Line of Best Fit

Learning Goals

- make a scatter plot
- look for a pattern
- use the pattern to draw a line of best fit
- use the line of best fit to make conclusions about the data
- discuss the role of the correlation coefficient in patterns

Oct 1-11:15 AM

Line of Best Fit

Use pg 148 to complete the activity.

Sabah is looking to buy a house that is
- detached,
- 2-stories,
- $300 000 or less,
- approximately 2000 sq. ft. of living space.

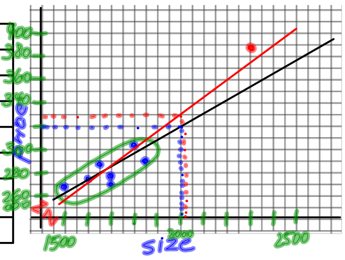
She has collected the price and square footage of houses that meet these criteria.

Using the grid below, plot the following data as a scatter plot.

Independent variable: size.
Dependent variable: price.

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



Estimate the price of a 2000 sf. house we interpolate

2000 sf → $320 000

2000 sf → $330 000

better for bigger ones

this is better for smaller house

Oct 27-12:43 PM

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 notice a new house that is 2300 sq. ft. and $384 900. Although this house does not meet her criteria, she adds it to her 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?

Oct 1-11:21 AM

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 relationship between two variables.

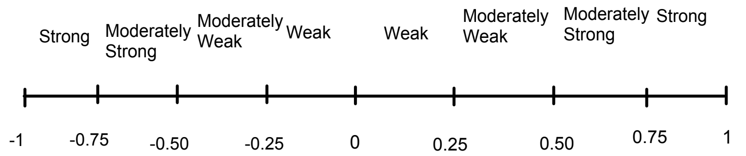
A good line of best fit:

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

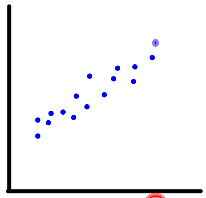
Correlation Coefficient

- The strength of the linear relationship is measured with the **correlation coefficient** (r).
- The correlation coefficient is a number from -1 to 1.
- If the number is positive, then this indicates positive correlation. If the number is negative, this indicates negative correlation.

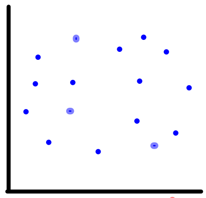
- The closer the number is to +1 or -1, the stronger the relationship. Conversely, the closer the number is to 0, the weaker the relationship.



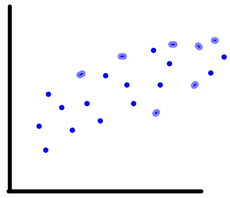
Oct 2-9:00 AM



r = 0.9
positve strong



r = 0.1
no correlation



0.5
positve weak

Coefficient of Determination

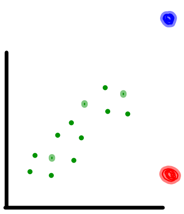
- The **coefficient of determination** (r^2) is an alternate measure of the strength of the linear relationship.
- Since it is squared, there are no negative values.
- If r^2=0.67, then 67% of the variance in the dependent variable is due to a change in the independent variable.

Oct 27-12:40 PM

Outliers

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

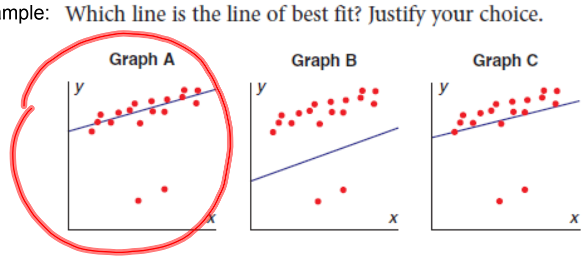
← not an outlier



← outlier

Outliers are the result of inaccurate measurements or special 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.



Oct 29-7:02 AM

Extrapolation and Interpolation

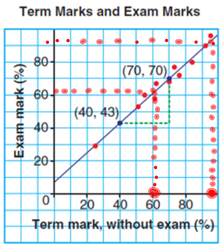
One of the most important reasons to make a scatter plot is to find the trend in the data so it can be used to make predictions.

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

Extrapolation: predictions are made about points beyond 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 60% 61%

b) Extrapolate to determine Dara's exam mark if her term mark was 98% 92%

c) Find the equation fo the line of best fit and use it to verify your answer for b)

$m = \frac{rise}{run}$

$= \frac{27}{30}$

$= 0.9$

slope → $y = mx + b$ ← y-int

$y = 0.9x + b$

$(70) = 0.9(70) + b$

$70 = 63 + b$

$7 = b$

$\therefore y = 0.9x + 7$

Now substitute $x = 98$ into your equation to determine what y is:

$y = 0.9(98) + 7$

$= 88.2 + 7$

$= 95.2$

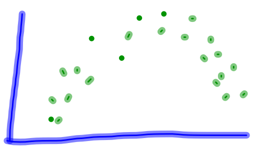
Therefore, Dara will likely score 95 % on the final exam

Oct 29-7:03 AM

RELIABILITY OF LINEAR MODELS

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

- no correlation
- Data is too clustered
- Data is not linear.
- Don't have a lot of points
- There are too many outliers



Oct 29-7:07 AM

Seatwork

pg 153 # 1-7

Oct 2-9:39 AM

Attachments

(5.5) Exponential Equations.doc