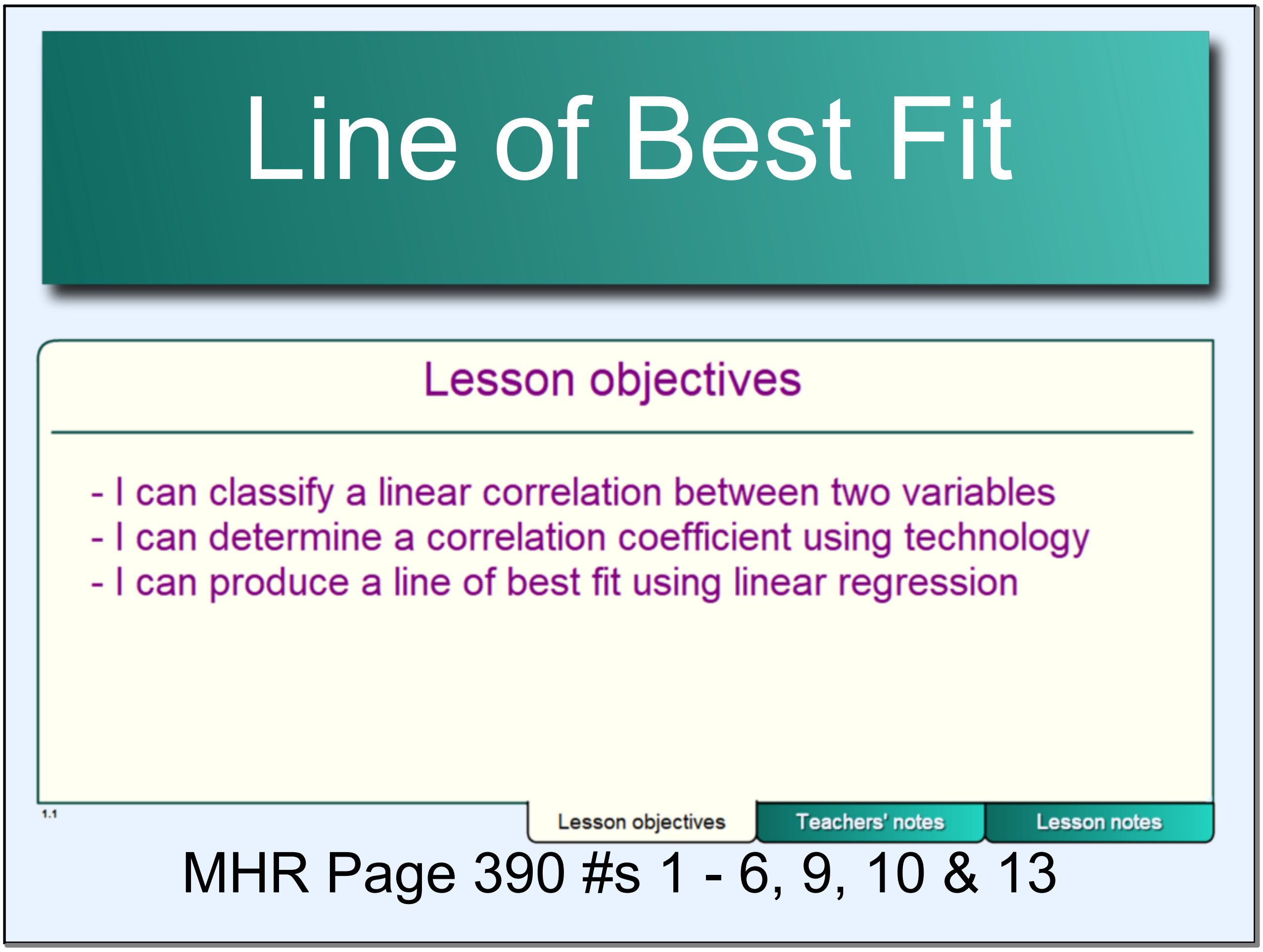


Line of Best Fit

Lesson objectives

- I can classify a linear correlation between two variables
- I can determine a correlation coefficient using technology
- I can produce a line of best fit using linear regression

1.1

Lesson objectives | Teachers' notes | Lesson notes

MHR Page 390 #s 1 - 6, 9, 10 & 13

Warm up

Some parents like to track their children's growth over time. One way to do this is to measure the child's height on every birthday and record it on a wall or door frame.

- What are some other ways to record these data?
- Do you think a child grows by the same amount every year?

