6.3 Modelling Data with a Line of Best Fit

GOAL

Determine the linear function that best fits a set of data, and use the function to solve a problem.

YOU WILL NEED

- graphing technology
- ruler
- metre stick
- graph paper

INVESTIGATE the Math

Nathan wonders whether he can predict the size of a person's hand span based on the person's height. His math class investigated this relationship and recorded measurements from 15 students in the tables below.

Height (cm)	Hand Span (cm)	Height (cm)	Hand Span (cm)
165.0	20.0	182.5	25.0
172.5	21.1	172.5	23.0
172.5	17.6	180.0	20.2
153.8	16.5	177.5	21.1
157.5	17.5	165.0	20.7
170.0	19.0	165.0	16.0
168.8	20.8	175.0	21.2
177.5	22.5		

Can you predict a classmate's hand span based on the classmate's height? Model the relationship between the data by writing a linear function that outputs a person's hand span when you input her or his height.

? What linear function best fits the hand-span and height data for high school students?

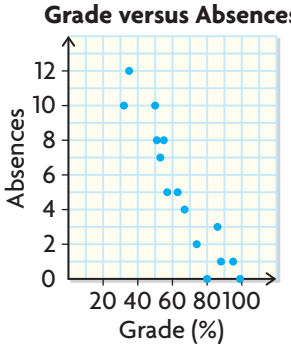
A. Choose the dependent variable, and describe the relationship in the data. What is a reasonable domain and range for this relationship? Explain.

B. Plot the data. Use a ruler to draw a line that approximates the trend in your scatter plot.

C. Use two points on your line to determine its equation. How does your equation compare with your classmates' equations?

EXPLORE...

- The **scatter plot** below compares students' absences from math class with the grade they obtained in the course.



Describe the characteristics of a polynomial function that might be used to model the data in the scatter plot.

Communication *Tip*

The independent variable is the variable that is being manipulated. The dependent variable is the variable that is being observed. The independent variable is always placed on the horizontal axis of a graph.

line of best fit

A straight line that best approximates the trend in a scatter plot.

regression function

A line or curve of best fit, developed through a statistical analysis of data.

D. Use technology to determine the **line of best fit** for the data. You determined an equation for a line that approximates the trend in the data in part C. How does your equation of the linear **regression function** compare with your equation from part C? How does your new equation compare with your classmates' new equations?

E. Collect 15 more data points from your classmates. Determine the equation of a new linear regression function for the combined data.

F. Use all three linear equations you determined to estimate a person's hand span after measuring her or his height. How do your estimates compare?

Reflecting

G. In part B, you drew a line that approximates the trend in the hand-span data. Explain the reasoning you used to draw the line.

H. You have three mathematical models for this relationship: a scatter plot, a line of best fit, and the equation of the line of best fit. Which mathematical model do you think provides a more reliable estimate of a person's hand span for their height? Explain.

I. You determined three different equations for the line of best fit. Which equation do you think is the best to use for estimating someone's hand span? Explain why.

APPLY the Math

EXAMPLE 1 | Using technology to determine a linear model for continuous data

The one-hour record is the farthest distance travelled by bicycle in 1 h. The table below shows the world-record distances and the dates they were accomplished.

Year	1996	1998	1999	2002	2003	2004	2007	2008	2009
Distance (km)	78.04	79.14	81.16	82.60	83.72	84.22	86.77	87.12	90.60

International Human Powered Vehicle Association

a) Use technology to create a scatter plot and to determine the equation of the line of best fit.

b) **Interpolate** a possible world-record distance for the year 2006, to the nearest hundredth of a kilometre.

c) Compare your estimate with the actual world-record distance of 85.99 km in 2006.

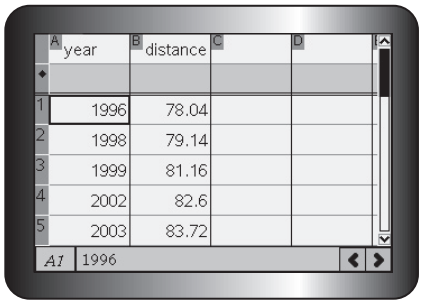
British Columbian Georgi Georgiev designed and built a human-powered bicycle, the Varna Tempest, which Canadian Sam Whittingham used to break the one-hour record.

interpolation

The process used to estimate a value within the domain of a set of data, based on a trend.

