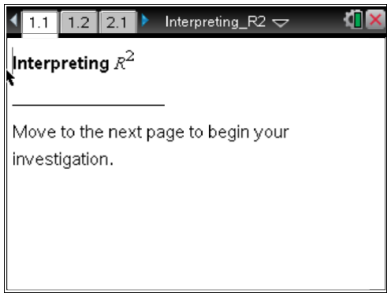
Interpreting R^2

Student Activity

Name _____

Class _____

Open the TI-Nspire document *Interpreting_R^2.tns*.



Statistical software often reports the equation of the least-squares regression line and a value called the coefficient of determination, or R^2. This activity explores what the R^2 value represents in terms of the relationship between the explanatory and the response variables.

Move to page 1.2.

Press ctrl ▶ and ctrl ◀ to navigate through the lesson.

Tip: If you have difficulty dragging a line, check to make sure that you have moved the cursor until it becomes a hand (✋) getting ready to grab the line. Also, be sure that the word *line* appears, not the word *text*. Then, press ctrl ✋ to grab the point and close the hand (✊).

A group of students at a particular high school were orally given the following instructions, with no additional information provided: "Write your name legibly, exactly twenty times. When you finish, note how long it took." The number of seconds that it took each student to complete this task was recorded. Page 1.2 displays a dot plot of the resulting data.

Suppose a student, say, Miss Terri, is to be selected at random from this group and you have been asked to predict the time that the Miss Terri took to complete the "name time" task. The "prediction" slider allows you to move a vertical line across the dot plot.

1. Position the line where you think it represents a good prediction of Miss Terri's time to complete the "name time" task. Explain how you selected where to place the line.

2. Compare your predicted value to that of a classmate. Devise a method for determining which prediction is better. Explain your thinking.

In the field of statistics, it is common to compare predicted values of some variable to the actual observed values from a particular sample. We often consider, or measure, how far the prediction "missed." This leads to the general concept of a **residual**, namely, the difference "Observed minus Predicted." Each observation in a data set has its own residual value which can be positive, negative, or zero.

3. Sketch the dot plot and prediction line you placed on Page 1.2 to mark your prediction of the time for the randomly-selected student. Suppose Miss Terri actually had a time larger than your prediction. Sketch an arrow to represent the residual for a chosen point higher than your prediction. Then suppose Miss Terri's actual time was less than your prediction. Sketch another arrow to represent this residual. Using your sketch, explain the meaning of the signs of residuals.

4. Based on your sketch and discussion with classmates, suggest a single calculation that can be used as a "score" to measure how well your prediction estimates the time for any subject in the sample. State whether, when using your method, a large "score" or a small score indicates a better prediction. Explain your reasoning.

Move to page 2.1.

It is possible that you and your classmates suggested methods for "scoring" predictions that somehow involved looking at "misses"—that is, residuals. The remainder of this activity explores ways to visualize residuals and the information they can convey.

5. A dot plot can be presented in several different ways. For example, the right panel of Page 2.1 shows a vertical dot plot of the same "time" data used on Page 1.2. Explain the meaning of the horizontal line in the right panel and why it might be useful.

6. Click once on the arrow in the right panel. Describe what happens to the plots, and explain the meanings of the segment and the point that were added.

7. Click the arrow in the right panel five more times to complete the display in the left panel. Notice that the residuals really are just numbers, so they may be treated as a data set in their own right. Compare the dot plot of the residual data on the left to the plot of the "time" data on the right. Be sure to compare centers, shapes, and spreads as you would for any other data sets.

The predictor used in the plot on Page 2.1 is the mean of the data. Using the mean guarantees that the mean of the residuals (how far the prediction misses) is exactly zero. There are several techniques to measure the reliability, or goodness, of a prediction. However, for the remainder of this activity, *variance* will be used. Recall that variance is an average of the squares of the distance each observation is from the mean, which is just the average of the squares of the residuals. Thus this measure utilizes every observed time's distance from the prediction.

8. If all of the students' times in the original data set had been identical, predicting the time for Miss Terri would be trivial. Remember, all of the students who took part in the "name time" activity received exactly the same instructions. Why do you think their writing times were <u>not</u> all exactly the same? Does one explanation seem more important than any other?

Move to page 3.1.

9. One possible explanation for the variability in the name-time data is that students' names were of different lengths. Better predictions might be possible if we consider the length of each student name. Click the arrow to include lengths of names (in number of letters) in the plot.
   a. Describe the plot in the right panel.

   b. How did the residuals change when you added the new variable? Why?

10. When you created the scatterplot, the displayed prediction remained the same. It's still the sample mean of the times. The prediction does not take into consideration any of the names' lengths.
    a. To use length of name in the prediction process, predicted time needs to change as length of name changes. What would you need to do to the graph in order to include that idea in the prediction line?

    b. The prediction line in the plot is "pinned" so that it contains the centroid, (mean letters, mean time), but it is free to rotate about that point. Grab the left end of the line, and rotate the line a small amount. Describe what happens to the residuals and what it means about predictions.

    c. What does the number labeled s^2 near the bottom of the residual plot indicate, and why is it important in thinking about predictions?

11. Based on your observations in question 10, what do you think would be the best line to use in predicting times from name lengths?

Move to page 4.1.

12. Click the arrow on Page 4.1 to toggle between two predictors—the mean of the times and the least-squares regression line (LSRL). Describe how the dot plot of the residuals changes, and record the values of the displayed numbers below.

    description: _____

    linear correlation coefficient (R): _____

    square of linear correlation coefficient (R^2): _____

    variance of residuals using mean of times: _____

    variance of residuals using LSRL: _____

Comment on the change in variability of predictions between the two methods.

#1

#2

Move to page 4.2.

One way to measure improvement in prediction is with the amount of decrease in variability, that is, the difference in s^2, when changing methods. But is a decrease in s^2 by, say, 500 units, an impressive gain? Not if the original value of s^2 was 1,000,000. A better measure is the **relative change**, defined as

$$\frac{\text{amount of change}}{\text{original amount}}$$

13. Use the calculator application on Page 4.2 to compute the relative decrease in variability of predictions when changing from predictions based on mean of times to predictions based on the LSRL. Then compare that relative decrease to other values you recorded in question 12.

This value, the relative decrease in variability of predictions using an equation based on a related variable instead of using just the mean of the response variable, is known as the **coefficient of determination** for that prediction equation. For the "name time" situation of this lesson, 64% of the variation in time is accounted for (removed) by the linear model relating time to length of name.