Does A Correlation Exist?

About the Lesson

In this activity, students will:

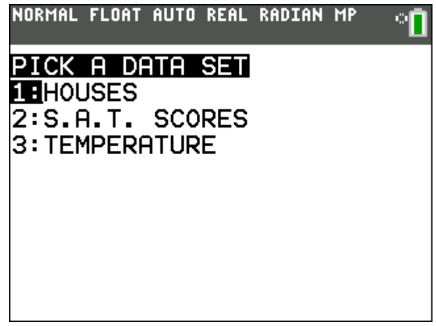
- Find the linear regression equation and calculate the correlation coefficient of three data sets.
- Use a linear regression equation to analyze the relationship and predict the values of two variables.
- Describe the real-world significance of the slope and *y*-intercept in a linear regression for a data set.
- Identify and distinguish between correlation and causation in a data set.

Vocabulary

- correlation coefficient
- linear regression
- outliers and influential points

Teacher Preparation and Notes

- Students should have prior experience using the TI-84 Plus calculator to create scatter plots. In addition, they should have had some discussion about correlation. They should also know how to solve an equation using the **solve** command or be prepared to demonstrate the tech tip.
- Be sure students are successful with the first set of data. The following problems can then be done more independently in a small group in class, as homework, or as an assessment.
- The program will automatically turn the diagnostic on so that the correlation coefficient will be displayed when the linear regression is calculated. After the activity students can change the mode from G-T, Graph-Table, back to FULL. To do this, press [MODE], select FULL, and press [ENTER].



Tech Tips:

- This activity includes screen captures taken from the TI-84 Plus C Silver Edition. It is also appropriate for use with the TI-84 Plus family with the latest TI-84 Plus operating system (2.55MP) featuring MathPrint ™ functionality. Slight variations to these directions given within may be required if using other calculator models.
- Access free tutorials at http://education.ti.com/calculators/pd/US/Online-Learning/Tutorials
- Any required calculator files can be distributed to students via handheld-to-handheld transfer.

Compatible Devices:

- TI-84 Plus Family
- TI-84 Plus C Silver Edition

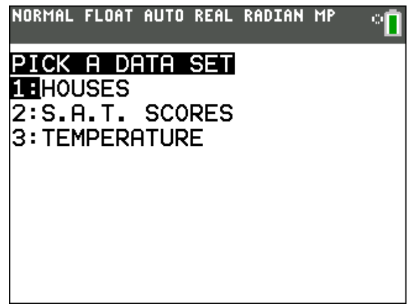
Associated Materials:

- DoesACorrelationExist_Student.pdf
- DoesACorrelationExist_Student.doc
- CORLATE.8xp (TI-84 Plus program)

Students will consider and discuss the relationship between data from three sets of lists. Then they will plot the data with the help of the program **CORLATE**. For each data set the students will create a scatter plot, and then they will use the Home screen to explore the data and answer questions. Depending on previous discussion and lessons, it may be important to help guide the discussion for this first problem.

> **Teacher Tip:** While students are discussing their answer to question 1 send the file CORLATE.

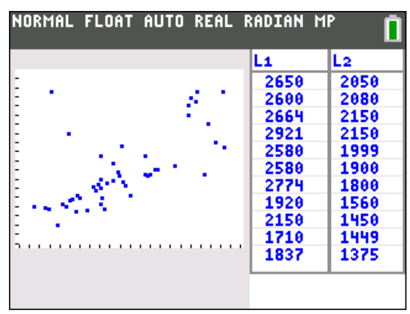
Problem 1 – Home Price vs. Square Footage



Students will discuss the first question with a partner and record their thoughts before looking at the data.

1. How do you think the selling price of a house relates to the amount of area of the house or square footage? Do you think there is any correlation? What are the variables? Which variable is the independent variable? Which variable is the dependent variable? What else might the price of a house depend upon?

   Answer: Answers will vary, but students should realize that the larger the house, the more expensive it is. The dependent variable is the price of the house since it depends on the area of the house, which is the independent variable. Factors, or variables, that could contribute to the cost of the house include not only the size of the house, but the size of the property that the house sits on, the location of the house, and the quality of the house.



Next students will run the program **CORLATE** and select option 1, **HOUSES**. The area measured in square feet of a house is in **L1**. The selling price of the corresponding house (given in hundreds of dollars) is in **L2**. Once students have graphed the data, they should consider if they still agree with initial prediction and answer the following questions:

2. Explain the meaning of the point (2650,2050). Include units.

   Answer: A house that has 2,650 square feet is priced at $205,000.

