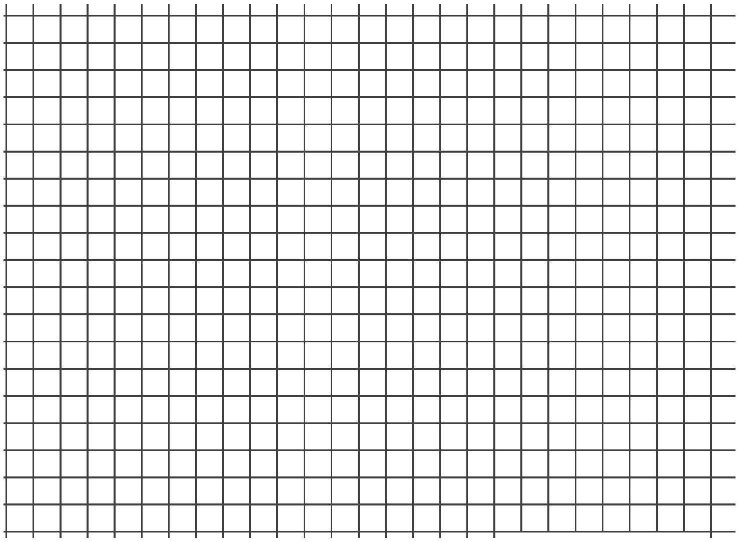
(3.3) Line of Best Fit

I am looking to buy a house that is detached, 2-stories, $300 000 or less, and approximately 2000 sq. ft. of living space. I've collected the price and square footage of houses that meet these criteria.

Using the grid below, plot the following data as a scatter plot.
Independent variable: ___________________________.
Dependent variable: ____________________________.
Make your house size scale up to 2500 sq. ft. and your price scale up to $400 000.

Size (sq. ft.)	Price ($)
1700	271 900
1850	289 900
1600	277 900
1650	289 900
1700	279 000
1800	294 900
1550	269 900



Draw a line of best fit that best describes the trend in the data on your graph.

Use your line of best fit to estimate the price of a 2000 sq. ft. house.

I noticed a new house that is 2300 sq. ft. and $384 900. Although this house does not meet my criteria, I add it to my collection anyways.

Plot this point on your graph and with a different colour, draw a new line of best fit.

Use your new line to estimate the cost of a 2000 sq. ft. house.

Compare your two estimates. What do you notice?

Lines of Best Fit

A **Line of Best Fit** (aka Trend Line, or Regression Line) is a line drawn through a set of data that best represents the ______________________________ between two variables.

A good line of best fit:

- Is as close as possible to all data points.
- Follows the trend of the data points.

Correlation Coefficient

- The strength of the linear relationship is measured with the **correlation coefficient** (r).
- The correlation coefficient is a number from ____ to ____.
- If the number is positive, then this indicates _______________ correlation. If the number is negative, this indicates _______________ correlation.
- The closer the number is to +1 or –1, the _______________ the relationship. Conversely, the closer the number is to 0, the _______________ the relationship.

Negative				**Positive**			
Strong	**Moderately Strong**	**Moderately Weak**	**Weak**	**Weak**	**Moderately Weak**	**Moderately Strong**	**Strong**

-1 -0.75 -0.50 -0.25 0 +0.25 +0.50 +0.75 +1

Coefficient of Determination

- The **coefficient of determination** (r^2) is an alternate measure of the strength of the linear relationship.
- Since it is squared, there are no negative values.
- If r^2=0.67, then 67% of the variance in the dependent variable is due to a change in the independent variable.

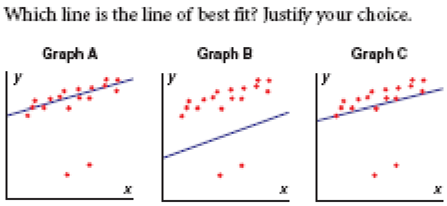
Outliers

An **outlier** is a point or points in a data set that lie(s) __________________ the trend of the other points.

Outliers are the result of ___________________ measurements or _________________ cases. For example, an antique car’s price would be an outlier in a set of data comparing price and age of cars.

Since outliers are so unusual or inaccurate, we **ignore** these points when we create a line of best fit. This prevents the outliers from significantly changing the trend line and the predictions made from it.
(NOTE: we usually circle the outliers and label them as such on a scatter plot)

Example:


Which line is the line of best fit? Justify your choice.

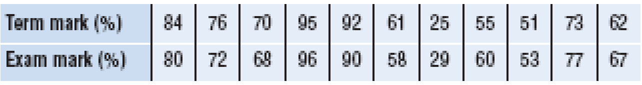
Extrapolation and Interpolation

One of the most important reasons to make a scatter plot is to find the trend in the data so it can be used to make predictions.

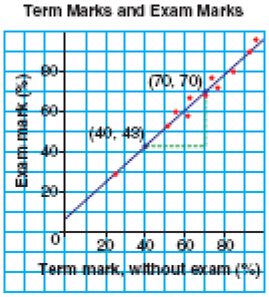
Interpolation: predictions are made from the _______________ points on a scatter plot.

Extrapolation: predictions are made about points ____________ those on the scatter plot.

Example: The following data was obtained from a grade 12 math class last semester. The scatter plot compares term marks with exam marks.



Term mark (%)	84	76	70	95	92	61	25	55	51	73	62
Exam mark (%)	80	72	68	96	90	58	29	60	53	77	67



a) Interpolate to determine Brent’s exam mark if his term mark was 55%. ________

b) Extrapolate to determine Dara’s exam mark if her term mark was 68%.

You will need to find the equation of the line of best fit: y=mx + b

m=

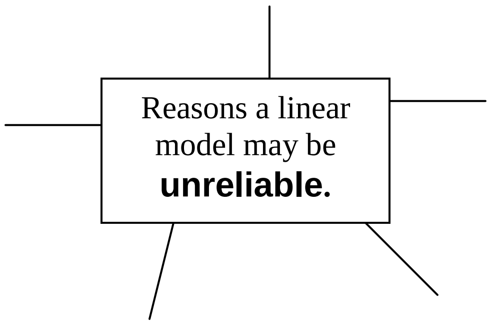
y = ____x + b

(   ) = ____ (   ) + b

Now substitute x = 68 into your equation to determine what y is:

Therefore, Dara will likely score _____% on the final exam.

Reliability of Linear Models



Reasons a linear model may be **unreliable.**

TEACHER NOTES:
Linear models can be unreliable if:
-They are based on too few data points (sample is too small)
-They are based on data that is too clustered.
-There does not appear to be a correlation
-There are too many outliers.
-The relationship is not linear (ie. may be quadratic or exponential instead).