Scatter Plots and Linear Correlation

Does ice cream consumption cause crime? Are older people paid more? In cottage country, are the sales of small businesses affected by the amount of precipitation? These questions deal with possible *relationships* between two variables. Often the answers are not clear-cut, but in *Unit 7: Two-Variable Statistics,* we will investigate methods for **detecting relationships** between variables, for **developing mathematical models** of these relationships, and for **making predictions** using these models.

A. Scatter Plots

Scatter plot:

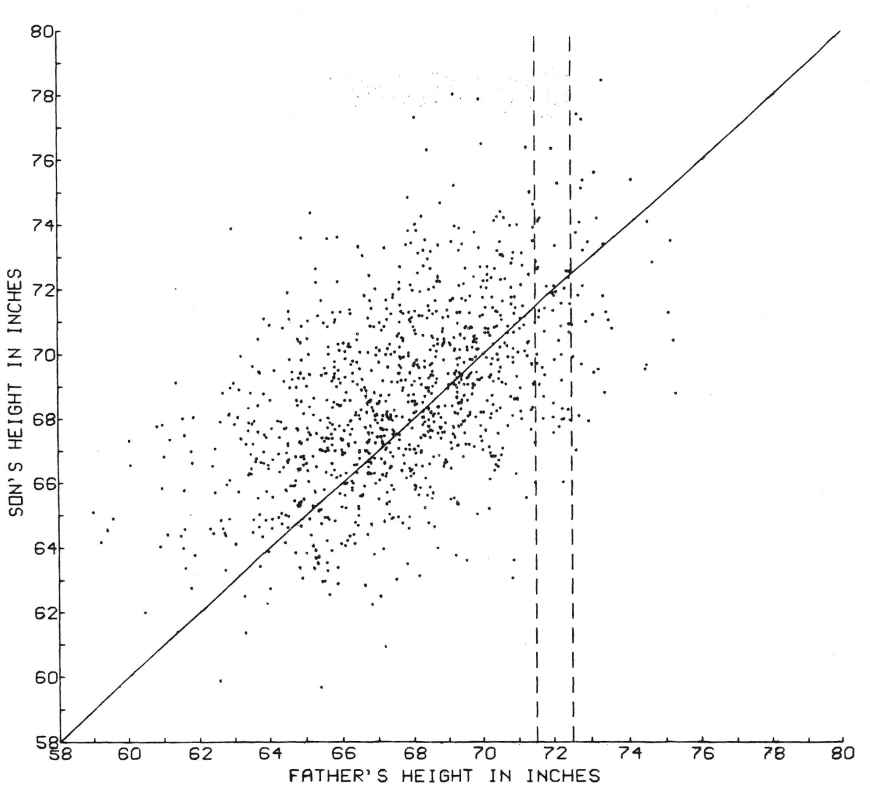
Independent Variable:

Dependent Variable:

Linear Correlation:

Line of Best Fit:

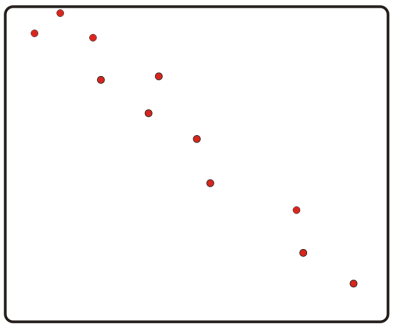
Ex. 1: Statisticians in Victorian England were fascinated by the strength of resemblance between children and their parents, and they gathered huge amounts of data on the subject. One study by Sir Francis Galton (1822 – 1911) and his disciple Karl Pearson (1857 – 1936) compared data on the heights of 1078 fathers and their sons at maturity, one son per father. These numbers would be impossible to grasp as a list, but a scatter plot can be very descriptive. (Each point on the diagram represents one father-son pair.)



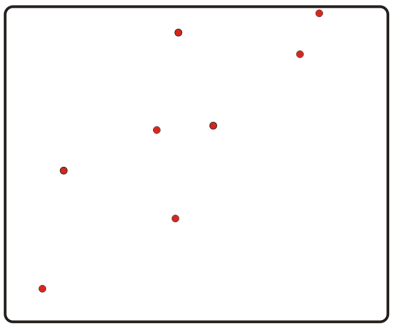
B. Classifying Linear Correlations

Variables are said to have a **linear correlation** if changes in one variable tend to be **proportional** to changes in the other. **The stronger the correlation, the more closely the data points cluster around the line of best fit.** Linear correlations can be classified according to their *direction* (positive, negative) and their *strength* (none [0], weak, moderate, strong, perfect [1]). Positive correlations mean the data cloud slopes up – as one variable increases, so does the other; negative correlations mean the data cloud slopes down – as one variable increases, the other decreases. A **perfect correlation** means that *all points* fall exactly on the line of best fit.

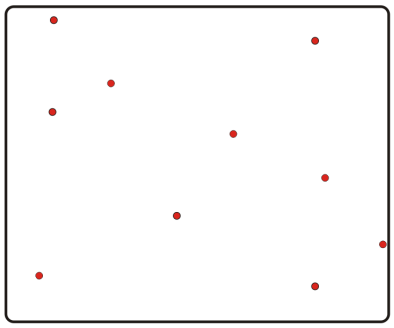
Ex. 2: Classify the relationship between the variables for the data shown in the following scatter plots (none, linear, non-linear; if linear, positive or negative; weak, moderate, strong or perfect).