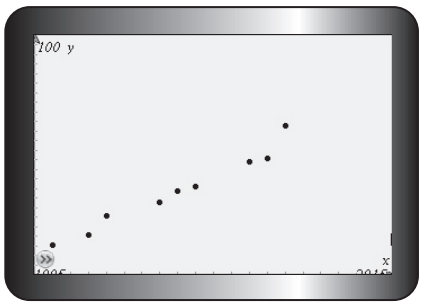
Carmen's Solution: Using a graphing calculator

a) I entered the data into my graphing calculator.



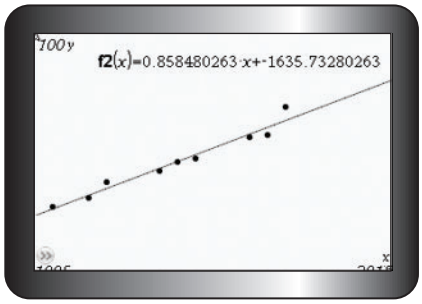
	A year	B distance	C	D
1	1996	78.04		
2	1998	79.14		
3	1999	81.16		
4	2002	82.6		
5	2003	83.72		

A1 1996



I reasoned that year should be the independent variable, since there is a world-record distance for each year. I set my calculator to graph the years on the *x*-axis and the distances on the *y*-axis.

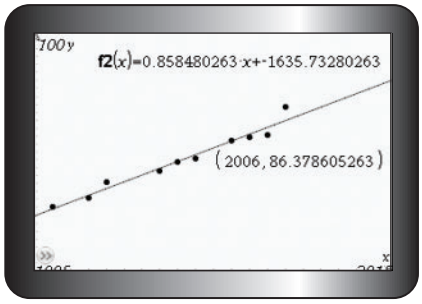
I changed my settings to display the *x*-axis from 1995 to 2015 and the *y*-axis from 75 to 100.



Based on the scatter plot, the data appears to be linear. I determined the equation of the line of best fit for the data using linear regression. Linear regression gave me the equation of a line that balances points on either side of the line.

The line of best fit for this data is
$y = 0.858...x - 1635.732...$
where y represents the distance in kilometres and x represents the year.

b)



I traced to the year 2006 on the graph on my calculator. I obtained 86.378... km from the graph.

The value of 86.38 km is a reasonable estimate, based on the other world-record distances.

The world record in 2006 may be about 86.38 km.

c) $86.38 - 85.99 = 0.39$
My estimate is 0.39 km greater than the actual world-record distance for 2006.

I subtracted the world-record distance from my estimate.

Sandra's Solution: Using a spreadsheet

a) I used a spreadsheet to determine the equation of the linear regression function.

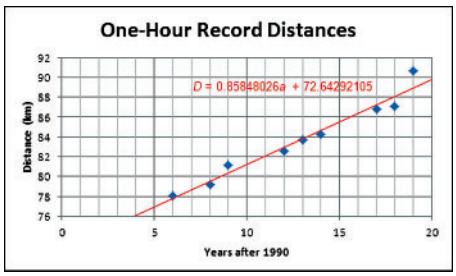
	A	B
1	Years after 1990	Distance (km)
2	6	78.04
3	8	79.14
4	9	81.16
5	12	82.60
6	13	83.72
7	14	84.22
8	17	86.77
9	18	87.12
10	19	90.60

I input the data into the cells. I decided to enter the year data as "Years after 1990" because these values were easier to enter into the spreadsheet.

The distance records change depending on the year in which they were broken, not on the distance achieved, so time is the independent variable.

I created a scatter plot with time along the horizontal axis.

I added the line of best fit to my scatter plot.



I chose a scale that would fit all of my data points.

I noticed that the trend in the scatter plot is somewhat linear.

I chose linear regression in the spreadsheet program to draw the line of best fit and determine the equation of the linear regression function.

The equation that represents the trend is

$D = 0.858...a + 72.642...$

where D represents the distance and a represents the year.

b) $D = 0.858...a + 72.642...$
$D = 0.858...(16) + 72.642...$
$D = 86.378...$
The world record in 2006 may be about 86.38 km.

I used the equation to interpolate a distance for the year 2006. I substituted 16 for *a* because 2006 is 16 years after 1990.

c) $86.38 - 85.99 = 0.39$
My estimate is 0.39 km greater than the actual world-record distance in 2006.

I subtracted the world-record distance from my estimate.

Your Turn

If there had been a world-record distance record in the year 2000, what would you expect this distance to have been?

EXAMPLE 2 Using linear regression to solve a problem that involves discrete data

Matt buys T-shirts for a company that prints art on T-shirts and then resells them. When buying the T-shirts, the price Matt must pay is related to the size of the order. Five of Matt's past orders are listed in the table below.

