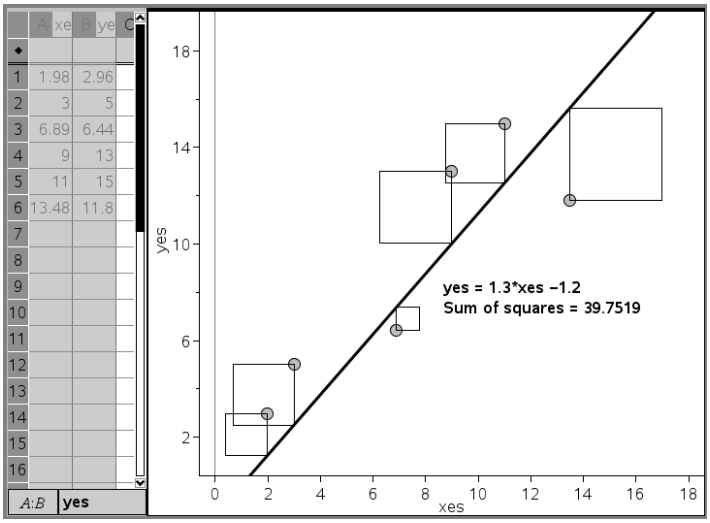
Opening the Black Box: **LinReg** Explained



John Hanna

T³ - Teachers Teaching with Technology

jehanna@optonline.net

www.johnhanna.us

Presented at the NCTM 2008 Annual Meeting

Salt Lake City, Utah

April 9-12, 2008

Opening the 'Black Box': LinReg Exposed

John Hanna
T³ - Teachers Teaching with Technology
April, 2006

Many teachers and students use the ***LinReg*** function of the Texas Instruments graphing calculators without ever delving into the ***whys*** of the algorithm. This document will explain the underlying algebra of the ***LinReg*** function and provide a graphical demonstration of the appropriateness of the algebraic results compared with the ***LinReg*** function. I also provide several Cabri II Plus dynamic geometry files for exploring the principles geometrically.

Here are some 'data points' stored in the lists L1 and L2:

L1	L2
2	1
5	4
8	6
3	2
1	6
4	9

Our mission is to determine the **Least Squares Line** for this dataset. This is the algorithm that ***Linear Regression (LinReg)*** implements. The algorithm finds the line that minimizes the *sum of the squares of the residuals* (the vertical distances from a line to the actual data points), or ***SSR***. See the Cabri II Plus file ***Least Squares Line*** which provides a geometric playground to explore the principle of minimizing this value.

The **Least Squares Line** is one of several *best-fit-lines* defined by mathematicians. It is not the only *best-fit-line.* There is not really a proof that this is the 'best line'. For example, the TI graphing calculators also have a ***Med-Med*** function on the Stat/Calc menu that produces another *best-fit-line*.

Part 1

We first *assume* (more about this later) that the least squares line passes through ***(xbar, ybar)***, the centroid* of the data points. We can find these coordinates by calculating:

$$\text{mean}(L1) \to xbar \qquad \text{mean}(L2) \to ybar$$

$$\frac{23}{6} \quad \text{and} \quad \frac{14}{3}$$

Now we define two functions:

Define the equation of the line through **(*xbar, ybar*).** This line has *two* independent variables; ***m*** is the slope of the line and, of course, ***x***.

$y(m, x) := m \cdot (x - xbar) + ybar$

"Done"

Define the '*sum of squared residuals*' (SSR) function , the sum of the squares of the vertical distances between the line defined above and the data points as a function of its slope:

$$SSE(m) := \sum_{i=1}^{\dim(L1)} \left(\left(y\left(m, L1_{[i]}\right) - L2_{[i]} \right)^2 \right)$$

"Done"

Here is the result of simplifying that '*sum of squared residuals* ' function:

$SSE(m)$

$$\frac{185 \cdot m^2}{6} - \frac{64 \cdot m}{3} + \frac{130}{3}$$

Notice that this function is merely quadratic in ***m***. Our goal is to 'minimize' this function. *(We could find the minimum of this function by simply writing it in 'vertex form' where the x-coordinate would be **-b/2a**, but I don't think there's a simple way of extracting the values of **a** and **b** from the quadratic above)* The following math box gives the value for ***m*** that produces the minimum value of **SSR*(m)*** and we store this result in the variable ***slope***:

$\text{right}(\text{fMin}(SSE(m), m)) \rightarrow slope$

$$\frac{64}{185}$$

$y(slope, x)$

Now let's look at our '**Least Squares Line**': $\frac{64 \cdot x}{185} + \frac{618}{185}$

==========

Since our line above might use fractions, we convert to decimals here to compare our line to the **LinReg** line