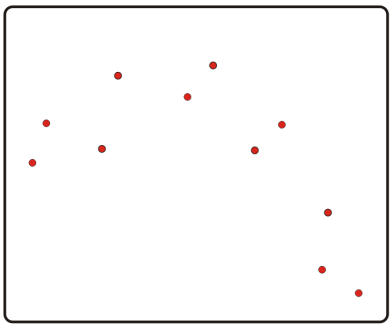
a)_____



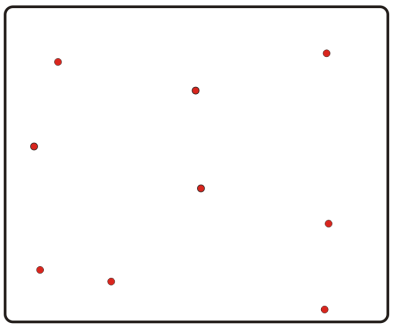
b)_____



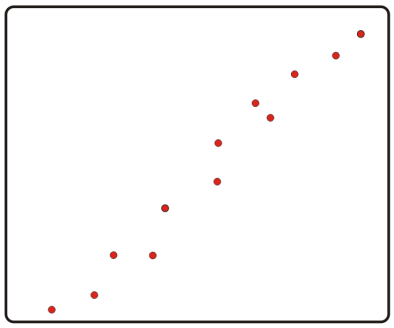
c)_____



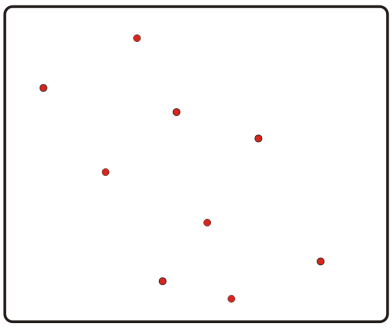
d)_____



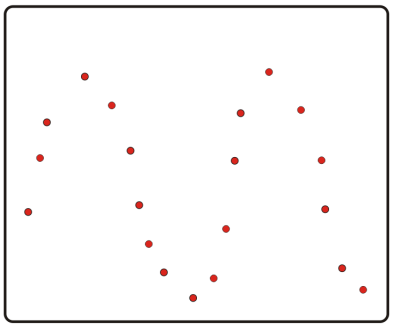
e)_____



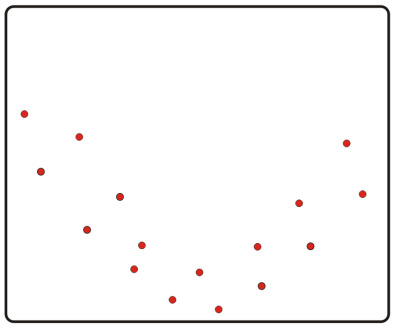
f)_____



g)_____



h)_____



i)_____

C. The Correlation Coefficient

The scatter plot can only give a rough indication of the association between two variables. A more precise way to measure **linear correlation** is to calculate the **correlation coefficient**, ***r***, which is a mathematical measurement of the correlation between two sets of data. It is a pure number, without units. It does not depend on the units or scale chosen for either variable. There are several ways to calculate r, depending on the data you have available and the summary statistics you have calculated already. We will not derive the expressions here although your text does a good job on p. 161 – 163. Karl Pearson, who also invented the term *standard deviation*, developed these formulas. He is considered to be a key figure in the development of modern statistics.

Here, *x* represents individual values of the variable *X* and *y* represents individual values of the variable *Y* and *n* takes its usual meaning, the number of values in the data sets *X* and *Y*. We define the correlation coefficient, *r*:

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

We can determine the relative strength of a correlation, based on the value of *r*, using the scale below:

Whenever possible, look at the scatter plot to check for **outliers** and **non-linear association**. **Outliers** can have a significant impact on the calculation of the correlation coefficient. The correlation coefficient only measures **linear association**, rather than association in general. (More on other types of association later!)

---

*IMPORTANT: correlation ≠ causation

Just because an *r* value is high (close to 1 or -1) does not mean that one variable *necessarily causes* the other to occur! Correlation measures association, but association is not the same as causation! There could be a number of other reasons for a high *r* value, including a common cause for both or some other circumstance (like poor sampling or other confounding factors)!

Ex. 3: "Ice cream sales" and "crime rate" have a very high correlation. Explain why this could be.

Ex. 4: Calculate the correlation coefficient for the following data. Then classify the linear correlation.

x	y	xy	x^2	y^2
54.5	186			
55.0	178			
55.9	172			
56.3	150			
58.4	127			
59.2	112			
60.2	102			
62.3	83			
$\sum x =$	$\sum y =$	$\sum xy =$	$\sum x^2 =$	$\sum y^2 =$

CW: p. 168 "Communicate Your Understanding" # 1-3 ; p. 168 "Practice" # 1, 2, 6bc (by hand), 10

Scatter Plots and Linear Correlation In Excel

Using Technology to Create Scatter Plots and Calculate the Correlation Coefficient, *r*

Ex. 1: Use an Excel spreadsheet to 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, *r*. What is the type and strength of the linear correlation between these two variables?

	A	B	C
1	**Vehicle**	**Horsepower**	**Fuel Consumption (L/100km)**
2	Chevrolet Cruise	105	6.7
3	Jeep Wrangler	170	17.8
4	Toyota RAV 4	124	10.2
5	Honda Shadow Motorcycle	17	3.4
6	Lamborghini Diablo	296	8.4

Steps:

1. Enter the data into columns in Excel (similar to the chart above), making horsepower the independent variable (x) [in cells B2:B6] and fuel consumption the dependent variable (y) [in cells C2:C6].
2. Select the values for both columns and then click on "Insert Chart" and choose "XY Scatter" from the chart options.
3. Add appropriate axis labels and chart title to the scatter plot.
4. Right click on any data point in the chart and select "Add Trendline". Ensure that the type of trend line you choose for this sample is "Linear". Then, from your options, click the radio buttons next to "Display equation on chart" and "Display R-squared value on chart".
5. In the worksheet, you can square-root the R^2 value to get the absolute value of the correlation coefficient, *r*, using the following formula for our dataset: =sqrt(0.17415)
6. To determine the sign of *r*, you look at the slope of the line on the scatter plot: a positive slope gives a positive *r* value; a negative slope gives a negative *r* value.
7. To find the correlation coefficient, *r*, without square-rooting the R^2 value from the scatterplot, type the following formula into a cell to find the correlation coefficient for these two variables: =CORREL(B2:B6,C2:C6) and you get the value of *r* **and** its sign! ☺
8. Use the sign and the value of *r* to describe the type and strength of the linear correlation between these two variables.

CW: p. 168-170 #3, 5, 6 (use Excel for all)

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?

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.

Linear Regression

Regression is an analytic technique for **modeling** the relationship between a dependent variable and an independent variable. **Linear regression** is used when the two variables have a **linear association** (*i.e.* changes in one variable tend to be proportional to changes in the other). Using linear regression, we are able to create an equation for the **regression line** or **line of best fit** (LOBF), rather than simply estimating it by drawing it through a series of points.