Number of Shirts	Cost per Shirt ($)
500	3.25
700	1.95
200	5.20
460	3.51
740	1.69

Matt has misplaced the information from his supplier about price discounts on bulk orders. He would like to get the price per shirt below $1.50 on his next order.

a) Use technology to create a scatter plot and determine an equation for the linear regression function that models the data.

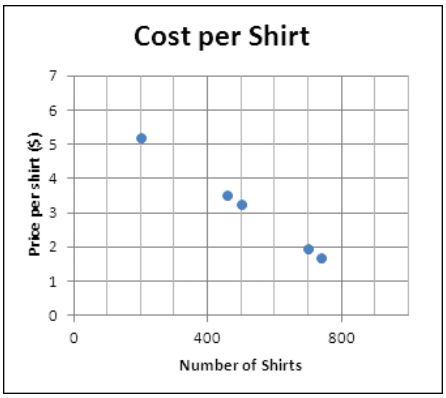
b) What do the slope and *y*-intercept of the equation of the linear regression function represent in this context?

c) Use the linear regression function to **extrapolate** the size of order necessary to achieve the price of $1.50 per shirt.

extrapolation
The process used to estimate a value outside the domain of a set of data, based on a trend.

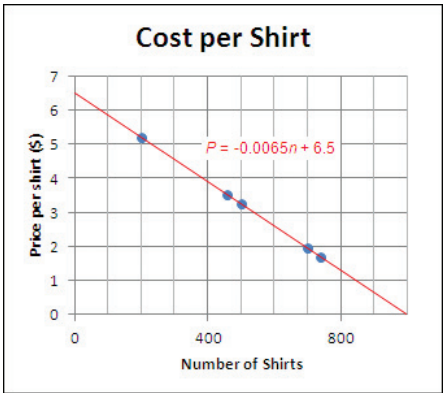
Matt's Solution

a)



I entered the data table into a spreadsheet and used the spreadsheet to create a scatter plot.

Since the price per shirt depends upon the number of shirts ordered, price is the dependent variable. It goes along the vertical axis.



I graphed the line of best fit and obtained the equation of the linear regression function.

Let P represent the price per shirt, and let n represent the number of shirts ordered:

$P = -0.0065n + 6.5$

b) The slope is -0.0065. It represents a drop in price of \$0.0065 per additional shirt ordered.

The y-intercept is 6.5.

The slope represents the rise over the run. On this graph, the rise is the price per shirt and the run is the number of shirts. Therefore, the negative slope means that the price drops for every additional shirt ordered.

The y-intercept is the point on the line where the number of shirts ordered is zero.

c)

$$P = -0.0065n + 6.5$$
$$1.50 = -0.0065n + 6.5$$
$$-5.00 = -0.0065n$$
$$\frac{-5.00}{-0.0065} = n$$
$$769.230... = n$$

I need to order 770 shirts to get a price less than \$1.50.

I substituted \$1.50 for P in the equation to extrapolate the number of shirts needed for the order.

I cannot order a fraction of a shirt, so I must round up.

Your Turn

Create a problem about Matt ordering T-shirts that you could solve using interpolation. Solve your problem.

In Summary

Key Ideas

- A scatter plot is useful when looking for trends in a given set of data.
- If the points on a scatter plot seem to follow a linear trend, then there may be a linear relationship between the independent variable and the dependent variable.

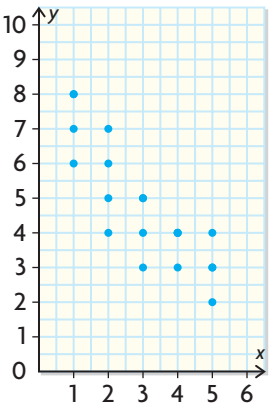
Need to Know

- If the points on a scatter plot follow a linear trend, technology can be used to determine and graph the equation of the line of best fit.
- Technology uses linear regression to determine the line of best fit. Linear regression results in an equation that balances the points in the scatter plot on both sides of the line.
- A line of best fit can be used to predict values that are not recorded or plotted. Predictions can be made by reading values from the line of best fit on a scatter plot or by using the equation of the line of best fit.

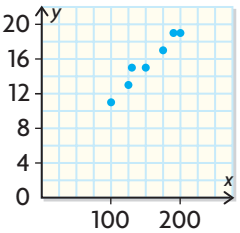
CHECK Your Understanding

1. Use a clear ruler to help you estimate the slope and y-intercept for a line that best approximates the data in each scatter plot below.

a)



b)



2. Determine the equation of each line in question 1.

3. Determine the independent and dependent variables for each relationship. Justify your reasoning.

a) The distance travelled in a car is related to the average speed of the car.

