard deviations. Each standard deviation has the units of its respective variable. So, the units of the slope are a ratio, too, and are always expressed in units of  $y$  per unit of  $x$ .



## JUST CHECKING

A scatterplot of sales per month (in thousands of dollars) vs. number of employees for all the outlets of a large computer chain shows a relationship that is straight, with only moderate scatter and no outliers. The correlation between *Sales* and *Employees* is 0.85, and the equation of the least squares model is:

$$\widehat{Sales} = 9.564 + 122.74 \text{ Employees}$$

- 6 What does the slope of 122.74 mean?
- 7 What are the units of the slope?
- 8 The outlet in Edmonton has 10 more employees than the outlet in Calgary. How much more *Sales* do you expect it to have?

### RONA

Summary statistics:

$$\text{Sales: } \bar{y} = 1049.15; s_y = 288.4$$

$$\text{Improvements: } \bar{x} = 8833.4; s_x = 2175.3$$

$$\text{Correlation} = 0.885$$

$$\text{So, } b_1 = r \frac{s_y}{s_x} = (0.885) \frac{288.4}{2175.3}$$

$$= 0.117$$

(\$M Sales per \$M Improvement expenditures)

And

$$b_0 = \bar{y} - b_1 \bar{x} = 1049.15 - (0.117)8833.4 = 15.64$$

The equation from the computer output has slope 0.117 and intercept 12.13. The differences are due to rounding error. We've shown the calculation using rounded summary statistics, but if you are doing this by hand, you should always keep all digits in intermediate steps.

How do we find the intercept? If you had to predict the  $y$ -value for a data point whose  $x$ -value was average, what would you say? The best fit line predicts  $\bar{y}$  for points whose  $x$ -value is  $\bar{x}$ . Putting that into our equation and using the slope we just found gives:

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

and we can rearrange the terms to find:

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

It's easy to use the estimated linear model to predict RONA *Sales* for any amount of national spending on residential *Renovations*. For example, in the second quarter of 2012 the total was \$12 268(M). To estimate RONA *Sales*, we substitute this value for  $x$  in the model:

$$\widehat{Sales} = 12.13 + 0.117 \times 12\,268 = 1447.5$$

Sales actually were 1417.1 (\$M), so the residual of  $1417.1 - 1447.5 = -30.4$  (\$M) tells us how much worse RONA did than the model predicted.

**Least squares lines are commonly called regression**

**lines.** Although this name is an accident of history (as we'll soon see), "regression" almost always means "the linear model fit by least squares." Clearly, regression and correlation are closely related. We'll need to check the same condition for regression as we did for correlation:

1. **Quantitative Variables Condition**
2. **Linearity Condition**
3. **Outlier Condition**

A little later in the chapter we'll add two more.

## Understanding Regression from Correlation

The slope of a regression line depends on the units of both  $x$  and  $y$ . Its units are the units of  $y$  per unit of  $x$ . The units are expressed in the slope because  $b_1 = r \frac{s_y}{s_x}$ .

The correlation has no units, but each standard deviation contains the units of its respective variable. For our regression of RONA *Sales* on home *Renovations*,

the slope was millions of dollars of sales *per* million dollars of renovation expenditure.

It can be useful to see what happens to the regression equation if we were to standardize both the predictor and response variables and regress  $z_y$  on  $z_x$ . For both these standardized variables, the standard deviation is 1 and the means are zero. That means that the slope is just  $r$ , and the intercept is 0 (because both  $\bar{y}$  and  $\bar{x}$  are now 0).

This gives us the simple equation for the regression of standardized variables:

$$\hat{z}_y = rz_x.$$

Although we don't usually standardize variables for regression, it can be useful to think about what this means. Thinking in  $z$ -scores is a good way to understand what the regression equation is doing. The equation says that for every standard deviation we deviate from the mean in  $x$ , we predict that  $y$  will be  $r$  standard deviations away from the mean in  $y$ .

Let's be more specific. For the RONA example, the correlation is 0.885. So, we know immediately that:

$$\hat{z}_{Sales} = 0.885 z_{Renovations}.$$

## 6.7 Regression to the Mean



Pavel L Photo and Video/Shutterstock



Library of congress

Sir Francis Galton was the first to speak of "regression," although others had fit lines to data by the same method.

Suppose you were told that a new male student was about to join the class and you were asked to guess his height in inches. What would be your guess? A good guess would be the mean height of male students. Now suppose you are also told that this student had a grade point average (GPA) of 3.9—about 2 SDs above the mean GPA. Would that change your guess? Probably not. The correlation between GPA and height is near 0, so knowing the GPA value doesn't tell you anything and doesn't move your guess. (And the standardized regression equation,  $\hat{z}_y = rz_x$ , tells us that as well, since it says that we should move  $0 \times 2$  SDs from the mean.)

On the other hand, if you were told that, measured in centimetres, the student's height was 2 SDs above the mean, you'd know his height in inches. There's a perfect correlation between *Height* in inches and *Height* in centimetres ( $r = 1$ ), so you know he's 2 SDs above mean height in inches as well.

What if you were told that the student was 2 SDs above the mean in shoe size? Would you still guess that he's of average height? You might guess that he's taller than average, since there's a positive correlation between height and shoe size. But would you guess that he's 2 SDs above the mean? When there was no correlation, we didn't move away from the mean at all. With a perfect correlation, we moved our guess the full 2 SDs. Any correlation between these extremes should lead us to move somewhere between 0 and 2 SDs above the mean. (To be exact, our best guess would be to move  $r \times 2$  SDs away from the mean.)

Notice that if  $x$  is 2 SDs above its mean, we won't ever move more than 2 SDs away for  $y$ , since  $r$  can't be bigger than 1.0. So each predicted  $y$  tends to be closer to its mean (in standard deviations) than its corresponding  $x$  was. **This property of the linear model is called regression to the mean. This is why the line is called the regression line.**

### More on Regression to the Mean

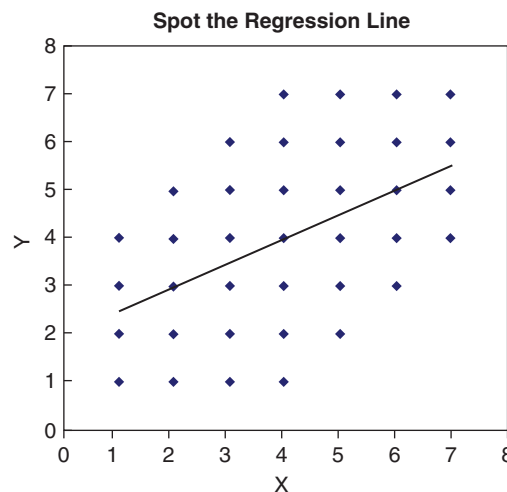
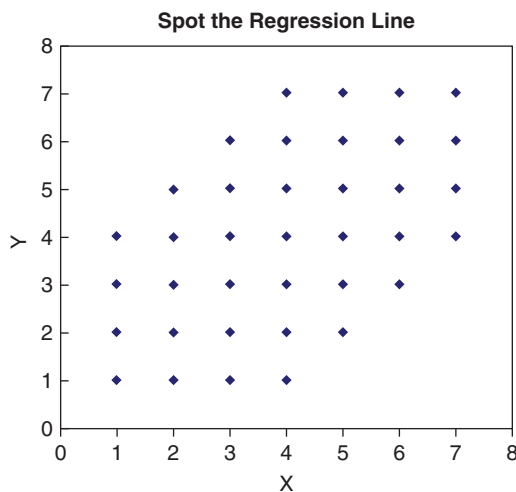
Misinterpretation of "regression to the mean" is a phenomenon that still plagues decision-makers in countless areas. Stephen Senn wrote, "A Victorian eccentric [Francis Galton] . . . made an important discovery of a phenomenon that is so trivial that all should be capable of learning it and so deep that many scientists spend their whole career being fooled by it."

The previous illustrations show that unless the correlation between  $X$  and  $Y$  is perfect, predictions of  $Y$  from  $X$  will always appear to be less dramatic than one

might expect. That's because, "A point that is 1 SD above the mean in the X-variable is, on average,  $r$  SDs above the mean in the Y-variable. Similarly, a point that is 1 SD below the mean in X is, on average,  $r$  SDs below the mean in Y."

Suppose you missed writing the final exam in a course and wanted to try to predict what grade you might have received based on how you did on the midterm exam in the same course. Results from the rest of the class showed that the midterm had a mean of 75% and SD of 10%, while the final exam had a mean of 70%, also with a SD of 10%. The correlation between midterm and final exam grades was 0.75; that is, students who did well on the midterm generally did well on the final. Suppose you performed exceedingly well on the midterm and received a grade of 95%. That means you were 2 SDs above the average. According to the rule of regression, you would be predicted to get  $r \times 2$  SDs above the average on the final exam, which would be a grade of 85% (i.e.,  $70\% + (0.75 \times 2 \times 10\%) = 85\%$ ). That's still a very good grade, but not as high, even relatively, as your midterm grade. Why not? Because the correlation is not perfect. There are many other explanations for, or predictors of high final grades, such as long hours of studying!

Here is another way to state the main idea of regression: "For each value of X, the regression line passes through average value of Y." Look at the following graph. We know that the regression line will pass through the point of averages  $(\bar{x}, \bar{y})$ , which is right in the centre of the scatterplot. Can you visualize the slope of the regression line? The answer may surprise you. It is not the line that goes through the main diagonal or "principal axis" of the ellipse. It goes through the average Y at each X. The second graph shows where the regression line lies.



Thus for  $X = 7$ , which is 3 units above the average, the regression line predicts a value of 5.5 for Y, which is 1.5 units above average. That's because the correlation here is only 0.5.

Regression to the mean is the tendency for a very high value of a random quantity whose values cluster around an average, to be followed by a value closer to that average, and similarly for a very low value. It is a natural effect. Misinterpretations of this effect lead to regression fallacies, of which there are countless examples. Here are a few.

- ◆ The "sophomore jinx" in professional sports: a player who has a brilliant rookie (first-year) performance is not quite as brilliant the second year.



- ◆ The “cover of Sports Illustrated jinx”: an athlete who makes the cover of Sports Illustrated magazine experiences a drop in performance (because he/she only made the cover due to a much better-than-average performance!)
- ◆ The sequel to a movie is rarely ever as good as the original: a really bad movie will not likely have a sequel, so we rarely experience a series of movies that gets better.
- ◆ The “reward and punishment” child-rearing fallacy: Psychologists Amos Tversky and Daniel Kahneman write, “Behaviour is most likely to improve after punishment and deteriorate after reward. Consequently . . . one is most often rewarded for punishing others, and most often punished for rewarding them.” Reward a child for angelic behaviour and the next time the behaviour is not exemplary as the child returns to his/her average behaviour. Similarly, punish a child for devilish behaviour and the next time the behaviour is not so bad. It isn’t really because of the punishment; it is regression to the mean.
- ◆ The effect of red-light cameras on accidents: when you put cameras in a high-accident-rate intersection the accident rate will decline; at the same time, when you take a camera away from a low-accident intersection the rate will increase!

Regression to the mean was the brilliant observation of the great Victorian era scientist Sir Francis Galton. He was also an explorer, geographer, weather-forecaster, travel-writer, and inventor. He devised the fingerprint classification system for identification used to this day. He had a famous cousin, but it would be accurate to call Galton, “Charles Darwin’s smarter cousin!” We celebrated this giant of statistics in 2011, which was the 100th anniversary of Galton’s death and the 125th anniversary of his landmark paper that introduced the term “regression to mediocrity” (his term for regression to the mean).

In the paper, Galton related the heights of sons to the height of their fathers. He found that the slope of his line was less than 1. That is, sons of tall fathers were tall, but not as much above their average as their fathers had been above their average. Similarly, sons of short fathers were short, but generally not as far from their mean as their fathers. Galton interpreted the slope correctly, calling it “regression” (i.e., moving back) toward the mean height. The name stuck.

Regression to the mean is very often the explanation for many phenomena that so-called experts attribute to something real, not just chance. Don’t let yourself be fooled by it!

## MATH BOX

Where does the equation of the line of best fit come from? To write the equation of any line, we need to know a point on the line and the slope. It’s logical to expect that an average  $x$  will correspond to an average  $y$ , and, in fact, the line does pass through the point  $(\bar{x}, \bar{y})$ . (This is not hard to show as well.)

To think about the slope, we look once again at the  $z$ -scores. We need to remember a few things.

1. The mean of any set of  $z$ -scores is 0. This tells us that the line that best fits the  $z$ -scores passes through the origin  $(0, 0)$ .
2. The standard deviation of a set of  $z$ -scores is 1, so the variance is also 1.

This means that  $\frac{\sum(z_y - \bar{z}_y)^2}{n - 1} = \frac{\sum(z_y - 0)^2}{n - 1} = \frac{\sum z_y^2}{n - 1} = 1$  a fact that

will be important soon.

