3.2 Linear Regression

Regression is an analytic technique for determining the relationship between a dependent variable and an independent variable. When the two variables have a linear correlation, you can develop a simple mathematical model of the relationship between the two variables by finding a line of best fit. You can then use the equation for this line to make predictions by **interpolation** (estimating between data points) and **extrapolation** (estimating beyond the range of the data).

INVESTIGATE & INQUIRE: Modelling a Linear Relationship

A university would like to construct a mathematical model to predict first-year marks for incoming students based on their achievement in grade 12. A comparison of these marks for a random sample of first-year students is shown below.

Grade 12 Average	85	90	76	78	88	84	76	96	86	85
First-Year Average	74	83	68	70	75	72	64	91	78	86

1. a) Construct a scatter plot for these data. Which variable should be placed on the vertical axis? Explain.

   b) Classify the linear correlation for this data, based on the scatter plot.

2. a) Estimate and draw a line of best fit for the data.

   b) Measure the slope and y-intercept for this line, and write an equation for it in the form $y = mx + b$.

3. Use this linear model to predict

   a) the first-year average for a student who had an 82 average in grade 12

   b) the grade-12 average for a student with a first-year average of 60

4. a) Use software or the linear regression instruction of a graphing calculator to find the slope and y-intercept for the line of best fit. (Note that most graphing calculators use a instead of m to represent slope.)

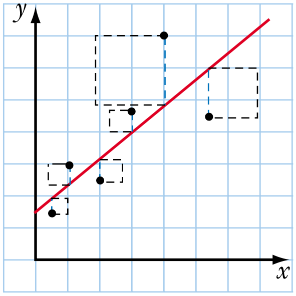
   b) Are this slope and y-intercept close to the ones you measured in question 2? Why or why not?

c) Estimate how much the new values for slope and y-intercept will change your predictions in question 3. Check your estimate by recalculating your predictions using the new values and explain any discrepancies.

5. List the factors that could affect the accuracy of these mathematical models. Which factor do you think is most critical? How could you test how much effect this factor could have?

It is fairly easy to "eyeball" a good estimate of the line of best fit on a scatter plot when the linear correlation is strong. However, an analytic method using a **least-squares fit** gives more accurate results, especially for weak correlations.

Consider the line of best fit in the following scatter plot. A dashed blue line shows the **residual** or vertical deviation of each data point from the line of best fit. The residual is the difference between the values of y at the data point and at the point that lies on the line of best fit and has the same x-coordinate as the data point. Notice that the residuals are positive for points above the line and negative for points below the line. The boxes show the squares of the residuals.



For the line of best fit in the least-squares method,

- the sum of the residuals is zero (the positive and negative residuals cancel out)
- the sum of the squares of the residuals has the least possible value

Although the algebra is daunting, it can be shown that this line has the equation

$$y = ax + b, \text{ where } a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \text{ and } b = \overline{y} - a\overline{x}$$

Recall from Chapter 2 that $\overline{x}$ is the mean of x and $\overline{y}$ is the mean of y. Many statistics texts use an equation with the form $y = a + bx$, so you may sometimes see the equations for a and b reversed.

Example 1 Applying the Least-Squares Formula

This table shows data for the full-time employees of a small company.

a) Use a scatter plot to classify the correlation between age and income.

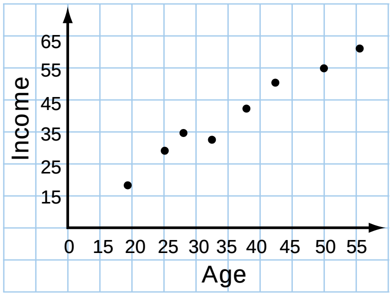
b) Find the equation of the line of best fit analytically.

c) Predict the income for a new employee who is 21 and an employee retiring at age 65.

Age (years)	Annual Income ($000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35

Solution

a) The scatter plot suggests a strong positive linear correlation between age and income level.



b) To determine the equation of the line of best fit, organize the data into a table and compute the sums required for the formula.