A. Using the Least Squares Method to Find the LOBF

Drawing in a line of best fit usually involves some compromise: we move the line closer to some points while increasing its distance from other points; we strive to have a similar number of points above the line as below the line; and we take into consideration the relative distances from each point to the line. The least squares method formalizes this and gives us a precise method for determining the line.

A **residual** is

- Residuals are **positive**
- Residuals are **negative**

For the LOBF in *the Least Squares Method*:

-
-

$$y = mx + b, \quad \text{where} \quad m = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \quad \text{and} \quad b = \bar{y} - m\bar{x}$$

When we have the equation of the regression line, we can estimate values and make predictions by:

Interpolating

Extrapolating

Ex. 2: A university would like to construct a mathematical model to predict first year marks for incoming students based on their grade 12 marks.

Grade 12 Mark (%), x	**1st Year Average (%), y**	**xy**	**x^2**
85	74		
90	83		
76	68		
78	70		
88	75		
84	72		
76	64		
96	91		
86	78		
85	86		
$\sum x =$	$\sum y =$	$\sum xy =$	$\sum x^2 =$

a) Use the given formulas to manually calculate the equation of the line of best fit.
b) Use technology to create a scatter plot of the data and then calculate the equation of best fit.
c) Use the equation to predict the first year average for a student who had an average of 82% in grade 12.

CW: p. 180 # 1, 6, 7 (by hand); p. 181 # 8, 11, 13 (Excel)

Practise

A

1. Identify any outliers in the following sets of data and explain your choices.

a)

X	25	34	43	55	92	105	16
Y	30	41	52	66	18	120	21

b)

X	5	7	6	6	4	8
Y	304	99	198	205	106	9

2. a) Perform a linear regression analysis to generate the line of best fit for each set of data in question 1.
   b) Repeat the linear regressions in part a), leaving out any outliers.
   c) Compare the lines of best fit in parts a) and b).

Apply, Solve, Communicate

B

3. Use the formula for the method of least squares to verify the slope and intercept values you found for the data in the investigation on page 171. Account for any discrepancies.

4. Use software or a graphing calculator to verify the regression results in Example 1.

5. Application The following table lists the heights and masses for a group of fire-department trainees.

Height (cm)	Mass (kg)
177	91
185	88
173	82
169	79
188	87
182	85
175	79

   a) Create a scatter plot and classify the linear correlation.
   b) Apply the method of least squares to generate the equation of the line of best fit.
   c) Predict the mass of a trainee whose height is 165 cm.
   d) Predict the height of a 79-kg trainee.
   e) Explain any discrepancy between your answer to part d) and the actual height of the 79-kg trainee in the sample group.

6. A random survey of a small group of high-school students collected information on the students' ages and the number of books they had read in the past year.

Age (years)	Books Read
16	5
15	3
18	8
17	6
16	4
15	4
14	5
17	15

   a) Create a scatter plot for this data. Classify the linear correlation.
   b) Determine the correlation coefficient and the equation of the line of best fit.
   c) Identify the outlier.
   d) Repeat part b) with the outlier excluded.
   e) Does removing the outlier improve the linear model? Explain.
   f) Suggest other ways to improve the model.
   g) Do your results suggest that the number of books a student reads depends on the student's age? Explain.

7. Application Market research has provided the following data on the monthly sales of a licensed T-shirt for a popular rock band.

Price ($)	Monthly Sales
10	2500
12	2200
15	1600
18	1200
20	800
24	250

a) Create a scatter plot for these data.

b) Use linear regression to model these data.

c) Predict the sales if the shirts are priced at $19.

d) The vendor has 1500 shirts in stock and the band is going to finish its concert tour in a month. What is the maximum price the vendor can charge and still avoid having shirts left over when the band stops touring?

8. Communication MDM Entertainment has produced a series of TV specials on the lives of great mathematicians. The executive producer wants to know if there is a linear correlation between production costs and revenue from the sales of broadcast rights. The costs and gross sales revenue for productions in 2001 and 2002 were as follows (amounts in millions of dollars).

2001		2002	
Cost ($M)	Sales ($M)	Cost ($M)	Sales ($M)
5.5	15.4	2.7	5.2
4.1	12.1	1.9	1.0
1.8	6.9	3.4	3.4
3.2	9.4	2.1	1.9
4.2	1.5	1.4	1.5

a) Create a scatter plot using the data for the productions in 2001. Do there appear to be any outliers? Explain.

b) Determine the correlation coefficient and the equation of the line of best fit.

c) Repeat the linear regression analysis with any outliers removed.

d) Repeat parts a) and b) using the data for the productions in 2002.

e) Repeat parts a) and b) using the combined data for productions in both 2001 and 2002. Do there still appear to be any outliers?

f) Which of the four linear equations do you think is the best model for the relationship between production costs and revenue? Explain your choice.

g) Explain why the executive producer might choose to use the equation from part d) to predict the income from MDM's 2003 productions.

9. At Gina's university, there are 250 business students who expect to graduate in 2006.

Chapter Problem

a) Model the relationship between the total number of graduates and the number hired by performing a linear regression on the data in the table on page 157. Determine the equation of the line of best fit and the correlation coefficient.

b) Use this linear model to predict how many graduates will be hired in 2006.

c) Identify any outliers in this scatter plot and suggest possible reasons for an outlier. Would any of these reasons justify excluding the outlier from the regression calculations?

d) Repeat part a) with the outlier removed.

e) Compare the results in parts a) and d). What assumptions do you have to make?

10. Communication Refer to Example 2, which describes population data for wolves and rabbits in a wildlife reserve. An alternate theory has it that the rabbit population depends on the wolf population since the wolves prey on the rabbits.

a) Create a scatter plot of rabbit population versus wolf population and classify the linear correlation. How are your data points related to those in Example 2?

b) Determine the correlation coefficient and the equation of the line of best fit. Graph this line on your scatter plot.

c) Is the equation of the line of best fit the inverse of that found in Example 2? Explain.

d) Plot both populations as a time series. Can you recognize a pattern or relationship between the two series? Explain.

e) Does the time series suggest which population is the dependent variable? Explain.

11. The following table lists the mathematics of data management marks and grade 12 averages for a small group of students.