3. The correlation is  $r = \frac{\sum z_x z_y}{n - 1}$ , also important soon.

Remember that our objective is to find the slope of the best fit line. Because it passes through the origin, the equation of the best fit line will be of the form  $\hat{z}_y = m z_x$ . We want to find the value for  $m$  that will minimize the sum of the squared errors. Actually we’ll divide that sum by  $n - 1$  and minimize this mean squared error (MSE). Here goes:

$$\text{Minimize:} \quad \text{MSE} = \frac{\sum(z_y - \hat{z}_y)^2}{n - 1}$$

$$\text{Since } \hat{z}_y = m z_x: \quad \text{MSE} = \frac{\sum(z_y - m z_x)^2}{n - 1}$$

$$\text{Square the binomial:} \quad = \frac{\sum(z_y^2 - 2mz_xz_y + m^2z_x^2)}{n-1}$$

$$\text{Rewrite the summation:} \quad = \frac{\sum z_y^2}{n-1} - 2m \frac{\sum z_xz_y}{n-1} + m^2 \frac{\sum z_x^2}{n-1}$$

$$4. \text{ Substitute from (2) and (3):} \quad = 1 - 2mr + m^2$$

This last expression is a quadratic. A parabola in the form  $y = ax^2 + bx + c$  reaches its minimum at its turning point, which occurs when  $x = \frac{-b}{2a}$ . We can minimize the mean of squared errors by

$$\text{choosing } m = \frac{-(-2r)}{2(1)} = r.$$

The slope of the best fit line for  $z$ -scores is the correlation,  $r$ . This fact leads us immediately to two important additional results:

A slope with value  $r$  for  $z$ -scores means that a difference of 1 standard deviation in  $z_x$  corresponds to a difference of  $r$  standard deviations in  $\hat{z}_y$ . Translate that back to the original  $x$  and  $y$  values: “Over one standard deviation in  $x$ , up  $r$  standard deviations in  $\hat{y}$ .”

$$\text{The slope of the regression line is } b = \frac{rs_y}{s_x}.$$

We know choosing  $m = r$  minimizes the sum of the squared errors (SSE), but how small does that sum get? Equation (4) told us that the mean of the squared errors is  $1 - 2mr + m^2$ . When  $m = r$ ,  $1 - 2mr + m^2 = 1 - 2r^2 + r^2 = 1 - r^2$ . This is the percentage of variability *not* explained by the regression line. Since  $1 - r^2$  of the variability is *not* explained, the percentage of variability in  $y$  that *is* explained by  $x$  is  $r^2$ . This important fact will help us assess the strength of our models.

And there’s still another bonus. Because  $r^2$  is the percent of variability explained by our model,  $r^2$  is at most 100%. If  $r^2 \leq 1$ , then  $-1 \leq r \leq 1$ , proving that correlations are always between  $-1$  and  $+1$ .

### Why $r$ for correlation?

In his original paper on correlation, Galton used  $r$  for the “index of correlation”—what we now call the correlation coefficient. He calculated it from the regression of  $y$  on  $x$  or of  $x$  on  $y$  after standardizing the variables, just as we have done. It’s fairly clear from the text that he used  $r$  to stand for (standardized) regression.

## 6.8 Checking the Model

The linear regression model is perhaps the most widely used model in all of Statistics. It has everything we could want in a model: two easily estimated parameters, a meaningful measure of how well the model fits the data, and the ability to predict new values. It even provides a self-check in plots of the residuals to help us avoid all kinds of mistakes. Most models are useful only when specific assumptions are true. Of course, assumptions are hard—often impossible—to check. That’s why we *assume* them. But we should check to see whether the assumptions are *reasonable*. Fortunately, we can often check *conditions* that provide information about the assumptions. For the linear model, we start by checking the same ones we checked earlier in this chapter for using correlation.

### Linear Regression Conditions

- ◆ **Quantitative Data Condition:** Linear models only make sense for quantitative data. Don’t be fooled by categorical data recorded as numbers. You probably don’t want to predict area codes from credit card account numbers.

**Make a Picture**

Check the scatterplot. The shape must be linear, or you can't use regression for the variables in their current form. And watch out for outliers. A useful rule of thumb is that the assumptions are probably reasonable if the scatterplot has an approximate oval shape. That doesn't check for independence but it is a quick check on the other assumptions.

- ◆ **Linearity Assumption:** The regression model *assumes* that the relationship between the variables is, in fact, linear. If you try to model a curved relationship with a straight line, you'll usually get what you deserve. We can't ever verify that the underlying relationship between two variables is truly linear, but an examination of the scatterplot will let you decide whether the *Linearity Assumption* is reasonable. The **Linearity Condition** we used for correlation is designed to do precisely that and is satisfied if the scatterplot looks reasonably straight. If the scatterplot is not straight enough, stop. You can't use a linear model for just *any* two variables, even if they are related. The two variables must have a *linear* association, or the model won't mean a thing. Some nonlinear relationships can be saved by re-expressing the data to make the scatterplot more linear.
- ◆ **Outlier Condition:** Watch out for outliers. The linearity assumption also requires that no points lie far enough away to distort the line of best fit. Check to make sure no point needs special attention. Outlying values may have large residuals, and squaring makes their influence that much greater. Outlying points can dramatically change a regression model. Unusual observations can even change the sign of the slope, misleading us about the direction of the underlying relationship between the variables.
- ◆ **Independence Assumption:** Another assumption that is usually made when fitting a linear regression is that the residuals are independent of each other. We don't strictly need this assumption to fit the line, but to generalize from the data it's a crucial assumption and one that we'll come back to when we discuss inference. As with all assumptions, there's no way to be sure that *Independence Assumption* is true. However we could check that the cases are a random sample from the population.

We can also check displays of the regression residuals for evidence of patterns, trends, or clumping, any of which would suggest a failure of independence. In the special case when we have a time series, a common violation of the Independence Assumption is for the errors to be correlated with each other (autocorrelation). The error our model makes today may be similar to the one it made yesterday. We can check this violation by plotting the residuals against time (usually  $x$  for a time series) and looking for patterns.

When our goal is just to explore and describe the relationship, independence isn't essential (and so we won't insist that the conditions relating to it be formally checked). However, when we want to go beyond the data at hand and make inferences for other situations (in Chapter 14) this will be a crucial assumption, so it's good practice to think about it even now, especially for time series.

- ◆ **Residuals:** We always check conditions with a scatterplot of the data, but we can learn even more after we've fit the regression model. There's extra information in the residuals that we can use to help us decide how reasonable our model is and how well the model fits. So, we plot the residuals and check the conditions again.

The residuals are the part of the data that *hasn't* been modelled. We can write

$$\text{Data} = \text{Predicted} + \text{Residual}$$

or, equivalently,

$$\text{Residual} = \text{Data} - \text{Predicted}$$

Or, as we showed earlier, in symbols:

$$e = y - \hat{y}.$$

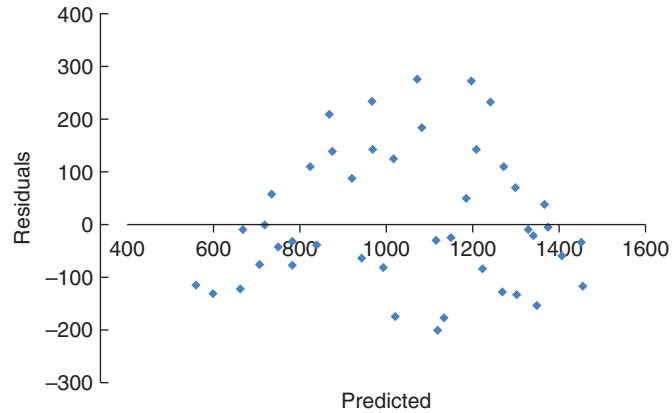
A scatterplot of the residuals versus the  $x$ -values should be a plot without patterns. It shouldn't have any interesting features—no direction, no shape. It should

**Why  $e$  for residual?**

The easy answer is that  $r$  is already taken for correlation, but the truth is that  $e$  stands for "error." It's not that the data point is a mistake but that statisticians often refer to variability not explained by a model as error.

stretch horizontally, showing no bends, and it should have no outliers. If you see nonlinearities, outliers, or clusters in the residuals, find out what the regression model missed.

Let's examine the residuals from our regression of RONA Sales on residential Renovations expenditures.<sup>5</sup>



**Figure 6.6** Residuals of the regression model predicting RONA Sales from residential *Renovation* expenditures 2002–2012.

The residual plot is suitably boring. The only noticeable feature is that at the lower end, the residuals are smaller and mostly negative. The residuals are smaller in the earlier years because the sales are lower in the early years so the error in the predictions must be smaller. The residuals are mostly negative because the speed of growth increased after 2003.

This is a good time to point out a common mistake in the interpretation of residual plots, namely, looking too hard for patterns or unusual features. The residual plots are designed to show major departures from the assumptions. Don't fall victim to over-interpreting them.

Not only can the residuals help check the conditions, but they can also tell us how well the model performs. The better the model fits the data, the less the residuals will vary around the line. The standard deviation of the residuals,  $s_e$ , gives us a measure of how much the points spread around the regression line. Of course, for this summary to make sense, the residuals should all share the same underlying spread. So we must *assume* that the standard deviation around the line is the same wherever we want the model to apply.

◆ **Equal Spread Condition:** This new assumption about the standard deviation around the line gives us a new condition, called the *Equal Spread Condition*. The associated question to ask is does the plot have a consistent spread or does it fan out? We check to make sure that the spread of the residuals is about the same everywhere. We can check that either in the original scatterplot of  $y$  against  $x$  or in the scatterplot of residuals (or, preferably, in both plots). **We estimate the standard deviation of the residuals in almost the way you'd expect:**

$$s_e = \sqrt{\frac{\sum e^2}{n-2}}$$

### Equal Spread Condition

This condition requires that the scatter is about equal for all  $x$ -values. It's often checked using a plot of residuals against predicted values. The underlying assumption of equal variance is also called *homoscedasticity*.

The term comes from the Greek words “*homos*” meaning “same” and “*skedastikos*” meaning able to scatter. So *homoscedasticity* means “same scatter.”

<sup>5</sup> Most computer statistics packages plot the residuals as we did in Figure 6.6, against the predicted values, rather than against  $x$ . When the slope is positive, the scatterplots are virtually identical except for the axes labels. When the slope is negative, the two versions are mirror images. Since all we care about is the patterns (or, better, lack of patterns) in the plot, either plot is useful.

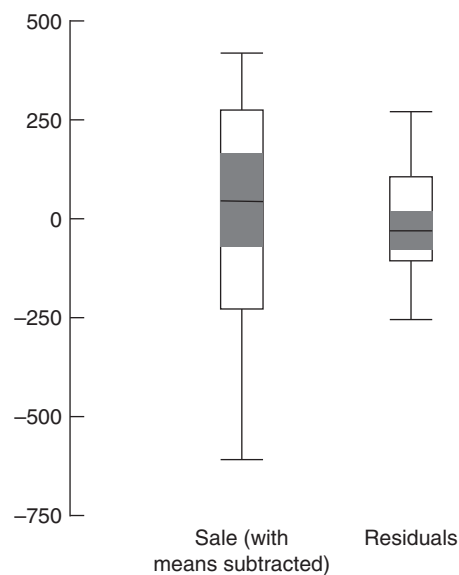
We don't need to subtract the mean of the residuals because  $\bar{e} = 0$ . Why divide by  $n - 2$  rather than  $n - 1$ ? We used  $n - 1$  for  $s$  when we estimated the mean. Now we're estimating both a slope and an intercept. Looks like a pattern—and it is. We subtract one more for each parameter we estimate.

If we predict RONA Sales in the third quarter of 2008 when home *Renovation* totalled 10 737.5 \$M, the regression model gives a predicted value of 1272.7 \$M. The actual value was about 1381.7 \$M. So our residual is  $1272.7 - 1381.7 = 109.0$ . The value of  $s_e$  from the regression is 135.6, so our residual is only  $109.0/135.6 = 0.80$  standard deviations away from the actual value. That's a fairly typical size for a residual because it's within 2 standard deviations.

## 6.9 Variation in the Model and $R^2$

The variation in the residuals is the key to assessing how well the model fits. Let's compare the variation of the response variable with the variation of the residuals. *Sales* has a standard deviation of 288.4 (\$M). The standard deviation of the residuals is only 134.0 (\$M). If the correlation were 1.0 and the model predicted the *Sales* values perfectly, the residuals would all be zero and have no variation. We couldn't possibly do any better than that.

On the other hand, if the correlation were zero, the model would simply predict 1049.2 (\$M) (the mean) for all menu items. The residuals from that prediction would just be the observed *Sales* values minus their mean. These residuals would have the same variability as the original data because, as we know, just subtracting the mean doesn't change the spread.



**Figure 6.7** Compare the variability of Sales with the variability of the residuals from the regression. The means have been subtracted to make it easier to compare spreads. The variation left in the residuals is unaccounted for by the model, but it's less than the variation in the original data.

How well does the regression model do? Look at the boxplots. The variation in the residuals is smaller than in the data, but bigger than zero. That's nice to know, but how much of the variation is still left in the residuals? If you had to put a



Is a correlation of 0.80 twice as strong as a correlation of 0.40? Not if you think in terms of  $R^2$ . A correlation of 0.80 means an  $R^2$  of  $0.80^2 = 64\%$ . A correlation of 0.40 means an  $R^2$  of  $0.40^2 = 16\%$ —only a quarter as much of the variability accounted for. A correlation of 0.80 gives an  $R^2$  *four* times as strong as a correlation of 0.40 and accounts for four times as much of the variability.