3. Choose the type of correlation (circle your answer).

a. positive negative

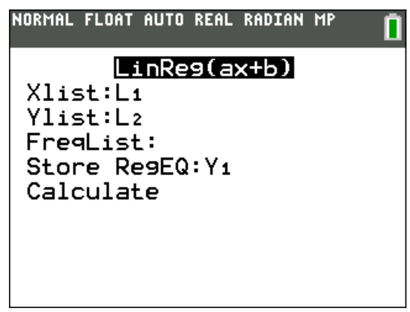
b. very strong moderately strong moderately weak very weak

Answer: Students should circle positive and moderately strong.

4. Predict the value of the correlation coefficient to one or two decimals. Explain your reasoning.

Answer: Answer will vary, but answers should agree with what was circled in question 3. Students should predict a value larger than 0.5 but smaller than 0.9.

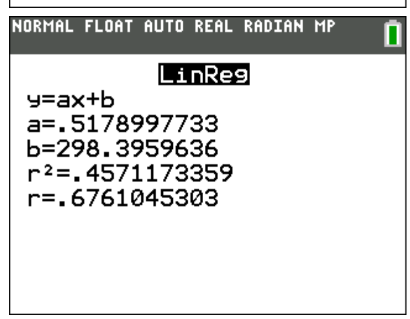
> **Teacher Tip:** Having students predict the correlation coefficient increases student engagement. This also fosters quality discussion and deeper understanding. Students who are new to the concept of correlation coefficient may not be comfortable making a prediction, but by the end of this activity, they should improve.



On the Home screen, have students calculate the linear regression equation. Students should press [STAT], move the cursor to the CALC menu, and select **LinReg(ax+b)**. Then, have students enter the list of their independent variable, the list of their dependent variable, and store the regression equation in **Y1** using [ALPHA] [F4].

5. What is the regression equation?

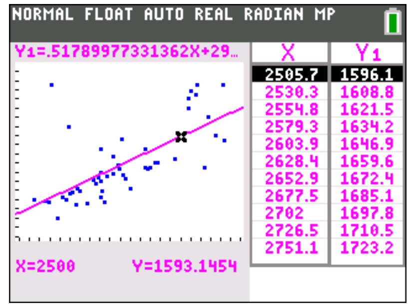
Answer: $y = 0.518x + 298$



6. What is the correlation coefficient, *r*? How does the coefficient compare with your description of the correlation? How does your prediction compare?

Answer: 0.676 is the correlation coefficient. Students should also describe how this compares to their prediction.

Press [GRAPH] to return to the scatter plot. The regression equation will be graphed with the plot. Press [TRACE], and then use the down arrow [▼] to view the equation in the top-left corner.

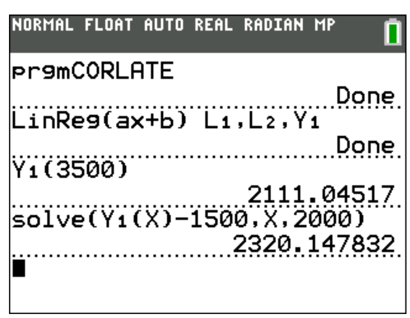


7. What is the sign of the slope? How does this relate to the sign of the correlation coefficient? What is the meaning of the slope in the context of the data? Also explain the *y*-intercept in the context of the data.

   <u>Answer:</u> The slope is positive. There is a positive correlation between the variables. The price of the house increases about \$52 for every square foot. The slope is the average rate; it is \$52/sq ft. Also ask students if they understand the significant of the *y*-intercept. The *y*-intercept is approximately \$29,800, and this would relate to the price of a house with zero square feet. Although some may think a zero square feet house does not have any real-world significance, because that house cannot exist, it is significant because there is still a cost to land even without a house on it. It can be understood as the cost of the land.

> **Teacher Tip:** Encourage a class discussion about the significance of the slope and *y*-intercept in the linear regression equation. Guide students to state that the slope is the average rate of change, meaning it is the amount that the price of a house changes as the square footage of the house changes. If students have trouble with this concept, have them explicitly plug 1 square foot into the linear equation and then calculate the price of the house. Then, have them plug 2 square feet into the equation and calculate the price of the house. Have students determine the difference in the prices of the two houses. They can repeat this process for 3 and 4 square feet. Students should realize that the price of a house increases \$52 for every increase of a square foot. This is the significance of the slope in the equation.
>
> Students should also realize that the *y*-intercept corresponds to the cost of property with a house of zero square feet. Discuss why. Property still has a cost even if there is not a house on the land.

