Key

Math 1

~~U3 L1 E3~~ Exploration

2-7 Linear Regression, Correlation, Causation

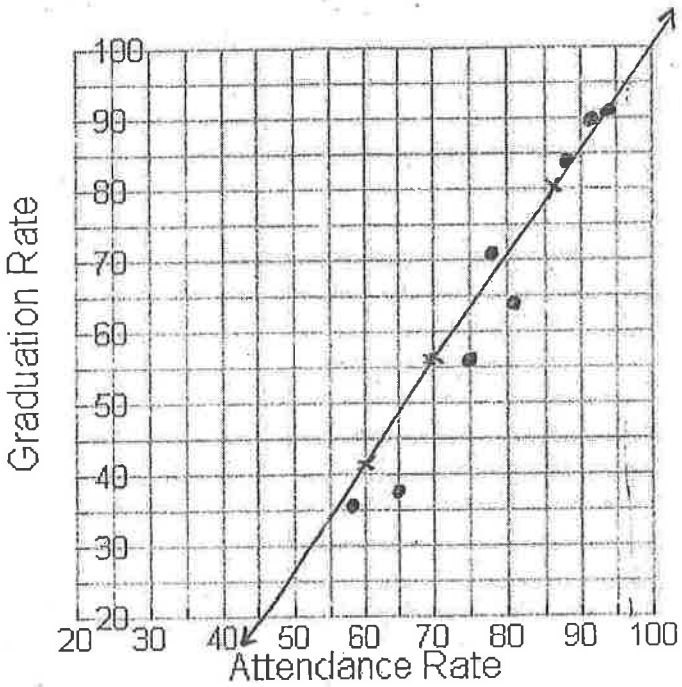
Name Key

Date

Below is data for the attendance rate and graduation rate of several high schools (both are percentages). Use this data to answer questions 1-5.

Attendance Rate	76	65	81	78	88	94	58	92
Graduation Rate	56	37	64	71	84	91	36	90

1. Make a scatterplot of the data on the below coordinate plane. Remember, in a scatterplot you DO NOT CONNECT THE DOTS.



2. While this data is clearly not perfectly linear, does it seem to have a linear trend? Explain.

Overall, it looks like a line (linear).

3. One line suggested to fit this data passes through the points (60, 41) and (86, 80) [*Note: These points are NOT part of the data set, they are just points that the linear model passes through*]. Graph these two points and use a straightedge to draw a line that passes through the two points.

✓

The line you drew in number (3) is what is called a ***linear model****. Linear model is just a fancy math term for a line that models the trend in the data – it does not pass through all of the points in the data set (and it may in fact not pass through any points in the data set), but it seems to be reasonably close to the data.*

4. Use your linear model (the line you drew) to estimate what the graduation rate will be for a school that has an attendance rate of 70%. Show your work on the graph.

About a 56% graduation rate.

5. Use your linear model to estimate attendance rate needed to have a graduation rate of 95%. Show your work on the graph.

About a 97% attendance rate.

6. Use the points (60, 41) and (86, 80) to find the equation of your linear model. Show your work.

$m = \frac{80-41}{86-60} = \frac{39}{26} = 1.5$

$y - 41 = 1.5(x - 60)$

$y = 1.5x - 90 + 41$

$\boxed{y = 1.5x - 49}$

7. What is the slope of your linear model? Explain what the slope means ***in the context of this data set.***

Slope = 1.5

Graduation rate increases by 1.5% for every 1% increase in Attendance rate.

8. What is the y-intercept of your linear model? Explain what the y-intercept means ***in the context of this data set.***

y-int. = 49

Graduation rate is -49% for someone who never attends school (0% attendance rate)

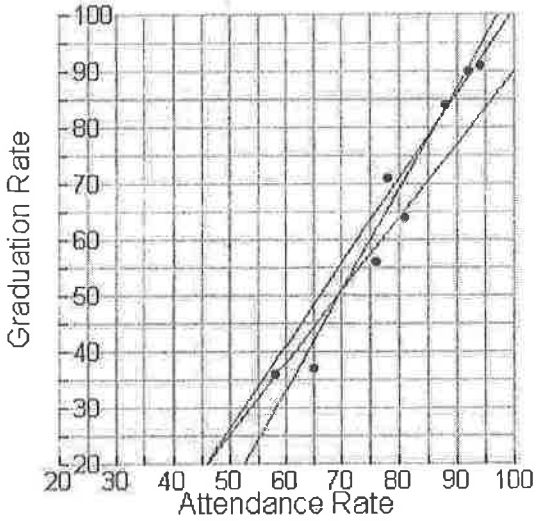
9. Use the equation of your linear model to predict the graduation rate of a school that has a 70% attendance rate. Show your work below. How close was your estimation from number (4)?