### Some Extreme Tales

One major company developed a method to differentiate between proteins. To do so, they had to distinguish between regressions with  $R^2$  of 99.99% and 99.98%. For this application, 99.98% was not high enough.

The president of a financial services company reports that although his regressions give  $R^2$  below 2%, they are highly successful because those used by his competition are even lower.

number between 0% and 100% on the fraction of the variation left in the residuals, what would you say?

All regression models fall somewhere between the two extremes of zero correlation and perfect correlation. We'd like to gauge where our model falls. Can we use the correlation to do that? Well, a regression model with correlation  $-0.5$  is doing as well as one with correlation  $+0.5$ . They just have different directions. But if we *square* the correlation coefficient, we'll get a value between 0 and 1, and the direction won't matter. The squared correlation,  $r^2$ , gives the fraction of the data's variation accounted for by the model, and  $1 - r^2$  is the fraction of the original variation left in the residuals. For the RONA *Sales* model,  $r^2 = 0.885^2 = 0.783$  and  $1 - r^2$  is 0.216, so only 21.6% of the variability in *Sales* has been left in the residuals.

All regression analyses include this statistic, although by tradition, it is written with a capital letter,  $R^2$ , and pronounced "R-squared." An  $R^2$  of 0 means that none of the variance in the data is in the model; all of it is still in the residuals. It would be hard to imagine using that model for anything.

Because  $R^2$  is a fraction of a whole, it is often given as a percentage.<sup>6</sup> For the RONA *Sales* data,  $R^2$  is 78.3%.

When interpreting a regression model, you need to report what  $R^2$  means. According to our linear model, 78.3% of the variability in RONA *Sales* is accounted for by variation in residential *Renovations* expenditures.

- ◆ **How can we see that  $R^2$  is really the fraction of variance accounted for by the model?** It's a simple calculation. The variance of *Sales* is  $288.4^2 = 83\,175$ . If we treat the residuals as data, the variance of the residuals is 17\,961.<sup>7</sup> As a fraction of the variance of *Sales*, that's 0.216 or 21.6%. That's the fraction of the variance that is *not* accounted for by the model. The fraction that *is* accounted for is  $100\% - 21.6\% = 78.3\%$ , just the value we got for  $R^2$ .

### How Big Should $R^2$ Be?

The value of  $R^2$  is always between 0% and 100%. But what is a "good"  $R^2$  value? The answer depends on the kind of data you are analyzing and on what you want to do with it. Just as with correlation, there is no value for  $R^2$  that automatically determines that the regression is "good." Data from scientific experiments often



### JUST CHECKING

Let's go back to our regression of sales (\$000) on number of employees again.

$$\widehat{Sales} = 9.564 + 122.74 \text{ Employees}$$

The  $R^2$  value is reported as 71.4%.

- 9 What does the  $R^2$  value mean about the relationship of *Sales* and *Employees*?
- 10 Is the correlation of *Sales* and *Employees* positive or negative? How do you know?
- 11 If we measured the *Sales* in thousands of euros instead of thousands of dollars, would the  $R^2$  value change? How about the slope?

<sup>6</sup> By contrast, we usually give correlation coefficients as decimal values between  $-1.0$  and  $1.0$ .

<sup>7</sup> This isn't quite the same as squaring  $s_e$  which we discussed previously, but it's very close.

**Sum of Squares**

The sum of the squared residuals  $\sum(y - \hat{y})^2$  is sometimes written as SSE (sum of squared errors). If we call  $\sum(y - \bar{y})^2$  SST (for total sum of squares) then

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

have  $R^2$  in the 80% to 90% range and even higher. Data from observational studies and surveys, though, often show relatively weak associations because it's so difficult to measure reliable responses. An  $R^2$  of 30% to 50% or even lower might be taken as evidence of a useful regression. The standard deviation of the residuals can give us more information about the usefulness of the regression by telling us how much scatter there is around the line.

As we've seen, an  $R^2$  of 100% is a perfect fit, with no scatter around the line. The  $s_e$  would be zero. All of the variance would be accounted for by the model with none left in the residuals. This sounds great, but it's too good to be true for real data.<sup>8</sup>

## 6.10 Reality Check: Is the Regression Reasonable?

Statistics don't come out of nowhere. They are based on data. The results of a statistical analysis should reinforce common sense. If the results are surprising, then either you've learned something new about the world or your analysis is wrong.

Whenever you perform a regression, think about the coefficients and ask whether they make sense. Is the slope reasonable? Does the direction of the slope seem right? The small effort of asking whether the regression equation is plausible will be repaid whenever you catch errors or avoid saying something silly or absurd about the data. It's too easy to take something that comes out of a computer at face value and assume that it makes sense.

Always be skeptical and ask yourself if the answer is reasonable.

### GUIDED EXAMPLE

### Home Size and Price

Real estate agents know the three most important factors in determining the price of a house are *location, location, and location*. But what other factors help determine the price at which a house should be listed? Number of bathrooms? Size of the yard? A student amassed publicly available data on thousands of homes. We've drawn a random sample of

1057 homes to examine house pricing. Among the variables she collected were the total living area (in square feet), number of bathrooms, number of bedrooms, size of lot (in acres), and age of house (in years). We will investigate how well the size of the house, as measured by living area, can predict the selling price.



**Setup** State the objective of the study.  
Identify the variables and their context.

We want to find out how well the living area of a house can predict its selling price.

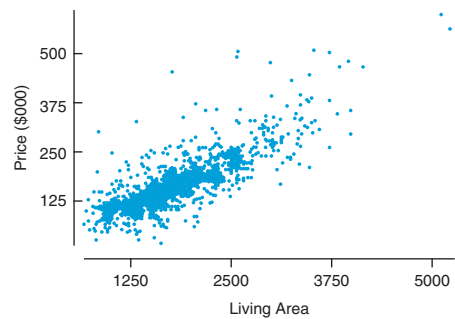
We have two quantitative variables: the living area (in square feet) and the selling price (\$). These data come from public records in 2006.

<sup>8</sup> If you see an  $R^2$  of 100%, it's a good idea to investigate what happened. You may have accidentally regressed two variables that measure the same thing.

**Model** We need to check the same conditions for regression as we did for correlation. To do that, make a picture. Never fit a regression without looking at the scatterplot first.

Check the Linearity, Equal Spread, and Outlier Conditions.

✓ **Quantitative Variables Condition**



✓ **Linearity Condition** The scatterplot shows two variables that appear to have a fairly strong positive association. The plot appears to be fairly linear.

✓ **Equal Spread Condition** The scatterplot shows a consistent spread across the x-values.

✓ **Outlier Condition** There appear to be a few possible outliers, especially among large, relatively expensive houses. A few smaller houses are expensive for their size. We will check their influence on the model later.

We have two quantitative variables that appear to satisfy the conditions, so we will model this relationship with a regression line.



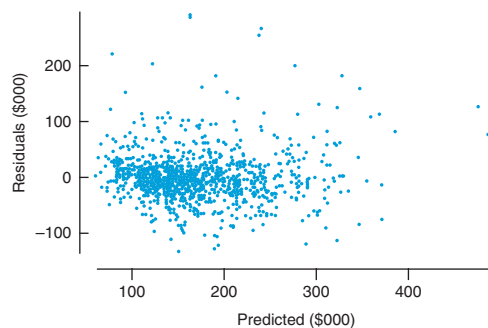
**Mechanics** Find the equation of the regression line using a statistics package. Remember to write the equation of the model using meaningful variable names.


Once you have the model, plot the residuals and check the Equal Spread Condition again.

Our software produces the following output.

```

Dependent variable is: Price
1057 total cases
R squared = 62.43%
s = 57930 with 1000 - 2 = 998 df
Variable      Coefficient
Intercept     6378.08
Living Area   115.13
    
```



		<p>The residual plot appears generally patternless. The few relatively expensive small houses are evident, but setting them aside and refitting the model did not change either the slope or intercept very much so we left them in. There is a slight tendency for cheaper houses to have less variation, but the spread is roughly the same throughout.</p>
	<p><b>Conclusion</b> Interpret what you have found in the proper context.</p>	<p><b>MEMO:</b></p> <p><b>Re: Report on housing prices.</b></p> <p>We examined how well the size of a house could predict its selling price. Data were obtained from recent sales of 1057 homes. The model is:</p> $\widehat{\text{Price}} = \$6376.08 + 115.13 + \text{Living Area}$ <p>In other words, from a base of \$6376.08, houses cost about \$115.13 per square foot.</p> <p>This model appears reasonable from both a statistical and real estate perspective. Although we know that size is not the only factor in pricing a house, the model accounts for 62.4% of the variation in selling price.</p> <p>As a reality check, we checked with several real estate pricing sites (<a href="http://www.realestateabc.com">www.realestateabc.com</a>, <a href="http://www.zillow.com">www.zillow.com</a>) and found that houses in this region were averaging \$100 to \$150 per square foot, so our model is plausible.</p> <p>Of course, not all house prices are predicted well by the model. We computed the model without several of these houses, but their impact on the regression model was small. We believe that this is a reasonable place to start to assess whether a house is priced correctly for this market. Future analysis might benefit by considering other factors.</p>



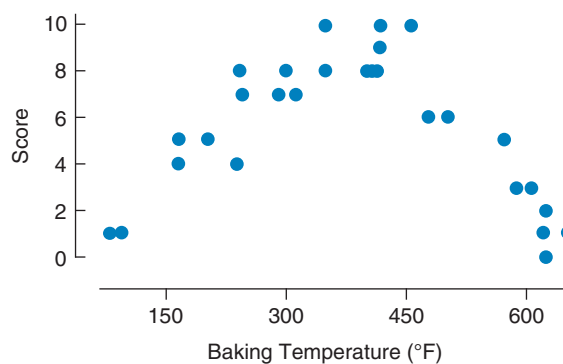
## WHAT CAN GO WRONG?

- **Don't say "correlation" when you mean "association."** How often have you heard the word "correlation"? Chances are pretty good that when you've heard the term, it's been misused. It's one of the most widely misused Statistics terms, and given how often Statistics are misused, that's saying a lot. One of the problems is that many people use the specific term *correlation* when they really mean the more general term *association*. Association is a deliberately vague term used to describe the relationship between two variables.

Correlation is a precise term used to describe the strength and direction of a linear relationship between quantitative variables.

(continued)

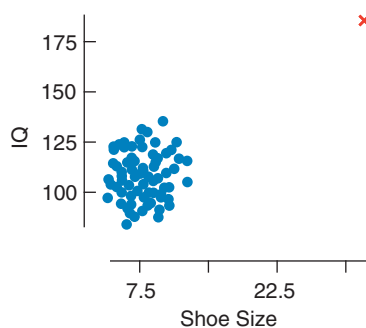
- **Don't correlate categorical variables.** Be sure to check the Quantitative Variables Condition. It makes no sense to compute a correlation of categorical variables.
- **Make sure the association is linear.** Not all associations between quantitative variables are linear. Correlation can miss even a strong nonlinear association. And linear regression models are never appropriate for relationships that are not linear. A company, concerned that customers might use ovens with imperfect temperature controls, performed a series of experiments<sup>9</sup> to assess the effect of baking temperature on the quality of brownies made from their freeze-dried reconstituted brownies. The company wants to understand the sensitivity of brownie quality to variation in oven temperatures around the recommended baking temperature of 325°F. The lab reported a correlation of  $-0.05$  between the scores awarded by a panel of trained taste-testers and baking temperature and a regression slope of  $-0.02$ , so they told management that there is no relationship. Before printing directions on the box telling customers not to worry about the temperature, a savvy intern asks to see the scatterplot.



**Figure 6.8** The relationship between brownie taste score and baking temperature is strong, but not linear.

The plot actually shows a strong association—but not a linear one. Don't forget to check the Linearity Condition.

- **Beware of outliers.** You can't interpret a correlation coefficient or a regression model safely without a background check for unusual observations. Here's an example. The relationship between IQ and shoe size among comedians shows a surprisingly strong positive correlation of 0.50. To check assumptions, we look at the scatterplot.



**Figure 6.9** IQ vs. Shoe Size.

<sup>9</sup> Experiments designed to assess the impact of environmental variables outside the control of the company on the quality of the company's products were advocated by the Japanese quality expert Dr. Genichi Taguchi starting in the 1980s in the United States.

From this “study,” what can we say about the relationship between the two? The correlation is 0.50. But who *does* that point in the upper right-hand corner belong to? The outlier is Bozo the Clown, known for his large shoes and widely acknowledged to be a comic “genius.” Without Bozo the correlation is near zero.

Even a single unusual observation can dominate the correlation value. That’s why you need to check the Unusual Observations Condition.

- **Don’t confuse correlation with causation.** Once we have a strong correlation, it’s tempting to try to explain it by imagining that the predictor variable has *caused* the response to change. Putting a regression line on a scatterplot tempts us even further. Humans are like that; we tend to see causes and effects in everything. Just because two variables are related does not mean that one *causes* the other.

**Does cancer cause smoking?** Even if the correlation of two variables is due to a causal relationship, the correlation itself cannot tell us what causes what.

Sir Ronald Aylmer Fisher (1890–1962) was one of the greatest statisticians of the twentieth century. Fisher testified in court (paid by the tobacco companies) that a causal relationship might underlie the correlation of smoking and cancer:

“Is it possible, then, that lung cancer . . . is one of the causes of smoking cigarettes? I don’t think it can be excluded . . . the pre-cancerous condition is one involving a certain amount of slight chronic inflammation . . .

