3.3 Line of Best Fit

Learning Goals: I am learning to…
- ☐ Analyze two-variable data using a line of best fit
- ☐ Determine the line of best fit for a given set of data and use interpolation and extrapolation to analyze the data

What is a line of best fit?

A line of best fit (LOBF) is used when comparing two-variables in a scatter plot. It best represents a linear relationship and is a straight line.

A line of best fit will vary depending on the scatter plot. In general, it can be said that the weaker the correlation, the harder it is to make a line of best fit, since the relationship is loosely linear.

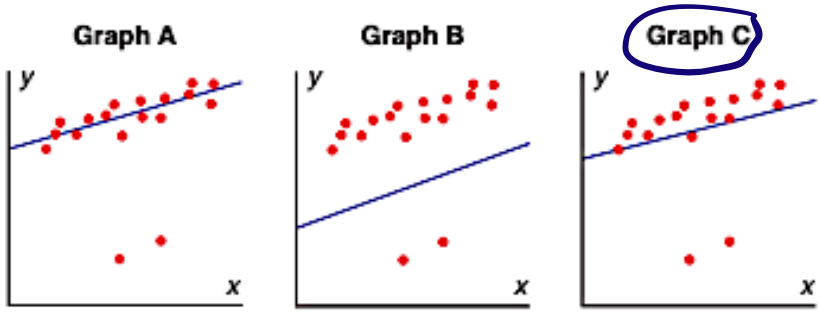
Outliers

An outlier on a LOBF represents a point that is far away from the other data points.

Outliers can occur for a number of reasons, including:
- Inaccurate measurements
- Anomalies in the data (i.e. You are surveying people about their height and you survey the world's tallest man)

The LOBF you make should reflect all points in your data set, including any outliers. The more outliers you have in a data set though, the greater the impact on your LOBF.

Example 1: Given the three scatter plots below, determine which has the best LOBF and explain why.



Graph A: Not good. The LOBF ignores the outliers.

Graph B: Not good. The LOBF is too close to the outliers.

Graph C: Best option. Considers the outliers, but also takes into account the main cluster of data.

MAP4C1 Unit 3: Two-Variable Data

Interpolating and Extrapolating

Once a line of best fit has been made, it can be used to **interpolate** or **extrapolate** values.

- **Estimating** values <u>within</u> the set of data is known as interpolating
- **Predicting** values <u>outside</u> the set of data is known as extrapolating

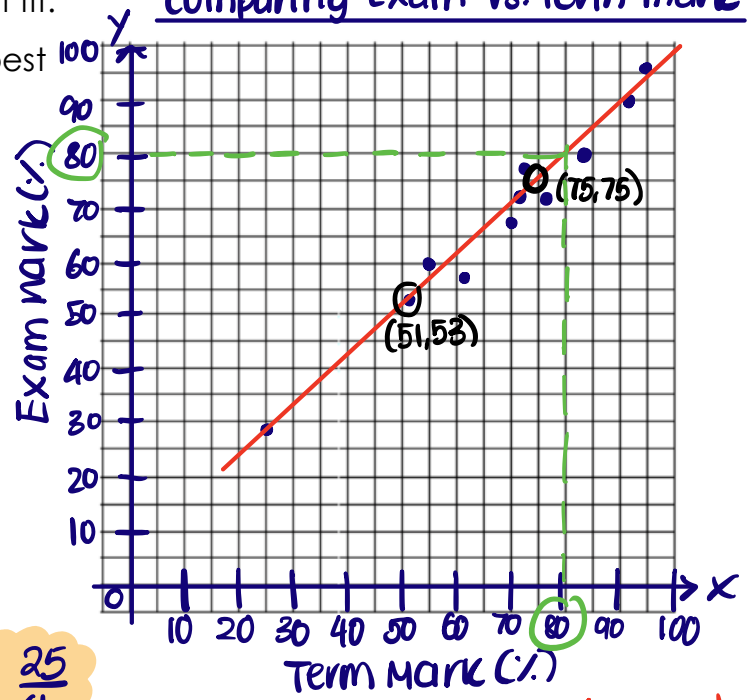
Example 2: The table below summarises a Grade 12 math class term marks with the final exam marks.

Term Mark (%) (x)	84	76	70	95	92	61	25	55	51	73	71
Exam Mark (%) (y)	80	72	68	96	90	58	29	60	53	77	73

(84, 80)

a) Graph the data and draw a line of best fit.

