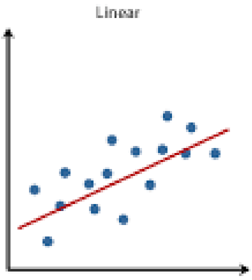
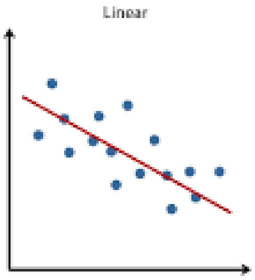
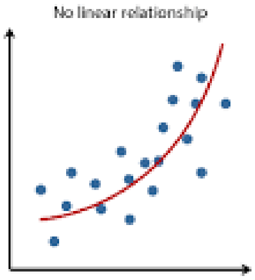
Chapter 2

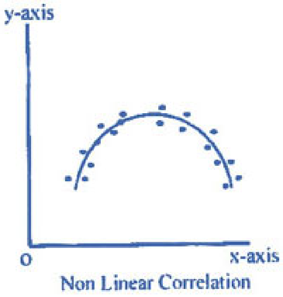
Section 2.1 Relationships

We study the relationship between two variables or more. These variables could be either quantitative or qualitative. (Common types of relationships-see images below)



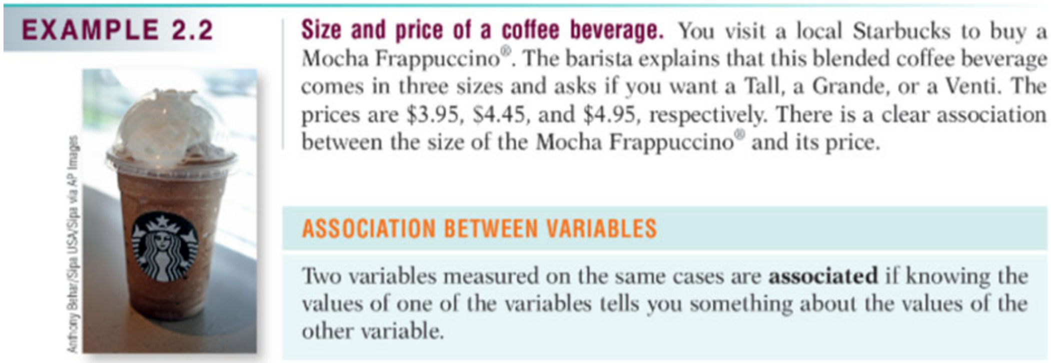






Response Variable (Dependent): It measures an outcome of a study. Explanatory Variable (Independent): Explains or causes changes in the response variables.

Example 2.2 page 80.



EXAMPLE 2.2

Anthony Behar/Sipa USA/Sipa via AP Images

Size and price of a coffee beverage. You visit a local Starbucks to buy a Mocha Frappuccino®. The barista explains that this blended coffee beverage comes in three sizes and asks if you want a Tall, a Grande, or a Venti. The prices are $3.95, $4.45, and $4.95, respectively. There is a clear association between the size of the Mocha Frappuccino® and its price.

ASSOCIATION BETWEEN VARIABLES

Two variables measured on the same cases are **associated** if knowing the values of one of the variables tells you something about the values of the other variable.

Goal: To show that changes in one or more Independent variables actually causes changes in the Dependent variable.

Section 2.2 Scatterplots

In regression, the regular plot of Y vs X is called “scatter Plot”. It allows us to study the association between two variables. The association could be positive or negative.

Positive Association: Two **variables** have a **positive association** when the values of one **variable** tend to increase as the values of the other **variable** increase. **(Grade on an Exam and Hours of studying)**

Negative Association: Two **variables** have a **negative association** when the values of one **variable** tend to decrease as the values of the other **variable** increase. **(Grade and Number of absences)**

Homework: 2.17, 2.25, 2.27, and 2.32 pages 97-99.

Section 2.3 Linear Correlation(r)

The linear correlation coefficient, **r**, measures the strength of the linear association between two quantitative variables, positive or negative.