A slight cause of irritation . . . is commonly accompanied by pulling out a cigarette, and getting a little compensation for life’s minor ills in that way. And . . . is not unlikely to be associated with smoking more frequently.”

Ironically, the proof that smoking indeed is the cause of many cancers came from experiments conducted following the principles of experiment design and analysis that Fisher himself developed.

In 2012, the prestigious *New England Journal of Medicine* published research that found countries with higher chocolate consumption win more Nobel prizes. The also-prestigious British journal *Practical Neurology* followed this up with a study that showed a nearly identical link between milk consumption and Nobel prize success. Putting the two results together must mean that chocolate milk (or milk chocolate) is the ultimate brain food! Both studies were, of course, tongue-in-cheek but were undertaken to emphasize the difference between correlation and causation, and that this difference is often overlooked.

A much less serious organization, the Church of the Flying Spaghetti Monster (FSM) published a graph showing a strong negative correlation between the world population of pirates and average global temperatures over the past 200 years (see [www.venganza.org](http://www.venganza.org)). According to the FSM founder, “Global warming, earthquakes, hurricanes, and other natural disasters are a direct effect of the shrinking numbers of pirates since the 1800s.” We point out that in recent years the pirate population has begun increasing again (e.g., off the coast of Somalia), and global temperatures are decreasing (which is why the term climate change has superseded the term global warming). Of course, the real reason for global warming is the end of the Cold War!

Scatterplots, correlation coefficients, and regression models *never* prove causation. This is, for example, partly why it took so long for the government to require warning labels on cigarettes. Although there was plenty of evidence that increased smoking was *associated* with increased levels of lung cancer, it took years to provide evidence that smoking actually *causes* lung cancer. (The tobacco companies used this to great advantage.)

- **Watch out for lurking variables.** A scatterplot of the damage (in dollars) caused to a house by fire would show a strong correlation with the number of firefighters at the scene. Surely the damage doesn’t cause firefighters. And firefighters actually do cause damage, spraying water all around and chopping holes, but does that mean we shouldn’t call the fire department? Of course not. There is an underlying variable that leads to both more damage and more firefighters—the size of the blaze. A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a *lurking variable*. You can often debunk claims made about data by finding a lurking variable behind the scenes.
- **Don’t fit a straight line to a nonlinear relationship.** Linear regression is suited only to relationships that are, in fact, linear.

(continued)



- **Beware of extraordinary points.** Data values can be extraordinary or unusual in a regression in two ways. They can have  $y$ -values that stand off from the linear pattern suggested by the bulk of the data. These are what we have been calling outliers; although with regression, a point can be an outlier by being far from the linear pattern even if it is not the largest or smallest  $y$ -value. Points can also be extraordinary in their  $x$ -values. Such points can exert a strong influence on the line. Both kinds of extraordinary points require attention.
- **Don't extrapolate far beyond the data. A linear model will often do a reasonable job of summarizing a relationship in the range of observed  $x$ -values.** Once we have a working model for the relationship, it's tempting to use it. But beware of predicting  $y$ -values for  $x$ -values that lie too far outside the range of the original data. The model may no longer hold there, so such extrapolations too far from the data are dangerous.
- **Don't choose a model based on  $R^2$  alone.** Although  $R^2$  measures the *strength* of the linear association, a high  $R^2$  does not demonstrate the *appropriateness* of the regression. A single unusual observation, or data that separate into two groups, can make the  $R^2$  seem quite large when, in fact, the linear regression model is simply inappropriate. Conversely, a low  $R^2$  value may be due to a single outlier. It may be that most of the data fall roughly along a straight line, with the exception of a single point. Always look at the scatterplot.
- **Beware of the dangers of computing correlation aggregated across different groups.** In Chapter 4 we discussed Simpson's Paradox, and how absurd results can occur when measurements from different groups are combined. The reversals that can happen with categorical data and percentages can also happen with quantitative data and correlation. For example, income is generally accepted to be positively correlated with education. But if a scatterplot were prepared using two groups of people, National Hockey League players and university professors, a negative correlation would be seen. NHL players have much higher salaries, and much lower education than professors!



## ETHICS IN ACTION

An ad agency hired by a well-known manufacturer of dental hygiene products (electric toothbrushes, oral irrigators, etc.) put together a creative team to brainstorm ideas for a new ad campaign. Trisha Simes was chosen to lead the team as she has had the most experience with this client to date. At their first meeting, Trisha communicated to her team the client's desire to differentiate themselves from their competitors by not focusing their message on the cosmetic benefits of good dental care. As they brainstormed ideas, one member of the team, Brad Jonns, recalled a recent CNN broadcast that reported a "correlation" between flossing teeth and reduced risk of heart disease. Seeing potential in promoting the health benefits of proper dental care, the team agreed to pursue this idea further. At their next meeting several team members commented on how surprised they were to find so many articles, medical, scientific, and popular, that seemed to claim good dental hygiene resulted in good health. One member noted that he found articles that linked gum disease not only to heart attacks and strokes but to diabetes and even cancer. Although Trisha puzzled over why their client's competitors had not yet capitalized on these research findings, her team was on a roll and had already begun to focus on designing the campaign around this core message.

**ETHICAL ISSUE** *Correlation does not imply causation. The possibility of lurking variables is not explored. For example, it is likely that those who take better care of themselves would floss regularly and also have less risk of heart disease (related to ASA Ethical Guidelines which can be found at <http://www.amstat.org/about/ethicalguidelines.cfm>).*

**ETHICAL SOLUTION** *Refrain from implying cause and effect from correlation results.*

Jill Hathway is looking for a career change and is interested in starting a franchise. After spending the last 20 years working

as a mid-level manager for a major corporation, Jill wants to indulge her entrepreneurial spirit and strike out on her own. She is considering a franchise in the health and fitness industry, including *Pilates One*, for which she requested a franchise packet. Included in the packet information were data showing how various regional demographics (age, gender, income) related to franchise success (revenue, profit, return on investment). *Pilates One* is a relatively new franchise with only a few scattered locations. Nonetheless, the company reported various graphs and data analysis results to help prospective franchisers in their decision-making process. Jill was particularly interested in the graph and the regression analysis that related the proportion of women over the age of 40 within a 30-kilometre radius of a *Pilates One* location to return on investment for the franchise. She noticed that there was a positive relationship. With a little research, she discovered that the proportion of women over the age of 40 in her city was higher than for any other *Pilates One* location (attributable, in part, to the large number of retirees relocating to her city). She then used the regression equation to project return on investment for a *Pilates One* located in her city and was very pleased with the result. With such objective data, she felt confident that *Pilates One* was the franchise for her.

**ETHICAL ISSUE** *Pilates One is reporting analysis based on only a few observations. Jill is extrapolating beyond the range of  $x$ -values (related to ASA Ethical Guidelines which can be found at <http://www.amstat.org/about/ethicalguidelines.cfm>).*

**ETHICAL SOLUTION** *Pilates One should include a disclaimer that the analysis was based on very few observations and that the equation should not be used to predict success at other locations or beyond the range of  $x$ -values used in the analysis.*

## WHAT HAVE WE LEARNED?

In previous chapters we learned how to listen to the story told by data from a single variable. Now we've turned our attention to the more complicated (and more interesting) story we can discover in the association between two quantitative variables.

We've learned to begin our investigation by looking at a scatterplot. We're interested in the *direction* of the association, the *form* it takes, and its *strength*.

We've learned that, although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.

- The sign of the correlation tells us the direction of the association.
- The magnitude of the correlation tells us of the *strength* of a linear association. Strong associations have correlations near  $+1$  or  $-1$ , and very weak associations have correlations near 0.

- Correlation has no units, so shifting or scaling the data, standardizing, or even swapping the variables has no effect on the numerical value.

We've learned that to use correlation we have to check certain conditions for the analysis to be valid.

- Before finding or talking about a correlation, we'll always check the Linearity Condition.
- And, as always, we'll watch out for unusual observations.

We've learned not to make the mistake of assuming that a high correlation or strong association is evidence of a cause-and-effect relationship. Beware of lurking variables!

We've learned that when the relationship between quantitative variables is linear, a linear model can help summarize that relationship and give us insights about it.

- The regression (best fit) line doesn't pass through all the points, but it is the best compromise in the sense that the sum of squares of the residuals is the smallest possible.

We've learned several things the correlation,  $r$ , tells us about the regression:

- The slope of the line is based on the correlation, adjusted for the standard deviations of  $x$  and  $y$ . We've learned to interpret that slope in context.
- For each SD that a case is away from the mean of  $x$ , we expect it to be  $r$  SDs in  $y$  away from the  $y$  mean.
- Because  $r$  is always between  $-1$  and  $+1$ , each predicted  $y$  is fewer SDs away from its mean than the corresponding  $x$  was, a phenomenon called *regression to the mean*.
- The square of the correlation coefficient,  $R^2$ , gives us the fraction of the variation of the response accounted for by the regression model. The remaining  $1 - R^2$  of the variation is left in the residuals.

## Terms

### Association

- **Direction:** A **positive** direction or association means that, in general, as one variable increases, so does the other. When increases in one variable generally correspond to decreases in the other, the association is **negative**.
- **Form:** The form we care about most is **linear**, but you should certainly describe other patterns you see in scatterplots.
- **Strength:** A scatterplot is said to show a strong association if there is little scatter around the underlying relationship.

### Correlation coefficient

A numerical measure of the direction and strength of a linear association.

$$r = \frac{\sum z_x z_y}{n - 1}$$

### Explanatory or independent variable ( $x$ -variable)

The variable that accounts for, explains, predicts, or is otherwise responsible for the  $y$ -variable.

### Intercept

The intercept,  $b_0$ , gives a starting value in  $y$ -units. It's the  $\hat{y}$  value when  $x$  is 0.

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

<b>Least squares</b>	A criterion that specifies the unique line that minimizes the variance of the residuals or, equivalently, the sum of the squared residuals. The resulting line is called the <b>Least squares line</b> .
<b>Linear model (Line of best fit)</b>	The linear model of the form $\hat{y} = b_0 + b_1x$ fit by least squares. Also called the regression line. To interpret a linear model, we need to know the variables and their units.
<b>Lurking variable</b>	A variable other than $x$ and $y$ that simultaneously affects both variables, accounting for the correlation between the two.
<b>Outlier</b>	A point that does not fit the overall pattern seen in the scatterplot.
<b>Predicted value</b>	The prediction for $y$ found for each $x$ -value in the data. A predicted value, $\hat{y}$ , is found by substituting the $x$ -value in the regression equation. The predicted values are the values on the fitted line; the points $(x, \hat{y})$ lie exactly on the fitted line.
<b>Regression line</b>	The particular linear equation that satisfies the least squares criterion, often called the line of best fit.
<b>Regression to the mean</b>	Because the correlation is always less than 1.0 in magnitude, each predicted $y$ tends to be fewer standard deviations from its mean than its corresponding $x$ is from its mean.
<b>Residual</b>	The difference between the actual data value and the corresponding value predicted by the regression model—or, more generally, predicted by any model.
<b>Response or dependent variable (<math>y</math>-variable)</b>	The variable that the scatterplot is meant to explain or predict.
<b><math>R^2</math></b>	<ul style="list-style-type: none"> <li>• The square of the correlation between <math>y</math> and <math>x</math></li> <li>• The fraction of the variability of <math>y</math> accounted for by the least squares linear regression on <math>x</math></li> <li>• An overall measure of how successful the regression is in linearly relating <math>y</math> to <math>x</math></li> </ul>
<b>Scatterplot</b>	A graph that shows the relationship between two quantitative variables measured on the same cases.
<b>Standard deviation of the residuals</b>	<p><math>s_e</math> is found by:</p> $s_e = \sqrt{\frac{\sum e^2}{n - 2}}$
<b>Slope</b>	<p>The slope, <math>b_1</math>, is given in <math>y</math>-units per <math>x</math>-unit. Differences of one unit in <math>x</math> are associated with differences of <math>b_1</math> units in predicted values of <math>y</math>:</p> $b_1 = r \frac{s_y}{s_x}$

## Skills



- Recognize when interest in the pattern of a possible relationship between two quantitative variables suggests making a scatterplot.

- Be able to identify the roles of the variables and to place the response variable on the  $y$ -axis and the explanatory variable on the  $x$ -axis.
- Know the conditions for correlation and how to check them.
- Know that correlations are between  $-1$  and  $+1$  and that each extreme indicates a perfect linear association.
- Understand how the magnitude of the correlation reflects the strength of a linear association as viewed in a scatterplot.
- Know that the correlation has no units.
- Know that the correlation coefficient is not changed by changing the centre or scale of either variable.
- Understand that causation cannot be demonstrated by a scatterplot or correlation.
- Know how to identify response ( $y$ ) and explanatory ( $x$ ) variables in context.
- Understand how a linear equation summarizes the relationship between two variables.
- Recognize when a regression should be used to summarize a linear relationship between two quantitative variables.
- Know how to judge whether the slope of a regression makes sense.
- Examine a scatterplot of your data for violations of the Linearity, Equal Spread, and Outlier Conditions that would make it inappropriate to compute a regression.
- Understand that the least squares slope is easily affected by extreme values.
- Define residuals as the differences between the data values and the corresponding values predicted by the line, and that the Least Squares Criterion finds the line that minimizes the sum of the squared residuals.



