3.1 Scatter Plots and Linear Correlation

Does smoking cause lung cancer? Is job performance related to marks in high school? Do pollution levels affect the ozone layer in the atmosphere? Often the answers to such questions are not clear-cut, and inferences have to be made from large sets of data. Two-variable statistics provide methods for detecting relationships between variables and for developing mathematical models of these relationships.

The visual pattern in a graph or plot can often reveal the nature of the relationship between two variables.

INVESTIGATE & INQUIRE: Visualizing Relationships Between Variables

A study examines two new obedience-training methods for dogs. The dogs were randomly selected to receive from 5 to 16 h of training in one of the two training programs. The dogs were assessed using a performance test graded out of 20.

Rogers Method		Laing System	
Hours	Score	Hours	Score
10	12	8	10
15	16	6	9
7	10	15	12
12	15	16	7
8	9	9	11
5	8	11	7
8	11	10	9
16	19	10	6
10	14	8	15

1. Could you determine which of the two training systems is more effective by comparing the mean scores? Could you calculate another statistic that would give a better comparison? Explain your reasoning.
2. Consider how you could plot the data for the Rogers Method. What do you think would be the best method? Explain why.
3. Use this method to plot the data for the Rogers Method. Describe any patterns you see in the plotted data.
4. Use the same method to plot the data for the Laing System and describe any patterns you see.
5. Based on your data plots, which training method do you think is more effective? Explain your answer.

6. Did your plotting method make it easy to compare the two sets of data? Are there ways you could improve your method?

7. a) Suggest factors that could influence the test scores but have not been taken into account.

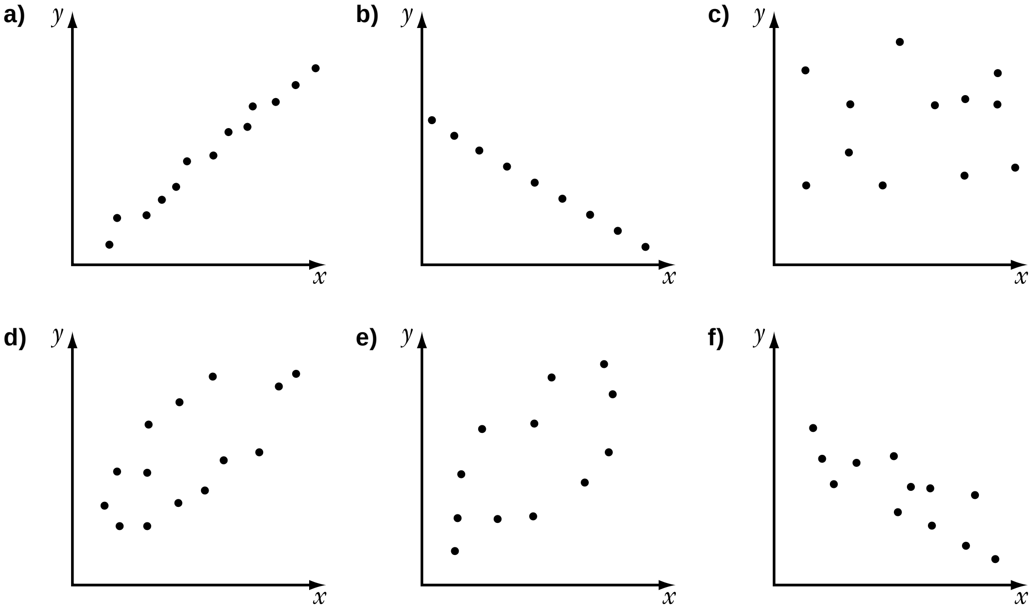
   b) How could these factors affect the validity of conclusions drawn from the data provided?