Rules for interpreting r:

a. The value of r always falls between –1 and 1. A positive value of r indicates positive correlation and a negative value of r indicates negative correlation.

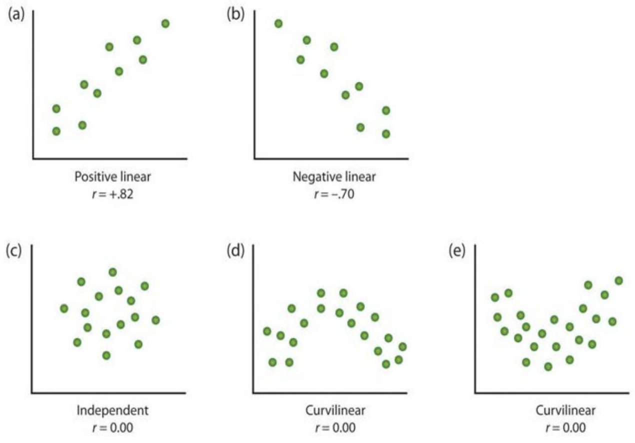
b. The closer r is to 1, the stronger the positive correlation and the closer r is to –1, the stronger the negative correlation. Values of r closer to zero indicate no linear association.

c. The larger the absolute value of r, the stronger the relationship between the two variables.

d. r measures only the strength of linear relationship between two variables.

e. Changing the unit of measurement on the variables the value of r remains the same.

Below are some images noting the degree of linear relationship(r).



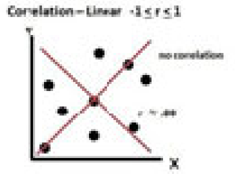
Positive linear $r = +.82$

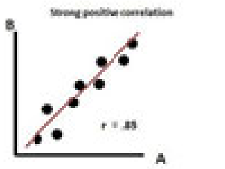
Negative linear $r = -.70$

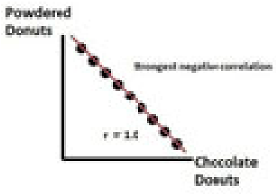
Independent $r = 0.00$

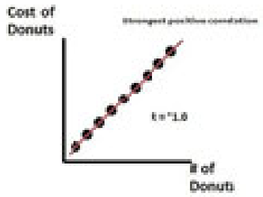
Curvilinear $r = 0.00$

Curvilinear $r = 0.00$









Formula for computing r: $r = \dfrac{ss_{xy}}{\sqrt{ss_x ss_y}}$ where

$$ss_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} x_i y_i - n\overline{X}\overline{Y} \quad ;$$

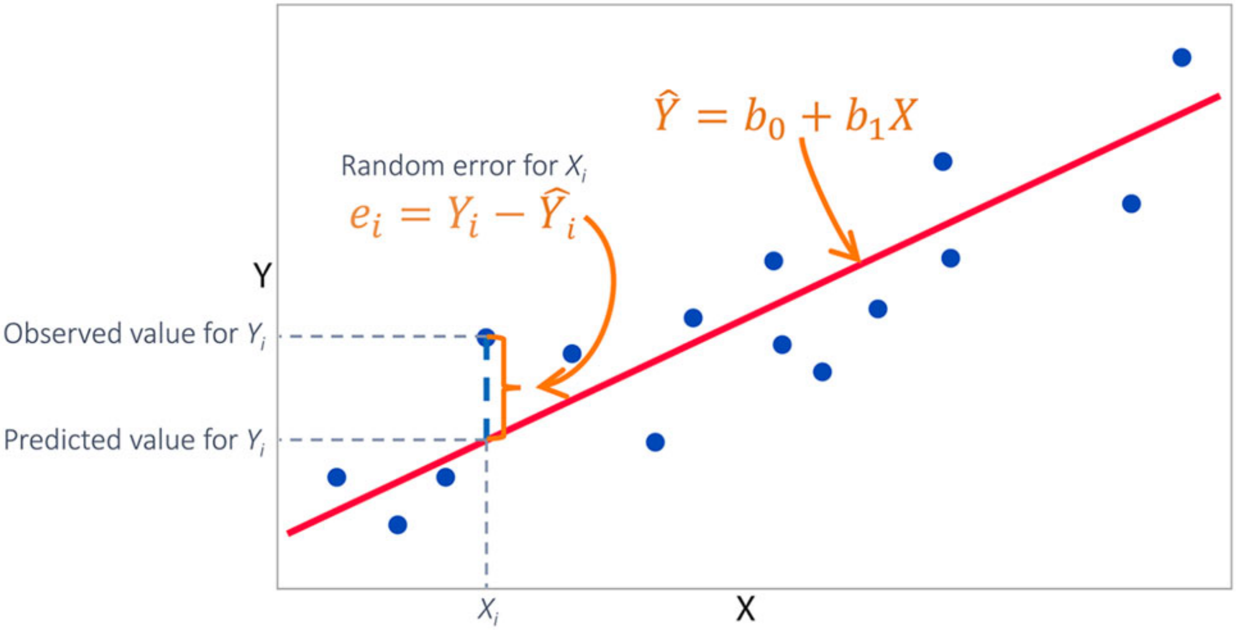
$$ss_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\left(\overline{X}\right)^2 \quad ; \quad ss_y = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\left(\overline{Y}\right)^2$$

