LinReg Exposed!

John Hanna, March, 2014

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 dynamic constructions for exploring the principles geometrically.

To the right are some 'data points' stored in the lists **L1** and **L2.** 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 errors* (the vertical distances from a line to the actual data points, AKA "residuals"), or SSE.

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, there is also a Med−Med function on graphing calculators that produces another *best−fit−line*.

	A l1	B l2	C
=			
1	2	1	
2	5	4	
3	8	6	
4	3	2	
5	1	6	
6	4	9	
7			
8			

A1 2

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

$\textbf{xbar}:=\text{mean}(\textbf{l1}) \blacktriangleright \frac{23}{6}$ and $\textbf{ybar}:=\text{mean}(\textbf{l2}) \blacktriangleright \frac{14}{3}$

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

$\textbf{y}(m,x):=m\cdot(x-\textbf{xbar})+\textbf{ybar} \blacktriangleright \textit{Done}$

Define the 'sum of squared errors' (SSE) 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:

$$\textbf{sse}(m):=\sum_{i=1}^{\dim(\textbf{l1})}\left(\left(\textbf{y}(m,\textbf{l1}[i])-\textbf{l2}[i]\right)^2\right) \blacktriangleright \textit{Done}$$

$$\textbf{sse}(m) \blacktriangleright \frac{185\cdot m^2}{6}-\frac{64\cdot m}{3}+\frac{130}{3}$$

Notice that this function is *merely* quadratic in **m**. Our goal is to 'minimize' this function. There are several ways to do this. I choose...

$$\text{completeSquare}(\textbf{sse}(m),m) \blacktriangleright \frac{185\cdot\left(m-\frac{64}{185}\right)^2}{6}+\frac{7334}{185}$$ is the *Vertex Form* of the function so the minimum occurs at

$$\textbf{slope}:=\frac{64}{185} \blacktriangleright \frac{64}{185}$$

Now let's look at our 'Least Squares Line':

$\mathbf{y}(\mathbf{slope},x) \blacktriangleright \frac{64 \cdot x}{185}+\frac{618}{185}$

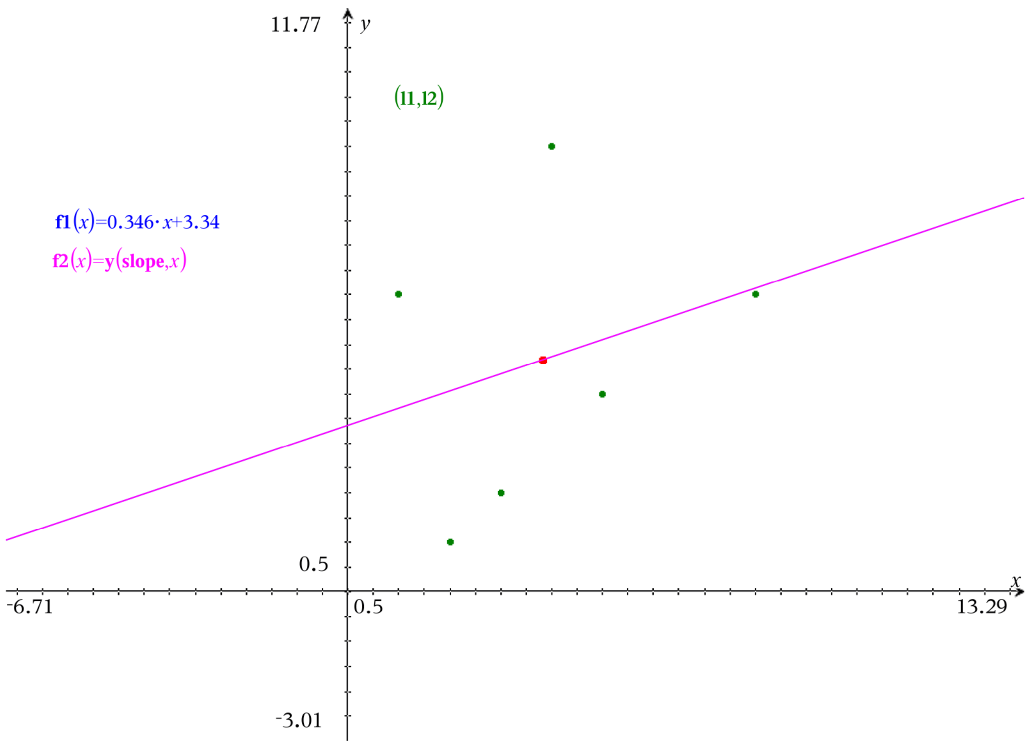
Since our line above might use fractions, we convert to decimals here to compare our line to the LinReg line

$\text{approx}\left(\mathbf{y}(\mathbf{slope},x)\right) \blacktriangleright 0.345946 \cdot x+3.34054$