- Be able to make a scatterplot by hand (for a small set of data) or with technology.
- Know how to compute the correlation of two variables.
- Know how to read a correlation table produced by a statistics program.
- Know how to find the slope and intercept values of a regression.
- Be able to use regression to predict a value of  $y$  for a given  $x$ .
- Know how to compute the residual for each data value and how to compute the standard deviation of the residuals.
- Be able to evaluate the Equal Spread Condition with a scatterplot of the residuals after computing the regression.



- Be able to describe the direction, form, and strength of a scatterplot.
- Be prepared to identify and describe points that deviate from the overall pattern.
- Be able to use correlation as part of the description of a scatterplot.
- Be alert to misinterpretations of correlation.
- Understand that finding a correlation between two variables does not indicate a causal relationship between them. Beware the dangers of suggesting causal relationships when describing correlations.

- Write a sentence explaining what a linear equation says about the relationship between  $y$  and  $x$ , basing it on the fact that the slope is given in  $y$ -units per  $x$ -unit.
- Understand how the correlation coefficient and the regression slope are related. Know that  $R^2$  describes how much of the variation in  $y$  is accounted for by its linear relationship with  $x$ .
- Be able to describe a prediction made from a regression equation, relating the predicted value to the specified  $x$ -value.

## TECHNOLOGY HELP: Correlation and Regression

All statistics packages make a table of results for a regression. These tables may differ slightly from one package to another, but all are essentially the same—and all include much more than we need to know for now. Every computer regression table includes a section that looks something like this:

Standard dev of residuals ( $s_e$ ): In some packages this is called Root MSE. In Excel it is misnamed as Standard Error.

$R$  squared

The “dependent,” response, or  $y$ -variable

Dependent variable is: Sales					
R squared = 69.0%					
$s = 9.277$					
Variable	Coefficient	SE(Coeff)	t-ratio	P-value	
Intercept	6.83077	2.664	2.56	0.0158	
Shelf Space	0.971381	0.1209	8.04	$\leq 0.0001$	

The “independent,” predictor, or  $x$ -variable

The slope

The intercept

We'll deal with all of these later in the book. You may ignore them for now.

The slope and intercept coefficient are given in a table such as this one. Usually the slope is labelled with the name of the  $x$ -variable, and the intercept is labelled “Intercept” or “Constant.” So the regression equation shown here is

$$\widehat{\text{Sales}} = 6.83077 + 0.97138 \text{ Shelf Space.}$$

### EXCEL

To make a scatterplot with the Excel Chart Wizard,

- Click on the **Chart Wizard** Button in the menu bar. Excel opens the Chart Wizard's Chart Type Dialog window.
- Make sure the **Standard Types** tab is selected, and select **XY (Scatter)** from the choices offered.
- Specify the **scatterplot without** lines from the choices offered in the Chart subtype selections. The **Next** button takes you to the Chart Source Data dialog.
- If it is not already frontmost, click on the **Data Range** tab, and enter the data range in the space provided.
- By convention, we always represent variables in columns. The Chart Wizard refers to variables as Series. Be sure the **Column** option is selected.
- Excel places the leftmost column of those you select on the  $x$ -axis of the scatterplot. If the column you wish to see on the  $x$ -axis is not the leftmost column in your spreadsheet, click on the **Series** tab and edit the specification of the individual axis series.
- Click the **Next** button. The Chart Options dialog appears.
- Select the **Titles** tab. Here you specify the title of the chart and names of the variables displayed on each axis.



- Type the chart title in the **Chart title:** edit box.
- Type the x-axis variable name in the **Value (X) Axis:** edit box. Note that you must name the columns correctly here. Naming another variable will not alter the plot, only mislabel it.
- Type the y-axis variable name in the **Value (Y) Axis:** edit box.
- Click the **Next** button to open the chart location dialog.
- Select the **As new sheet:** option button.
- Click the **Finish** button.

Often, the resulting scatterplot will require rescaling. By default, Excel includes the origin in the plot even when the data are far from zero. You can adjust the axis scales. To change the scale of a plot axis in Excel,

- Double-click on the axis. The **Format Axis Dialog** appears.
- If the **scale tab** is not the frontmost, select it.
- Enter new minimum or new maximum values in the spaces provided. You can drag the dialog box over the scatterplot as a straightedge to help you read the maximum and minimum values on the axes.
- Click the **OK** button to view the rescaled scatterplot.
- Follow the same steps for the x-axis scale.

Compute a correlation in Excel with the **CORREL** function from the drop-down menu of functions. If **CORREL** is not on the menu, choose **More Functions** and find it among the statistical functions in the browser.

In the dialog box that pops up, enter the range of cells holding one of the variables in the space provided.

Enter the range of cells for the other variable in the space provided. To calculate a regression, make a scatterplot of the data. With the scatterplot front-most, select **Add Trendline . . .** from the **Chart** menu. Click the **Options** tab and select **Display Equation on Chart**. Click **OK**.

## EXCEL 2007

To make a scatterplot in Excel 2007:

- Select the columns of data to use in the scatterplot. You can select more than one column by holding down the control key while clicking.
- In the Insert tab, click on the **Scatter** button and select the **Scatter with only Markers** chart from the menu.

To make the plot more useful for data analysis, adjust the display as follows:

- With the chart selected, click on the **Gridlines** button in the Layout tab to cause the Chart Tools tab to appear.
- Within Primary Horizontal Gridlines, select **None**. This will remove the gridlines from the scatterplot.
- To change the axis scaling, click on the numbers of each axis of the chart, and click on the **Format Selection** button in the Layout tab.
- Select the **Fixed** option instead of the Auto option, and type a value more suited for the scatterplot. You can use the popup dialog window as a straightedge to approximate the appropriate values.

Excel 2007 automatically places the leftmost of the two columns you select on the x-axis, and the rightmost one on the y-axis. If that's not what you'd prefer for your plot, you'll want to switch them.

To switch the X- and Y-variables:

- Click the chart to access the **Chart Tools** tabs.
- Click on the **Select Data** button in the Design tab.

- In the popup window's Legend Entries box, click on **Edit**.
- Highlight and delete everything in the Series X Values line, and select new data from the spreadsheet. (Note that selecting the column would inadvertently select the title of the column, which would not work well here.)
- Do the same with the Series Y Values line.
- Press **OK**, then press **OK** again.

To calculate a correlation coefficient:

- Click on a blank cell in the spreadsheet.
- Go to the **Formulas** tab in the Ribbon and click **More Functions** → **Statistical**.
- Choose the **CORREL** function from the drop-down menu of functions.
- In the dialog that pops up, enter the range of one of the variables in the space provided.
- Enter the range of the other variable in the space provided.
- Click **OK**.

### Comments

The correlation is computed in the selected cell. Correlations computed this way will update if any of the data values are changed. Before you interpret a correlation coefficient, always make a scatterplot to check for nonlinearity and outliers. If the variables are not linearly related, the correlation coefficient cannot be interpreted.



## MINI CASE STUDY PROJECTS



Isak55/Shutterstock

### Fuel Efficiency

With the ever-increasing price of gasoline, both drivers and auto companies are motivated to raise the fuel efficiency of cars. There are some simple ways to increase fuel efficiency: avoid rapid acceleration, avoid driving over 90 kph, reduce idling, and reduce the vehicle's weight. An extra 50 kilograms can reduce fuel efficiency by up to 2%. A marketing executive is studying the relationship between the fuel efficiency of cars (as measured in litres per 100 kilometres [L/100km]) and their weight to design a new compact car campaign. In the data set **ch06\_MCSP\_Fuel\_Efficiency\_Canada.xlsx** you'll find data on the variables below.<sup>10</sup>

- Model of Car
- Engine Size (L)
- Cylinders
- MSRP (Manufacturer's Suggested Retail Price in \$)
- City (L/100 km)
- Highway (L/100 km)
- Weight (kilograms)
- Type and Country of manufacturer

Describe the relationship of weight, MSRP, and engine size with fuel efficiency (both city and highway) in a written report. Be sure to plot the residuals.

### Energy Use at YVR

In 2013, for the fourth year in a row, the Vancouver International Airport Authority (YVR) was named the Best Airport in North America (and 8th overall worldwide) by Skytrax World Airport Awards. The operation of an airport is a complex undertaking. Budget planning requires being able to forecast costs of energy to operate the airport. With a clear idea of needs, it may be easier to negotiate favourable contracts with energy suppliers.

Earlier in this Chapter, we looked at the scatterplot of energy use versus number of passengers. Now we examine additional factors, and their relationship with energy use. The data file **ch06\_MCSP\_Energy\_Use\_YVR.xlsx** has the following variables on a monthly basis from January 1997 to December 2010.

- Date (month and year)
- Energy Use (thousands of kWh = kilowatt hours)
- MeanTemp = Mean monthly temperature (degrees Celsius)
- TotalArea = Total Area of all terminals (sq. m.)
- Pax\_Domestic = Domestic passengers (000s)
- Pax\_US = U.S (Trans-border) passengers (000s)
- Pax\_Intl = International passengers (000s)
- Pax\_Total = Total passengers (000s)

Describe the relationships between *Energy Use* and each of *MeanTemp*, *TotalArea*, and *Pax\_Total* (i.e., three separate relationships) in a written report. Based on correlations and linear regression, which data provide the best prediction of *Energy Use*?

### Cost of Living

The Mercer Human Resource Consulting website ([www.mercerhr.com](http://www.mercerhr.com)) lists prices of certain items in selected cities around the world. They also report an overall cost-of-living index for each city compared to the costs of hundreds of items in New York City. For example, London at 110.6 is 10.6% more expensive than New York. You'll find the 2006

<sup>10</sup> Data are from the 2004 model year and were compiled from [www.Edmonds.com](http://www.Edmonds.com).

data for 16 cities in the data set **ch06\_MCSP\_Cost\_of\_Living.xls**. Included are the 2006 cost of living index, cost of a luxury apartment (per month), price of a bus or subway ride, price of a compact disc, price of an international newspaper, price of a cup of coffee (including service), and price of a fast-food hamburger meal. All prices are in U.S. dollars.

Examine the relationship between the overall cost of living and the cost of each of these individual items. Verify the necessary conditions and describe the relationship in as much detail as possible. (Remember to look at direction, form, and strength.) Identify any unusual observations.

Based on the correlations and linear regressions, which item would be the best predictor of overall cost in these cities? Which would be the worst? Are there any surprising relationships? Write a short report detailing your conclusions.

## Canadian Banks

The Canadian Bankers Association works on behalf of 54 domestic banks, foreign bank subsidiaries, and foreign bank branches operating in Canada and provides a centralized contact to all banks on matters relating to banking in Canada. The CBA advocates for effective public policies that contribute to a sound, successful banking system, and promotes financial literacy to help Canadians make informed financial decisions. The CBA is involved in financial data collection and analysis, consumer protection efforts, fighting bank fraud, and developing industry consensus on issues impacting banks in Canada.

In addition to policies on financial issues, management must make decisions on infrastructure—how many branches should be opened, and how many bank machines, or automatic teller machines (ABMs) to dispense cash, should be maintained. Each year, the CBA compiles data on the number of ABMs and number of Canadian branches in each province and in the territories for the major banks in Canada: BMO, Royal Bank, TD, Scotiabank, CIBC, HSBC, Laurentian, and National Bank.

The data file **ch06\_MCSP\_Canadian\_Banks.xlsx** has 2011 data on number of branches, number of ABMs, as well as the provincial population (in 000s) and the provincial GDP (in \$ millions). The latter two come from Statistics Canada data.

Prepare four scatterplots of: *Branches* against *Population* and *GDP*, and *ABMs* against *Population* and *GDP*. Construct a correlation table from all four variables. Why are the correlation coefficients so high? Which is the better predictor of branches—population or GDP? Which is the better predictor of ABMs—population or GDP? Compute two linear regression equations, one for predicting number of branches from population, the other for predicting number of ABMs from population. Examine the residuals to determine which provinces are “underserved;” that is, have fewer branches and fewer ABMs than would be predicted from your models. Write a short report summarizing your findings.

### MyStatLab Students! Save time, improve your grades with MyStatLab.

The Exercises marked in red can be found on MyStatLab. You can practice them as often as you want, and most feature step-by-step guided solutions to help you find the right answer. You'll find a personalized Study Plan available to you too! Data Sets for exercises marked **T** are also available on MyStatLab for formatted technologies.

## EXERCISES

**1. Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction and form. **LO1**

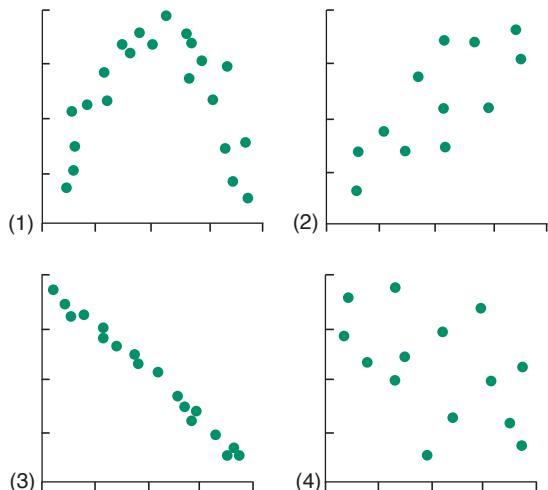
- Cellphone bills: number of text messages, cost.
- Automobiles: Fuel efficiency (L/100 km), sales volume (number of autos).
- For each week: Ice cream cone sales, air conditioner sales.
- Product: Price (\$), demand (number sold per day).