Note: The book is using different notation

Homework: 2.41, 2.46, and 2.58 pages 105-106

Section 2.4 Least – Squares Regression

The **least square method** is a procedure that is used to find the line that provides the best approximation for the relationship between x and y by minimizing the error(Residual). We refer to this equation of the line developed using the least square method as the **regression line**.



Regression Line: $\hat{Y} = a + bx$ **where**

a = y-intercept of the line
b = slope of the line
$\hat{Y}$ = estimated value of the dependent variable

Least Square Method : The values of a and b can be computed using the following equations.

$$b = \frac{SS_{xy}}{SS_x} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\left(\overline{X}\right)^2} \quad \text{and} \quad a = \overline{Y} - b\overline{X}$$

where n = total number of observations.

r² – Square = $100r^2$

The square of the correlation, r^2, is percentage of the variation explained in the y variable (dependent) using one or more x variables (independent). The prediction of y using one or more x variables is good, if the r^2 – Square is 75% or higher.

Note: Talk about

(a) Restriction on the range of the x values on the prediction
(b) The interpretation of the y-intercept and slope

Example: Real Life Application

Dr. XXXXX

I'm a VSU alumnus-class of '95. My sole proprietorship business (me) needs a solution to a math problem, and since my BS was in Psychology, I'm unqualified and was hoping you could help.

Much thanks in advance,
Mr. XXXXXXXXXXX

Problem:

Below is a table. Left column is the length of an auger (it moves cement powder through a tube with a motorized "screw) . Right column is HP required to maintain a certain production "constant" at the respective length. I need to know the equation (if it exists) to obtain the HP given ANY length (eg. 12' or 28') . Length range is 10' - 40' . MS Excel showed me that the graph is a mild "S" shape, so I knew I was in trouble, given that anything beyond linear relationships is a nightmare for me.

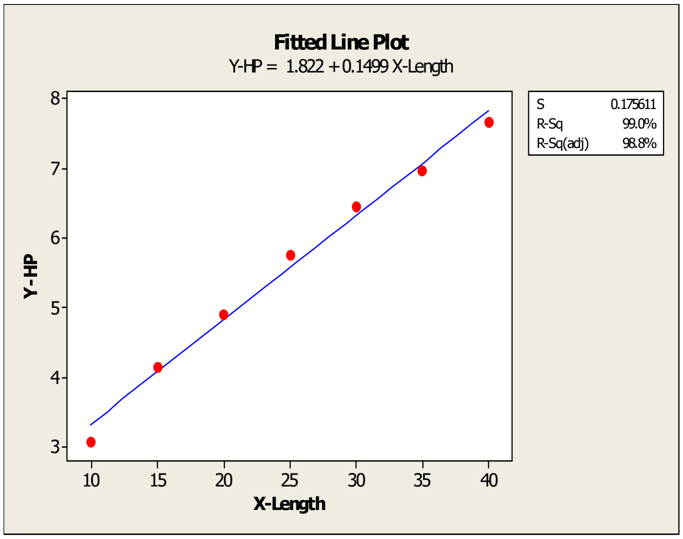
Data	*Length*	*HP*
1	10	3.08
2	15	4.14
3	20	4.91
4	25	5.76
5	30	6.45
6	35	6.98
7	40	7.67