Have students use the regression equation or the table feature to determine the following predictions.



8. Predict the price of a house that has 3,500 sq. ft.

   <u>Answer:</u> \$211,100

9. Predict the number of square feet for a house costing $150,000.

 Answer: Students can set the regression equation equal to 1500 and solve for *x*. They should obtain a value of approximately 2320 square feet.

10. Predict the price of a house with 50,000 sq. ft. Does this prediction seem reasonable based on the data given? Explain.

 Answer: $\mathbf{Y_1}$(50 000) = $2,619,800. Based on the model this prediction seems reasonable although the houses may be substantially more or less expensive based on other variables.

> **Tech Tip:** Students can check their answers using the **solve** command from the Catalog. This can provide an opportunity to discuss an appropriate domain and range within which to make predictions. Student can see the following syntax tip if they press [+] when the 'solve(' is selected in the catalog:
>
> **solve(expr, variable, guess, {lower bound, high boundary})**.

11. Predict the number of square feet for a house costing $5.2 million. Does this prediction seem reasonable based on the given data? Explain.

 Answer: Using solve($\mathbf{Y_1}(x)$ – 52000,*x*,guess), where guess is any value, results in 99,829, or about a 100,000 sq ft house. Again, since the price is in hundreds of dollars, 5,200,000 would be 52,000. Students can also set the regression equation equal to 52,000 and solve for *x*, which gives the solution of 99,810. (Note that these values are different due to the rounded numbers of the regression equation.) This is reasonable based on the given data. This does not mean, however, that a house with 100,000 square feet will always be exactly $5.2 million. (This can be confirmed from an Internet search of real estate search of million dollar homes.)

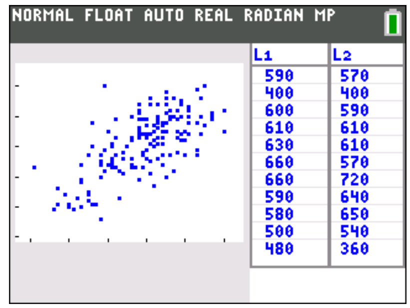
> **Teacher Tip:** Additional question that can be used to guide discussion is: Does there have to be an independent and dependent variable in each relation? The next set of data will help students consider that there does not.
>
> How does one know if the correlation is positive or negative? What would the scatter plot look like if the correlation coefficient is 1? –1? As students progress through the three data sets in this activity they should recognize that a negative correlation describes a data set where one variable increases as the other decreases. The closer the correlation coefficient is the absolute value of 1, the more closely the data can be modeled by a linear regression.

Problems 2 and 3

These problems have different sets of data. The student worksheet asks the student to make predictions about correlation and to make interpolations or extrapolations based on regression equations. It does not make sense to extrapolate to values greater than 800 for the SAT data.

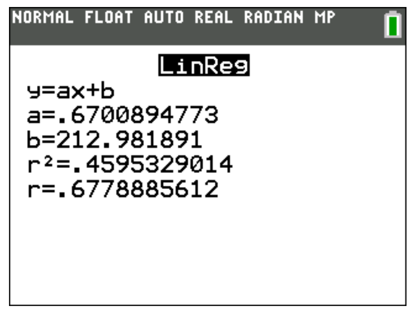
In addition to discussing the relationship between the variables in the data sets, students will also consider if one variable is dependent on the other. (That is, students will consider correlation vs. causation.)

Have students discuss the next question before looking at the data. Students will be analyzing Math and Verbal scores from male and female students who took the SAT exam.



12. Do you think students who score well on the Verbal section of the SAT exam also score well on the Math section? Discuss and record your thoughts on which variable is the independent and dependent variable. Do you think there will be a correlation?

    <u>Answer:</u> Students may realize that a student who does well in one section often does well in both sections. However, there is no clear dependent or independent variable. Verbal scores do not depend on Math scores or vice versa.



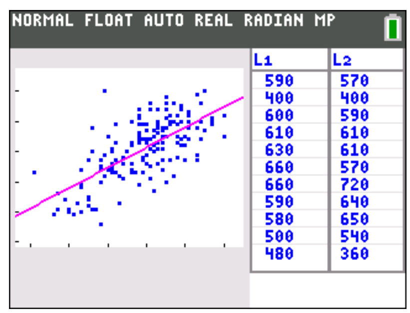
Have students run the program **CORLATE** and select option 2, **S.A.T. SCORES**. The Verbal scores for a sample of 162 students are located in **L1**. The Math scores for those students are in **L2**.

