Math Objectives

- Students will recognize that bivariate data can be transformed to reduce the curvature in the graph of a relationship between two variables.
- Students will use scatterplots, residual plots, and correlation coefficients of different transformations of bivariate data to determine which transformation is more effective at eliminating the curvature.
- Students will make predictions about the response variable using a transformed data model.
- Students will construct viable arguments (CCSS Mathematical Practices).
- Students will model with mathematics (CCSS Mathematical Practices).

Vocabulary

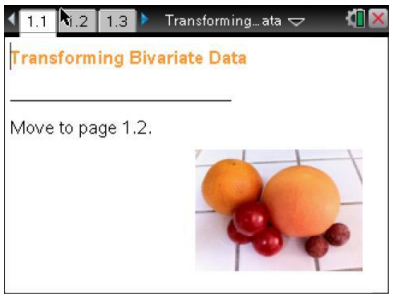
- bivariate data
- correlation coefficient
- exponential function
- least-squares regression line
- log transformation
- log-log transformation
- power function
- quadratic function
- residual plot
- scatterplot
- square root transformation

About the Lesson

- This lesson involves square root, semi-log, and log-log transformations of curved bivariate data using given data sets
- As a result, students will:
  - Observe scatterplots, residual plots, and correlation coefficients of bivariate data.
  - Transform data using square root, semi-log, and log-log transformations.
  - Determine which transformation is more effective at reducing the curvature in the graph of two variables.
  - Use the least-squares regression line based on the transformed data to make predictions.

TI-Nspire™ Navigator™ System

- Transfer a File.
- Use Quick Poll to assess students' understanding.



TI-Nspire™ Technology Skills:

- Download a TI-Nspire document
- Open a document
- Move between pages
- Grab and drag a point

Tech Tips:

- Make sure the font size on your TI-Nspire handhelds is set to Medium.
- You can hide the function entry line by pressing ctrl G.

Lesson Files:

Student Activity

Transforming_Bivariate_Data_Student.pdf

Transforming_Bivariate_Data_Student.doc

TI-Nspire document

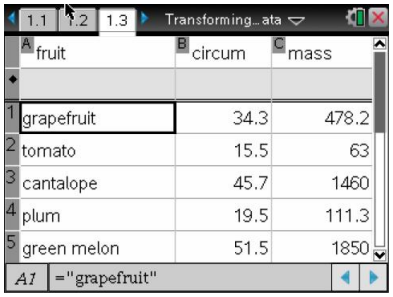
Transforming_Bivariate_Data.tns

Visit www.mathnspired.com for lesson updates and tech tip videos.

Discussion Points and Possible Answers

Analyzing the relationship between two variables, usually an explanatory variable and a response variable, is an important statistical task, and statisticians have developed tools to do this task. The theory and assumptions necessary to use these tools, however, are based on the use of linear relationships between the variables. Unfortunately many relationships are not linear. This activity explores techniques for transforming non-linear relationships into linear relationships.

Move to page 1.3.

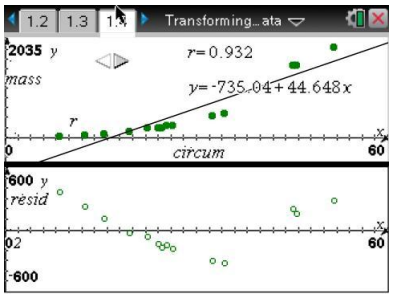


The data on Page 1.3 are the type, the circumference in centimeters, and the mass in grams of 14 approximately spherical pieces of fruit.

1. Use the bar on the right to scroll through the measurements. What do you think a scatterplot of these bivariate data will look like?

   Sample Answer: Looking at just the first few pairs of values, it does not appear that the scatterplot will be linear—there will likely be some kind of curve to these data because the change in the number of grams of mass for a given change in the circumference does not seem to be constant.

Move to page 1.4.



Page 1.4 displays a scatterplot of the fruit data with the least-squares regression line and the correlation coefficient. Beneath the scatterplot is the residual plot.

2. a. Describe the shape of the distribution of fruit mass versus fruit circumference with respect to shape. Does the distribution surprise you given what you saw in the spreadsheet? Explain your thinking.

   Sample Answer: The distribution of fruit mass versus fruit circumference is fairly strongly curved—possibly a quadratic or an exponential curve. This is not surprising as the trend in the data did not seem to follow a linear pattern.

c. What is the correlation coefficient? What does that suggest about the linear relationship between fruit circumference and fruit mass?

Sample Answer: An *r*-value of 0.9325 is fairly high. This might suggest that there is a linear relationship between the variables.