	X	Y	X*Y	X^2	Y^2
	10	3.08	30.8	100	9.4864
	15	4.14	62.1	225	17.1396
	20	4.91	98.2	400	24.1081
	25	5.76	144	625	33.1776
	30	6.45	193.5	900	41.6025
	35	6.98	244.3	1225	48.7204
	40	7.67	306.8	1600	58.8289
sum=	175	38.99	1079.7	5075	233.0635

X-bar=	25
Y-bar=	5.57

SSxy=	104.95
SSy=	15.8892
SSx=	700

b=	0.149929
a=	1.821786
r=	0.995136
100*r^2=	0.990296



Homework: 2.80, 2.85, 286, and 2.87 Use Minitab When You Can pp. 121-122.

2.5 Cautions about Regression and Correlation

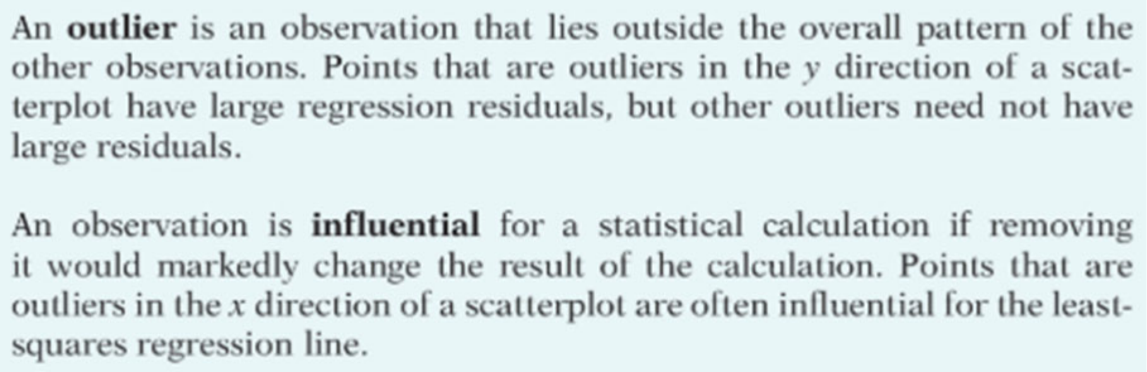
Residuals: A residual is the difference between an observed value of y and the value predicted by the regression line. That is,

residual = observed y – predicted $e_i = Y_i - \hat{Y}_i$.

Residual Plots: Plot the residuals versus x variable(s)

It helps us assess the fit of a regression line. (See figures on Notes page 8)

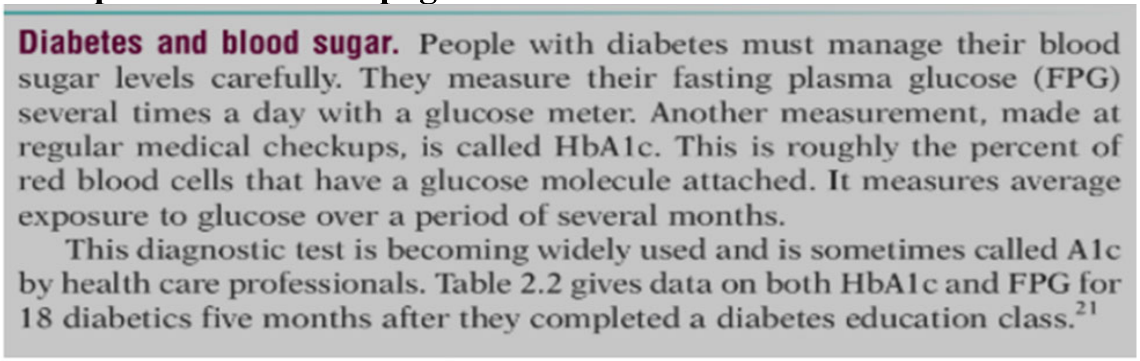
Outliers and influential Observations



An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the *y* direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