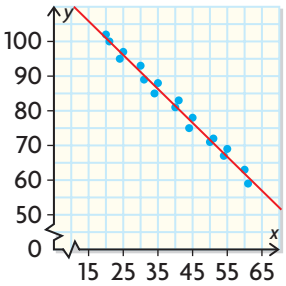
b) The size of a family is related to the number of cellphones in the family.

c) The number of people in a cafeteria is related to the time of day.

d) The number of hours of daylight is related to the time of year.

Yellowknife, Northwest Territories, enjoys the sunniest summers in Canada. There are approximately 1037 h of sunshine during June, July, and August.

PRACTISING



4. A line of best fit has been drawn for the scatter plot at the left.

a) Describe the characteristics of the line of best fit.

b) Use the line of best fit to estimate the value of y when x is 47. Is this interpolation or extrapolation? Explain.

c) Use the line of best fit to estimate the value of x when y is 70. Is this interpolation or extrapolation? Explain.

d) Use the line of best fit to estimate the value of y when x is 15. Is this interpolation or extrapolation? Explain.

5. The world-record times for women's 3000 m speed skating, from 1981 to 2006, are given in the table below.

Skater	Time (min)	Date
Gabi Schönbrunn	4:21.70	March 28, 1981
Andrea Schöne	4:20.91	March 23, 1984
Karin Kania	4:18.02	March 21, 1986
Yvonne van Gennip	4:16.85	March 19, 1987
Gabi Zange	4:16.76	December 5, 1987
Yvonne van Gennip	4:11.94	February 23, 1988
Gunda Kleemann	4:10.80	December 9, 1990
Gunda Niemann	4:09.32	March 25, 1994
Gunda Niemann-Stirnemann	4:07.80	December 7, 1997
Claudia Pechstein	4:07.13	December 13, 1997
Gunda Niemann-Stirnemann	4:05.08	March 14, 1998
Gunda Niemann-Stirnemann	4:01.67	March 27, 1998
Gunda Niemann-Stirnemann	4:00.51	January 30, 2000
Gunda Niemann-Stirnemann	4:00.26	February 17, 2001
Claudia Pechstein	3:59.27	March 2, 2001
Claudia Pechstein	3:57.70	February 10, 2002
Cindy Klassen	3:53.34	March 18, 2006

International Skating Union

Cindy Klassen at the Vancouver Winter Olympics, 2010

a) Create a scatter plot to compare the world-record time with the year in which it was set.

b) Describe the characteristics of the trend between the variables.

c) Determine the equation of the linear regression function that models the data. What do the slope and y-intercept of the equation of the linear regression function represent in this context?

d) Cindy Klassen earned her first world record in 2005, but her time is missing from the table. Interpolate her world-record time in 2005.

e) Research Cindy's actual world-record time in 2005. How close was your estimate?

6. The world-record time for the men's 100 m sprint was 10.00 s in 1960. The table below shows the world-record times since 1960.

Years after 1960	0	8	23	31	36	39	45	48	49
Time (s)	10.00	9.95	9.93	9.86	9.84	9.79	9.77	9.72	9.58

a) Create a scatter plot to display the data.
b) Describe the characteristics of the trend in the data.
c) Determine the equation of the linear regression function that models the data. What do the slope and y-intercept of the equation represent in this context?
d) Interpolate a possible world-record time for 2007.
e) Asafa Powell, from Jamaica, accomplished a world-record time on September 9, 2007. Research his time, and determine the difference between this actual time and your estimate.

Asafa Powell during his record-breaking run at the International Association of Athletics Federation's Grand Prix in Rieti, Italy

7. Average daily temperatures for the month of July, from 14 weather stations in British Columbia, are listed in the table below.

Weather Station	Latitude of Weather Station (°)	Mean Temperature for July (°C)
Keremeos	49.2	20.9
Dease Lake	58.4	12.8
Hixon	53.5	16.5
Heffley Creek	50.9	16.8
Lake Cowichan	48.8	17.5
McBride	53.4	15.1
North Vancouver	49.3	16.8
Nakusp	50.3	18.3
Prince George	54.1	15.5
Vanderhoof	54.0	16.3
Likely	52.6	15.4
Bullmoose	55.1	13.4
Fort St. John	56.2	15.7
Todagin Ranch	57.6	11.6

Environment Canada

a) What relationship do you expect to see between the latitudes of the weather stations and the mean temperatures? Plot the data to verify your prediction.
b) Determine the equation of the linear regression function that models the data.
c) According to the linear regression function, what average temperature would you expect at a latitude of 52.0° N?
d) At what latitude would you expect a mean temperature of 18 °C for July?