Mathematics of Data Management Mark	Grade 12 Average
74	77
81	87
66	68
53	67
92	85
45	55
80	76

a) Using Fathom™ or *The Geometer's Sketchpad®*,

i) create a scatter plot for these data

ii) add a moveable line to the scatter plot and construct the geometric square for the deviation of each data point from the moveable line

iii) generate a dynamic sum of the areas of these squares

iv) manoeuvre the moveable line to the position that minimizes the sum of the areas of the squares.

v) record the equation of this line

b) Determine the equation of the line of best fit for this set of data.

c) Compare the equations you found in parts a) and b). Explain any differences or similarities.

12. Application Use E-STAT or other sources to obtain the annual consumer price index figures from 1914 to 2000.

a) Download this information into a spreadsheet or statistical software, or enter it into a graphing calculator. (If you use a graphing calculator, enter the data from every third year.) Find the line of best fit and comment on whether a straight line appears to be a good model for the data.

b) What does the slope of the line of best fit tell you about the rate of inflation?

c) Find the slope of the line of best fit for the data for just the last 20 years, and then repeat the calculation using only the data for the last 5 years.

d) What conclusions can you make by comparing the three slopes? Explain your reasoning.

Non-Linear Regression

Recall that **linear regression** is a technique for finding the equation of the **line of best fit** (LOBF) when two variables have a **linear association** (*i.e.* changes in one variable tend to be *proportional* to changes in the other). **Non-linear regression** is an analytic technique for finding equations that model the ***curve* of best fit** when two variables have a **non-linear association**.

For linear associations, we used the correlation coefficient, *r*, to describe how closely the LOBF models the data. For **non-linear associations**, we will instead use **r^2, the coefficient of determination,** to determine how closely a curve fits the data. Calculations for curves are more complicated than those for straight lines, so we will be using technology to assist us. Note that because r has been squared, $0 \le r^2 \le 1$.

A. Exponential Regression

Exponential regressions produce equations of the form $y = ab^x$ or $y = ae^{kx}$, where $e = 2.71828\ldots$ is an irrational number commonly used as the base for exponents and logarithms. These two forms are equivalent, and it is straightforward to convert from one to the other using the equation $b = e^k$. In the formula, ***a*** represents the initial value and ***b*** represents the factor by which the initial value changes for each increase in ***x***.

***Equation of the Curve of Best-Fit for Exponential Regression*:**

$y = ab^x$ or $y = ae^{kx}$, where $e = 2.71828\ldots$ and $b = e^k$

Ex. 1: A lab technician monitors the growth of a bacterial culture by scanning it every hour and estimating the number of bacteria. The initial population is unknown.

Time (h)	0	1	2	3	4	5	6	7
Population	1	10	21	43	82	168	320	475

a) Enter the data into the technology of your choice and create a scatter plot. Sketch the result on the axes below.

b) In Excel, use **linear regression** to determine the correlation coefficient and the equation of the **line of best fit**. Graph the LOBF on your Excel scatter plot and sketch the result below.

r =

y =

c) What do you observe?

d) For the same data, use **exponential regression** to find the coefficient of determination and the equation of the **exponential curve of best fit**. Graph the curve of best fit on your Excel scatter plot and sketch the result on the same axes as in part b).

r^2 =

y =

e) What do you observe?

f) Use the equation to estimate the population of the bacterial sample at 9h.

We will now investigate two other types of regression curves: power regression and polynomial regression. It is important to note that sometimes more than one type of regression curve can provide a good fit for the data. To be an effective model, however, the curve must be useful for **extrapolating** beyond the data.

B. Power Regression

Equation of the Curve of Best-Fit for a Power Regression: $y = ax^b$

Ex. 2: For a physics project, a group of students videotape a ball dropped from the top of a 4-m high ladder, which they have marked every 10 cm. During playback, they stop the videotape every tenth of a second and compile the following table for the distance the ball travelled.

Time (s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Distance (m)	0.05	0.2	0.4	0.8	1.2	1.7	2.4	3.1	3.9	4.9

a) Enter the data into Excel and create a scatter plot. Sketch the result on the axes below.

b) Perform a **linear regression** and an **exponential regression**. Graph the LOBF and the COBF in Excel and sketch the results on the axes.

c) What do you observe?

d) Use a **power regression** to find a curve of best fit for the data. (In *Excel*, use the **Chart** feature to add a trend line, but select **Power** this time). Sketch the result.

e) Does the power regression curve fit the data more closely than the exponential model? Describe what you observe.

f) Determine the coefficient of determination and the equation of the curve of best fit.

$r^2 =$

$y =$

g) Use this equation to predict how far the ball will fall in 5 s.

C. Polynomial Regression

Polynomial regression curves include quadratic curves ($y = ax^2 + bx + c$), cubic curves ($y = ax^3 + bx^2 + cx + d$), quartic curves, and other polynomial functions of degree n.

Equation of the Curve of Best-Fit for a Polynomial Regression of Degree n:

$$y = ax^n + bx^{n-1} + cx^{n-2} + \ldots , \quad \text{where } a \neq 0.$$

Ex. 3: Suppose that the lab technician from Example 1 took additional measurements of the bacterial culture.

Time (h)	8	9	10	11	12	13	14
Population	630	775	830	980	1105	1215	1410

a) Enter the previous data and the new data into Excel and create another scatter plot. Sketch the result below.

b) Add a trend line using an **exponential model**, as we did for the previous data. Sketch the results.

c) How effective is this exponential model now?

d) Use a **polynomial regression** with *degree*/*order* = 2 to find a curve of best fit. Sketch the result.

r^2 =

y =

e) Now try a **polynomial regression** with *degree*/*order* = 3 to find a curve of best fit.

r^2 =

y =

f) Keep on increasing the degree/order of the polynomial and note what happens to r^2 each time.

One interesting property of polynomial regression is that you can get a curve that fits the data perfectly as long as the degree of your polynomial function is one less than the number of your data points; *i.e.* you have *n* data points and you choose a polynomial regression with degree (*n* – 1). However, these polynomials are not always the best models for the data. Often they can give inaccurate predictions outside of the data range. *Extrapolating* to an initial state or final state (looking at "end behaviour") may help to determine which model is most suitable.