In data analysis, you are often trying to discern whether one variable, the **dependent** (or **response**) **variable**, is affected by another variable, the **independent** (or **explanatory**) **variable**. Variables have a **linear correlation** if changes in one variable tend to be proportional to changes in the other. Variables *X* and *Y* have a **perfect positive** (or **direct**) **linear correlation** if *Y* increases at a constant rate as *X* increases. Similarly, *X* and *Y* have a **perfect negative** (or **inverse**) **linear correlation** if *Y* decreases at a constant rate as *X* increases.

A **scatter plot** shows such relationships graphically, usually with the independent variable as the horizontal axis and the dependent variable as the vertical axis. The **line of best fit** is the straight line that passes as close as possible to all of the points on a scatter plot. The stronger the correlation, the more closely the data points cluster around the line of best fit.

Example 1 Classifying Linear Correlations

Classify the relationship between the variables *X* and *Y* for the data shown in the following diagrams.



Solution

a) The data points are clustered around a line that rises to the right (positive slope), indicating definitely that Y increases as X increases. Although the points are not perfectly lined up, there is a *strong positive linear correlation* between X and Y.

b) The data points are all exactly on a line that slopes down to the right, so Y decreases as X increases. In fact, the changes in Y are *exactly* proportional to the changes in X. There is a *perfect negative linear correlation* between X and Y.

c) No discernible linear pattern exists. As X increases, Y appears to change randomly. Therefore, there is *zero linear correlation* between X and Y.

d) A definite positive trend exists, but it is not as clear as the one in part a). Here, X and Y have a *moderate positive linear correlation*.

e) A slight positive trend exists. X and Y have a *weak positive linear correlation*.

f) A definite negative trend exists, but it is hard to classify at a glance. Here, X and Y have a *moderate or strong negative linear correlation*.

As Example 1 shows, a scatter plot often can give only a rough indication of the correlation between two variables. Obviously, it would be useful to have a more precise way to measure correlation. Karl Pearson (1857–1936) developed a formula for estimating such a measure. Pearson, who also invented the term *standard deviation*, was a key figure in the development of modern statistics.

The Correlation Coefficient

To develop a measure of correlation, mathematicians first defined the **covariance** of two variables in a sample:

$$s_{XY} = \frac{1}{n-1}\sum(x-\overline{x})(y-\overline{y})$$

where n is the size of the sample, x represents individual values of the variable X, y represents individual values of the variable Y, $\overline{x}$ is the mean of X, and $\overline{y}$ is the mean of Y.

Recall from Chapter 2 that the symbol $\sum$ means "the sum of." Thus, the covariance is the sum of the *products* of the deviations of x and y for all the data points divided by $n - 1$. The covariance depends on how the deviations of the two variables are related. For example, the covariance will have a large positive value if both $x - \overline{x}$ and $y - \overline{y}$ tend to be large at the same time, and a negative value if one tends to be positive when the other is negative.

The **correlation coefficient**, ***r***, is the covariance divided by the product of the standard deviations for X and Y:

$$r = \frac{s_{XY}}{s_X \times s_Y}$$

where s_X is the standard deviation of X and s_Y is the standard deviation of Y.

This coefficient gives a quantitative measure of the strength of a linear correlation. In other words, the correlation coefficient indicates how closely the data points cluster around the line of best fit. The correlation coefficient is also called the **Pearson product-moment coefficient of correlation (PPMC)** or **Pearson's *r***.

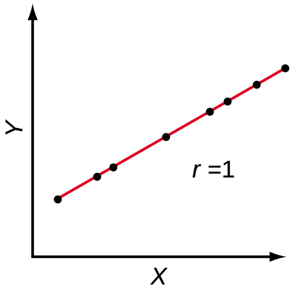
The correlation coefficient always has values in the range from −1 to 1. Consider a perfect positive linear correlation first. For such correlations, changes in the dependent variable Y are directly proportional to changes in the independent variable X, so $Y = aX + b$, where a is a positive constant. It follows that