Definitions

Line of Best Fit

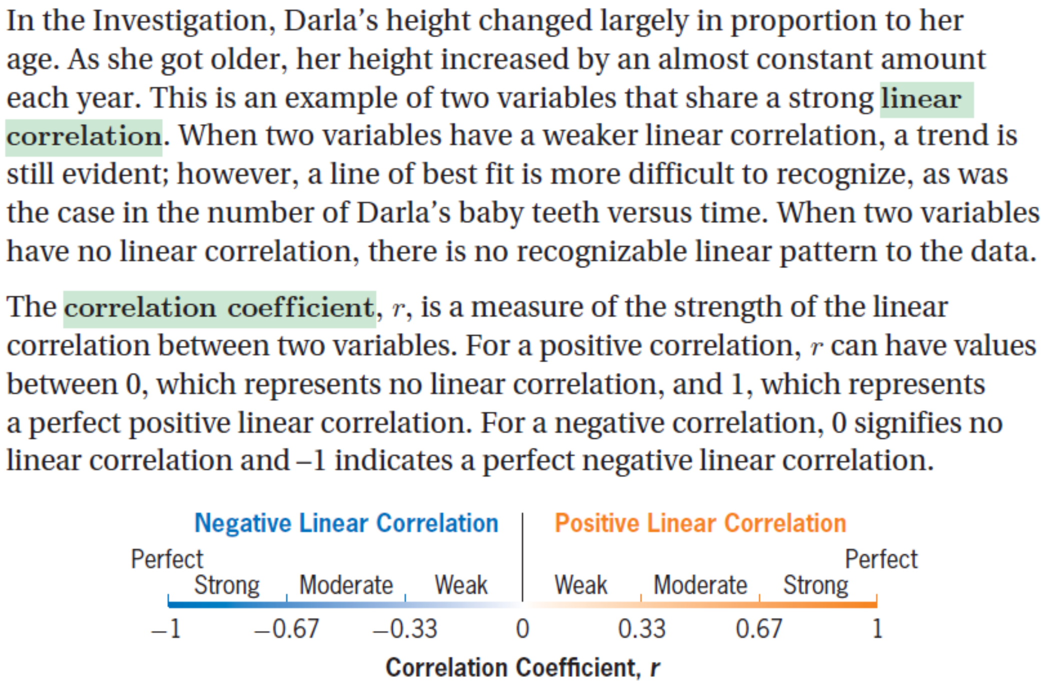
- A straight line that represents a trend in the scatter plot as long as the pattern is more or less linear
- Should pass through as many points as possible, with about half the points above and below the line
- A solid line represents continuous data that are constantly changing
- A dashed line represents discrete data that change only in steps

Linear Correlation

- A relationship in which a change in one variable tends to correspond to a proportional change in another variable

Correlation Coefficient

- A measure of how well a linear model fits a two-variable set of data
- Values of r between -1 and 0 indicate a negative correlation, so the line of best fit has a negative slope
- An r value of 0 indicates that there is no linear correlation
- Values of r between 0 and 1 indicate a positive correlation, so the line of best fit has a positive slope



In the Investigation, Darla's height changed largely in proportion to her age. As she got older, her height increased by an almost constant amount each year. This is an example of two variables that share a strong **linear correlation**. When two variables have a weaker linear correlation, a trend is still evident; however, a line of best fit is more difficult to recognize, as was the case in the number of Darla's baby teeth versus time. When two variables have no linear correlation, there is no recognizable linear pattern to the data.

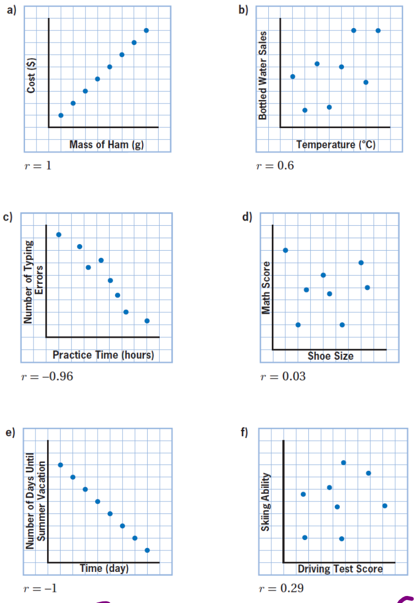
The **correlation coefficient**, r, is a measure of the strength of the linear correlation between two variables. For a positive correlation, r can have values between 0, which represents no linear correlation, and 1, which represents a perfect positive linear correlation. For a negative correlation, 0 signifies no linear correlation and −1 indicates a perfect negative linear correlation.