CW: p. 191 # 1, 2c, 3, 5, 7, 9 (Excel for all)

Key Concepts

- Some relationships between two variables can be modelled using non-linear regressions such as quadratic, cubic, power, polynomial, and exponential curves.
- The coefficient of determination, r^2, is a measure of how well a regression curve fits a set of data.
- Sometimes more than one type of regression curve can provide a good fit for data. To be an effective model, however, the curve must be useful for extrapolating beyond the data.

Communicate Your Understanding

1. A data set for two variables has a linear correlation coefficient of 0.23. Does this value preclude a strong correlation between the variables? Explain why or why not.
2. A best-fit curve for a set of data has a coefficient of determination of $r^2 = 0.76$. Describe some techniques you can use to improve the model.

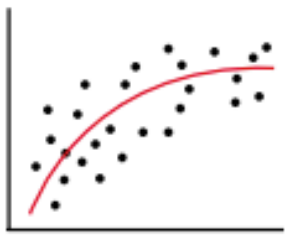
Practise

A

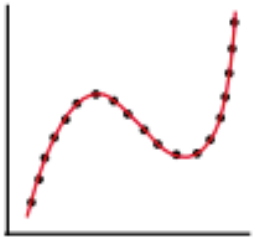
1. Match each of the following coefficients of determination with one of the diagrams below.

a) 0 **b)** 0.5 **c)** 0.9 **d)** 1

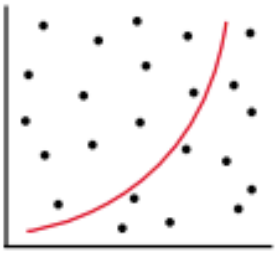
i)



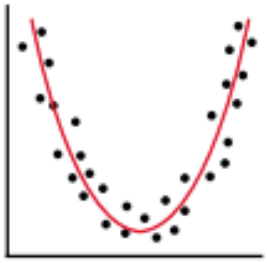
ii)



iii)



iv)



2. For each set of data use software or a graphing calculator to find the equation and coefficient of determination for a curve of best fit.

a)

x	y
−2.8	0.6
−3.5	−5.8
−2	3
−1	6
0.2	4
1	1
−1.5	5
1.4	−3.1
0.7	3
−0.3	6.1
−3.3	−3.1
−4	−7
2	−5.7

b)

x	y
−2.7	1.6
−3.5	−3
−2.2	3
−0.5	−0.5
0	1.3
0.6	4.7
−1.8	1.7
−3.8	−7
−1.3	0.6
0.8	7
0.5	2.7
−1	1.5
−3	−1.1

c)

x	y
1.1	2.5
3.5	11
2.8	8.6
2.3	7
0	1
3.8	14
1.4	4.2
−4	0.2
−1.3	0.6
3	12
4.1	17
2.2	5
−2.7	0.4

Apply, Solve, Communicate

B

3. The heights of a stand of pine trees were measured along with the area under the cone formed by their branches.

Height (m)	Area (m^2)
2.0	5.9
1.5	3.4
1.8	4.8
2.4	8.6
2.2	7.3
1.2	2.1
1.8	4.9
3.1	14.4

a) Create a scatter plot of these data.

b) Determine the correlation coefficient and the equation of the line of best fit.

c) Use a power regression to calculate a coefficient of determination and an equation for a curve of best fit.

d) Which model do you think is more accurate? Explain why.

e) Use the more accurate model to predict

i) the area under a tree whose height is 2.7 m

ii) the height of a tree whose area is 30 m^2

f) Suggest a reason why the height and circumference of a tree might be related in the way that the model in part d) suggests.

4. Application The biologist Max Kleiber (1893–1976) pioneered research on the metabolisms of animals. In 1932, he determined the relationship between an animal's mass and its energy requirements or basal metabolic rate (BMR). Here are data for eight animals.

Animal	Mass (kg)	BMR (kJ/day)
Frog	0.018	0.050
Squirrel	0.90	1.0
Cat	3.0	2.6
Monkey	7.0	4.0
Baboon	30	14
Human	60	25
Dolphin	160	44
Camel	530	116

a) Create a scatter plot and explain why Kleiber thought a power-regression curve would fit the data.

b) Use a power regression to find the equation of the curve of best fit. Can you rewrite the equation so that it has exponents that are whole numbers? Do so, if possible, or explain why not.

c) Is this power equation a good mathematical model for the relationship between an animal's mass and its basal metabolic rate? Explain why or why not.

d) Use the equation of the curve of best fit to predict the basal metabolic rate of

i) a 15-kg dog

ii) a 2-tonne whale

5. Application As a sample of a radioactive element decays into more stable elements, the amount of radiation it gives off decreases. The level of radiation can be used to estimate how much of the original element remains. Here are measurements for a sample of radium-227.

Time (h)	Radiation Level (%)
0	100
1	37
2	14
3	5.0
4	1.8
5	0.7
6	0.3

a) Create a scatter plot for these data.

b) Use an exponential regression to find the equation for the curve of best fit.

c) Is this equation a good model for the radioactive decay of this element? Explain why or why not.

d) A half-life is the time it takes for half of the sample to decay. Use the regression equation to estimate the half-life of radium-227.

6. a) Create a time-series graph for the mean starting salary of the graduates who find jobs. Describe the pattern that you see.

b) Use non-linear regression to construct a curve of best fit for the data. Record the equation of the curve and the coefficient of determination.

c) Comment on whether this equation is a good model for the graduates' starting salaries.

7. An engineer testing the transmitter for a new radio station measures the radiated power at various distances from the transmitter. The engineer's readings are in microwatts per square metre.

Distance (km)	Power Level ($\mu W/m^2$)
2.0	510
5.0	78
8.0	32
10.0	19
12.0	14
15.0	9
20.0	5

a) Find an equation for a curve of best fit for these data that has a coefficient of determination of at least 0.98.

b) Use the equation for this curve of best fit to estimate the power level at a distance of

i) 1.0 km from the transmitter

ii) 4.0 km from the transmitter

iii) 50.0 km from the transmitter

8. Communication *Logistic* curves are often a good model for population growth. These curves have equations with the form $y = \frac{c}{1 + ae^{-bx}}$, where a, b, and c are constants.

