Math 1 Name\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2-7 Linear Regression, Correlation, Causation** Date\_\_\_\_\_\_\_\_

* *I can create and use a line of best fit both by hand (using two points) and using technology (linear or exponential).*
* *I can explain the meaning (using appropriate units) of the slope and y-intercept of a linear function in a real-world situation.*
* *I can find and interpret a correlation coefficient.*
* *I can distinguish between correlation and causation.*

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.

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.

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

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

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

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

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)?

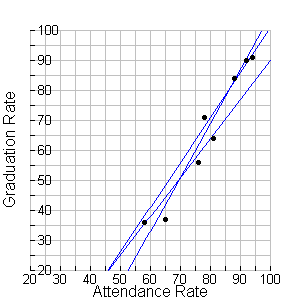
10. How accurate do you think your prediction from number (9) will be? Explain.

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***. 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.

12. Write the linear regression equation (round to nearest thousandth) that we found in Number 11.

13. What is the slope of the regression equation (round to nearest thousandth)? What does the slope mean in the context of this data set?

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

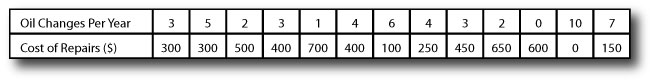
15. Use your linear regression equation to predict the graduation rate if the attendance rate is 80%. Predict if the attendance rate is 96%.

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

17. Find and interpret the correlation coefficient.

***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.



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.

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

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*.

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

23. Find and interpret the correlation coefficient.

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

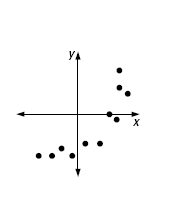
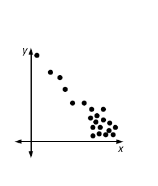
25. Predict the *Cost of Repairs* if someone gets 6 oil changes per year. Why isn’t this value the same as the value for 6 oil changes in the above table? Explain.

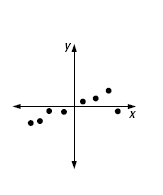
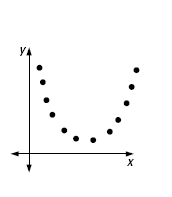
The below data is the shoe size and ACT score (out of 36) for a math class.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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?

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.

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



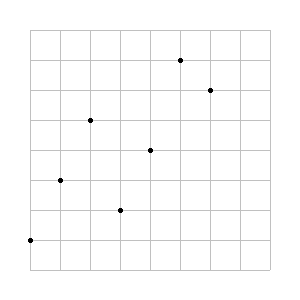
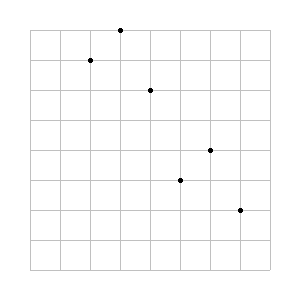
28. 29. 30. 31.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

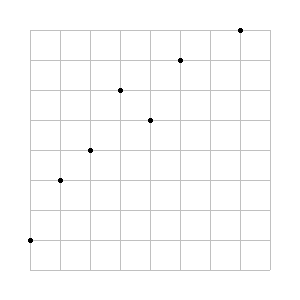
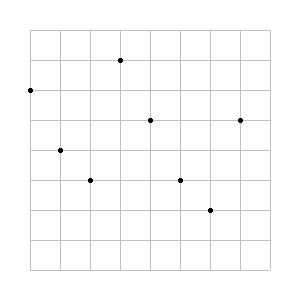
32. *r* = 0.9 33. *r*  = 0.7 34. *r* = -0.8 35. *r*  = -0.2

\_\_\_\_\_ \_\_\_\_\_ \_\_\_\_\_ \_\_\_\_\_



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.

**Lurking Variable:**

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 | Recorded Crimes per 100,000  Inhabitants per Year |
| 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 |

1. 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. Round to the nearest tenth.
2. Find and interpret the correlation coefficient.
3. Interpret the slope of the regression line in the context of the data.
4. What is the association (correlation or causation) between the amount of ice cream consumption and recorded crime?
5. Is there a lurking variable? If so, give a possible lurking variables.