$$\begin{aligned} s_{XY} &= \frac{1}{n-1}\sum(x-\bar{x})(y-\bar{y}) \\ &= \frac{1}{n-1}\sum(x-\bar{x})[(ax+b)-(a\bar{x}+b)] \\ &= \frac{1}{n-1}\sum(x-\bar{x})(ax-a\bar{x}) \\ &= \frac{1}{n-1}\sum a(x-\bar{x})^2 \\ &= a\frac{\Sigma(x-\bar{x})^2}{n-1} \\ &= as_X^2 \end{aligned}$$

$$\begin{aligned} s_Y &= \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} \\ &= \sqrt{\frac{\Sigma[(ax+b)-(a\bar{x}+b)]^2}{n-1}} \\ &= \sqrt{\frac{\Sigma(ax-a\bar{x})^2}{n-1}} \\ &= \sqrt{\frac{a^2\Sigma(x-\bar{x})^2}{n-1}} \\ &= a\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \\ &= as_X \end{aligned}$$

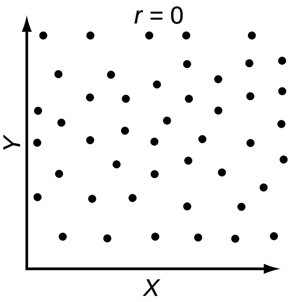
Substituting into the equation for the correlation coefficient gives

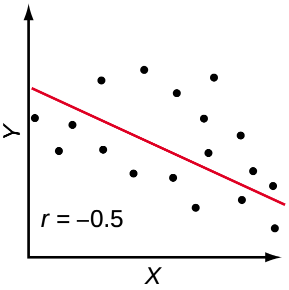
$$\begin{aligned} r &= \frac{s_{XY}}{s_X s_Y} \\ &= \frac{as_X^2}{s_X(as_X)} \\ &= 1 \end{aligned}$$



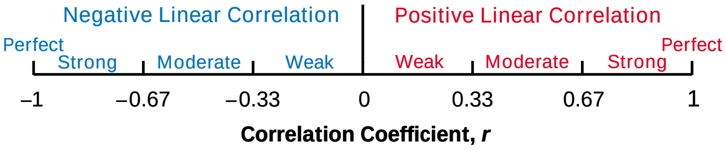
Similarly, $r = -1$ for a perfect negative linear correlation.

For two variables with no correlation, Y is equally likely to increase or decrease as X increases. The terms in $\sum(x - \overline{x})(y - \overline{y})$ are randomly positive or negative and tend to cancel each other. Therefore, the correlation coefficient is close to zero if there is little or no correlation between the variables. For moderate linear correlations, the summation terms partially cancel out.





The following diagram illustrates how the correlation coefficient corresponds to the strength of a linear correlation.


Negative Linear Correlation
Positive Linear Correlation
Perfect
Strong
Moderate
Weak
Weak
Moderate
Strong
Perfect
–1
–0.67
–0.33
0
0.33
0.67
1
Correlation Coefficient, r

Using algebraic manipulation and the fact that $\sum x = n\overline{x}$, Pearson showed that

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where n is the number of data points in the sample, x represents individual values of the variable X, and y represents individual values of the variable Y. (Note that $\sum x^2$ is the sum of the squares of all the individual values of X, while $(\sum x)^2$ is the square of the sum of all the individual values.)

Like the alternative formula for standard deviations (page 150), this formula for r avoids having to calculate all the deviations individually. Many scientific and statistical calculators have built-in functions for calculating the correlation coefficient.

It is important to be aware that increasing the number of data points used in determining a correlation improves the accuracy of the mathematical model. Some of the examples and exercise questions have a fairly small set of data in order to simplify the computations. Larger data sets can be found in the e-book that accompanies this text.

Example 2 Applying the Correlation Coefficient Formula

A farmer wants to determine whether there is a relationship between the mean temperature during the growing season and the size of his wheat crop. He assembles the following data for the last six crops.