Consider the following data for the bacterial culture in Example 1:

Time (h)	0	1	2	3	4	5
Population	?	10	21	43	82	168
Time (h)	6	7	8	9	10	11
Population	320	475	630	775	830	980
Time (h)	12	13	14	15	16	17
Population	1105	1215	1410	1490	1550	1575
Time (h)	18	19	20			
Population	1590	1600	1600			

a) Use software or a graphing calculator to find the equation and coefficient of determination for the logistic curve that best fits the data for the bacteria population from 1 to 20 h.

b) Graph this curve on a scatter plot of the data.

c) How well does this curve appear to fit the entire data set? Describe the shape of the curve.

d) Write a brief paragraph to explain why you think a bacterial population may exhibit this type of growth pattern.

Cause and Effect

A. Causal Relationships

Ex. 1: Before the invention of the Salk vaccine against polio, investigators looked at the relationship between the incidence of polio and the number of soft drinks sold. For each week of the year, they tabulated the number of soft drinks sold and the number of new polio cases reported. These data points showed strong positive correlation: during weeks when more soft drinks were sold, there were more new cases of polio; when fewer drinks were sold, there were fewer such cases.

Can we conclude that soft drinks cause polio?

Ex. 2: For a sample of American men aged 18 to 54, what would you expect the correlation to be between income and blood pressure?

A positive correlation is observed between blood pressure and income. Does this point to a causal relationship, or can it be explained some other way?

Correlation does not *necessarily* imply **causation**!

Correlations can also result from: **common-cause** factors, **reverse cause-and-effect** relationships, **accidental** relationships, and **presumed** relationships!

1. Cause-and-Effect Relationship:

2. Common-Cause Factor:

3. Reverse Cause-and-Effect Relationship:

4. Accidental Relationship:

5. Presumed Relationship:

Ex. 3: Classify the relationships in the following situations:

a. The rate of a chemical reaction increases with temperature.

b. Leadership ability has a positive correlation with academic achievement.

c. The prices of motorcycles and butter have a strong positive correlation.

d. Sales of cell phones had a strong negative correlation with ozone levels in the atmosphere over the past decade.

e. Traffic congestion has a strong correlation with the number of urban expressways.

B. Extraneous Variables

With the variety of causal relationships (cause-and-effect, common-cause, reverse cause-and-effect, accidental, and presumed) we noticed that often several types of relationships might be involved in the same situation. In order to appreciate the relationship that exists, we also must consider **extraneous variables** (or **confounding factors**).

- **Extraneous Variables:**

Ex: 1: You would expect there to be a strong positive correlation between your mid-term mark and your final mark for this course. What are a few of the possible extraneous variables that could affect this relationship?

C. Experimental Design

In order to minimize the effects of extraneous variables, researchers will often compare an **experimental group** to a **control group**.

- **Experimental Group:**

- **Control Group:**

Any *differences* in the dependent variable for the two groups can then be attributed to the changes in the independent variable.

Ex. 2: A medical researcher wants to test a new drug believed to help smokers overcome the addictive effects of nicotine. Fifty people who want to quit smoking volunteer for the study. The researcher carefully divides the volunteers into two groups, each with an equal number of moderate and heavy smokers. One group is given nicotine patches with the new drug, while the second group uses ordinary nicotine patches. Fourteen people in the first group quit smoking, as do 9 people in the second group.

a) Identify the experimental group, the control group, the independent variable, and the dependent variable.
b) Can the researcher conclude that the new drug is effective?
c) What further study should the researcher do?

a) *Experimental Group:* *Control Group:*

Independent Variable: *Dependent Variable:*

b) Conclusions from the study:

i.

ii.

c) Further steps

i. Should the participants be made aware of which drug is being administered to them?

ii. Should the researchers know who has the drug and who doesn't?

iii. *Placebo* effect:

CW: p. 199-201 # 1 – 6, 10, 14

- Use sampling methods that hold the extraneous variables constant.
- Conduct similar investigations with different samples and check for consistency in the results.
- Remove, or account for, possible common-cause factors.

The later chapters in this book introduce probability theory and some statistical methods for a more quantitative approach to determining cause-and-effect relationships.

Project Prep

In your statistics project, you may wish to consider cause-and-effect relationships and extraneous variables that could affect your study.

Key Concepts

- Correlation does not necessarily imply a cause-and-effect relationship. Correlations can also result from common-cause factors, reverse cause-and-effect relationships, accidental relationships, and presumed relationships.
- Extraneous variables can invalidate conclusions based on correlational evidence.
- Comparison with a control group can help remove the effect of extraneous variables in a study.

Communicate Your Understanding

1. Why does a strong linear correlation not imply cause and effect?
2. What is the key characteristic of a reverse cause-and-effect relationship?
3. Explain the difference between a common-cause factor and an extraneous variable.
4. Why are control groups used in statistical studies?

Practise

A

1. Identify the most likely type of causal relationship between each of the following pairs of variables. Assume that a strong positive correlation has been observed with the first variable as the independent variable.
   a) alcohol consumption, incidence of automobile accidents
   b) score on physics examination, score on calculus examination
   c) increase in pay, job performance
   d) population of rabbits, consumer price index
   e) number of scholarships received, number of job offers upon graduation
   f) coffee consumption, insomnia
   e) funding for athletic programs, number of medals won at Olympic games

2. For each of the following common-cause relationships, identify the common-cause factor. Assume a positive correlation between each pair of variables.
   a) number of push-ups performed in one minute, number of sit-ups performed in one minute
   b) number of speeding tickets, number of accidents
   c) amount of money invested, amount of money spent

Apply, Solve, Communicate

3. A civil engineer examining traffic flow problems in a large city observes that the number of traffic accidents is positively correlated with traffic density and concludes that traffic density is likely to be a major cause of accidents. What alternative conclusion should the engineer consider?

B

4. Communication An elementary school is testing a new method for teaching grammar. Two similar classes are taught the same material, one with the established method and the other with the new method. When both classes take the same test, the class taught with the established method has somewhat higher marks.
   a) What extraneous variables could influence the results of this study?
   b) Explain whether the study gives the school enough evidence to reject the new method.
   c) What further studies would you recommend for comparing the two teaching methods?

5. Communication An investor observes a positive correlation between the stock price of two competing computer companies. Explain what type of causal relationship is likely to account for this correlation.