$y = 1.5(70) - 49 = \boxed{56\%}$ Same as my estimate! :)

10. How accurate do you think your prediction from number (9) will be? Explain.

It should be fairly accurate since the line I drew fits the data pretty well.

The linear model that you used to answer the above questions is one of an infinite number of lines that we could use to model this data. Below is a scatterplot of our data and several lines that look like they are good fits for the data. The question then becomes, which line is *the line of best fit*. In Math 2, you will study how this line of best fit is developed. For this course though, we just need to understand that there is a line of best fit, and our calculator can tell us the equation of the line of best fit. We will use our calculator to come up with this equation below.



11. The line of best fit is also called the **linear regression equation**. In your notebook, take notes on how to find the linear regression equation for a data set as we do it together as a class.

Data in a list. MENU → 4 → 1 → 3 (or 4) → Pick x & y list → ENTER

12. Write the linear regression equation that we found in number (11) below.

$y = 1.693x - 67.60$

13. What is the slope of the regression equation (*round to 3 decimal places*)? What does the slope mean in the context of this data set?

$m = 1.693$ Grad. rate increases by 1.693% for every 1% increase in attendance rate.

14. What is the y-intercept of the regression equation? What does the y-intercept mean in the context of this data set?

$b = -67.60$ -67.6% grad. rate (meaning 0%) for someone with 0% attendance rate.

15. Use your linear regression equation to predict the graduation rate if the attendance rate is 80%. Predict if the attendance rate is 96%.

$y = 1.693 \cdot 80 - 67.6 = \boxed{67.84\%}$

$y = 1.693 \cdot 96 - 67.6 = \boxed{94.928\%}$

16. Put the regression equation into your calculator and make a table. Predict the attendance rate if the graduation rate is 84.77.

Approximately 90% attendance rate.

17. Find and interpret the correlation coefficient.

$r = 0.975$ This means that att. rate & grad rate have a high linear correlation. Strong, linear pattern.

Oil Changes and Engine Repairs

The table below displays data that relate the number or oil changes per year and the cost of engine repairs. The activity which follows uses these data to introduce students to modeling with a linear function. To predict the cost of repairs from the number of oil changes, use the number of oil changes as the x variable and engine-repair cost as the y variable.

Oil Changes Per Year	3	5	2	3	1	4	6	4	3	2	0	10	7
Cost of Repairs ($)	300	300	500	400	700	400	100	250	450	650	600	0	150

18. Make a scatterplot of this data on your calculator. Use *Oil Changes Per Year* as the independent variable.

19. Does this data appear to have a linear trend? Explain.

Yes, resembles a line.

20. Find the linear regression equation for this data. Write it below. Round to three decimal places.

$y = -73.070x + 650.269$

21. What is the slope of the linear regression equation? What does it mean in the context of this data set? Explain specifically what it means that the slope is *negative*.

$M = -73.07$ (slope) For each additional oil change, the cost of repairs decreases by $73.07.

22. What is the y-intercept of the regression equation? What does it mean in the context?

y-int = 650.269

Cost of repairs will be $650.27 if no oil changes completed.

23. Find and interpret the correlation coefficient.

$r = -0.914$

The data has a fairly strong, negative, linear relationship.
↳ Negative, linear pattern.

24. Predict the *Cost of Repairs* if someone gets 8 oil changes per year. Show your work.

$y = -73.07 \cdot 8 + 650.269$

$= \$65.71$

25. Predict the *Cost of Repairs* if someone gets 6 oil changes per year. Why isn't this value the same as the value for 6 oil changes in the above table? Explain.