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

Example 1

Strength of Correlation

Do the graphs below show a linear correlation? Describe the correlation for each relationship.



a) Perfect linear correlation $(r = 1)$. As mass of ham increases the cost also increases by a proportional amount.

b) As temperature increases sales of bottled water also increase. Not an obvious correlation $(r = 0.6)$.

c) More time practising results in fewer typing errors. There is a strong negative linear correlation $(r = -0.96)$.

d) Shoe size and math score seem unrelated. No obvious pattern to the data and $r = 0.03$

e) There is a perfect negative correlation $(r = -1)$. Each day that goes by means one less day to wait until summer vacation.

f) There would appear to be a slight increase in skiing ability as driving test scores increase. It is a weak positive linear correlation $(r = 0.29)$.

Your Turn

Sketch a scatter plot relating two variables that have:

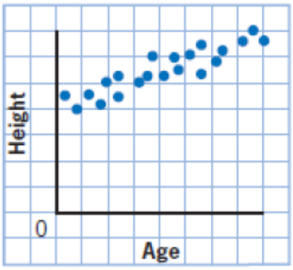
a) a strong positive correlation

b) a moderate weak correlation

c) no correlation

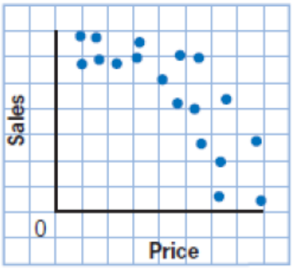
Indicate variables that could be correlated in this way for each case.

a)



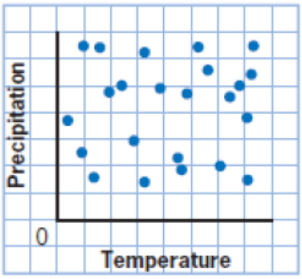
As age increases, height increases.

b)

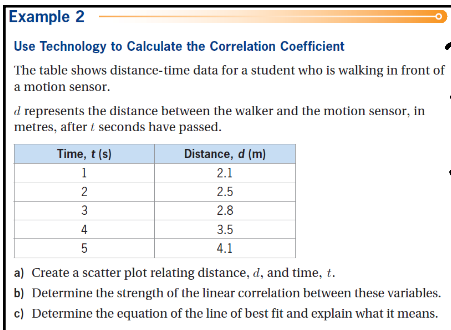


As prices increase, number of sales decrease

c)



The amount of precipitation is unaffected by temperature



Example 2

Use Technology to Calculate the Correlation Coefficient

The table shows distance-time data for a student who is walking in front of a motion sensor.

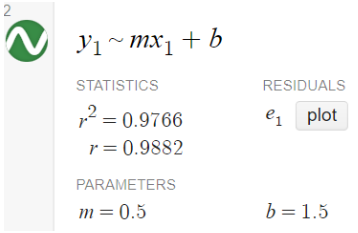
d represents the distance between the walker and the motion sensor, in metres, after t seconds have passed.

Time, t (s)	Distance, d (m)
1	2.1
2	2.5
3	2.8
4	3.5
5	4.1

a) Create a scatter plot relating distance, d, and time, t.

b) Determine the strength of the linear correlation between these variables.

c) Determine the equation of the line of best fit and explain what it means.



2

$y_1 \sim mx_1 + b$

STATISTICS

$r^2 = 0.9766$

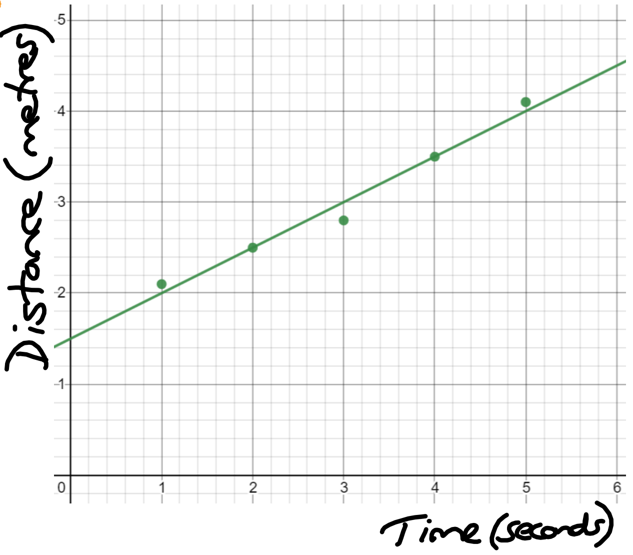
$r = 0.9882$

RESIDUALS

e_1 plot

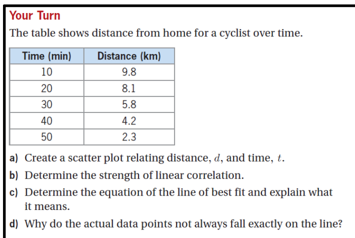
PARAMETERS

$m = 0.5$ $b = 1.5$



b) Strength of correlation ⇒ $r = 0.9882$
very strong.

c) Line of best fit "$y = mx + b$"
is $d = 0.5t + 1.5$ Distance between the walker and the sensor is increasing by 0.5 m/s. The walker started 1.5m from the sensor.



