IM2 Problem Set 4.2 - Graphing Bivariate Data and Scatter Plots with Technology

BIG PICTURE of this UNIT:	• How do we analyze and then make conclusions from a data set when we collect information on two variables? • How do I describe and analyze bivariate data? • What functions can I use to model bivariate data sets? • How do I decide on the validity/reliability of my data? Of my analysis? Of my conclusions? Of my decision?

PART 1 - Graphing Bivariate Data - DESMOS

Albany and Sydney are about the same distance from the equator. Make a scatter plot with Albany's temperature as the independent variable. Name the type of correlation. Then sketch a line of best fit and find its equation.

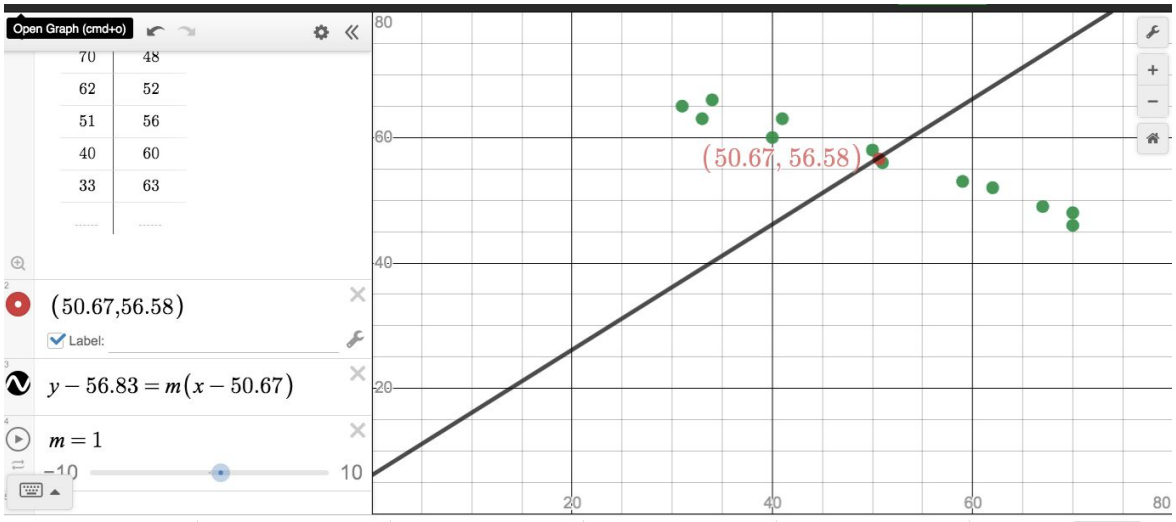
Average Minimum Temperature (°F)		
Month	**Albany**	**Sydney**
Jan	31	65
Feb	34	66
Mar	41	63
Apr	50	58
May	59	53
Jun	67	49
Jul	70	46
Aug	70	48
Sep	62	52
Oct	51	56
Nov	40	60
Dec	33	63

a. Enter the data as a table

b. Set your view window - set the x_{min}/x_{max} and y_{min}/y_{max}

c. Determine the mean of the Albany temperatures as well as the mean of the Sydney temperatures. Graph and label this "mean point" (which should be (50.67, 56.58))

d. Type in the equation of the line of best fit, using point-slope form ⇒ so type the equation as follows ⇒ y - 56.83 = m(x - 50.67). Add a slider for m.