$y = 73.07 \cdot 6 + 650.269$

$= \$211.85$

This equation represents the "best fit" line which does not go through every data point. This is why our answer is diff from the table.

The below data is the shoe size and ACT score (out of 36) for a math class.

Shoe Size	6	6.5	6.5	7	7	8	8.5	8.5	8.5	9	9.5
ACT score	23	32	24	20	33	19	21	32	27	28	27

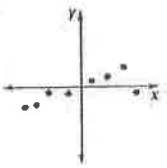
26. Make a scatterplot of the data on your calculator. Does this data look linear?

Not in the least.

27. Find the linear regression equation and write it below. Round to three decimal places. Do you think the regression equation will make accurate predictions for this data set? Support your answer based on your coefficient correlation.

$y = 0.256x + 24.023$. This will not make accurate predictions since it does not fit our data very well, as made obvious by a low correlation of $r = 0.195$

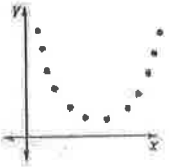
In numbers 28-31, state whether the given association is *positive, negative,* or *approximately zero.*

28. 

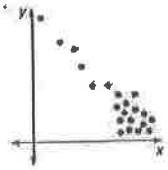
positive

29. 

positive

30. 

zero

31. 

negative

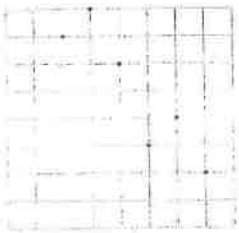
In numbers 32-35, match the *r* values with the appropriate graphs.

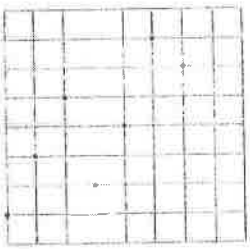
32.) $r = 0.9$ IV

33.) $r = 0.7$ II

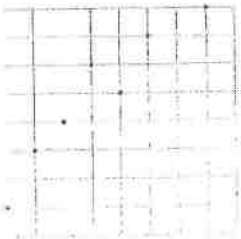
34.) $r = -0.8$ I

35.) $r = -0.2$ III

I. 

II. 

III. 

IV. 

36. When comparing two variables, it is important to distinguish between correlation and causation. **Correlation** is a relationship between two variables. **Causation** is the same as cause and effect. For example, there is a strong association between the number of sunglasses purchased and the amount of ice cream purchased in a day. Is this an example of correlation, or causation? Explain.

Correlation! These two things have nothing to do with each other. One does not cause the other to occur.

Lurking Variable: a variable in the background that makes it look like two other variables affect each other, when they really do not.

37. The 12 countries listed below have the highest per person ice cream consumption of any countries in the world. As shown in the following table and scatter plot, there is an association between the number of recorded crimes and ice cream consumption.

Country	Ice Cream Consumption Per Person (in liters) per Year (X)	Recorded Crimes per 100,000 Inhabitants per Year (Y)
New Zealand	26.3	12,591
United States	22.5	9,622
Canada	17.8	8,705
Australia	17.8	6,161
Switzerland	14.4	4,769
Sweden	14.2	13,516
Finland	13.9	7,273
Denmark	9.2	1,051
Italy	8.2	4,243
France	5.4	6,765
Germany	3.8	8,025
China	1.8	131

a. Using Ice Cream Consumption as the independent variable, make a scatter plot of the data on your calculator, then find and graph the regression line. Write the equation for the line below in function notation.

$f(x) = 342.7x + 2,468.7$

b. Find and interpret the correlation coefficient.

$r \approx 0.637$ There is a moderate, positive association between ice cream consumption + number of recorded crimes.

c. Interpret the slope of the regression line in the context of the data.

Slope $= \frac{342.7}{1}$ There are about 343 recorded crimes for each liter of ice cream consumed per person.

d. What is the association (correlation or causation) between the amount of ice cream consumption and recorded crime?

More ice cream consumed = more crimes committed?? What???

e. Is there a lurking variable? If so, give a possible lurking variables.

Yes! Temperature! More ice cream eaten when it's warmer. More windows & doors to houses left open, too.