Your Turn

The table shows distance from home for a cyclist over time.

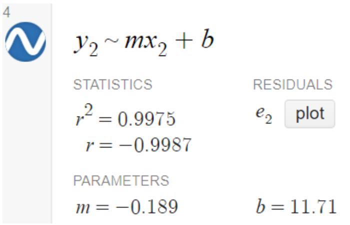
Time (min)	Distance (km)
10	9.8
20	8.1
30	5.8
40	4.2
50	2.3

a) Create a scatter plot relating distance, d, and time, t.

b) Determine the strength of linear correlation.

c) Determine the equation of the line of best fit and explain what it means.

d) Why do the actual data points not always fall exactly on the line?



4

$y_2 \sim mx_2 + b$

STATISTICS

$r^2 = 0.9975$

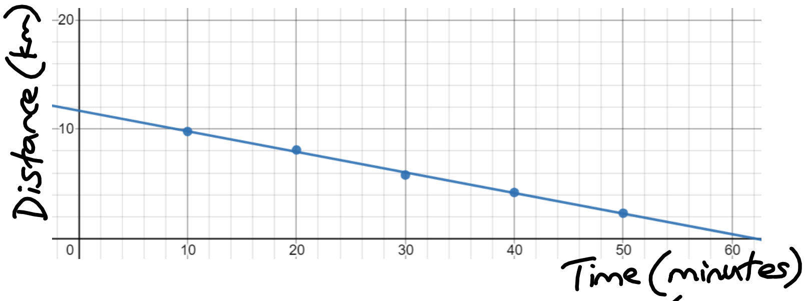
$r = -0.9987$

RESIDUALS

e_2 plot

PARAMETERS

$m = -0.189$ $b = 11.71$



b) Very strong negative correlation ($r = -0.9987$)

c) Line of best fit "$y = mx + b$"
$d = -0.189t + 11.71$
The distance from home is decreasing at a rate of 0.189 km/m. The cyclist started 11.71 km from home.

d) Data points are not exactly on the line because the cyclist's speed is unlikely to be constant. "m" represents the average speed.

Key Concepts

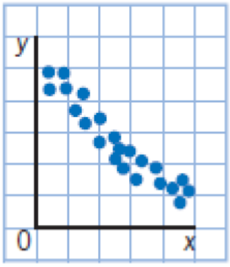
- When a change in one variable is accompanied by a proportional change in another variable, the variables share a linear correlation.
- The correlation coefficient, r, is a measure of the strength of linear correlation between two variables. The value of r, which can be between –1 and 1, gives an indication of how closely the data points relate to the line of best fit.
- The line of best fit can be used to model a linear correlation.
- Linear regression is the mathematical process that determines the line of best fit.

R1. Two variables, X and Y, share a strong negative correlation.

a) Sketch what a scatter plot of Y versus X could look like.

b) Describe in words the correlation between X and Y.

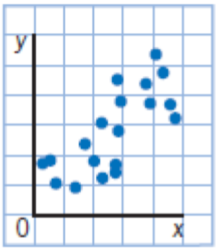
a)



b) The data is not very spread for a strong, negative correlation. The correlation coefficient, r, is somewhere between -0.67 and -1.

R2. Repeat R1, assuming that X and Y have a moderate positive correlation.

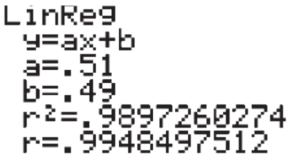
a)



b) The data is somewhat spread for a moderate, positive correlation. The correlation coefficient, r, is somewhere between 0.33 and 0.67.

R3. A student walking in front of a motion sensor generates distance-time data, where distance is in metres and time is in seconds. A linear regression on the data produces the information shown.

Describe everything you can about the motion of this walker and the relationship between distance and time.



```
LinReg
 y=ax+b
 a=.51
 b=.49
 r²=.9897260274
 r=.9948497512
```

The student started 0.49 m (b) from the sensor. They walked away from the sensor at a rate (speed) of 0.51 m/s (a). The data collected shows a very strong (almost perfect) positive correlation (r = 0.995).