NUMB3RS Activity: Less is More
Episode: “Hardball”

Topic: Least Squares Regression **Grade Level:** 9 - 10

Objective: In this activity, students will explore the concept of a line of best fit by applying a trial-and-error approach to a given data set. Students will then learn about the basis of how a least-squares regression line is formed by analyzing a geometric model.

Time: 20 - 25 minutes

Materials: TI-83 Plus/TI-84 Plus graphing calculator with the Cabri™ Jr. application (Version 2.00), TI-Navigator™ system, activity settings file *Hardball.act*, lists *Homeruns* and *Strikeouts*, lists *Year* and *TotalHR*, and the Cabri Jr. application variable *LEASTSQ.8xv*.
To download these files, go to **http://education.ti.com/exchange** and search for “8021.”

Introduction

In “Hardball,” an amateur mathematician named Oswald Kittner creates a mathematical process that uses baseball statistics to identify players that are abusing steroids. The FBI, with Charlie’s assistance, is called in to protect Oswald from being the target of murder for his findings.

The “homerun explosion” and subsequent allegations of steroid abuse have become the subject of extensive media coverage over the last several years. The proliferation in total home runs first became noticeable after 1981, the last time fewer than 3,000 total home runs were hit in one season. In fact, the 1980s became known as the decade of the home run, hitting its peak in 1987 with 4,458 home runs. This trend has continued until the present day, reaching its highest mark in 2000, when 5,693 total home runs were hit.

People associated with the game of baseball are aware that power hitters also tend to strike out with greater frequency. In this activity, students will analyze a scatter plot of strikeouts versus home runs. The data in this activity was taken from a systematic sample of players during the 2006 major league baseball season.

In Part I of this activity, students will use Activity Center to experiment with different values of slope and *y*-intercept until they find a line that appears to fit the data well. Emphasis will be placed on thinking about “best-fit” informally.

In Part II of this activity, students will use the Cabri Jr. application to gain an understanding of the premise behind the least squares regression equation and how it can be modeled geometrically. The extension to this activity gives students an opportunity to calculate the least squares regression equation and the correlation coefficient. Students will also have an opportunity to study the *nonlinear* growth of home run totals from 1901 to present.

The homerun data came from **http://www.baseball-almanac.com/hitting/hihr6.shtml**.

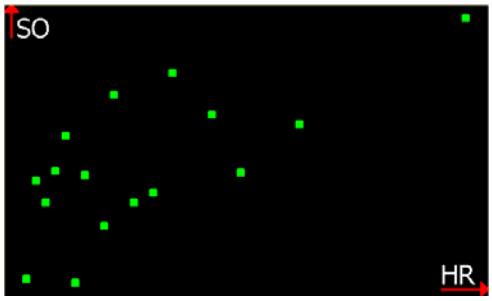
Steve Ouellette, Walpole High School, Walpole, MA

Part I: What is the Line of Best Fit?

1. a. Launch TI-Navigator on the computer and press **Begin Class** to start the session.

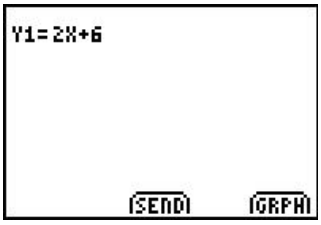
b. Have students log into NavNet on their calculators. You will want to log into NavNet as a teacher as well.

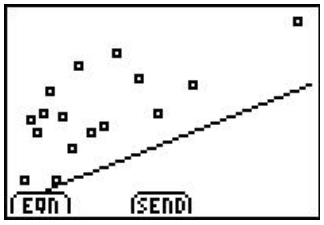
2. a. Load the ***Hardball.act*** activity settings file into Activity Center and click the "List – Graph" tab. Students will see a scatter plot of home runs (HR) versus strikeouts (SO) as shown below. They will also see the list data to the left of the scatter plot. Point out that there appears to be moderate positive correlation between home runs and strikeouts.



3. a. Press **Start Activity** and tell students to press **1:Activity Center**.

b. Explain to students that you want them to use a trial-and-error approach to try to find a line of best fit for the data. For example, in the first screen below, a student has entered the equation **Y1=2X+6**. By pressing GRPH, students can see the graph of the equation along with the scatter plot as shown in the second screen below. Students should press EQN to try another equation if necessary. They should then press SEND when they are ready to submit their graph to Activity Center.





c. Have a discussion with students about the results after all responses have been submitted to Activity Center. Part of this discussion should include gathering students' ideas about what comprises a "best-fit" line.

d. The regression equation produced by the calculator that models this data is $y = 2.09x + 53.76$. Enter this equation into your calculator and send it to Activity Center. The graph will appear green to differentiate it from the student graphs. Explain to students that this line is called the *Least Squares Regression* equation. Continue the discussion of a best-fit line, asking students if they agree that the regression equation is in fact the best line to model the data set.

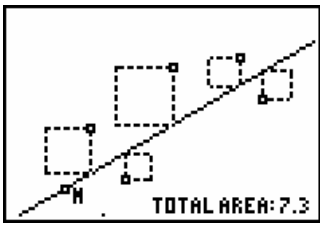
e. Press **Stop Activity**.

Part II: A Geometric Model of the Least Squares Regression Equation

1. a. Use **Send to Class** to send the Cabri™Jr. sketch *Leastsq.8xv* to the students.