Age, x	Income, y	x^2	xy
33	33	1089	1089
25	31	625	775
19	18	361	342
44	52	1936	2288
50	56	2500	2800
54	60	2916	3240
38	44	1444	1672
29	35	841	1015
$\sum x = 292$	$\sum y = 329$	$\sum x^2 = 11\ 712$	$\sum xy = 13\ 221$

Substitute these totals into the formula for a.

$$\begin{aligned} a &= \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\ &= \frac{8(13\ 221) - (292)(329)}{8(11\ 712) - (292)^2} \\ &= \frac{9700}{8432} \\ &\doteq 1.15 \end{aligned}$$

To determine b, you also need the means of x and y.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{292}{8} = 36.5 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{329}{8} = 41.125 \qquad \begin{aligned} b &= \bar{y} - a\bar{x} \\ &= 41.125 - 1.15(36.5) \\ &= -0.85 \end{aligned}$$

Now, substitute the values of a and b into the equation for the line of best fit.

$$\begin{aligned} y &= ax + b \\ &= 1.15x - 0.85 \end{aligned}$$

Therefore, the equation of the line of best fit is $y = 1.15x - 0.85$.

c) Use the equation of the line of best fit as a model.

For a 21-year-old employee,

$$\begin{aligned} y &= ax + b \\ &= 1.15(21) - 0.85 \\ &= 23.3 \end{aligned}$$

For a 65-year-old employee,

$$\begin{aligned} y &= ax + b \\ &= 1.15(65) - 0.85 \\ &= 73.9 \end{aligned}$$

Therefore, you would expect the new employee to have an income of about \$23 300 and the retiring employee to have an income of about \$73 900. Note that the second estimate is an extrapolation beyond the range of the data, so it could be less accurate than the first estimate, which is interpolated between two data points.

Note that the slope a indicates only how y varies with x on the line of best fit. The slope does not tell you anything about the strength of the correlation between the two variables. It is quite possible to have a weak correlation with a large slope or a strong correlation with a small slope.

Example 2 Linear Regression Using Technology

Researchers monitoring the numbers of wolves and rabbits in a wildlife reserve think that the wolf population depends on the rabbit population since wolves prey on rabbits. Over the years, the researchers collected the following data.

Year	1994	1995	1996	1997	1998	1999	2000	2001
Rabbit Population	61	72	78	76	65	54	39	43
Wolf Population	26	33	42	49	37	30	24	19

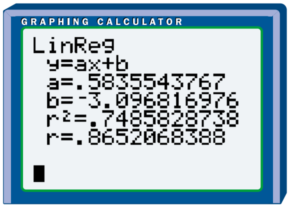
a) Determine the line of best fit and the correlation coefficient for these data.

b) Graph the data and the line of best fit. Do these data support the researchers' theory?

Solution 1 Using a Graphing Calculator

a) You can use the calculator's **linear regression instruction** to find both the line of best fit and the correlation coefficient. Since the theory is that the wolf population depends on the rabbit population, the rabbit population is the independent variable and the wolf population is the dependent variable.

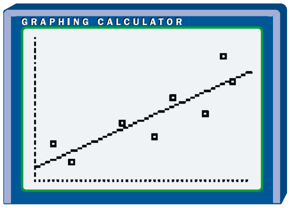
Use the STAT EDIT menu to enter the rabbit data into list L1 and the wolf data into L2. Set **DiagnosticOn**, and then use the STAT CALC menu to select **LinReg(ax+b)**.



```
LinReg
 y=ax+b
 a=.5835543767
 b=⁻3.096816976
 r²=.7485828738
 r=.8652068388
```

The equation of the line of best fit is $y = 0.58x - 3.1$ and the correlation coefficient is 0.87.

b) Store the equation for the line of best fit as a function, Y1. Then, use the STAT PLOT menu to set up the scatter plot. By displaying both Y1 and the scatter plot, you can see how closely the data plots are distributed around the line of best fit.



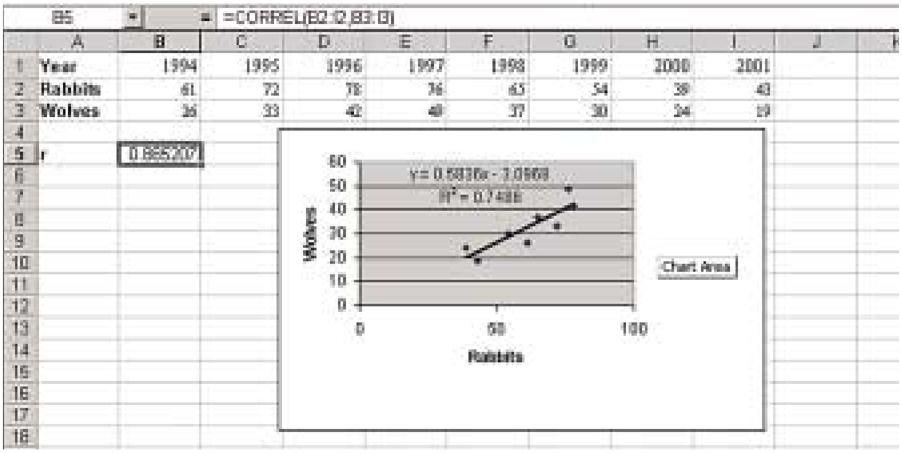
The correlation coefficient and the scatter plot show a strong positive linear correlation between the variables. This correlation supports the researchers' theory, but does not prove that changes in the rabbit population are the cause of the changes in the wolf population.