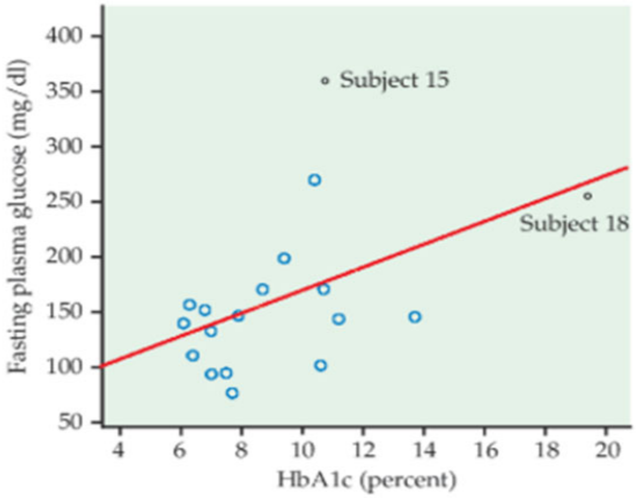
An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the *x* direction of a scatterplot are often influential for the least-squares regression line.

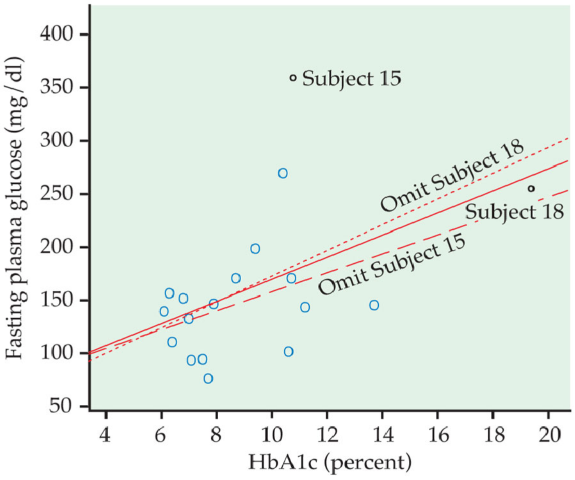
Example 2.28 and 2.29 pages 127-128.



Diabetes and blood sugar. People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA1c. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months.

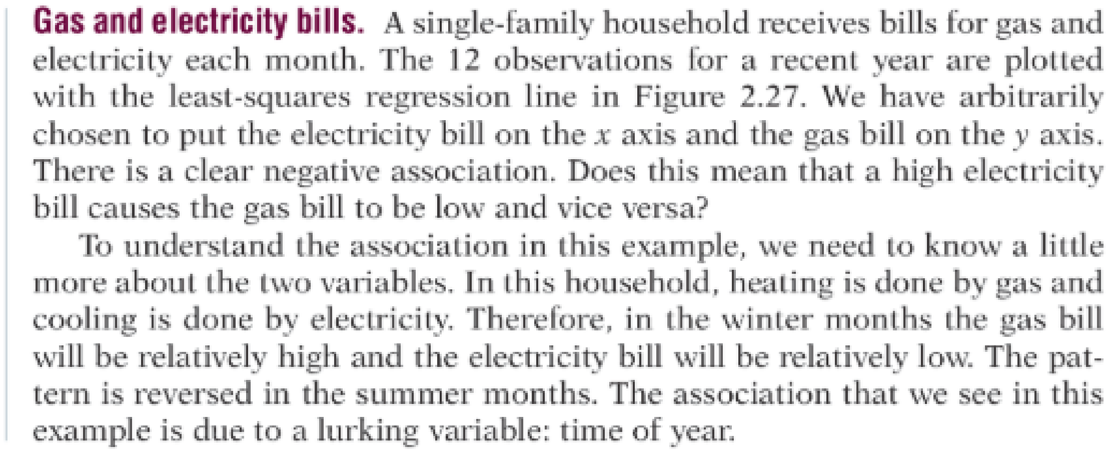
This diagnostic test is becoming widely used and is sometimes called A1c by health care professionals. Table 2.2 gives data on both HbA1c and FPG for 18 diabetics five months after they completed a diabetes education class.[21]