8. Call-Me Cellular is rolling out a new marketing plan that is aimed at families. Before finalizing the plan structure, the marketing department needs to know how the number of cellphones in a family is related to the size of the family. The marketing department conducted a survey and collected the following data:

Number of Cellphones	2	2	3	4	1	0	2	2	1	2	1	4	1	3	4	1	0	1
Family Size	5	4	4	5	3	2	4	4	2	2	1	6	4	5	7	2	2	1

a) Describe the relationship between the variables in the data.
b) Create a scatter plot, and determine the equation of a linear regression function that models the data.
c) The average family in Canada has three people. How many cellphones would you expect to find in an average Canadian family? Explain.

9. Devin is on a budget. He is trying to decide how many graduation events he can afford to attend this year. He interviewed 15 of his older brother's friends to see how much money they spent, compared to the number of events they attended. The data he collected is recorded in the table below.

Number of Events	1	3	4	2	1	3	3	5	4	3	2	4	3	2	2
Money Spent ($)	200	1300	1500	400	150	1100	900	1500	1450	1100	100	1100	800	300	600

Devin estimates that he has about $750 to spend on graduation events. Use linear regression to estimate how many events Devin should attend.

10. A large bicycle retailer is planning to open another store. The location that is being considered has 2000 sq ft of floor space. The retailer needs to know the number of bikes that would be needed to stock a store of this size. Use the data from the retailer's other locations to estimate the number of bikes that would be needed.

Number of Bikes	62	58	204	50	190	75	60	60
Floor Space (sq ft)	1500	1250	3200	1200	3000	1600	1550	1300

11. According to Statistics Canada, the life expectancy for Canadians has been increasing over the past few decades.

	Life Expectancy (years)	
Years	**Male**	**Female**
1920 to 1922	59	61
1930 to 1932	60	62
1940 to 1942	63	66
1950 to 1952	66	71
1960 to 1962	68	74
1970 to 1972	69	76
1980 to 1982	72	79
1990 to 1992	75	81
2000 to 2002	77	82

a) Create two scatter plots: one for the male data and one for the female data.

b) Determine the equation of a linear regression function that models each set of data.

c) Use your equations to estimate the life expectancy of males and females in 2010.

12. Mountain Madness Adrenaline Adventures runs mountain-bike tours in Canmore, Alberta. The manager noticed that the number of tours run in a season is related to the value of the Canadian dollar. The following table shows historical data for the average value of the Canadian dollar, as well as the number of tours run in each season.

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010
Dollar Value ($US)	0.65	0.72	0.76	0.82	0.89	0.95	0.99	0.89	0.96
Number of Tours	18	14	14	13	12	10	8	12	9

Economists predicted that the value of the Canadian dollar would be 1.00 during the summer of 2011. How many tours should Mountain Madness Adrenaline Adventures have expected to run during the summer of 2011? Support your answer with visuals and an analysis of the data.

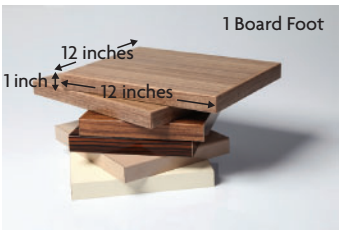
Closing

13. Suppose that a set of data follows a linear trend.

a) What methods could you use to estimate the value of the dependent variable for the data? Give an example.

b) How could you determine the value of the independent variable if you know the value of the dependent variable?

Extending



A board foot is a measure of volume. One board foot is equal to 1 ft by 1 ft by 1 in. of lumber.

14. Martha has been contracted to build several wood cabinets, with proportional door areas, over the summer. She has made a list of the number of board feet of oak that she will need to build each cabinet, based on the width of the cabinet.

Width of Cabinet (in.)	Board Feet Required (bd ft)
12	8.4
15	13.1
20	23.3
25	36.5
30	52.5

a) Determine the equation for the linear regression function that models this situation.

b) Use your regression equation to determine the number of board feet that Martha will need to build a cabinet with a width of 5 in. and a similar area. Explain why extrapolating does not always make sense.

c) Does the data appear linear? Explain.

d) What might be a better regression model to use in this situation? Explain why.