6. Application A random survey of students at Statsville High School found that their interest in computer games is positively correlated with their marks in mathematics.
   a) How would you classify this causal relationship?
   b) Suppose that a follow-up study found that students who had increased the time they spent playing computer games tended to improve their mathematics marks. Assuming that this study held all extraneous variables constant, would you change your assessment of the nature of the causal relationship? Explain why or why not.

7. a) The net assets of Custom Industrial Renovations Inc., an industrial construction contractor, has a strong negative linear correlation with those of MuchMega-Fun, a toy distributor. How would you classify the causal relationship between these two variables?
   b) Suppose that the two companies are both subsidiaries of Diversified Holdings Ltd., which often shifts investment capital between them. Explain how this additional information could change your interpretation of the correlation in part a).

8. Communication Aunt Gisele simply cannot sleep unless she has her evening herbal tea. However, the package for the tea does not list any ingredients known to induce sleep. Outline how you would conduct a study to determine whether the tea really does help people sleep.

9. Find out what a *double-blind* study is and briefly explain the advantages of using this technique in studies with a control group.

10. a) The data on page 157 show a positive correlation between the size of the graduating class and the number of

Chapter problem

graduates hired. Does this correlation mean that increasing the number of graduates causes a higher demand for them? Explain your answer.

b) A recession during the first half of the 1990s reduced the demand for business graduates. Review the data on page 157 and describe any trends that may be caused by this recession.

ACHIEVEMENT CHECK

Knowledge/ Understanding	Thinking/Inquiry/ Problem Solving	Communication	Application

11. The table below lists numbers of divorces and personal bankruptcies in Canada for the years 1976 through 1985.

Year	Divorces	Bankruptcies
1976	54 207	10 049
1977	55 370	12 772
1978	57 155	15 938
1979	59 474	17 876
1980	62 019	21 025
1981	67 671	23 036
1982	70 436	30 643
1983	68 567	26 822
1984	65 172	22 022
1985	61 976	19 752

a) Create a scatter plot and classify the linear correlation between the number of divorces and the number of bankruptcies.

b) Perform a regression analysis. Record the equation of the line of best fit and the correlation coefficient.

c) Identify an external variable that could be a common-cause factor.

d) Describe what further investigation you could do to analyse the possible relationship between divorces and bankruptcies.

12. Search the E-STAT, CANSIM II, or other databases for a set of data on two variables with a positive linear correlation that you believe to be accidental. Explain your findings and reasoning.

C

13. Use a library, the Internet, or other resources to find information on the Hawthorne effect and the placebo effect. Briefly explain what these effects are, how they can affect a study, and how researchers can avoid having their results skewed by these effects.

14. Inquiry/Problem Solving In a behavioural study of responses to violence, an experimental group was shown violent images, while a control group was shown neutral images. From the initial results, the researchers suspect that the gender of the people in the groups may be an extraneous variable. Suggest how the study could be redesigned to

a) remove the extraneous variable

b) determine whether gender is part of the cause-and-effect relationship

15. Look for material in the media or on the Internet that incorrectly uses correlational evidence to claim that a cause-and-effect relationship exists between the two variables. Briefly describe

a) the nature of the correlational study

b) the cause and effect claimed or inferred

c) the reasons why cause and effect was not properly proven, including any extraneous variables that were not accounted for

d) how the study could be improved

Critical Analysis

With an appreciation for the contents of this unit, you are now armed with knowledge to help you think critically about the information presented to you on a daily basis.

How easily do you accept statistical evidence from sources that could be biased or flawed?

On a scale from 1 – 10, (with 10 being very trustworthy and 1 being very untrustworthy), how would you rank the following:

- ✓ Newspapers
- ✓ Television advertisements
- ✓ Nightly news (CTV, Global, CBC)
- ✓ Investigative news programs (W5, 20/20, Dateline)

Although media is generally careful about how they present statistics, reporters, editors, or advertisers often face tight deadlines and lack the mathematical knowledge to thoroughly critique statistical material.

Lobby groups, advertisers and newsmakers like stats because it makes their claims sound more accurate and scientific. Unfortunately, these statistics are sometimes flawed by unintentional or deliberate bias. To judge the conclusions of a study properly, you need information about its sampling and analytical tools.

Ex. 1: The manager of a small business has declared that a specific aptitude test correctly predicts employee productivity and has made it essential for all potential employees to write the test before they are hired. As the person in charge of human resources and someone who is knowledgeable of statistical practices, you are not convinced the manager's claims are correct and you set out to test his hypothesis.

The manager had all 30 current employees write the test and then compared their scores to their productivity as measured by their most recent performance review. The data was ordered alphabetically by employee surname. To simplify the calculations, the manager selects a systematic sample using every seventh employee.

Test Score	98	57	82	76	65	72	91	87	81	39
Productivity	78	81	83	44	62	89	85	71	76	71

Test Score	50	75	71	89	82	95	56	71	68	77
Productivity	66	90	48	80	83	72	72	90	74	51

Test Score	59	83	75	66	48	61	78	70	68	64
Productivity	65	47	91	77	63	58	55	73	75	69

a) Using the sample the manager has selected (shaded cells) complete a linear regression analysis.

b) Examine the raw data. Create a scatter plot of all 30 data points and complete a regression using these data.

c) What can you conclude from these findings?

CW: Read Examples 2 and 3 in Section 3.5 of your text; complete p. 209 # 1 – 5, 8

Key Concepts

- Although the major media are usually responsible in how they present statistics, you should be cautious about accepting any claim that does not include information about the sampling technique and analytical methods used.
- Intentional or unintentional bias can invalidate statistical claims.
- Small sample sizes and inappropriate sampling techniques can distort the data and lead to erroneous conclusions.
- Extraneous variables must be eliminated or accounted for.
- A hidden variable can skew statistical results and yet still be hard to detect.

Communicate Your Understanding

1. Explain how a small sample size can lead to invalid conclusions.
2. A city councillor states that there are problems with the management of the police department because the number of reported crimes in the city has risen despite increased spending on law enforcement. Comment on the validity of this argument.
3. Give an example of a hidden variable not mentioned in this section, and explain why this variable would be hard to detect.