**2. Association, part 2.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction and form. **LO1**

- T-shirts at a store: price each, number sold.
- Real estate: house price, house size (square footage).
- Economics: Interest rates, number of mortgage applications.
- Employees: Salary, years of experience.

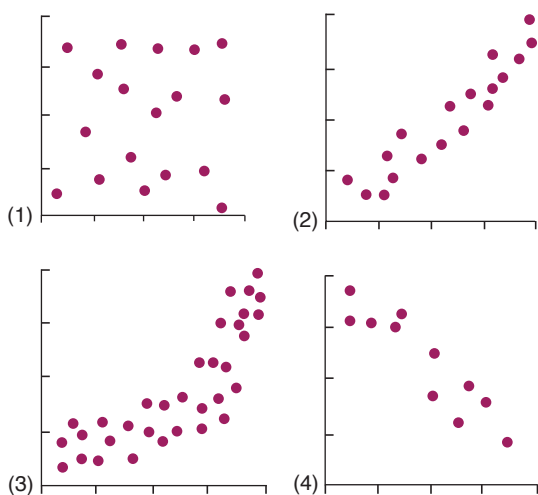
3. Scatterplots. Which of the scatterplots show: LO1

- a) Little or no association?
- b) A negative association?
- c) A linear association?
- d) A moderately strong association?
- e) A very strong association?



4. Scatterplots, part 2. Which of the scatterplots show: LO1

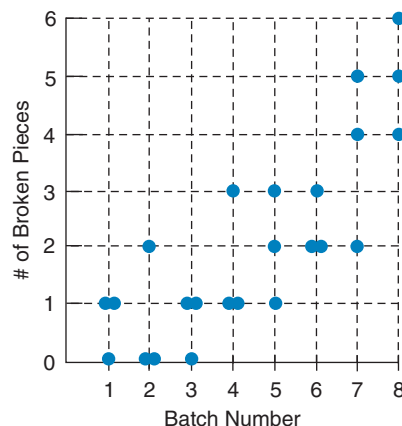
- a) Little or no association?
- b) A negative association?
- c) A linear association?
- d) A moderately strong association?
- e) A very strong association?



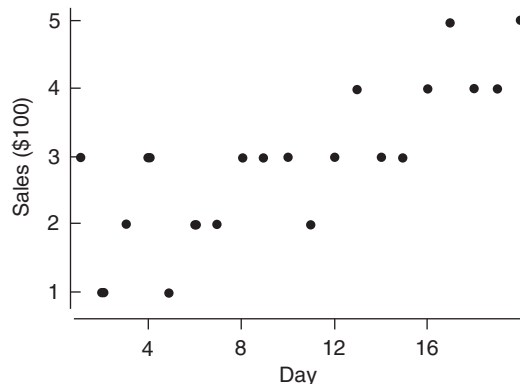
5. Manufacturing. A ceramics factory can fire eight large batches of pottery a day. Sometimes a few of the pieces break in the process. In order to understand the problem better, the factory records the number of broken pieces in each batch for three days and then creates the scatterplot shown. LO1

- a) Make a histogram showing the distribution of the number of broken pieces in the 24 batches of pottery examined.
- b) Describe the distribution as shown in the histogram. What feature of the problem is more apparent in the histogram than in the scatterplot?

c) What aspect of the company's problem is more apparent in the scatterplot?

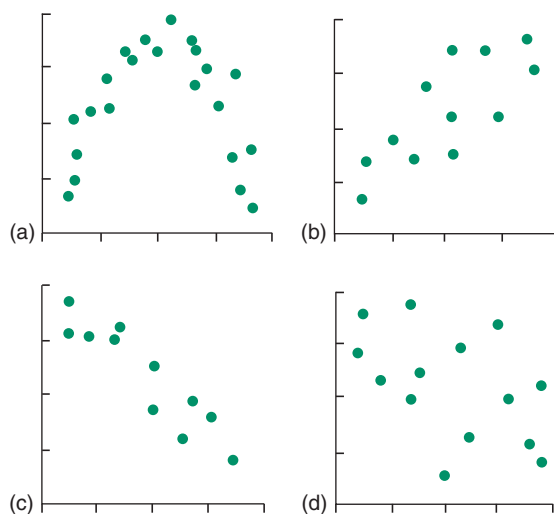


6. Coffee sales. Owners of a new coffee shop tracked sales for the first 20 days and displayed the data in a scatterplot (by day). LO1

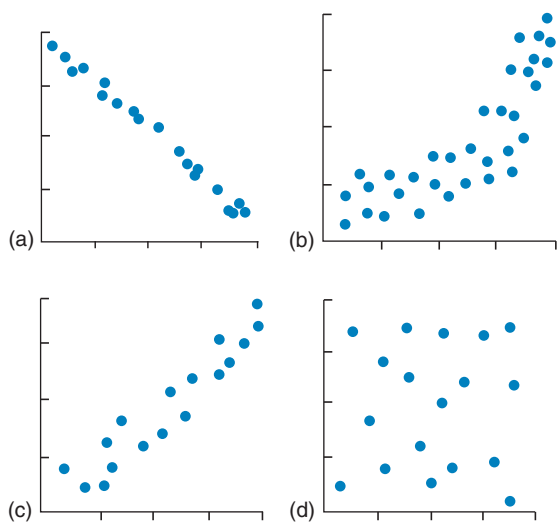


- a) Make a histogram of the daily sales since the shop has been in business.
- b) State one fact that is obvious from the scatterplot, but not from the histogram.
- c) State one fact that is obvious from the histogram, but not from the scatterplot.

7. Matching. Here are several scatterplots. The calculated correlations are  $-0.923$ ,  $-0.487$ ,  $0.006$ , and  $0.777$ . Which is which? LO1



8. **Matching, part 2.** Here are several scatterplots. The calculated correlations are  $-0.977$ ,  $-0.021$ ,  $0.736$ , and  $0.951$ . Which is which? **LO1**



**T 9. Pizza sales and price.** A linear model fit to predict weekly *Sales* of frozen pizza (in kilograms) from the average *Price* (\$/unit) charged by a sample of stores in 39 recent weeks is: **LO2**

$$\widehat{Sales} = 141\,865.53 - 24\,369.49 Price$$

- What is the explanatory variable?
- What is the response variable?
- What does the slope mean in this context?
- What does the  $y$ -intercept mean in this context? Is it meaningful?
- What do you predict the sales to be if the average price charged was \$3.50 for a pizza?
- If the sales for a price of \$3.50 turned out to be 60 000 kilograms, what would the residual be?

**T 10. Used Saab prices.** A linear model to predict the *Price* of a 2004 Saab 9-3 (in \$) from its *Mileage* (in miles) was fit to 38 cars that were available during the week of January 11, 2008 (Kelly's Blue Book, www.kbb.com). The model was: **LO2**

$$\widehat{Price} = 24\,356.15 - 0.0151 Mileage$$

- What is the explanatory variable?
- What is the response variable?
- What does the slope mean in this context?
- What does the  $y$ -intercept mean in this context? Is it meaningful?
- What do you predict the price to be for a car with 100 000 miles on it?
- If the price for a car with 100 000 miles on it was \$24 000, what would the residual be?

**T 11. Football salaries.** Is there a relationship between total team salary and the performance of teams in the National Football League (NFL)? For the 2006 season, a linear model predicting *Wins* (out of 16 regular season games) from the total team *Salary* (\$M) for the 32 teams in the league is: **LO2**

$$\widehat{Wins} = 1.783 + 0.062 Salary$$

- What is the explanatory variable?
- What is the response variable?
- What does the slope mean in this context?
- What does the  $y$ -intercept mean in this context? Is it meaningful?
- If one team spends \$10 million more than another on salary, how many more games on average would you predict them to win?
- If a team spent \$50 million on salaries and won eight games, would they have done better or worse than predicted?
- What would the residual of the team in part f be?

**T 12. Baseball salaries.** In 2007, the Boston Red Sox won the World Series and spent \$143 million on salaries for their players (fathom.info/salaryper). Is there a relationship between salary and team performance in Major League Baseball? For the 2007 season, a linear model fit to the number of *Wins* (out of 162 regular season games) from the team *Salary* (\$M) for the 30 teams in the league is: **LO2**

$$\widehat{Wins} = 70.097 + 0.132 Salary$$

- What is the explanatory variable?
- What is the response variable?
- What does the slope mean in this context?
- What does the  $y$ -intercept mean in this context? Is it meaningful?
- If one team spends \$10 million more than another on salaries, how many more games on average would you predict them to win?
- If a team spent \$110 million on salaries and won half (81) of their games, would they have done better or worse than predicted?
- What would the residual of the team in part f be?

**T 13. Pizza sales and price, revisited.** For the data in Exercise 9, the average *Sales* was 52,697 kilograms (SD = 10,261 kilograms), and the correlation between *Price* and *Sales* was  $-0.547$ .

If the *Price* in a particular week was 1 SD higher than the mean *Price*, how much pizza would you predict was sold that week? **LO2**

**T 14. Used Saab prices, revisited.** The 38 cars in Exercise 10 had an average *Price* of \$23 847 (SD = \$923), and the correlation between *Price* and *Mileage* was  $-0.169$ .

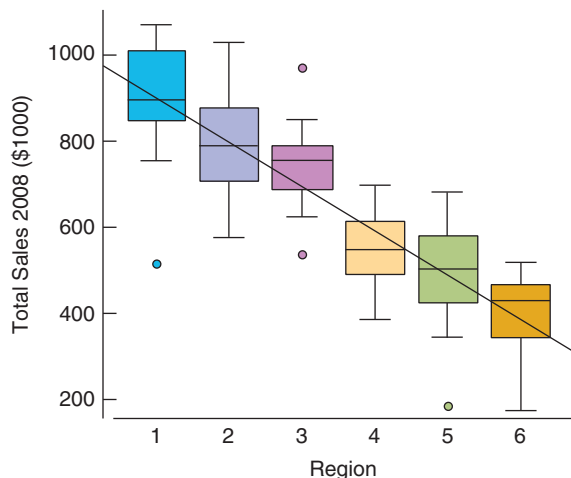
If the *Mileage* of a 2004 Saab was 1 SD below the average number of miles, what *Price* would you predict for it? **LO2**

**15. Packaging.** A CEO announces at the annual shareholders meeting that the new see-through packaging for the company's flagship product has been a success. In fact, he says, "There is a strong correlation between packaging and sales." Criticize this statement on statistical grounds. **LO1**

**16. Insurance.** Insurance companies carefully track claims histories so that they can assess risk and set rates appropriately. The Insurance Bureau of Canada reports that Honda Accords, Honda Civics, and Toyota Camrys are the cars most frequently reported stolen, while Ford Tauruses, Pontiac Vibes, and Buick LeSabres are stolen least often. Is it reasonable to say that there's a correlation between the type of car you own and the risk that it will be stolen? **LO1**



**17. Sales by region.** A sales manager for a major pharmaceutical company analyzes last year's sales data for her 96 sales representatives, grouping them by region (1 = East Coast U.S.; 2 = Mid West U.S.; 3 = West U.S.; 4 = South U.S.; 5 = Canada; 6 = Rest of World). She plots *Sales* (in \$1000) against *Region* (1–6) and sees a strong negative correlation. **LO1**



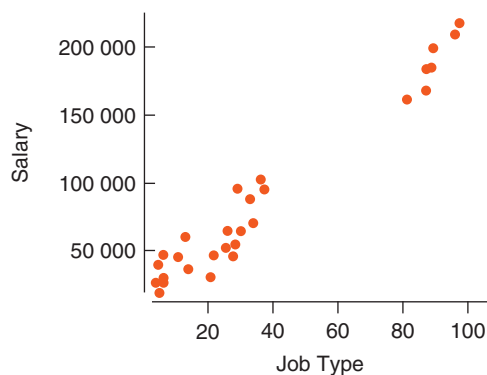
She fits a regression to the data and finds:

$$\widehat{\text{Sales}} = 1002.5 - 102.7 \text{ Region.}$$

The  $R^2$  is 70.5%.

Write a few sentences interpreting this model and describing what she can conclude from this analysis.

**18. Salary by job type.** At a small company, the head of human resources wants to examine salary to prepare annual reviews. He selects 28 employees at random with job types ranging from 01 = Stocking clerk to 99 = President. He plots *Salary* (\$) against *Job Type* and finds a strong linear relationship with a correlation of 0.96. **LO1**



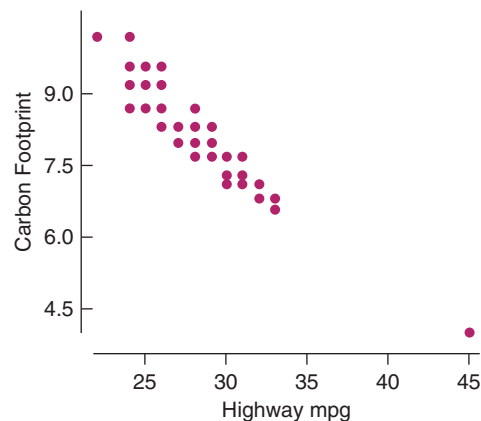
The regression output gives:

$$\widehat{\text{Salary}} = 15\,827.9 + 1939.1 \text{ Job Type}$$

Write a few sentences interpreting this model and describing what he can conclude from this analysis.