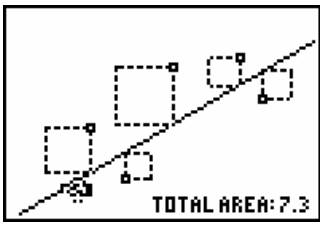
b. After students receive the file, instruct them to press **4:EXIT APP** to exit NavNet.

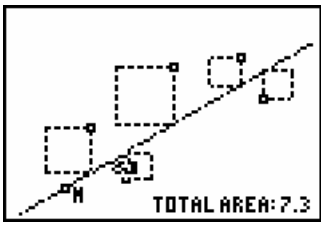
2. a. Have student open the Cabri Jr. application and open the sketch called *LEASTSQ*. Students will see the sketch below.



b. Explain to students that the line of best fit is called the *Least Squares Regression* equation because the sum of areas of the squares shown in the diagram must be minimized. Make sure that students understand that the side of each square is based on the vertical distance between each data point and the line. Also explain that it is important to square each of these vertical distances; otherwise positive and negative distances may “cancel” one another out when summed. The extension to this activity provides a link to an online example that reviews three different methods for finding a line of best fit, including a method using absolute values.

c. Instruct students to manipulate the line with the goal of making **TOTAL AREA** as small as possible. To manipulate the slope, move the cursor over to point **M** until it blinks, press the [ALPHA] key, and use the arrow keys ([▲], [▼], [◀], and [▶]). To manipulate the *y*-intercept, move the cursor on the line (but not near point **M**), press the [ALPHA] key, and press the arrow keys. Use **Screen Capture** to monitor student progress.





d. After some time, students should be able to obtain a minimum area of about 5.6. Instruct students to press [GRAPH] to access the **F5** menu, scroll down to **Coord. & Eq.**, press [ENTER], move the cursor to the line, and press [ENTER] again to reveal the equation of the line. Use **Screen Capture** again to view student results. Tell students that the *Least Squares Regression* line is approximately $y = 0.4x + 0.4$ and compare this equation with the equations found by the students.

e. Have students attempt to manipulate their line to match the least squares line and check (using **Screen Capture**) to see if they can adjust the line some more to further minimize the **TOTAL AREA** calculation (please note that this may be difficult due to limited screen resolution).

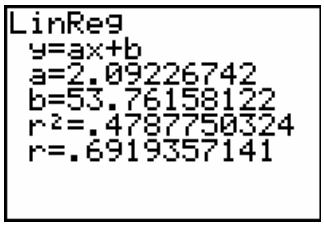
f. Students can also move the scatter plot points if you want them to repeat this experiment with a different scatter plot.

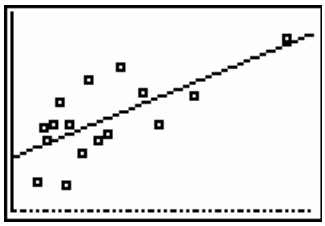
The goal of this activity is to give your students a short and simple snapshot into a very extensive math topic. TI and NCTM encourage you and your students to learn more about this topic using the extensions* provided *below and through your own independent research.

Extension

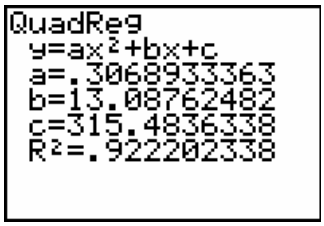
This activity is intended to provide students with a basic understanding of the premise behind how the graphing calculator determines a least squares regression equation. The first part of this extension is intended to teach students how to calculate the least square regression equation on their calculator.

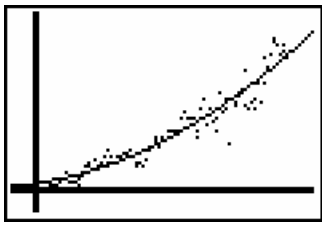
- Use **Send to Class** to send each student lists *Homeruns* and *Strikeouts*, which contain the data from Part I of this activity.
- Have students exit NavNet and run the command **DiagnosticOn**, found in the **Catalog** menu. This will display the *correlation coefficient r* and the *square of the correlation coefficient* r^2 when a linear regression is performed.
- Instruct students to perform a linear regression and graph the least squares regression equation along with a scatter plot of the data as shown below.





- Tell students that the closer $|r|$ is to 1, the better the fit. Furthermore, negative r values indicate negative correlation and positive r values indicate positive correlation. Finally, r^2, which is sometimes referred to as the *coefficient of determination*, helps explain how much of the behavior of the y-variable (strikeouts) can be attributed to the x-variable (homeruns). In the example above, 47.9% of the strikeout values can be explained by the regression line using homeruns at the predicting variable. The remaining 52.1% of the strikeout values is due to random chance or other variables other than homeruns.
- A graph of number of years since 1900 versus total home runs per season follows a *nonlinear* growth pattern as shown below.





This data is contained in the lists *Year* and *TotalHR* and can be sent down to students' calculators and analyzed in a similar manner. If students try an exponential regression, they will find that r^2 is 0.887 and therefore not as good a fit as the quadratic regression shown above.

- The least squares regression is the preferred method for determining a line of best fit. You can compare the least squares regression to two other methods at **http://standards.nctm.org/document/eexamples/chap7/7.4/index.htm**.
- If you would like to learn more about TI-Navigator™, visit **http://education.ti.com/navigator**.

Steve Ouellette, Walpole High School, Walpole, MA