> **Teacher Tip:** This would be a good time to remind the class that the shapes of the scatterplot and residual plot are more important than the *r*-value.

d. Would it be appropriate to use this least-squares regression line to make a prediction about the mass of a fruit based on its circumference? Explain?

Sample Answer: Both the scatterplot and the residual plot show a strong curved pattern. So regardless of the relatively high correlation coefficient, linear regression is not appropriate here, and predictions should not be made from the least-squares regression line. Because of the curvature in the data, predictions made at either end (very small or very large circumferences) might be too small, and predictions made for circumferences in the middle of the data might be too large. When there is a predictable error in the predictions, the model is not a good fit.

> **Teacher Tip:** A log-log transformation is a transformation that takes the log of both sides of the equation. If that transformation gives a linear result, then the original relationship, "untransformed" is the “independent” variable raised to a constant power, like $y = A \times x^p$.

The arrow will allow you to explore several transformations of the fruit data. For each transformation, you will see the scatterplot and residual plot of the transformed data, the equation of the least-squares regression line, and the correlation coefficient. The goal is to find a transformation that changes numbers in ways that makes the graph more linear. Using the arrow will display square root, semi-log, and log-log transformations.

3. Click on the arrow in the upper panel to take the square root of all of the fruit masses.
   a. How do the scatterplot and residual plot change with respect to shape? What is the new correlation coefficient?

Sample Answer: The scatterplot looks less curved than the scatterplot of the original data, but the residual plot still shows a curve. The new correlation coefficient is 0.9849.

b. How would you write the equation for the least-squares regression line for the transformed data? (Hint: Look at the variable names on each axis to see where you need to include "sqrt".)

Answer: predicted sqrt(mass) = -7.55 + 0.955 (circumference)

c. Do the transformed data have a linear relationship? Explain why or why not.

Sample Answer: No, even though the correlation coefficient is close to 1, the residual plot still shows a curved pattern, indicating there is not a linear relationship between fruit circumference and squareroot(fruitmass).

4. Another way to change the shape of the distribution is to take the common (base 10) or natural (base e) logarithm. Click on the right arrow again to take the common logarithm of the fruit mass values.
   a. How do the scatterplot and residual plot change with respect to shape? What is the new correlation coefficient?

Sample Answer: The scatterplot looks less curved than the scatterplot of the original data, but the residual plot still shows a curve. The new correlation coefficient is 0.9757, larger than that of the original data, but slightly smaller than the *r* for the square root transformation.

> **Teacher Tip:** You might ask students what difference they would expect to see if they used the natural logarithm, base *e*, rather than using the common logarithm base 10 for the transformation. It might be worthwhile to have them actually graph both transformations in a new file and compare the results, noticing that the linearity is not affected, just the coefficients.
>
> **Teacher Tip:** When data are plotted on along one axis using a linear scale, and a logarithmic scale is used on the other axis, it is sometimes called a semi-logarithmic transformation. (If both scales are logarithmic, you have a log-log transformation).

   b. Did this logarithmic transformation “fix” the problem of a non-linear relationship between fruit circumference and fruit mass? Explain your reasoning.

Sample Answer: No, even though the correlation coefficient is larger than that of the original data, the residual plot still shows a curved pattern (this time curved down), indicating there is not a linear relationship between fruit circumference and log(fruit mass).

5. a. Click on the right arrow again, and describe the new transformation.

Sample Answer: The transformation is taking the log(fruit circumference).

b. How do the scatterplot and residual plot change with respect to shape? What is the new correlation coefficient?

Sample Answer: Both the scatterplot and residual plot show strong curves, indicating a non-linear relationship between log(fruit circumference) and fruit mass. The correlation coefficient is also smaller, at $r = 0.8109$.

c. Do the transformed data have a linear relationship? Explain.

Sample Answer: No, the residual plot still shows a curved pattern, and the correlation coefficient is smaller than the one for the untransformed data, indicating that there is not a linear relationship between log(fruit circumference) and fruit mass.

6. Click on the right arrow again.

a. This is a log-log transformation used to check for a power relationship in the original data. Explain why the name makes sense.

Sample Answer: It makes sense to call the transformation log-log because you take the log of both sides of the equation.

b. How do the scatterplot and residual plot change with respect to shape? What is the new correlation coefficient? Did the log-log transformation "fix" the curve in the original data so that it would be possible to use this model for prediction – that is, do the transformed data have a linear relationship?

Sample Answer: The scatterplot looks much less curved than the scatterplot of the original data, and the residual plot has a random scatter. The new correlation coefficient is 0.9952. The log-log transformation transformed the values so that a linear model would be appropriate for prediction.

c. How would you write the equation for the least-squares regression line for the transformed data?