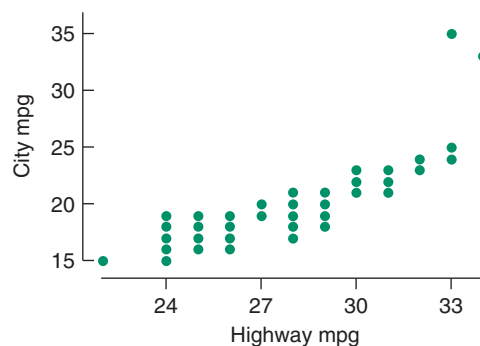
**T 19. Carbon footprint.** The scatterplot shows, for 2008 cars, the carbon footprint (tons of CO<sub>2</sub> per year) vs. the new Environ-

mental Protection Agency (EPA) highway mileage for 82 family sedans as reported by the U.S. government ([www.fueleconomy.gov/feg/findacar.shtml](http://www.fueleconomy.gov/feg/findacar.shtml)). The car with the highest highway mpg and lowest carbon footprint is the Toyota Prius. **LO1**



- The correlation is  $-0.947$ . Describe the association.
- Are the assumptions and conditions met for computing correlation?
- Using technology, find the correlation of the data when the Prius is not included with the others. Can you explain why it changes in that way?

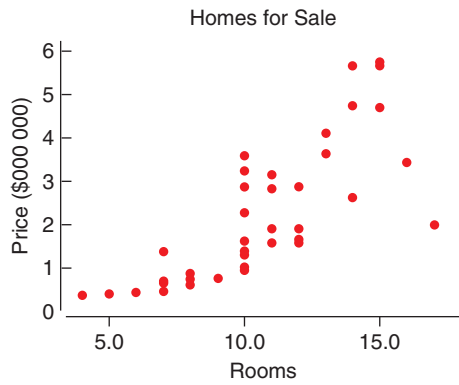
**T 20. EPA mpg.** In 2008, the EPA revised their methods for estimating the fuel efficiency (mpg) of cars—a factor that plays an increasingly important role in car sales. How do the new highway and city estimated mpg values relate to each other? Here's a scatterplot for 83 family sedans as reported by the U.S. government. These are the same cars as in Exercise 19 except that the Toyota Prius has been removed from the data and two other hybrids, the Nissan Altima and Toyota Camry, are included in the data (and are the cars with highest city mpg). **LO1**



- The correlation of these two variables is 0.823. Describe the association.
- If the two hybrids were removed from the data, would you expect the correlation to increase, decrease, or stay the same? Try it using technology. Report and discuss what you find.

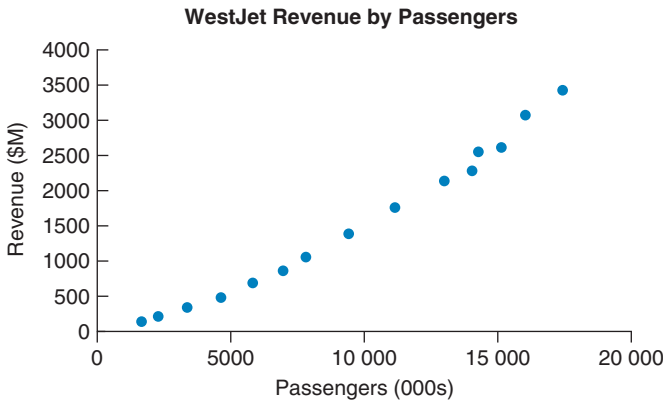
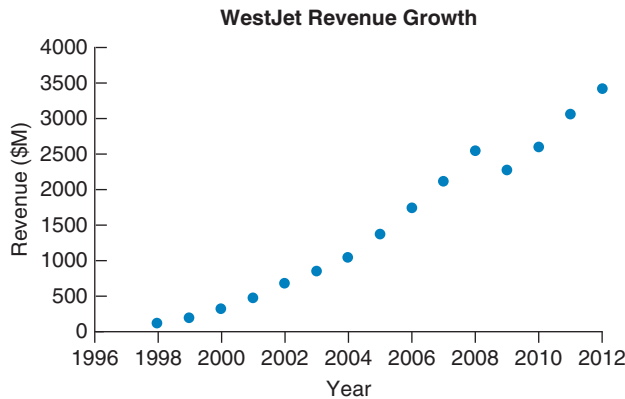
**T 21. Real estate.** Is the number of total rooms in the house associated with the price of a house? Here is the scatterplot of a random sample of homes for sale: **LO1**





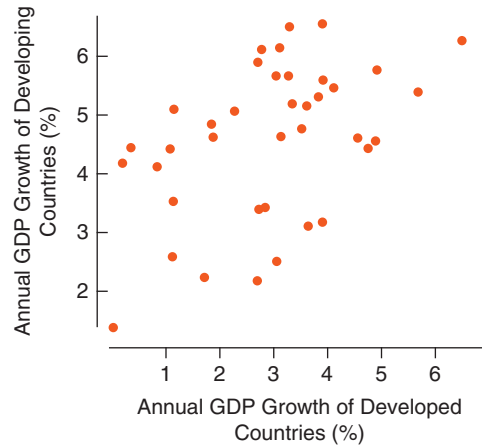
- a) Is there an association?
- b) Check the assumptions and conditions for correlation.

**22. WestJet.** WestJet is Canada’s second largest Canadian air carrier, and the ninth-largest in North America by passengers carried (over 17 million in 2012). Founded in 1996, WestJet is a public company with over 9000 employees, non-unionized, and not part of any airline alliance. Here are two scatterplots of data from 1998–2012. The first shows the growth in annual revenue over time. The second shows the relationship between annual revenue and number of passengers (called “segment guests”). Describe the relationships seen in the two scatterplots. Are they linear? Are there any unusual features or data points? **LO1**



**T 23. GDP growth.** Is economic growth in the developing world related to growth in the industrialized countries? Here’s a scatterplot of the growth (in % of Gross Domestic Product) of the developing countries vs. the growth of developed countries

for 180 countries as grouped by the World Bank. Each point represents one of the years from 1970 to 2007. The output of a regression analysis follows. **LO2**



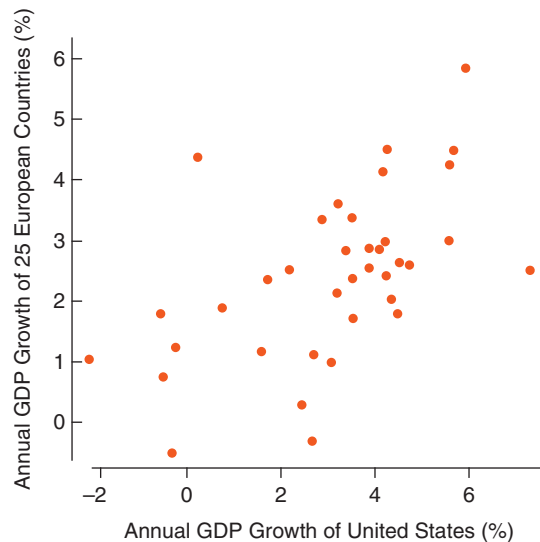
Dependent variable: GDP Growth Developing Countries  
 $R^2 = 20.81\%$   
 $s = 1.244$

Variable	Coefficient
Intercept	3.46
GDP Growth Developed Countries	0.433

- a) Check the assumptions and conditions for the linear model.
- b) Explain the meaning of  $R^2$  in this context.
- c) What are the cases in this model?

**T 24. European GDP growth.** Is economic growth in Europe related to growth in the United States? Here’s a scatterplot of the average growth in 25 European countries (in % of Gross Domestic Product) vs. the growth in the United States. Each point represents one of years from 1970 to 2007. **LO2**

Dependent variable: 25 European Countries GDP Growth  
 $R^2 = 29.65\%$   
 $s = 1.156$



Variable	Coefficient
Intercept	1.330
U.S. GDP Growth	0.3616

- a) Check the assumptions and conditions for the linear model.
- b) Explain the meaning of  $R^2$  in this context.

**T 25. GDP growth part 2.** From the linear model fit to the data on GDP growth of Exercise 23. **LO2**

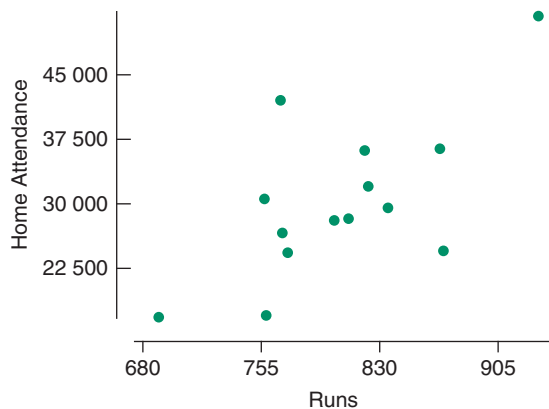
- a) Write the equation of the regression line.
- b) What is the meaning of the intercept? Does it make sense in this context?
- c) Interpret the meaning of the slope.
- d) In a year in which the developed countries grow 4%, what do you predict for the developing world?
- e) In 2007, the developed countries experienced a 2.65% growth, while the developing countries grew at a rate of 6.09%. Is this more or less than you would have predicted?
- f) What is the residual for this year?

**T 26. European GDP growth part 2.** From the linear model fit to the data on GDP growth of Exercise 24. **LO2**

- a) Write the equation of the regression line.
- b) What is the meaning of the intercept? Does it make sense in this context?
- c) Interpret the meaning of the slope.
- d) In a year in which the United States grows at 0%, what do you predict for European growth?
- e) In 2007, the United States experienced a 3.20% growth, while Europe grew at a rate of 2.16%. Is this more or less than you would have predicted?
- f) What is the residual for this year?

**T 27. Attendance 2006.** American League baseball games are played under the designated hitter rule, meaning that weak-hitting pitchers do not come to bat. Baseball owners believe that the designated hitter rule means more runs scored, which in turn means higher attendance. Is there evidence that more fans attend games if the teams score more runs? Data collected from American League games during the 2006 season have a correlation of 0.667 between *Runs Scored* and the number of people at the game (www.mlb.com). **LO1**

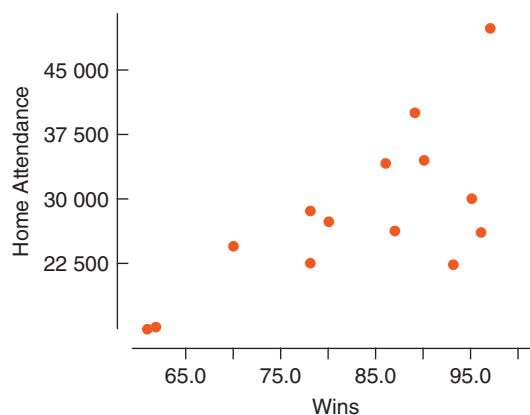
- a) Does the scatterplot indicate that it's appropriate to calculate a correlation? Explain.



- b) Describe the association between attendance and runs scored.
- c) Does this association prove that the owners are right that more fans will come to games if the teams score more runs?

**28. Second inning 2006.** Perhaps fans are just more interested in teams that win. The displays are based on American League teams for the 2006 season (espn.go.com). Are the teams that win necessarily those that score the most runs? **LO1**

	CORRELATION		
	Wins	Runs	Attend
Wins	1.000		
Runs	0.605	1.000	
Attend	0.697	0.667	1.000



- a) Do winning teams generally enjoy greater attendance at their home games? Describe the association.
- b) Is attendance more strongly associated with winning or scoring runs? Explain.
- c) How strongly is scoring more runs associated with winning more games?

**T 29. University tuition.** The data set provided contains the yearly tuitions in 2012–2013 for undergraduate programs in arts and humanities at 66 universities and colleges that are members of the AUCC (Association of Universities and Colleges of Canada. These data were originally used in Chapter 5, Exercises 3 and 52.) Tuition fees are different for Canadian and international students. Would you expect to find a relationship between the tuitions charged by universities and colleges for each type of student? **LO2**

- a) Use the data provided to make a scatterplot of the tuition for international students against the tuition charged for Canadian students. Describe the relationship.
- b) Is the direction of the relationship what you expected?
- c) What is the regression equation for predicting the tuition for an international student from the tuition for a Canadian student at the same university/college?
- d) Is a linear model appropriate?
- e) How much more do universities/colleges charge on average in yearly tuition for international students compared to Canadian students according to this model?
- f) What is the  $R^2$  value for this model? Explain what it says.

**T 30. NHL salaries.** In Exercise 11 you examined the relationship between total team salary and performance of teams in the National Football League (NFL). Here we will examine the relationship in a different professional sports league, the National Hockey League (NHL). In 2005–2006 the NHL instituted a salary cap, the total amount of money that teams are permitted to pay their players. The purpose is to keep teams in larger markets, and therefore with more revenue, from signing all the top players to extend their advantage over smaller-market teams. The data set provided has each team's total *Payroll* (\$M) and number of *Points* (based on victories) during the regular season for the 2011–2012 season. The cap that year was \$64.3 million. **L02**

- Use the data provided to make a scatterplot of *Points* versus *Payroll*. Describe the relationship.
- Is the direction of the relationship what you expected?
- Is a linear model appropriate?
- What is the regression equation for predicting *Points* from *Payroll*?
- What does the slope mean in this context?
- What does the  $y$ -intercept mean in this context? Is it meaningful?
- What is the  $R^2$  value for this model? Explain what it says.
- If one team spends \$10 million more than another on salary, how many more points on average would you predict them to get?