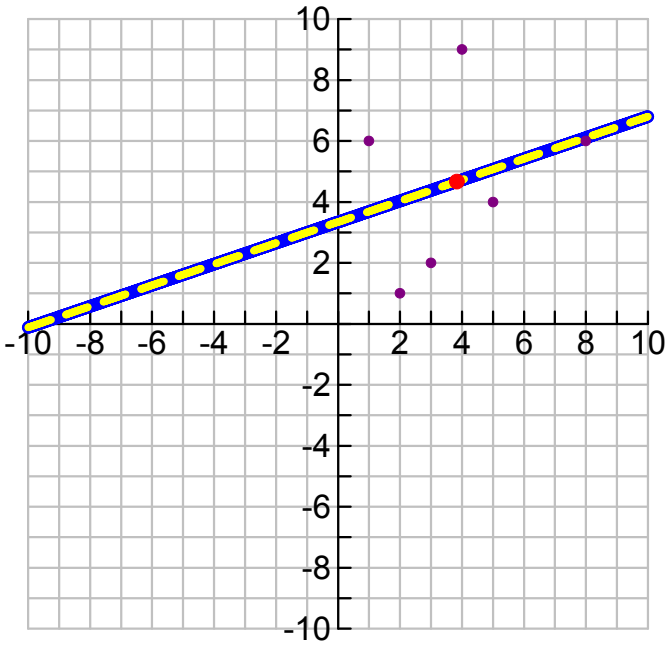
approx$(y(slope, x))$

$.345945945946 \cdot x + 3.34054054054$

Let's see how our line compares to the *Linear Regression* algorithm:

Linear Regression (ax+b)
regEQ(x) = .34594594595x + 3.3405405405

Not only does our algebra sometimes produce an equation with fraction (*exact*) coefficients, it may show even greater accuracy than the *Linear Regression* function in its decimal form! The graph below shows the data points (purple), the centroid* (red dot), and the two lines (one from the algebra and one from the Linear Regression) in blue and yellow:



Interactivity: Now go back to page 1, open the List Editor containing the data points, scroll back to here, and activate the List Editor while looking at this page. Change or add to the values in the two lists and watch the changes in the 'algebraic line' and the Linear Regression line.

Part 2:

Let's assume this time that we know the *slope* of the line (the variable ***slope*** from above) but do not assume that it contains (***xbar, ybar***)...

slope

$$\frac{64}{185}$$

...and want to find the *y-intercept* of the line instead. Our goal is to demonstrate that the resulting line will contain the centroid* of the data set (*xbar, ybar*).

First, define the line function:

define $y(x, b) = slope \cdot x + b$

"Done"

$y(x, b)$

$$\frac{64 \cdot x}{185} + b$$

, which looks like this:

Second, define a new *sum of squared residuals (SSRb)* function using this line:

define $\mathrm{SSRb}(b) = \sum_{i=1}^{\dim(l1)} \left(\left(y\left(L1_{[i]}, b\right) - L2_{[i]} \right)^2 \right)$

"Done"

Here's the *SSRb(b)* function simplified; another quadratic:

$\mathrm{SSRb}(b)$

$$6 \cdot b^2 - \frac{7416 \cdot b}{185} + \frac{3648334}{34225}$$

Again, we determine the value for *b* that produces the minimum for this function and store the value in the variable y_int:

$\mathrm{right}(\mathrm{fmin}(\mathrm{SSRb}(b), b)) \rightarrow y\_int$

$$\frac{618}{185}$$

And here is the equation of *this* **Least Squares Line**:

$y(x, y\_int)$

$$\frac{64 \cdot x}{185} + \frac{618}{185}$$

and here we confirm that (*xbar, ybar*) is on the line as it was before:

$y(xbar, y\_int)$ $\qquad$ $ybar$

$$\frac{14}{3} \qquad\qquad \frac{14}{3}$$

as it must be, **since it is the same equation**!

Notes:

CENTROID*: The term 'centroid' as used here is not fully accurate. If the points are assumed to be discrete 'atoms' (as they are in this application), then the average of the x's and the average of the y's produce a point which *does* represent the 'center of mass' property that the **centroid of a triangle** has, but if the data points represent the vertices of a laminar, or 'solid' sheet, like a polygon made out of cardboard, then the point (*xbar, ybar*) is not the 'centroid' (center of mass) in the traditional, triangular sense and should not be called the 'centroid'.

The choice of using (***xbar, ybar***) as a starting point for the 'Least Squares Line' is also motivated by Gloria Barrett, who examines the *sum-of-residuals* of various lines. This preliminary exploration leads to the conclusion that *any* line that contains the point (***xbar, ybar***) produces a *sum-of-residuals* value of 0. Hence, it is appropriate in our quest to find a 'best-fit-line' to first require that our line must satisfy this condition. Pretty reasonable, eh? See the Cabri file ***Least Squares Line - sum of residuals***. Here is the algebraic demonstration:

m

m See that the variable m is undefined.

We define the equation of the line through (***xbar, ybar***) again:

$y(m, x) := m \cdot (x - xbar) + ybar$

"Done"

and define the sum of residuals (**SR**) function:

$$\text{define SR}(x) = \sum_{i=1}^{\dim(L1)} \left(y(m, L1_{[i]}) - L2_{[i]}\right)$$

"Done"

$\text{SR}(x)$

Notice that **0** (regardless of m)!

$y(m, xbar)$ $ybar$

and the line's value at ***xbar*** $\frac{14}{3}$ is $\frac{14}{3}$.

References:

Vonder Embse, Charles, *Exploring Regression Concepts with the TI-92.* Central Michigan University, 2000

Barrett, Gloria, email, 2006