Solution 2 Using a Spreadsheet

Set up a table with the data for the rabbit and wolf populations. You can calculate the correlation coefficient with the **CORREL function**. Use the **Chart feature** to create a scatter plot.

In Corel® Quattro® Pro, you can find the equation of the **line of best fit** by selecting Tools/Numeric Tools/Regression. Enter the cell ranges for the data, and the program will display regression calculations including the constant (b), the x-coefficient (or slope, a), and r^2.

In Microsoft® Excel, you can find the equation of the **line of best fit** by selecting Chart/Add Trendline. Check that the default setting is Linear. Select the straight line that appears on your chart, then click Format/Selected Trendline/Options. Check the Display equation on chart box. You can also display r^2.



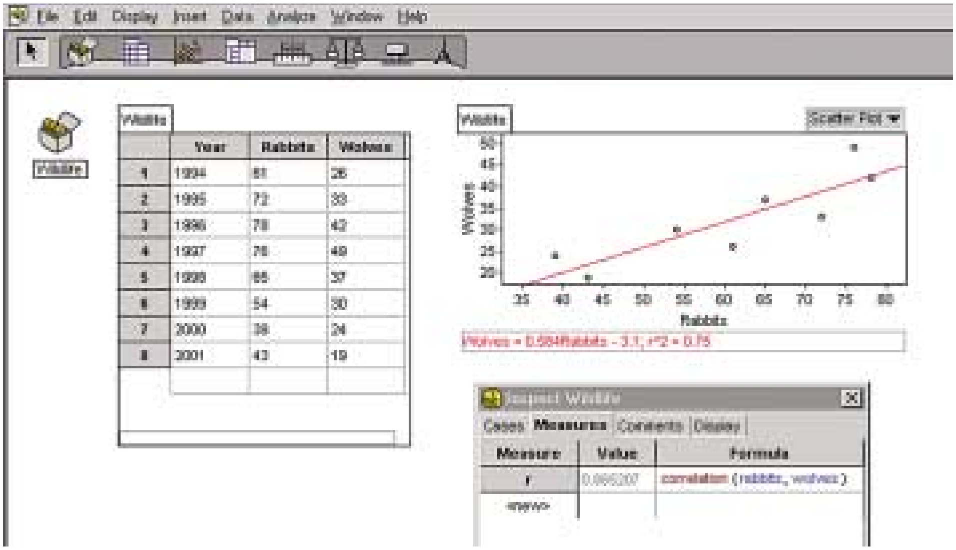
B5 =CORREL(B2:I2,B3:I3)

	A	B	C	D	E	F	G	H	I
1	Year	1994	1995	1996	1997	1998	1999	2000	2001
2	Rabbits	61	72	78	76	65	54	39	43
3	Wolves	26	33	42	48	37	30	24	19
4									
5	r	0.865207							

y = 0.5836x - 3.0969
R^2 = 0.7486

Solution 3 Using Fathom™

Drag a new **case table** to the workspace, create attributes for Year, Rabbits, and Wolves, and enter the data. Drag a new **graph** to the workspace, then drag the Rabbits attribute to the x-axis and the Wolves attribute to the y-axis. From the Graph menu, select Least Squares Line. Fathom™ will display r^2 and the equation for the **line of best fit**. To calculate the correlation coefficient directly, select Inspect Collection, click the Measures tab, then create a new measure by selecting Functions/Statistical/Two Attributes/correlation and entering Rabbits and Wolves as the attributes.



	Year	Rabbits	Wolves
1	1994	61	26
2	1995	72	33
3	1996	78	42
4	1997	76	48
5	1998	65	37
6	1999	54	30
7	2000	39	24
8	2001	43	19

Wolves = 0.584Rabbits - 3.1; r^2 = 0.75

Measure	Value	Formula
r	0.865207	correlation (rabbits, wolves)
<new>		

Project Prep

When analysing two-variable data for your statistics project, you may wish to develop a linear model, particularly if a strong linear correlation is evident.