Mean Temperature (°C)	Yield (tonnes/hectare)
4	1.6
8	2.4
10	2.0
9	2.6
11	2.1
6	2.2

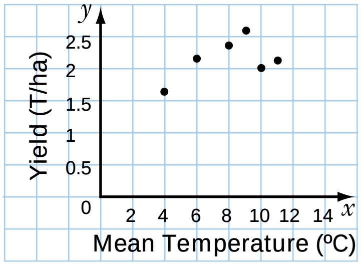
a) Does a scatter plot of these data indicate any linear correlation between the two variables?

b) Compute the correlation coefficient.

c) What can the farmer conclude about the relationship between the mean temperatures during the growing season and the wheat yields on his farm?

Solution

a) The farmer wants to know whether the crop yield depends on temperature. Here, temperature is the independent variable, X, and crop yield is the dependent variable, Y. The scatter plot has a somewhat positive trend, so there appears to be a moderate positive linear correlation.



b) To compute r, set up a table to calculate the quantities required by the formula.

Temperature, x	Yield, y	x^2	y^2	xy
4	1.6	16	2.56	6.4
8	2.4	64	5.76	19.2
10	2.0	100	4.00	20.0
9	2.6	81	6.76	23.4
11	2.1	121	4.41	23.1
6	2.2	36	4.84	13.2
$\sum x = 48$	$\sum y = 12.9$	$\sum x^2 = 418$	$\sum y^2 = 28.33$	$\sum xy = 105.3$

Now compute r, using the formula:

$$
\begin{aligned}
r &= \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} \\
&= \frac{6(105.3) - (48)(12.9)}{\sqrt{[6(418) - (48)^2][6(28.33) - (12.9)^2]}} \\
&= \frac{631.8 - 619.2}{\sqrt{(2508 - 2304)(169.98 - 166.41)}} \\
&= \frac{12.6}{26.99} \\
&= 0.467
\end{aligned}
$$

The correlation coefficient for crop yield versus mean temperature is approximately 0.47, which confirms a moderate positive linear correlation.

c) It appears that the crop yield tends to increase somewhat as the mean temperature for the growing season increases. However, the farmer cannot conclude that higher temperatures *cause* greater crop yields. Other variables could account for the correlation. For example, the lower temperatures could be associated with heavy rains, which could lower yields by flooding fields or leaching nutrients from the soil.

Data in Action

From 1992 to 2001, Canada produced an average of 27 million tonnes of wheat a year. About 70% of this crop was exported.

The important principle that a correlation does not prove the existence of a cause-and-effect relationship between two variables is discussed further in section 3.4.

Example 3 Using Technology to Determine Correlation Coefficients

Determine whether there is a linear correlation between horsepower and fuel consumption for these five vehicles by creating a scatter plot and calculating the correlation coefficient.

Vehicle	Horsepower, x	Fuel Consumption (L/100 km), y
Midsize sedan	105	6.7
Minivan	170	23.5
Small sports utility vehicle	124	5.9
Midsize motorcycle	17	3.4
Luxury sports car	296	8.4

Solution 1 Using a Graphing Calculator

Use the ClrList command to make sure lists L1 and L2 are clear, then enter the horsepower data in L1 and the fuel consumption figures in L2.

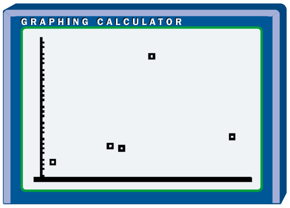
To display a scatter plot, first make sure that all functions in the Y= editor are either clear or turned off. Then, use STAT PLOT to select PLOT1.

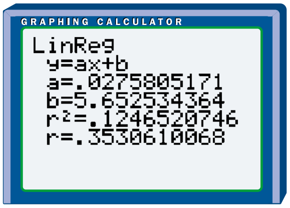
See Appendix B for more details on the graphing calculator and software functions used in this section.