13. Choose the type of correlation (circle your answer).

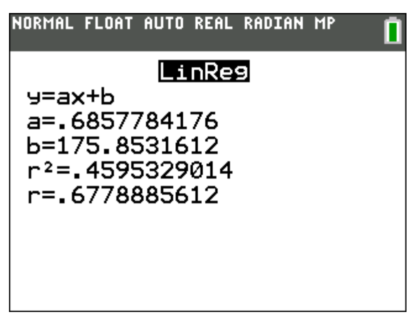
    a. positive negative

    b. Do you think the correlation is stronger, weaker, or about the same as the data set from Problem 1?

    <u>Answer:</u> The data show a positive correlation that looks similar to the data set from Problem 1.



Independent: L2; Dependent: L1



Find the linear regression equation.

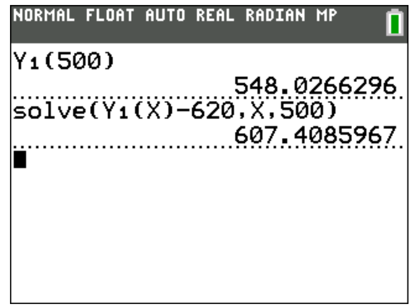
14. Record the regression equation and explain the meaning of the slope.

    <u>Answer:</u> $y = 0.67x + 213$. For every increase of 100 in the Verbal score, the Math section improves by 67.

15. What is the correlation coefficient? How does your prediction compare?

    <u>Answer:</u> 0.678, which is similar to data set from Problem 1.

> **Tech Tip:** Have students consider if the correlation would change if the variables were switched. This can be investigated easily by switching the variables and performing the calculations again. If student explore this they will discover that, while the regression equation changes, the correlation coefficient does not change. To confirm this, see the screen shot above.



Have students return to the scatter plot to view the regression equation. Students should use the regression equation to determine the following predictions.

16. Predict the Math score if the Verbal is 500.

    Answer: The Math score would be 548.

17. Predict the Verbal score if the Math score is 620.

    Answer: The Verbal score would be 607.

18. Is there a relationship between these two variables? Is one dependent on the other? Is there correlation and/or causation?

    Answer: Yes, there is a relationship. As the one increases, the other increases, however one is not dependent on the other. There is a correlation, but not necessarily causation.

Problem 3 – Latitudes vs. Temperatures in January

In this problem, students will investigate the temperature of locations at various latitudes on Earth in the month of January.

19. Do you think the latitude of a location is related to the temperature at that location? Discuss and record you thoughts. What is the independent and dependent variable? What are other variables that affect the temperature of a location?

    Answer: As you go further North, the temperature decreases. As the latitude increases (relative to the equator), the temperature decreases, but there are other factors to consider, such as mountains or elevation. Also ocean currents keep some coastal cities warmer – Anchorage, Alaska is an example.

20. Predict the type of correlation (circle your answer).

    a. positive negative

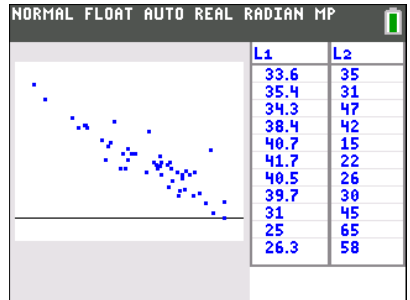
    b. very strong moderately strong moderately weak very weak

    Answer: Students should circle negative and very strong, but since this is a prediction before they even see the data, answers may vary.

Have students run the program **CORLATE** and select option 3, **TEMPERATURE**. The latitude (in degrees north of the equator) of 50 different locations is displayed in **L1**. The average minimum January temperature in °F for the 50 locations is in **L2**.

21. Predict the value of the correlation coefficient. Explain your reasoning.

    Answer: Students should predict a negative correlation coefficient with a magnitude greater than 0.8 because there is a strong correlation.

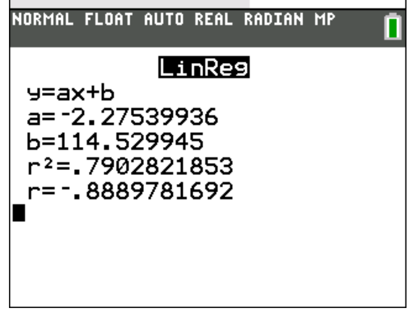


Find the linear regression equation and store the equation in $\mathbf{Y_1}$.