In Example 2, the sample size is small, so you should be cautious about making generalizations from it. Small samples have a greater chance of not being representative of the whole population. Also, outliers can seriously affect the results of a regression on a small sample.

Example 3 The Effect of Outliers

To evaluate the performance of one of its instructors, a driving school tabulates the number of hours of instruction and the driving-test scores for the instructor's students.

Instructional Hours	10	15	21	6	18	20	12
Student's Score	78	85	96	75	84	45	82

a) What assumption is the management of the driving school making? Is this assumption reasonable?

b) Analyse these data to determine whether they suggest that the instructor is an effective teacher.

c) Comment on any data that seem unusual.

d) Determine the effect of any outliers on your analysis.

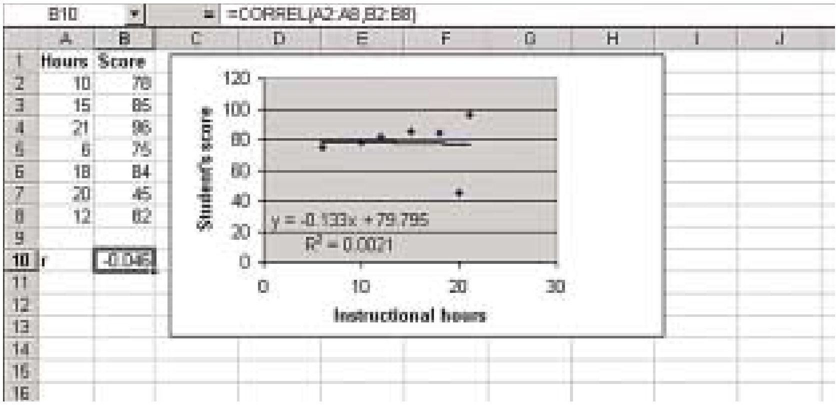
Solution

a) The management of the driving school is assuming that the correlation between instructional hours and test scores is an indication of the instructor's teaching skills. Such a relationship could be difficult to prove definitively. However, the assumption would be reasonable if the driving school has found that some instructors have consistently strong correlations between the time spent with their students and the students' test scores while other instructors have consistently weaker correlations.

b) The number of hours of instruction is the independent variable. You could analyse the data using any of the methods in the two previous examples. For simplicity, a spreadsheet solution is shown here.

Except for an obvious outlier at (20, 45), the scatter plot below indicates a strong positive linear correlation. At first glance, it appears that the number of instructional hours is positively correlated to the students' test scores. However, the linear regression analysis yields a line of best fit with the equation $y = -0.13x + 80$ and a correlation coefficient of -0.05.

These results indicate that there is virtually a zero linear correlation, and the line of best fit even has a negative slope! The outlier has a dramatic impact on the regression results because it is distant from the other data points and the sample size is quite small. Although the scatter plot looked favourable, the regression analysis suggests that the instructor's lessons had no positive effect on the students' test results.

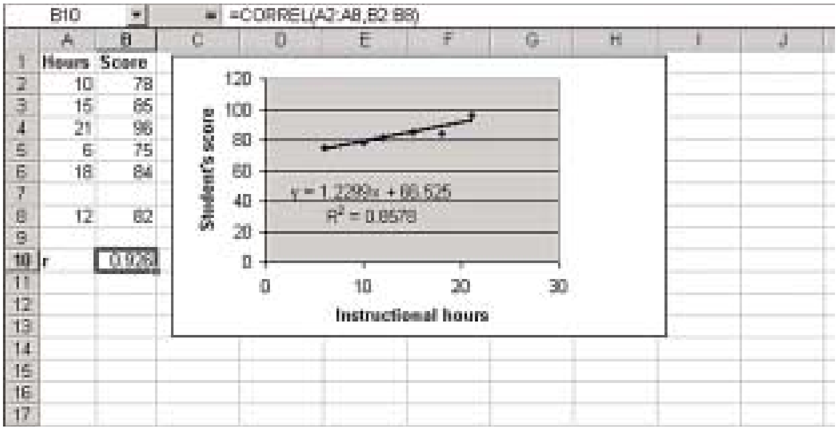


B10 =CORREL(A2:A8,B2:B8)

	A	B
1	Hours	Score
2	10	78
3	15	85
4	21	96
5	6	75
6	18	84
7	20	45
8	12	82
9		
10	r	-0.046

c) The fact that the outlier is substantially below all the other data points suggests that some special circumstance may have caused an abnormal result. For instance, there might have been an illness or emotional upset that affected this one student's performance on the driving test. In that case, it would be reasonable to exclude this data point when evaluating the driving instructor.

d) Remove the outlier from your data table and repeat your analysis.