Turn the plot on, select the scatter-plot icon, and enter L1 for XLIST and L2 for YLIST. (Some of these settings may already be in place.) From the ZOOM menu, select 9:ZoomStat. The calculator will automatically optimize the window settings and display the scatter plot.

To calculate the correlation coefficient, from the CATALOG menu, select DiagnosticOn, then select the LinReg(ax+b) instruction from the STAT CALC menu. The calculator will perform a series of statistical calculations using the data in lists L1 and L2. The last line on the screen shows that the correlation coefficient is approximately 0.353.

Therefore, there is a moderate linear correlation between horsepower and fuel consumption for the five vehicles.

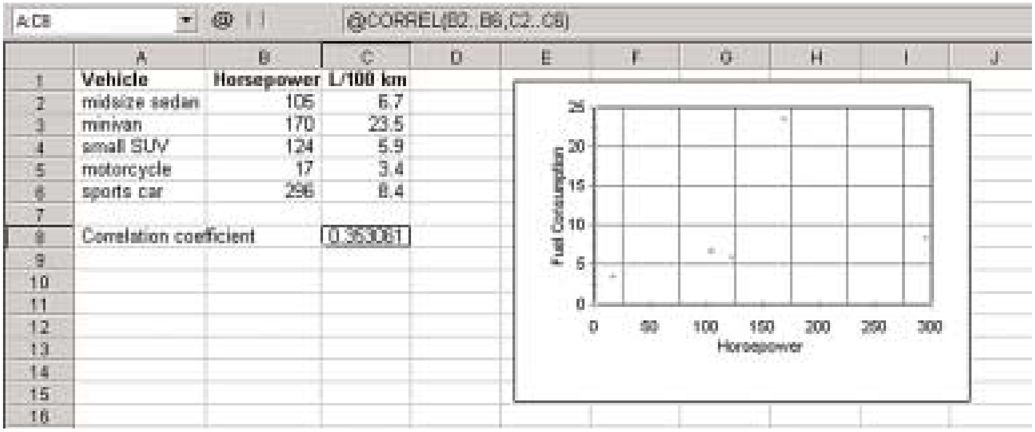




```
LinReg
 y=ax+b
 a=.0275805171
 b=5.652534364
 r²=.1246520746
 r=.3530610068
```

Solution 2 Using a Spreadsheet

Set up three columns and enter the data from the table above. Highlight the numerical data and use your spreadsheet's Chart feature to display a scatter plot. Both Corel® Quattro® Pro and Microsoft® Excel have a CORREL function that allows you to calculate the correlation coefficient easily. The scatter plot and correlation coefficient indicate a moderate correlation between horsepower and fuel consumption.



@CORREL(B2..B6,C2..C6)

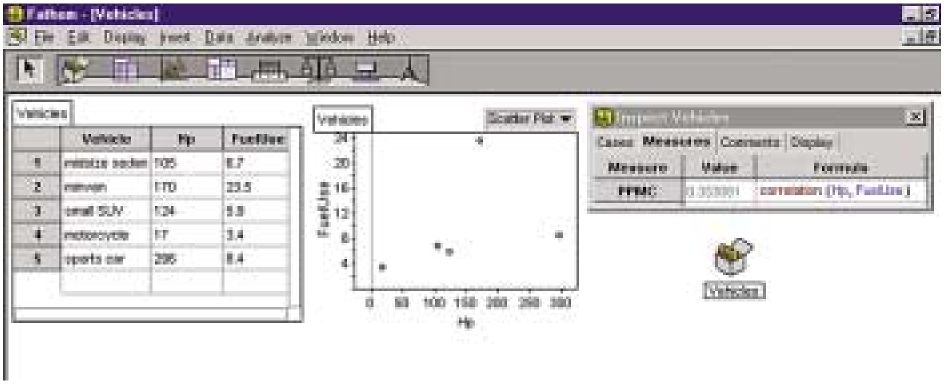
	A	B	C
1	Vehicle	Horsepower	L/100 km
2	midsize sedan	105	6.7
3	minivan	170	23.5
4	small SUV	124	5.9
5	motorcycle	17	3.4
6	sports car	296	8.4
7			
8	Correlation coefficient		0.353061

Solution 3 Using Fathom™

Create a new collection by setting up a case table with three attributes: Vehicle, Hp, and FuelUse. Enter the data for the five cases. To create a scatter plot, drag the graph icon onto the work area and drop the Hp attribute on the x-axis and the FuelUse attribute on the y-axis.