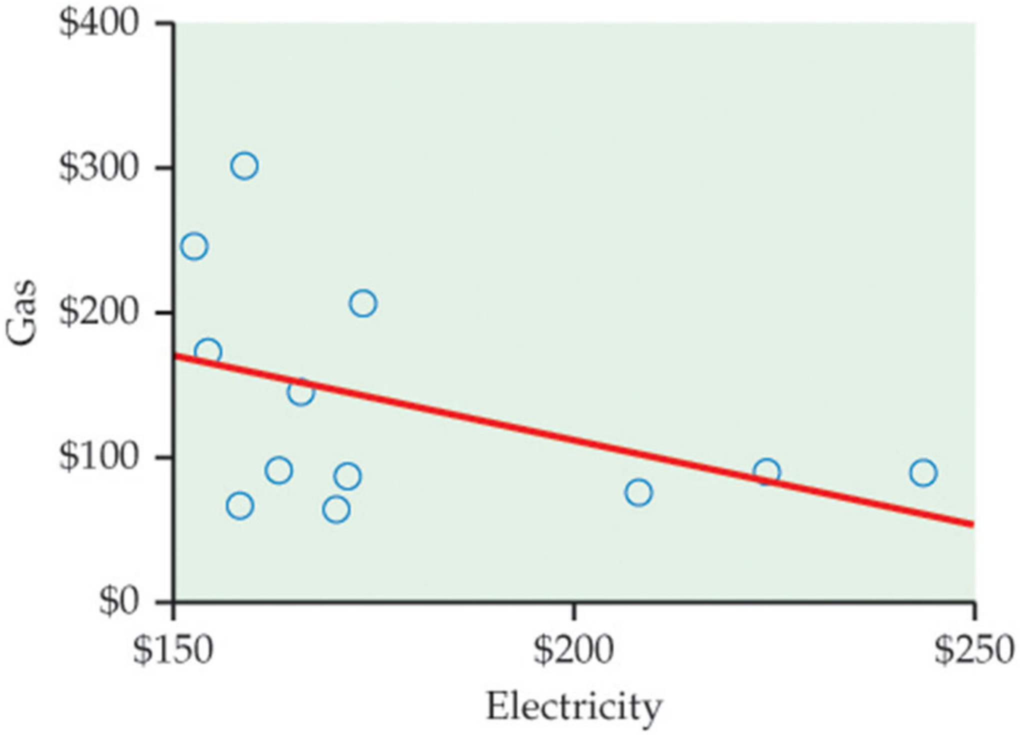
Lurking variable: A variable that it's not included but has an important effect on the relationship. Plot the residuals versus time, if you can, to find that out.

See example 2.31 page 130.