B10 =CORREL(A2:A8,B2:B8)

	A	B
1	Hours	Score
2	10	78
3	15	85
4	21	96
5	6	75
6	18	84
7		
8	12	82
9		
10	r	0.926

Notice that the line of best fit is now much closer to the data points and has a positive slope. The correlation coefficient, *r*, is 0.93, indicating a strong positive linear correlation between the number of instructional hours and the driver's test scores. This result suggests that the instructor may be an effective teacher after all. It is quite possible that the original analysis was not a fair evaluation. However, to do a proper evaluation, you would need a larger set of data, more information about the outlier, or, ideally, both.

As Example 3 demonstrates, outliers can skew a regression analysis, but they could also simply indicate that the data really do have large variations. A comprehensive analysis of a set of data should look for outliers, examine their possible causes and their effect on the analysis, and discuss whether they should be excluded from the calculations. As you observed in Chapter 2, outliers have less effect on larger samples.

Project Prep

If your statistics project involves a linear relationship that contains outliers, you will need to consider carefully their impact on your results, and how you will deal with them.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site and follow the links to learn more about linear regression. Describe an application of linear regression that interests you.

Key Concepts

- Linear regression provides a means for analytically determining a line of best fit. In the least-squares method, the line of best fit is the line which minimizes the sum of the squares of the residuals while having the sum of the residuals equal zero.
- You can use the equation of the line of best fit to predict the value of one of the two variables given the value of the other variable.
- The correlation coefficient is a measure of how well a regression line fits a set of data.
- Outliers and small sample sizes can reduce the accuracy of a linear model.

Communicate Your Understanding

1. What does the correlation coefficient reveal about the line of best fit generated by a linear regression?

2. Will the correlation coefficient always be negative when the slope of the line of best fit is negative? Explain your reasoning.

3. Describe the problem that outliers present for a regression analysis and outline what you could do to resolve this problem.

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.

Chapter Problem

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

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.

ACHIEVEMENT CHECK

Knowledge / Understanding	Thinking / Inquiry / Problem Solving	Communication	Application

13. The Worldwatch Institute has collected the following data on concentrations of carbon dioxide (CO_2) in the atmosphere.

Year	CO_2 Level (ppm)
1975	331
1976	332
1977	333.7
1978	335.3
1979	336.7
1980	338.5
1981	339.8
1982	341
1983	342.6
1984	344.3
1985	345.7
1986	347
1987	348.8
1988	351.4
1989	352.7
1990	354
1991	355.5
1992	356.2
1993	357
1994	358.8
1995	360.7

a) Use technology to produce a scatter plot of these data and describe any correlation that exists.

b) Use a linear regression to find the line of best fit for the data. Discuss the reliability of this model.

c) Use the regression equation to predict the level of atmospheric CO_2 that you would expect today.

d) Research current CO_2 levels. Are the results close to the predicted level? What factors could have affected the trend?

C

14. Suppose that a set of data has a perfect linear correlation except for two outliers, one above the line of best fit and the other an equal distance below it. The residuals of these two outliers are equal in magnitude, but one is positive and the other negative. Would you agree that a perfect linear correlation exists because the effects of the two residuals cancel out? Support your opinion with mathematical reasoning and a diagram.

15. **Inquiry / Problem Solving** Recall the formulas for the line of best fit using the method of least squares that minimizes the squares of vertical deviations.

a) Modify these formulas to produce a line of best fit that minimizes the squares of *horizontal* deviations.

b) Do you think your modified formulas will produce the same equation as the regular least-squares formula?

c) Use your modified formula to calculate a line of best fit for one of the examples in this section. Does your line have the same equation as the line of best fit in the example? Is your equation the inverse of the equation in the example? Explain why or why not.

16. **a)** Calculate the residuals for all of the data points in Example 3 on page 177. Make a plot of these residuals versus the independent variable, X, and comment on any pattern you see.

b) Explain how you could use such residual plots to detect outliers.