Answer: predicted log(mass) = -1.41 + 2.713 (log(circumference))

d. Use your equation to predict the mass of a spherical watermelon with a circumference of 44 cm and one with a circumference of 68 cm. Which prediction do you feel is more reliable, and why?

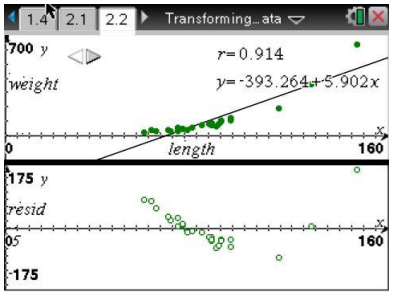
Sample Answer: using predicted log(mass) = -1.41 + 2.713 (log(circumference)),
predicted log(mass) = -1.41 + 2.713 (log(44)) = 3.04869 → mass = 1119 grams
predicted log(mass) = -1.41 + 2.713 (log(68)) = 3.5616 → mass = 3644 grams
The prediction for the watermelon with a circumference of 44 cm. seems more reliable because 44 cm. is within the range of circumferences in the data set. A circumference of 68 cm. is much larger than any of the values in the data set, and the regression model might not hold for fruit that large.

Move to page 2.1.

7. The data set on this page shows the lengths in inches and the weights in pounds for 25 alligators. What do you think a scatterplot of these bivariate data will look like?

Sample Answer: I do not think that the scatterplot of alligator data will be linear. I think there might be some kind of curvature.

Move to page 2.2.



Page 2.2 shows a scatterplot of the alligator weight vs. length with the least-squares regression line. The scatterplot on the lower screen is the residual plot.

8. Describe the distribution of alligator weight versus alligator length shown in the scatterplot with respect to shape. Does the distribution surprise you given what you saw in the spreadsheet?

Sample Answer: The distribution of alligator weight versus length is fairly strongly curved—possibly a quadratic or an exponential curve. The data values did not seem to follow a linear pattern, so this is not surprising.

9. a. Use the arrow in the upper panel to check four different transformations of the alligator data. Which transformation made the scatterplot of the data the most linear and the residual plot the most random? Explain your reasoning.

Sample Answer: Taking the logarithm of only the alligator weights produced the most linear scatterplot and random residual plot (i.e, (length, logweight). The scatterplot for the square root transformation looked linear, but its residual plot showed a strong curved pattern. Taking the logarithm of only the lengths produced a strongly curved pattern in both the scatterplot and residual plot. The log-log transformation (taking the logartithm of both lengths and weights) produced a scatterplot that looked linear, but the residual plot again showed some curvature.

b. How would you write the equation for the least-squares line for the transformation that seems best at eliminating the curvature in the alligator data?

Answer: predicted alligator (log(weight)) = 0.58 + 0.015 alligator length

c. Predict the weight of an alligator whose length is 140 cm.

Answer: Using the equation above, the weight of the alligator would be 7.98 pounds.

TI-Nspire Navigator Opportunity: *Quick Poll*
See Note 1 at the end of this lesson.

Wrap Up

Upon completion of the lesson, the teacher should ensure that students are able to understand:

- Bivariate data can be transformed in order to apply regression modeling procedures.
- Scatterplots and residual plots of different transformations of bivariate data can be used to determine which transformation is more effective at eliminating the curve.
- How to use models of transformed data to make predictions.

Assessment

Label each of the following as **sometimes**, **always**, or **never**. Be prepared to defend your answers.

1. The least-squares regression line can be used to make a reasonable prediction for a given explanatory variable.
   Answer: Sometimes true, if the relationship between the explanatory variable and either the response or the transformed response variable is approximately linear.

2. When statisticians analyze the relationship between two variables, it is important to have a linear relationship between the variables or transformed values of those variables.
   Answer: Always true because the statistical tools for analyzing the relationship are based on the assumption of linearity.

3. When you are transforming data to analyze the relationship, if you take the logarithm of the explanatory variable, you have to take the logarithm of the response variable.
   Sample Answer: False because you can take the logarithm of either (log transformations) or both variables (log-log transformations) as was done in the activity.

4. Square root transformations will reduce the curvature in bivariate data.
   Sample Answer: Sometimes true depending on the nature of the original relationship between the variables.

TI-Nspire Navigator

Note 1

Name of Feature: Quick Poll

Use Quick Poll to determine how well students understood the lesson. You might ask them to send responses to each of the questions in the Assessment above as well.

Data used in this activity:

Concept of using fruit (name of fruit, circumference in cm, mass in grams)--thanks to Tim Erickson of

http://www.eeps.com/zoo/index.html

Alligator data (length in inches and weight in pounds)--thanks to Richard Scheaffer