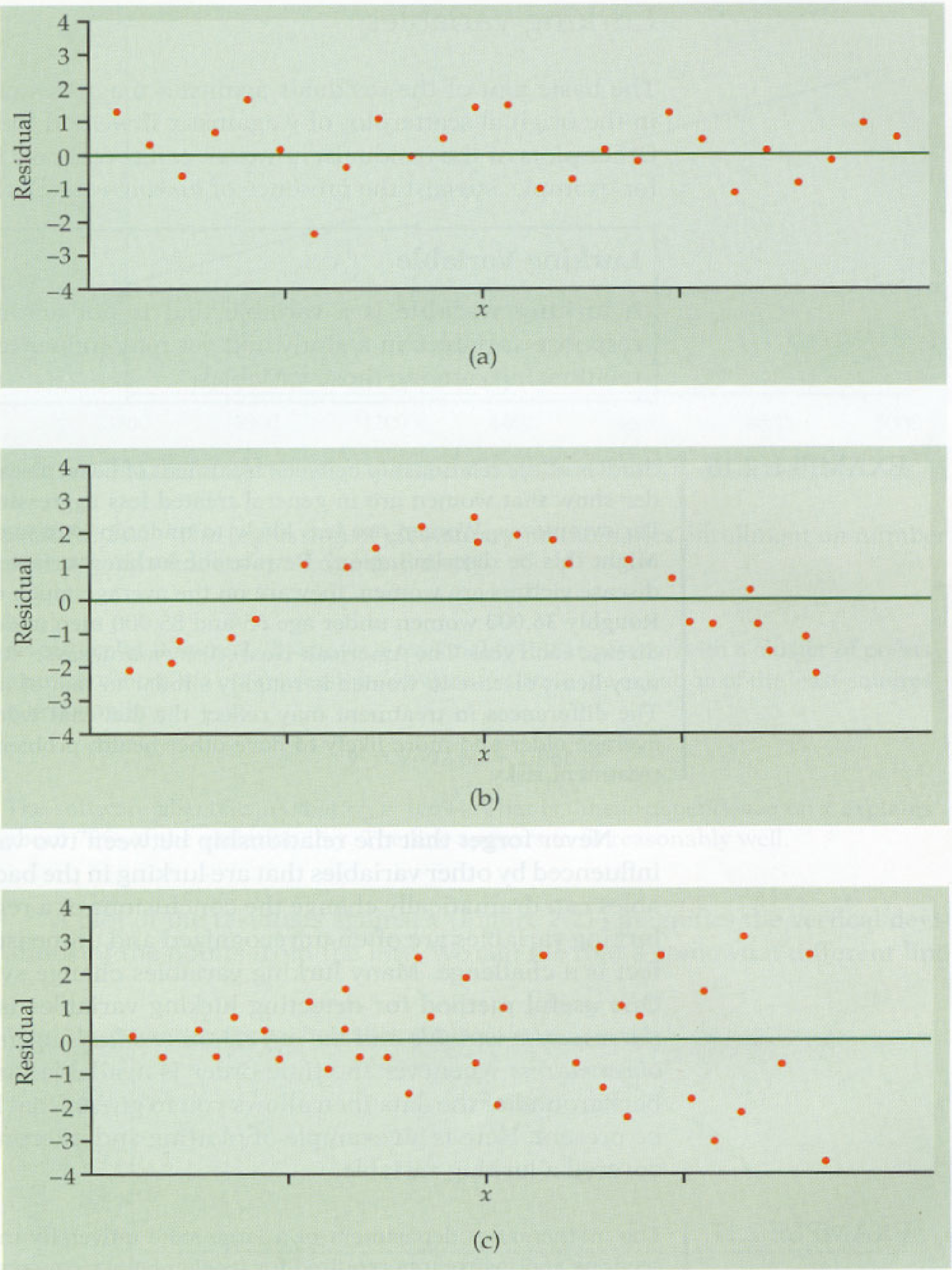


Gas and electricity bills. A single-family household receives bills for gas and electricity each month. The 12 observations for a recent year are plotted with the least-squares regression line in Figure 2.27. We have arbitrarily chosen to put the electricity bill on the x axis and the gas bill on the y axis. There is a clear negative association. Does this mean that a high electricity bill causes the gas bill to be low and vice versa?

To understand the association in this example, we need to know a little more about the two variables. In this household, heating is done by gas and cooling is done by electricity. Therefore, in the winter months the gas bill will be relatively high and the electricity bill will be relatively low. The pattern is reversed in the summer months. The association that we see in this example is due to a lurking variable: time of year.



Homework: 2.105, 2.107 page 135.



(a)

(b)

(c)