To calculate the **correlation coefficient**, right-click on the **collection** and select Inspect Collection. Select the Measures tab and name a new measure PPMC. Right-click this measure and select Edit Formula, then Functions/Statistical/Two Attributes/correlation. When you enter the Hp and FuelUse attributes in the correlation function, Fathom™ will calculate the correlation coefficient for these data.

Again, the scatter plot and correlation coefficient show a moderate linear correlation.



	Vehicle	Hp	FuelUse
1	midsize sedan	105	8.7
2	minivan	170	23.5
3	small SUV	124	5.8
4	motorcycle	17	3.4
5	sports car	295	8.4

Measure	Value	Formula
PPMC	0.353301	correlation (Hp, FuelUse)

Project Prep

For your statistics project, you may be investigating the linear correlation between two variables. A graphing calculator or computer software may be a valuable aid for this analysis.

Notice that the scatter plots in Example 3 have an outlier at (170, 23.5). Without this data point, you would have a strong positive linear correlation. Section 3.2 examines the effect of outliers in more detail.

Key Concepts

- Statistical studies often find linear correlations between two variables.
- A scatter plot can often reveal the relationship between two variables. The independent variable is usually plotted on the horizontal axis and the dependent variable on the vertical axis.
- Two variables have a linear correlation if changes in one variable tend to be proportional to changes in the other. Linear correlations can be positive or negative and vary in strength from zero to perfect.
- The correlation coefficient, r, is a quantitative measure of the correlation between two variables. Negative values indicate negative correlations while positive values indicate positive correlations. The greater the absolute value of r, the stronger the linear correlation, with zero indicating no correlation at all and 1 indicating a perfect correlation.
- Manual calculations of correlation coefficients can be quite tedious, but a variety of powerful technology tools are available for such calculations.

Communicate Your Understanding

1. Describe the advantages and disadvantages of using a scatter plot or the correlation coefficient to estimate the strength of a linear correlation.

2. a) What is the meaning of a correlation coefficient of
   i) −1?
   ii) 0?
   iii) 0.5?

   b) Can the correlation coefficient have a value greater than 1? Why or why not?

3. A mathematics class finds a correlation coefficient of 0.25 for the students' midterm marks and their driver's test scores and a coefficient of −0.72 for their weight-height ratios and times in a 1-km run. Which of these two correlations is stronger? Explain your answer.

Practise

A

1. Classify the type of linear correlation that you would expect with the following pairs of variables.
   a) hours of study, examination score
   b) speed in excess of the speed limit, amount charged on a traffic fine
   c) hours of television watched per week, final mark in calculus
   d) a person's height, sum of the digits in the person's telephone number
   e) a person's height, the person's strength

2. Identify the independent variable and the dependent variable in a correlational study of
   a) heart disease and cholesterol level
   b) hours of basketball practice and free-throw success rate
   c) amount of fertilizer used and height of plant
   d) income and level of education
   e) running speed and pulse rate

Apply, Solve, Communicate

B

3. For a week prior to their final physics examination, a group of friends collect data to see whether time spent studying or time spent watching TV had a stronger correlation with their marks on the examination.

Hours Studied	Hours Watching TV	Examination Score
10	8	72
11	7	67
15	4	81
14	3	93
8	9	54
5	10	66

   a) Create a scatter plot of hours studied versus examination score. Classify the linear correlation.
   b) Create a similar scatter plot for the hours spent watching TV.
   c) Which independent variable has a stronger correlation with the examination scores? Explain.

d) Calculate the correlation coefficient for hours studied versus examination score and for hours watching TV versus examination score. Do these answers support your answer to c)? Explain.

