Math Objectives

- Students will predict a response variable using the mean of that response variable and recognize why the mean of the response variable minimizes the typical “error” in the prediction if no other information is available.
- Students will recognize that R^2 measures the relative improvement in precision in predicting a response variable using an additional (explanatory) variable via the least-squares regression line (over the precision obtained using only the mean of the response variable).
- Students will look for and make use of structure (CCSS Mathematical Practices).

Vocabulary

- centroid
- coefficient of determination
- dot plot
- linear correlation coefficient
- linear regression
- mean
- residual
- scatterplot
- variance

About the Lesson

- This lesson involves predicting values of a particular variable.
- As a result, students will:
  - Examine both univariate and bivariate displays involving a particular variable in order to visualize the errors made by different methods of predicting that variable.
  - Consider various measures that might be used to quantify the precision of a particular predictor of values of a given variable.
  - Use percent decrease in the adopted measure of precision as a measure of improvement from one method of predicting to another.
  - Make and verify a conjecture describing the amount of improvement in the precision of predicting a response variable when comparing the least-squares prediction line to predictions using only the mean of the response variable.

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

Interpreting_R^2_Student.pdf

Interpreting_R^2_Student.doc

TI-Nspire document

Interpreting_R^2.tns

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

TI-Nspire™ Navigator™ System

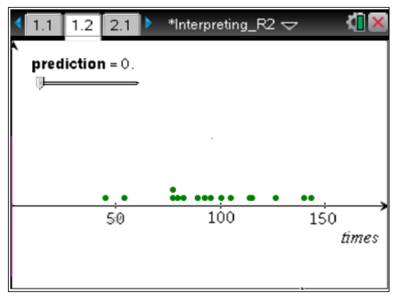
- Transfer a File.
- Use Screen Capture to compare different predictors.

Discussion Points and Possible Answers

> **Tech Tip:** If students experience difficulty dragging a line, check to make sure that they 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 ( ).

> **Teacher Tip:** Students should be familiar with variance, scatterplots, linear correlation, least-squares regression, and residuals.

Move to page 1.2.



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.

   <u>Sample Answers:</u> I would predict a time somewhere in the center of the dot plot so that no individual dot will be very far from my prediction—on average I would not be that far off. Maybe the mean or median would be a good prediction.

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

   <u>Sample Answers:</u> Student answers will vary. Some might say something like “If my largest miss is smaller than your largest miss, then my prediction is better.” Others might suggest methods based on calculating and comparing IQR, range, standard deviation, or variance. Still others might have more subjective methods, such as using the middle of the biggest cluster.

Teacher Tip: Discuss answers to question 2 with the class to ensure that students realize that "misses" are what matter in deciding between good and bad predictors.

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

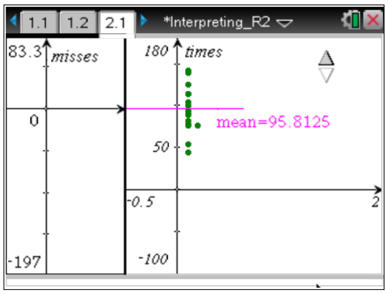
3. Sketch the dot plot and prediction line you placed on Page 1.2 to mark your prediction of the time for the randomly-selected student. Suppose Miss Terri actually had a time larger than your prediction. Sketch an arrow to represent the residual for a chosen point higher than your prediction. Then suppose 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.

   <u>Sample Answers:</u> Student sketch: Ideally arrows would go from prediction to data point. A positive residual means the Observed value is larger than the Predicted value, so the data point is to the right of the prediction. Similarly, a negative residual means the data point is to the left of the prediction.

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.

   <u>Sample Answers:</u> I would use the average absolute value (or average squared value) of all the residuals, with lower averages indicating better predictions. That would take every observation into consideration and would avoid having negative and positive residuals "cancel out."

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.

Sample Answers: It's the mean of the data (95.8125) and represents a predicted value for the time for the randomly-selected student, Miss Terri.

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.

   Sample Answers: A segment was drawn in the dot plot in the right panel from the prediction to one of the dots, and a dot was plotted in the left panel. The length of the segment represents the magnitude of the residual associated with that observation, and the dot in the left panel represents the residual's actual value.

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.

   Sample Answers: The spread and shape of the two plots are identical; but in the left panel, the center is now near 0 instead of near 96.

> **Teacher Tip:** It would be useful to start a class discussion of questions 1 and 4 at this point, setting up the choice of measures given below.

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.