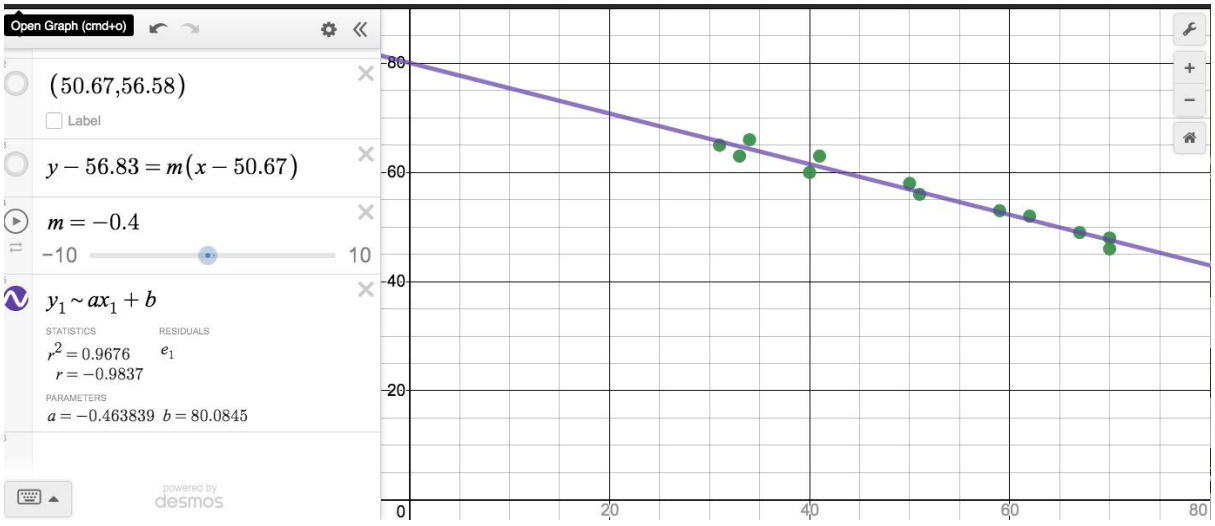


e. Adjust the slider for m as required. Record the equation that you feel best fits the data and models the trend in the data set.

f. Rewrite your equation in slope-intercept form.

PART 2 - Graphing Bivariate Data - DESMOS and Lines (Curves) of Best Fit

We can also ask DESMOS to calculate a line that best fits the data set as follows ⇒ type into DESMOS the equation $y_1 \sim ax_1 + b$. Record the values for a, b and for r. Compare these values to our mean line of best fit (from Steps e. and f. above)



PART 3 - PRACTICE Questions (Using DESMOS)

1. Anthropologists can use the femur, or thighbone, to estimate the height of a human being. The table shows the results of a randomly selected sample.

Femur Length and Height (cm)	
Length	Height
36	160
32	143
46	187
29	142
35	161
38	164
30	140
27	131

   a. Make a scatter plot of the data with femur length as the independent variable.
   b. Find the line of best fit and the correlation coefficient, r .
   c. Interpret the slope of the line of best fit in the context of the problem.
   d. The correlation coefficient should be $r \approx 0.986$. What type of correlation does it have?
   e. A man's femur is 41 cm long. Predict the person's height.
   f. If a person has a height of 200 cm, predict the length of their femur.
   g. How reliable is this prediction? Explain.

2. The gas mileage for randomly selected cars based upon engine horsepower is given in the table.

Gas Mileage and Horsepower of Cars										
Horsepower	175	255	140	165	115	120	190	180	110	125
Mileage (mi/gal)	22	13	25	18	32	28	15	21	35	30

   a. Make a scatter plot of the data with horse power as the independent variable.
   b. Find the line of best fit and the correlation coefficient, r .
   c. Interpret the slope of the line of best fit in the context of the problem.
   d. The correlation coefficient should be $r \approx -0.916$. What type of correlation does it have?
   e. Predict the gas mileage for a 210-horsepower engine.
   f. If a car has a gas mileage of 20 mi/gal, what is the predicted horsepower of its engine?
   g. How reliable is this prediction? Explain.
   h. Explain the terms interpolation and extrapolation.

3. The information for a data set on the number of grams of fat and the number of calories in sandwiches served at Dave's Deli is found on the table below. Use the equation of the line of best fit to predict the number of grams of fat in a sandwich with 420 Calories. How close is your answer to the value given in the table? How reliable is your answer?