4. **Application** Refer to the tables in the investigation on page 159.

a) Determine the correlation coefficient and classify the linear correlation for the data for each training method.

b) Suppose that you interchanged the dependent and independent variables, so that the test scores appear on the horizontal axis of a scatter plot and the hours of training appear on the vertical axis. Predict the effect this change will have on the scatter plot and the correlation coefficient for each set of data.

c) Test your predictions by plotting the data and calculating the correlation coefficients with the variables reversed. Explain any differences between your results and your predictions in part b).

5. A company studied whether there was a relationship between its employees' years of service and number of days absent. The data for eight randomly selected employees are shown below.

Employee	Years of Service	Days Absent Last Year
Jim	5	2
Leah	2	6
Efraim	7	3
Dawn	6	3
Chris	4	4
Cheyenne	8	0
Karrie	1	2
Luke	10	1

a) Create a scatter plot for these data and classify the linear correlation.

b) Calculate the correlation coefficient.

c) Does the computed r-value agree with the classification you made in part a)? Explain why or why not.

d) Identify any outliers in the data.

e) Suggest possible reasons for any outliers identified in part d).

6. **Application** Six classmates compared their arm spans and their scores on a recent mathematics test as shown in the following

Arm Span (m)	Score
1.5	82
1.4	71
1.7	75
1.6	66
1.6	90
1.8	73

a) Illustrate these data with a scatter plot.

b) Determine the correlation coefficient and classify the linear correlation.

c) What can the students conclude from their data?

Chapter Problem

7. **a)** Use data in the table on page 157 to create a scatter plot that compares the size of graduating classes in Gina's program to the number of graduates who found jobs.

b) Classify the linear correlation.

c) Determine the linear correlation coefficient.

8. **a)** Search sources such as E-STAT, CANSIM II, the Internet, newspapers, and magazines for pairs of variables that exhibit

i) a strong positive linear correlation

ii) a strong negative linear correlation

iii) a weak or zero linear correlation

b) For each pair of variables in part a), identify the independent variable and the dependent variable.

9. Find a set of data for two variables known to have a perfect positive linear correlation. Use these data to demonstrate that the correlation coefficient for such variables is 1. Alternatively, find a set of data with a perfect negative correlation and show that the correlation coefficient is –1.

10. **Communication**

a) Would you expect to see a correlation between the temperature at an outdoor track and the number of people using the track? Why or why not?

b) Sketch a typical scatter plot of this type of data.

c) Explain the key features of your scatter plot.

11. **Inquiry / Problem Solving** Refer to data tables in the investigation on page 159.

a) How could the Rogers Training Company graph the data so that their training method looks particularly good?

b) How could Laing Limited present the same data in a way that favours their training system?

c) How could a mathematically knowledgeable consumer detect the distortions in how the two companies present the data?

C

12. **Inquiry / Problem Solving**

a) Prove that interchanging the independent and dependent variables does not change the correlation coefficient for any set of data.

b) Illustrate your proof with calculations using a set of data selected from one of the examples or exercise questions in this section.

13. **a)** Search sources such as newspapers, magazines, and the Internet for a set of two-variable data with

i) a moderate positive linear correlation

ii) a moderate negative correlation

iii) a correlation in which $|r| > 0.9$

b) Outline any conclusions that you can make from each set of data. Are there any assumptions inherent in these conclusions? Explain.

c) Pose at least two questions that could form the basis for further research.

14. **a)** Sketch scatter plots of three different patterns of data that you think would have zero linear correlation.

b) Explain why r would equal zero for each of these patterns.

c) Use Fathom™ or a spreadsheet to create a scatter plot that looks like one of your patterns and calculate the correlation coefficient. Adjust the data points to get r as close to zero as you can.