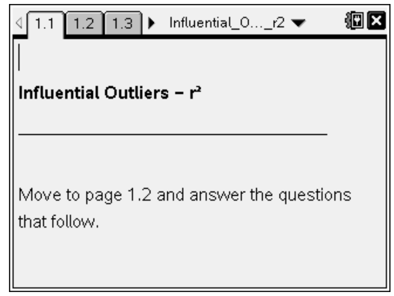
Influential Outliers – r^2

STUDENT ACTIVITY

In this lesson, students will identify outliers that are influential in respect to the variation of the data values. In addition, students will describe the role of the location of a point relative to the other data in determining whether that point has influence on the coefficient of determination.



- **Open the TI-Nspire document *Influential_Outliers_r2.tns*.**
- **Press ctrl ▶ and move to page 1.2 to begin the lesson.**

Page 1.2 displays a scatter plot of six points together with the associated linear regression line $\hat{y}^{1}$. The horizontal line $\bar{y}$ is the mean of the *y*-values of the data points.

The **explained** deviation is the vertical distance between the regression line and the mean of the *y*-values. It is represented by the bold line. The explained variation is the sum of the squares of the explained deviation values.

The **unexplained** deviation is the vertical distance between the data point and the regression line. It is represented by the dotted line. The unexplained variation is the sum of the squares of the unexplained deviation values.

1. Grab the rightmost point in the plot, (22, 25), and drag it vertically downward. As you move the point, observe the dotted and bold lines. What happens to the explained and unexplained deviations?

2. Drag the point in a circle around the other data points.

   a) When is the deviation for all of the points mostly explained?

   b) When is the deviation for all of the points mostly unexplained?

3. A point is called an **outlier** if it fails to fit the overall pattern of the set to which it belongs. How do your observations in Question 2 correspond to when the point is considered an outlier?

- **Move the point back to (22, 25).**

The **total variation** is equal to the explained variation plus the unexplained variation. This is shown in the bottom-left corner of the screen.

The **coefficient of determination**, r^2, is the explained variation divided by the total variation. This is shown in the bottom-right corner of the screen.

Currently, the value of r^2 is 1. This means that 100% of the total variation in *y* can be explained by the linear relationship between *x* and *y*.

4. Why do you think 100% of the variation can be explained?

____

____

5. What would it mean if 0% of the variation could be explained?

____

____

6. Drag the point in a circle around the other data points.

a) When is r^2 closest to a value of 1?

____

b) When is r^2 closest to a value of 0?

____

7. How do your observations from Question 6 correspond to when the point is considered an outlier?

____

____

8. Summarize in your own words the effect an outlier has on the coefficient of determination and the percent of total variation that can be explained by the linear relationship.

____

____