Dave's Deli Sandwiches Nutritional Information								
Fat (g)	5	9	12	15	12	10	21	14
Calories	360	455	460	420	530	375	580	390

4. A convenience store manager notices that sales of soft drinks are higher on hotter days, so she assembles the data in the table.

High Temperature (°F)	Number of cans sold
55	340
58	335
64	410
68	460
70	450
75	610
80	735
84	780

   a. Make a scatter plot of the data.
   b. Find and graph a linear regression equation that models the data.
   c. What does the slope mean in the context of this question?
   d. What would the y-intercept mean in the context of the data?
   e. Use the model to predict soft-drink sales if the temperature is 95° F.
   f. What does the model predict for the temperature if the number of cans sold was only 95?

5. The table gives the Olympic pole vault records in the twentieth century.

Year	Height (m)
1900	3.30
1904	3.50
1906	3.50
1908	3.71
1912	3.95
1920	4.09
1924	3.95
1928	4.20
1932	4.31
1936	4.35
1948	4.30
1952	4.55
1956	4.56
1960	5.10
1964	5.64
1968	5.40
1972	5.64
1976	5.64
1980	5.78
1984	5.75
1988	5.90
1992	5.87
1996	5.92
2000	5.90

   a. Find the regression line for the data.
   b. Make a scatter plot of the data on your calculator and graph the regression line. Does the regression line appear to be a suitable model for the data?
   c. Use the model to predict the record pole vault height for the 2004 Olympics. Find the actual record height and by whom. Is this a good prediction?
   d. Use the model to predict the record pole vault height for the 2012 Olympics. What was the actual gold metal height and by whom? Is this a good prediction?
   e. Use the model to predict the record pole vault height for the 2020 Olympics. Do you think the actual record in 2020 will be higher or lower than this prediction? Why?
   f. Find additional data and update your scatterplot and recalculate the equation of the line that best fits the data.

PART 3 - Graphing Bivariate Data - TI-84 and Lines (Curves) of Best Fit

We can do the same thing with our graphing calculators

https://youtu.be/AMx_SwQkn34

https://youtu.be/0as2Jh_eDwg

EXTENSION - Residuals

Residuals are defined as positive and negative deviations from the least squares line. Each residual is the difference between the observed y value and the corresponding predicted y value.

A residual is the difference between what is plotted in your scatter plot at a specific point, and what the regression equation predicts "should be plotted" at this specific point. It is the vertical distance from the actual plotted point to the point on the regression line. You can think of a residual as how far the data "fall" from the regression line (sometimes referred to as "observed error").

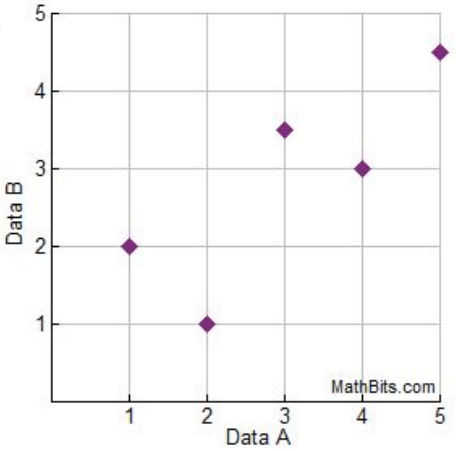
EXAMPLE

You are asked to find an equation to model the data in the set {(1,2), (2,1), (3,3½), (4,3), (5,4½)}.