22. Record the equation and explain the meaning of the slope and *y*-intercept.

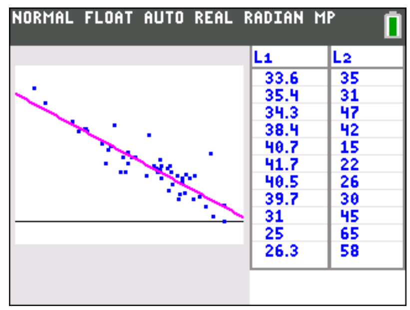
    Answer: $y = -2.28x + 115$. This means that for every degree north of the equator, the temperature drops about 2.3 ºF.

    The *y*-intercept implies that the temperature at the equator is approximately 115 ºF.



23. What is the correlation coefficient? How does your prediction compare? How does it compare with your description of the correlation?

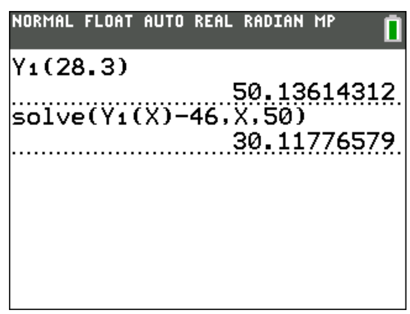
    Answer: –0.889. Students should also discuss how their prediction agreed with the actual correlation coefficient.



Use the regression equation to determine the following predictions:

24. Predict the average minimum January temperature for a city with latitude 28.3 degrees North.

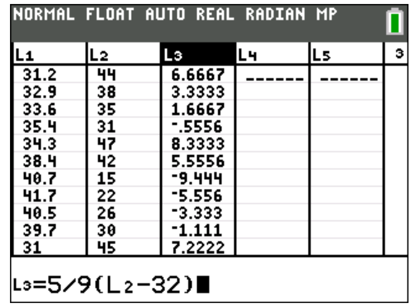
    Answer: 50.1 ºF



25. Predict the latitude for a city with an average minimum January temperature of 46°.

    Answer: 30.1 degrees North.

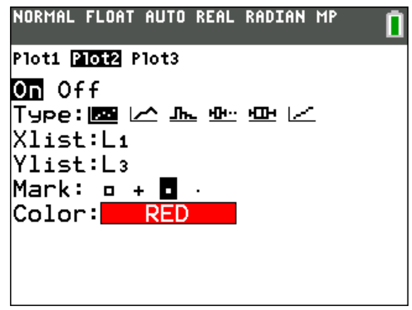
Now discuss and investigate what would happen if the temperatures were changed from Fahrenheit to Celsius.



26. If you know that 0 ºC is 32 ºF and 100 ºC is 212 ºF, what is the formula for the temperature in degrees Celsius as a function of the temperature in degrees Fahrenheit? Create a third list that converts the temperatures to Celsius by entering the formula in the top of **L3**.

    Answer: Since *C* = 5/9 (*F* – 62),

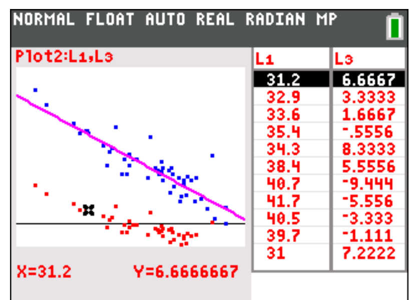
    students should enter 5/9(L2 – 32)

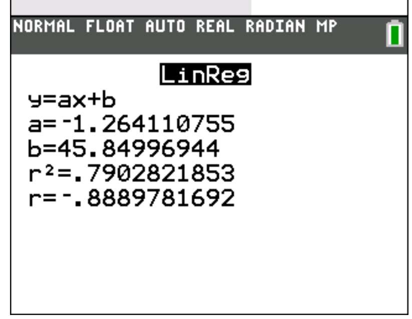


27. Use [2nd] [Y=] to add another [STAT PLOT] and find the new regression equation. Record the equation and correlation coefficient.

    Answer: $y = -1.26x + 45.8$.

    The correlation coefficient is still –0.889.





28. Describe what happened to the plot of Celsius vs. Latitude compared to the Fahrenheit vs. Latitude. Explain.

    Answer: The slope decreased and the data were shifted down vertically.