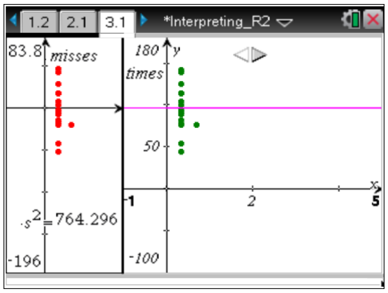
> **Teacher Tip:** Proving the claim that using the mean as the predictor forces the average residual to be zero can be a valuable algebraic exploration for interested students.

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?

   <u>Sample Answers:</u> Reasons might include length of name, which hand was used, whether the student numbered to 20 first or counted along the way, whether the student wrote long-hand or printed, etc. Length of name seems most important.

> **Teacher Tip:** Focus student attention here on variability. The reason the mean (or any other single number) cannot do very well as a predictor is that the data vary. The more variation, the poorer the prediction.

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.

   <u>Sample Answers:</u> It's a scatterplot of times versus length of name (letters). The line marking the mean of the times as the prediction and marking the corresponding residuals is still shown.

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

   <u>Sample Answers:</u> They did not change at all. That's because I did not change the prediction—it's still the mean of the times.

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?

    Sample Answers: I would need to make different predictions for different name lengths. Making the line slant along the trend of the plot instead of cutting across horizontally would be better.

    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.

    Sample Answers: As the line rotated, the residuals remain centered at zero, but their spread decreases. The typical misses are smaller than they were when predicting without using the lengths of names.

    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?

    Sample Answers: That number s^2 is the sample variance of the residuals. It indicates something about how large misses are, on average. Large s^2 means bigger misses, on average, than a small s^2.

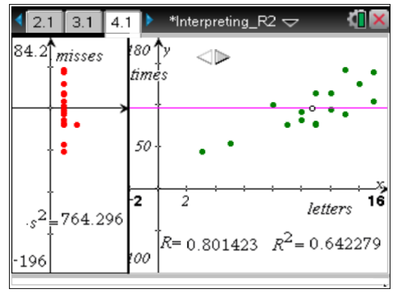
> **Teacher Tip:** Pinning the predictor at the centroid again forces the average residual to be zero. Proving this fact is a little more complicated, but it too can be a valuable algebraic exploration for interested students.

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

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?

    Sample Answers: : I think it might be the least-squares regression line. The least-squares line minimizes the sum of squared residuals, and s^2 is just that sum divided by (n-1).

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.

<u>Sample Answers:</u>

description: Changing from mean to the least squares regression line decreases the variability considerably

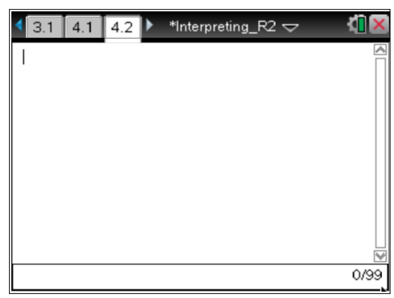
linear correlation coefficient (R): 0.801423

square of linear correlation coefficient (R^2): 0.642279

variance of residuals using mean of times: 764.296

variance of residuals using LSRL: 273.405

Move to page 4.2.



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.

<u>Sample Answers:</u> <u>:</u> To the precision of the displays, the relative decrease is 0.642279, which is exactly the same as R^2.

Teacher Tip: If the prediction equation is the LSRL, then the coefficient of determination is actually equal to r^2. For other prediction equations, the linear correlation coefficient, *r*, is meaningless, but the coefficient of determination's meaning is unchanged and is usually still denoted by R^2.

Wrap Up

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

- Using the mean to predict a response produces the best case error or minimum variability, as measured by variance, when no variable other than the response variable is available upon which to base predictions.
- Using the relationship between an explanatory variable to help make the prediction can reduce this error or variability.
- The coefficient of determination, R^2, is the relative decrease in variability, as measured by variance, in predicting a response variable when using the least-squares line, instead of just the mean of the response variable, as the predictor.

Assessment

Identify the following as sometimes true, always true, or never true. Be ready to justify your answer.

1. The mean is a good predictor of outcomes.

   Answer: Sometimes. It is the best you can have if you only know a set of outcomes and have no other data.

2. R^2 is the proportion of decrease in variability in predictions from the least-squares line determined by the relationship between an explanatory variable and the response variable as compared to predictions based only on the mean of the response variable.

   Answer: Always.

3. If $r^2 = 1$, then all of the variation in the responses are described completely by one explanatory variable.

   Answer: Always.

TI-Nspire Navigator

Note 1

Questions 2 & 10, Name of Feature: Screen Capture

A Screen Capture can be used to compare different prediction models across the class. This is appropriate not only with univariate models (at Question 2) but also with bivariate models (Question 10).