Comparing Exam vs. term mark



b) Determine an equation for the line of best fit.

$y = mx + b$ (m: slope, b: y-intercept)

① slope: $m = \frac{y_2 - y_1}{x_2 - x_1}$ (51,53) (75,75) $x_1\ y_1\ x_2\ y_2$

$m = \frac{75-53}{75-51}$

$m = \frac{22}{24}$

$m = \frac{11}{12}$

② find b: $y = \frac{11}{12}x + b$

sub (75, 75) as x & y

$75 = \frac{11}{12}(75) + b$

$75 = \frac{825}{12} + b$

$75 - \frac{825}{12} = b$

$\frac{25}{4} = b$

$y = \frac{11}{12}x + \frac{25}{4}$

*Equations of LOBF will vary!

c) Use the **graph** to approximate the exam mark for a student with a term mark of 80%

Interpolating → within the data.

Approx. 80% exam mark.

d) Use the **equation** to predict the exam mark for a student with a term mark of 10%

Extrapolation → outside the data.

$y = \frac{11}{12}x + \frac{25}{4}$ $x = 10$

$= \frac{11}{12}(10) + \frac{25}{4}$

$= \frac{55}{6} + \frac{25}{4}$

$= \frac{185}{12}$

$= 15.4\%$

∴ A term mark of 10% gives an exam mark of 15.4%.

MAP4C1 Unit 3: Two-Variable Data

Is a correlation linear?

Sometimes the relationship between two variables may not appear to be linear. We will look at these models later in the course.

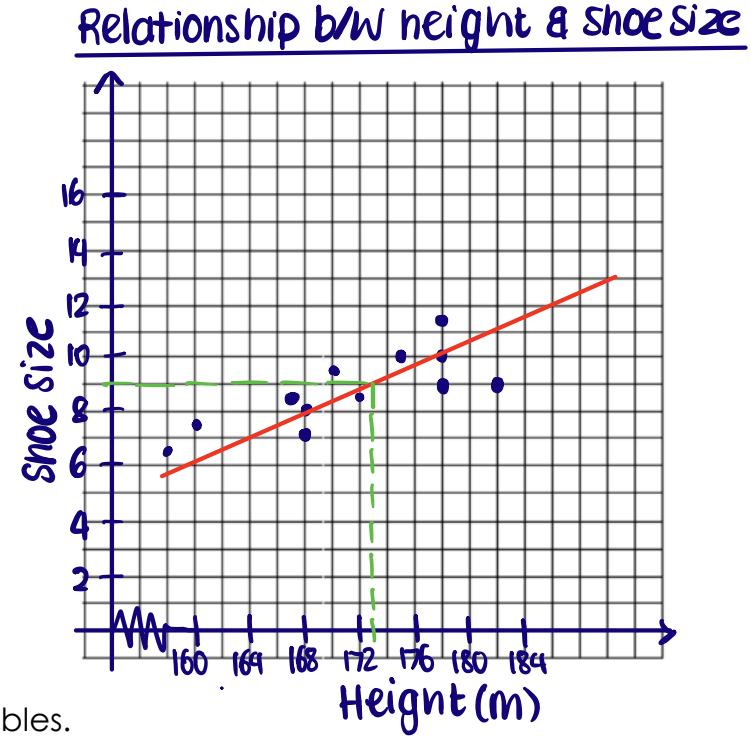
Linear models are not always the most reliable in the following situations:

- There is not enough data (too small sample size)
- The data is clustered together
- There does not appear to be any correlation
- There are multiple outliers
- The general shape of the data does not appear to be linear.

If a relationship between two variables does not appear to be linear, DO NOT force it to be. Simply state that the overall relationship is non-linear.

Practice! Graph the given set of data.

(x) Height (cm)	(y) Shoe Size
182	9
178	9
167	8.5
168	7
175	10
178	10
168	8
178	11.5
172	8.5
157	6.5
160	7.5
170	9.5



1. Describe the correlation between the variables.

   This is a moderate/weak positive correlation.

2. Draw a line of best fit.

3. What would you expect the shoe size of a person who is 173 cm tall?

   Approx size 9 (answers will vary depending on your LOBF)

4. How could you make the results more reliable?

   - Have a bigger sample size.
   - Use more specific samples
     - e.g. age, gender,
   - Use measurements rather than shoe size
     - varies store to store.