**31. Mutual funds.** As the nature of investing shifted in the 1990s (more day traders and faster flow of information using technology), the relationship between mutual fund monthly performance (*Return*) in percent and money flowing (*Flow*) into mutual funds (\$ million) shifted. Using only the values for the 1990s (we'll examine later years in later chapters), answer the following questions. (You may assume that the assumptions and conditions for regression are met.) **L02**

The least squares linear regression is:

$$\widehat{Flow} = 9747 + 771 \text{ Return.}$$

- Interpret the intercept in the linear model.
- Interpret the slope in the linear model.
- What is the predicted fund *Flow* for a month that had a market *Return* of 0%?
- If during this month, the recorded fund *Flow* was \$5 billion, what is the residual using this linear model? Did the model provide an underestimate or overestimate for this month?

**32. Online clothing purchases.** An online clothing retailer examined their transactional database to see if total yearly *Purchases* (\$) were related to customers' *Incomes* (\$). (You may assume that the assumptions and conditions for regression are met.) **L02**

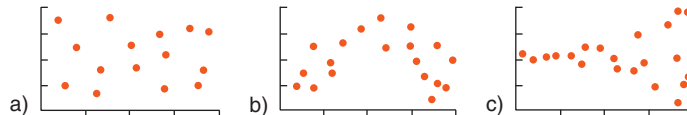
The least squares linear regression is:

$$\widehat{Purchases} = -31.6 + 0.012 \text{ Income.}$$

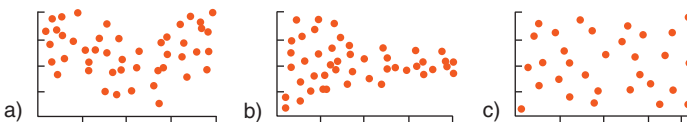
- Interpret the intercept in the linear model.
- Interpret the slope in the linear model.
- If a customer has an *Income* of \$20 000, what is his predicted total yearly *Purchases*?

d) This customer's yearly *Purchases* were actually \$100. What is the residual using this linear model? Did the model provide an underestimate or overestimate for this customer?

**33. Residual plots.** Tell what each of the following residual plots indicates about the appropriateness of the linear model that was fit to the data. **L03**



**34. Residual plots, again.** Tell what each of the following residual plots indicates about the appropriateness of the linear model that was fit to the data. **L03**



**T 35. Consumer spending.** An analyst at a large credit card bank is looking at the relationship between customers' charges to the bank's card in two successive months. He selects 150 customers at random, regresses charges in *March* (\$) on charges in *February* (\$), and finds an  $R^2$  of 79%. The intercept is \$730.20, and the slope is 0.79. After verifying all the data with the company's CPA, he concludes that the model is a useful one for predicting one month's charges from the other. Examine the data and comment on his conclusions. **L02**

**T 36. Insurance policies.** An actuary at a mid-sized insurance company is examining the sales performance of the company's sales force. She has data on the average size of the policy (\$) written in two consecutive years by 200 salespeople. She fits a linear model and finds the slope to be 3.00 and the  $R^2$  is 99.92%. She concludes that the predictions for next year's policy size will be very accurate. Examine the data and comment on her conclusions. **L02**

**37. What slope?** If you create a regression model for predicting the sales (\$ million) from money spent on advertising the prior month (\$ thousand), is the slope most likely to be 0.03, 300 or 3000? Explain. **L02**

**38. What slope, part 2?** If you create a regression model for estimating a student's business school GPA (on a scale of 1–5) based on his math SAT (on a scale of 200–800), is the slope most likely to be 0.01, 1, or 10? Explain. **L02**

**39. Misinterpretations.** An advertising agent who created a regression model using amount spent on *Advertising* to predict annual *Sales* for a company made these two statements. Assuming the calculations were done correctly, explain what is wrong with each interpretation. **L01**

- My  $R^2$  of 93% shows that this linear model is appropriate.
- If this company spends \$1.5 million on advertising, then annual sales will be \$10 million.

**40. More misinterpretations.** An economist investigated the association between a country's *Literacy Rate* and *Gross Domestic Product (GDP)* and used the association to draw the following

conclusions. Explain why each statement is incorrect. (Assume that all the calculations were done properly.) LO1

- The *Literacy Rate* determines 64% of the *GDP* for a country.
- The slope of the line shows that an increase of 5% in *Literacy Rate* will produce a \$1 billion improvement in *GDP*.

**41. Business admissions.** An analyst at a business school's admissions office claims to have developed a valid linear model predicting success (measured by starting salary (\$) at time of graduation) from a student's undergraduate performance (measured by GPA). Describe how you would check each of the four regression conditions in this context. LO3

**42. School rankings.** A popular magazine annually publishes rankings of business programs. The latest issue claims to have developed a linear model predicting the school's ranking (with "1" being the highest ranked school) from its financial resources (as measured by size of the school's endowment). Describe how you would apply each of the four regression conditions in this context. LO3

**T 43. Used BMW prices.** A business student needs cash, so he decides to sell his car. The car is a valuable BMW 840 that was only made over the course of a few years in the late 1990s. He would like to sell it on his own, rather than through a dealer so he'd like to predict the price he'll get for his car's model year. LO1

- Make a scatterplot for the data on used BMW 840's provided.
- Describe the association between year and price.
- Do you think a linear model is appropriate?
- Computer software says that  $R^2 = 57.4\%$ . What is the correlation between year and price?
- Explain the meaning of  $R^2$  in this context.
- Why doesn't this model explain 100% of the variability in the price of a used BMW 840?

**T 44. More used BMW prices.** Use the advertised prices for BMW 840s given in Exercise 43 to create a linear model for the relationship between a car's *Year* and its *Price*. LO2

- Find the equation of the regression line.
- Explain the meaning of the slope of the line.
- Explain the meaning of the intercept of the line.
- If you want to sell a 1997 BMW 840, what price seems appropriate?
- You have a chance to buy one of two cars. They are about the same age and appear to be in equally good condition. Would you rather buy the one with a positive residual or the one with a negative residual? Explain.

**T 45. Cost of living.** Mercer's *Worldwide Cost of Living Survey City Rankings* determine the cost of living in the most expensive cities in the world as an index. The survey covers 214 cities across five continents and measures the comparative cost of over 200 items in each location, including transport, food, clothing, household goods, and entertainment. The cost of housing is also included and, as it is often the biggest expense for expatriates, it plays an important part in determining where cities are ranked. New York is used as the base city and all cities are compared against it. Currency movements are measured against the U.S. dollar. The scatterplot shows the ranking (1 is the most expensive) of the top

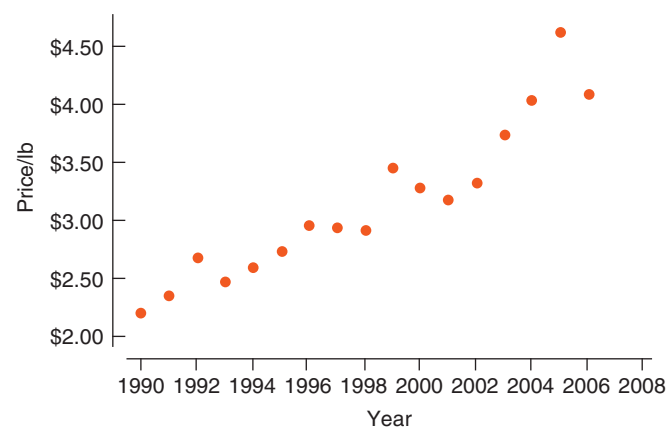


50 most expensive cities for 2012 plotted against the rankings those same cities had in the 2011. LO1

- Describe the association between the rankings in 2012 and 2011.
- The  $R^2$  for the regression equation is 0.590. Interpret the value of  $R^2$ .
- Using the data provided, find the correlation.
- Prepare a plot of the residuals. What does it say about the appropriateness of the linear model?

**T 46. Lobster prices.** Over the past few decades both the demand for lobster and the price of lobster have continued to increase. The scatterplot shows this increase in the *Price* of lobster (*Price/lb*) since 1990. LO1

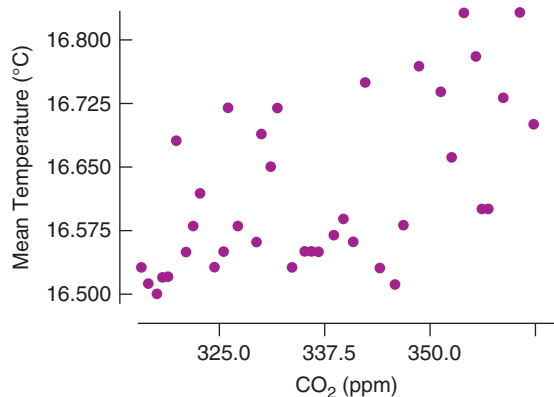
- Describe the increase in the *Price* of lobster since 1990.
- The  $R^2$  for the regression equation is 88.5%. Interpret the value of  $R^2$ .



- Find the correlation.
- Find the linear model and examine the plot of residuals versus predicted values. Is the Equal Spread Condition satisfied? (Use time starting at 1990 so that 1990 = 0.)

**47. El Niño.** Concern over the weather associated with El Niño has increased interest in the possibility that the climate on Earth is getting warmer. The most common theory relates an increase in atmospheric levels of carbon dioxide ( $\text{CO}_2$ ), a greenhouse gas, to increases in temperature. Here is a scatterplot showing the mean annual  $\text{CO}_2$  concentration in the atmosphere, measured in

parts per million (ppm) at the top of Mauna Loa in Hawaii, and the mean annual air temperature over both land and sea across the globe, in degrees Celsius (C). **LO2**

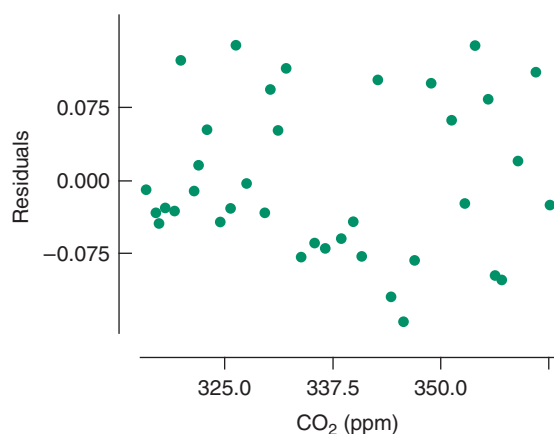


A regression predicting *Mean Temperature* from *CO<sub>2</sub>* produces the following output table (in part).

Dependent variable: Temperature  
R-squared = 33.4%

Variable	Coefficient
Intercept	15.3066
CO2	0.004

- What is the correlation between *CO<sub>2</sub>* and *Mean Temperature*?
- Explain the meaning of *R-squared* in this context.
- Give the regression equation.
- What is the meaning of the slope in this equation?
- What is the meaning of the intercept of this equation?
- Here is a scatterplot of the residuals vs. *CO<sub>2</sub>*. Does this plot show evidence of the violations of any of the assumptions of the regression model? If so, which ones?
- CO<sub>2</sub>* levels may reach 364 ppm in the near future. What *Mean Temperature* does the model predict for that value?



**48. Hospital beds.** An expert consultant in hospital resource planning states that the number of open beds that a hospital can

use effectively should be estimated by the number of FTEs (full-time equivalent employees) on staff. The consultant collected data on the number of open beds and number of FTEs for 12 hospitals, and computed the means and SDs as follows:

Number of open beds:	Mean = 50	SD = 20
Number of FTEs:	Mean = 140	SD = 40

She computed the least squares regression equation and found that for a hospital with 100 FTEs, the estimated number of open beds was 32. **LO2**

- Use this information to compute the value of the correlation coefficient.
- What is the regression equation she found?
- From the available data, what would you predict the number of open beds to be for a hospital with an unknown number of FTEs?
- What fraction of the variation in number of open beds is explained by the number of FTEs?
- Another expert consultant, this one in hospital administration, claims that the regression was done the wrong way around, and that the number of FTEs required in a hospital should be estimated from the number of open beds in the hospital. What would the value of the correlation coefficient be if the analysis were done this way?



### JUST CHECKING ANSWERS

- We know the scores are quantitative. We should check to see if the *Linearity Condition* and the *Outlier Condition* are satisfied by looking at a scatterplot of the two scores.
- It won't change.
- It won't change.
- They are more likely to do poorly. The positive correlation means that low closing prices for Intel are associated with low closing prices for Cypress.
- No, the general association is positive, but daily closing prices may vary.
- For each additional employee, monthly sales increase, on average, \$122 740.
- Thousands of \$ per employee.
- \$1 227 400 per month.
- Differences in the number of employees account for about 71.4% of the variation in the monthly sales.
- It's positive. The correlation and the slope have the same sign.
- $R^2$ , No. Slope, Yes.