Apply, Solve, Communicate

A

1. An educational researcher discovers that levels of mathematics anxiety are negatively correlated with attendance in mathematics class. The researcher theorizes that poor attendance causes mathematics anxiety. Suggest an alternate interpretation of the evidence.
2. A survey finds a correlation between the proportion of high school students who own a car and the students' ages. What hidden variable could affect this study?

B

3. A student compares height and grade average with four friends and collects the following data.

Height (cm)	Grade Average (%)
171	73
145	91
162	70
159	81
178	68

From this table, the student concludes that taller students tend to get lower marks.

a) Does a regression analysis support the student's conclusion?

b) Why are the results of this analysis invalid?

c) How can the student get more accurate results?

4. Inquiry/Problem Solving A restaurant chain randomly surveys its customers several times a year. Since the surveys show that the level of customer satisfaction is rising over time, the company concludes that its customer service is improving. Discuss the validity of the surveys and the conclusion based on these surveys.

5. Application A teacher offers the following data to show that good attendance is important.

Days Absent	Final Grade
8	72
2	75
0	82
11	68
15	66
20	30

A student with a graphing calculator points out that the data indicate that anyone who misses 17 days or more is in danger of failing the course.

a) Show how the student arrived at this conclusion.

b) Identify and explain the problems that make this conclusion invalid.

c) Outline statistical methods to avoid these problems.

6. Using a graphing calculator, Gina found the cubic curve of best fit for the salary data in the table on page 157. This curve has a coefficient of determination of 0.98, indicating an almost perfect fit to the data. The equation of the cubic curve is

$$\text{starting salary} = 0.0518y^3 - 310y^2 + 618\,412y - 411\,344\,091$$

where the salary is given in thousands of dollars and y is the year of graduation.

a) What mean starting salary does this model predict for Gina's class when they graduate in 2005?

b) Is this prediction realistic? Explain.

c) Explain why this model generated such an inaccurate prediction despite having a high value for the coefficient of determination.

d) Suggest methods Gina could use to make a more accurate prediction.

7. Communication Find a newspaper or magazine article, television commercial, or web page that misuses statistics of two variables. Perform a critical analysis using the techniques in this chapter. Present your findings in a brief report.

8. Application A manufacturing company keeps records of its overall annual production and its number of employees. Data for a ten-year period are shown below.

Year	Number of Employees	Production (000)
1992	158	75
1993	165	81
1994	172	84
1995	148	68
1996	130	58
1997	120	51
1998	98	50
1999	105	57
2000	110	62
2001	120	70

a) Create a scatter plot to see if there is a linear correlation between annual production and number of employees. Classify the correlation.

b) At some point, the company began to lay off workers. When did these layoffs begin?

c) Does the scatter plot suggest the presence of a hidden variable? Could the layoffs account for the pattern you see? Explain why or why not.

d) The company's productivity is its annual production divided by the number of

employees. Create a time-series graph for the company's productivity.

e) Find the line of best fit for the graph in part d).

f) The company has adopted a better management system. When do you think the new system was implemented? Explain your reasoning.

C

9. Search E-STAT, CANSIM II, or other sources for time-series data for the price of a commodity such as gasoline, coffee, or computer memory. Analyse the data and comment on any evidence of a hidden variable. Conduct further research to determine if there are any hidden variables. Write a brief report outlining your analysis and conclusions.

10. Inquiry/Problem Solving A study conducted by Stanford University found that behavioural counselling for people who had suffered a heart attack reduced the risk of a further heart attack by 45%. Outline how you would design such a study. List the independent and dependent variables you would use and describe how you would account for any extraneous variables.

Career Connection

Economist

Economists apply statistical methods to develop mathematical models of the production and distribution of wealth. Governments, large businesses, and consulting firms are employers of economists. Some of the functions performed by an economist include

- recognizing and interpreting domestic and international market trends
- using supply and demand analysis to assess market potential and set prices
- identifying factors that affect economic growth, such as inflation and unemployment
- advising governments on fiscal and monetary policies
- optimizing the economic activity of financial institutions and large businesses

Typically, a bachelor's degree in economics is necessary to enter this field. However, many positions require a master's or doctorate degree or specialized training. Since economists often deal with large amounts of data, a strong background in statistics and an ability to work with computers are definite assets.

An economist can expect to earn a comfortable living. Most employment opportunities for economists are in large cities. The current demand for economists is reasonably strong and likely to remain so for the foreseeable future, as governments and large businesses will continue to need the information and analysis that economists provide.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site and follow the links to learn more about a career as an economist and other related careers.

Review: Unit 7
Two Variable Statistics

1. The following data has been collected to determine is a relationship exists between the amount of snowfall in Toronto and the number of students who attend lecture at U of T.
   a. Complete the table below and use the results to calculate the correlation coefficient *by hand*. Do not make a scatter plot and do not use Excel. *(5 marks)*

Year	Snowfall, *x* (cm)	Number of students in class, *y*	x^2	y^2	xy
1995	173	182			
1996	165	190			
1997	152	207			
1998	184	180			
1999	178	184			
Totals					

   b. Explain what this correlation tells you about the relationship between the amount of snowfall and the number of students who attend lecture. *(2 marks)*

2. Future Shop kept track of the number of advertisements it placed in local newspapers and the number of stereo systems it sold each week. The data is shown in the chart below.

Week	1	2	3	4	5	6	7	8
Ads, x	6	5	3	2	1	4	3	2
Sales, y	20	15	12	8	6	7	9	7

   a. Determine the line of best fit by creating a chart in Excel and using the **formulas** learned in class. *(6 marks)*
   b. *Verify* your answer above by creating a scatter plot and displaying a trend line in Excel. Account for any differences. *(5 marks)*
   c. What is the correlation coefficient for this set of data? *(2 marks)*
   d. What does the correlation coefficient suggest about the effectiveness of the ads? *(2 marks)*

3. A car safety association conducted tests to measure the stopping distance of a new model of car and collected the following measurements.

Speed (km/h)	30	40	50	60	70	80	90	100
Stopping Distance (m)	19.2	22.2	24.8	27.1	29.5	31.6	33.2	35.0

   a. Make a scatter plot of these data. *(4 marks)*
   b. Use regression analysis to determine a suitable curve of best fit, its equation and the coefficient of determination. Explain the reasons you selected the model you have. *(6 marks)*
   c. Estimate the stopping distance for a speed of 140km/h using your selected model. *(3 marks)*