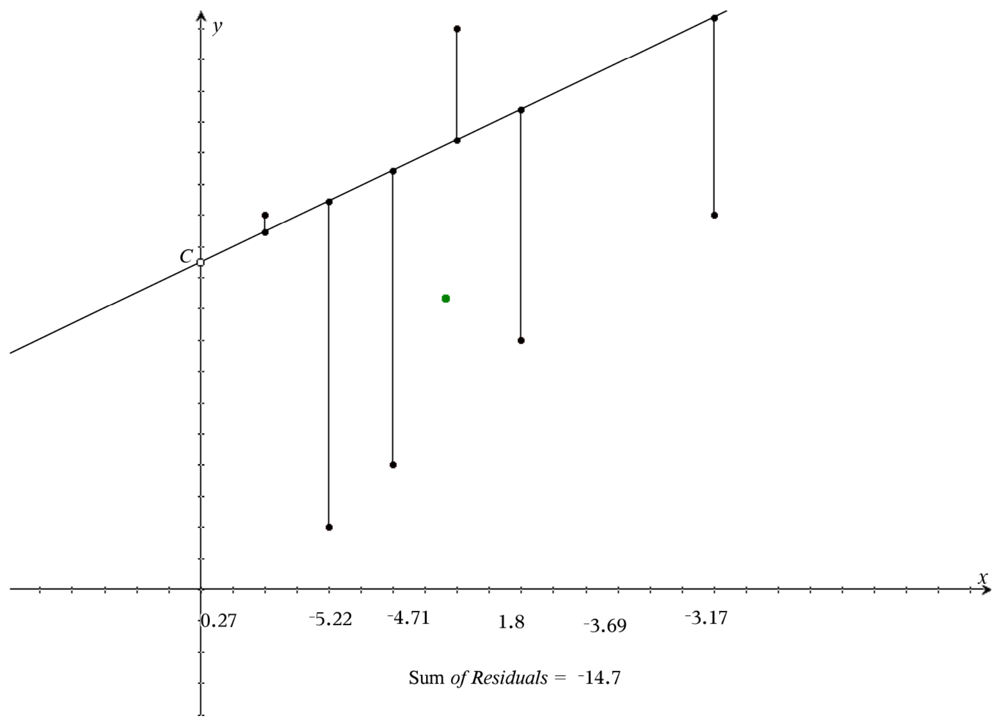
and compare it to

LinRegMx **l1**,**l2**,1: CopyVar **stat.RegEqn**,**f1**: **stat.RegEqn** $(x) \blacktriangleright 0.345946 \cdot x+3.34054$

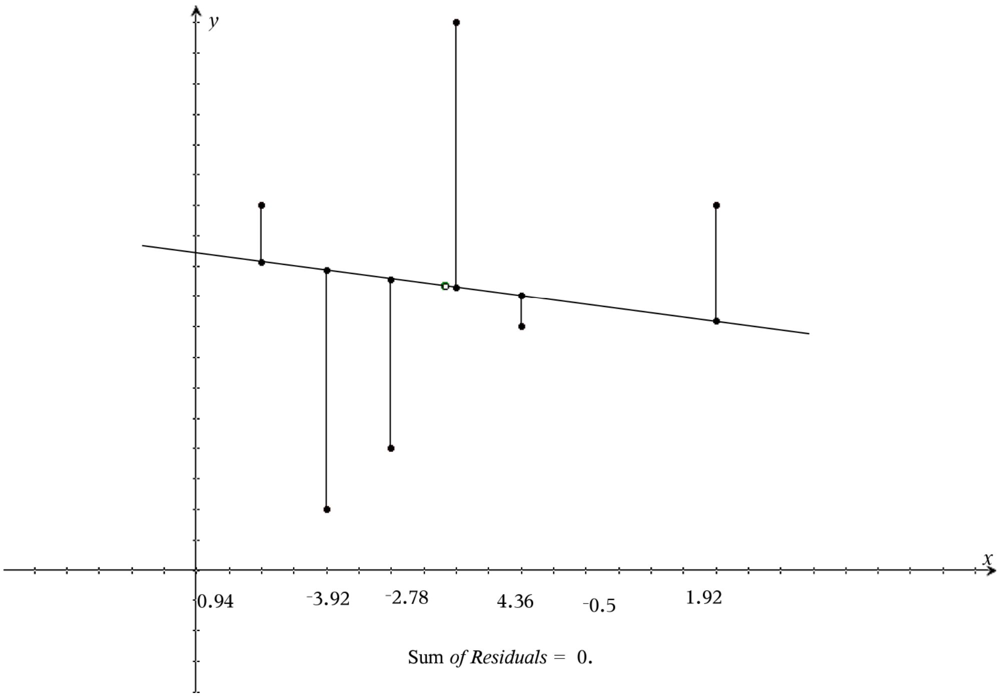
Happily, the results are the same☺



Why should the Least Squares Line go through (xbar, ybar)? Study the next two constructions...



The residuals are constructed and measured. Drag point C so that the Sum of Residuals is as small as possible. The green point is **(xbar, ybar).** Try rotating the line, too.



The line now goes through **(xbar, ybar).** Rotate the line and observe the calculated sum of residuals value.

Theorem: any line through **(xbar,ybar)** results in a Sum of Residuals = 0

Proof

$$\sum_{i=1}^{\dim(l1)} \left(y(m,l1[i])-l2[i]\right) = 0$$

References

Vonder Embse, Dr. CharlesCharles, Exploring Regression Concepts with the TI−92. Central Michigan University, 2000

Barrett, Gloria, email, 2006