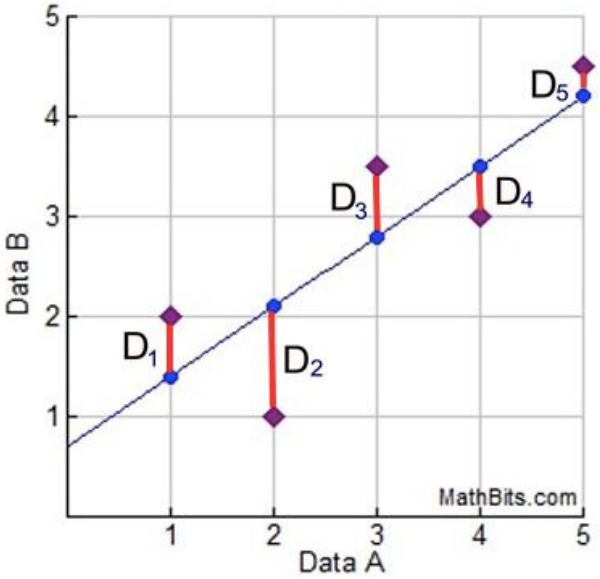
You prepare a scatter plot to see if you should be looking for a linear, quadratic or exponential regression equation. You decide to choose a linear regression, but you are not 100% sure of your choice.

You use your graphing calculator to find the linear regression equation, which is $y = 0.7x + 0.7$.

You graph the regression equation line on the scatter plot, as seen below.



The residuals are the red line segments, referenced by the letter "D" (for distance), vertically connecting the scatter plot points to the corrdinating points on the linear regression line.



Scatter Plot Points:
{(1,2), (2,1), (3,3½), (4,3), (5,4)}

Regression Points
{(1,1.4), (2,2.1), (3,2.8), (4,3.5), (5,4.2)}

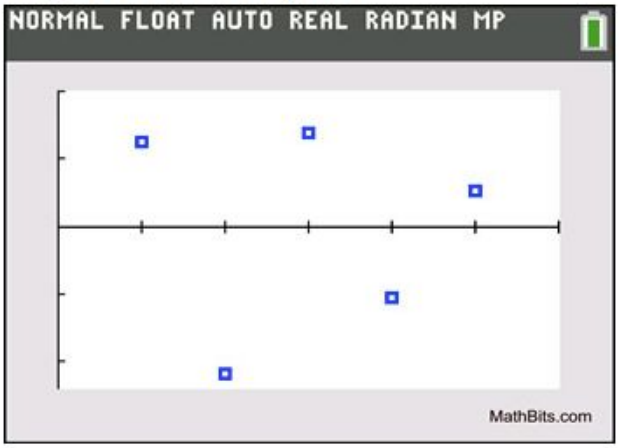
The Red Line Segments:
The red line segments represent the distances between the y-values of the actual scatter plot points, and the y-values of the regression equation at those points.

The lengths of the red line segments are called RESIDUALS.

You decide to plot the residuals to see if your choice of a linear regression model was appropriate for your data.

First, you must find the residuals.
Compute: scatter plot y-value minus regression line y-value for each point.

$D_1 = 2 - 1.4 = 0.6$
$D_2 = 1 - 2.1 = -1.1$
$D_3 = 3.5 - 2.8 = 0.7$
$D_4 = 3 - 3.5 = -0.5$
$D_5 = 4.5 - 4.2 = 0.3$



Now, plot the residuals.
(1,0.6), (2,-1.1), (3,0.7), (4,-0.5), (5,0.3)

As you examine the plots, you notice that the plots do not follow any pattern.

The plots are randomly placed above and below the horizontal axis. A linear model is an appropriate choice for this data.

PRACTICE:

Here are 2 data sets. For each data set, plot the points.

Data Set 1

x	0	1	1.5	2	3	3.5	4
y	-1	6.5	6	8.9	11	13	18

Data Set 2

x	0	1	1.5	2	3	3.5	4
y	1.05	2	3	3.9	7.9	12	15.8

a. For Data Set 1, determine the equation for the line of best fit and use this linear equation to determine the residuals. Prepare a residual plot.

b. For Data Set 2, determine the equation for the line of best fit and use this linear equation to determine the residuals. Prepare a residual plot.

c. Compare the residual plots. Which residual plot shows a "pattern"?