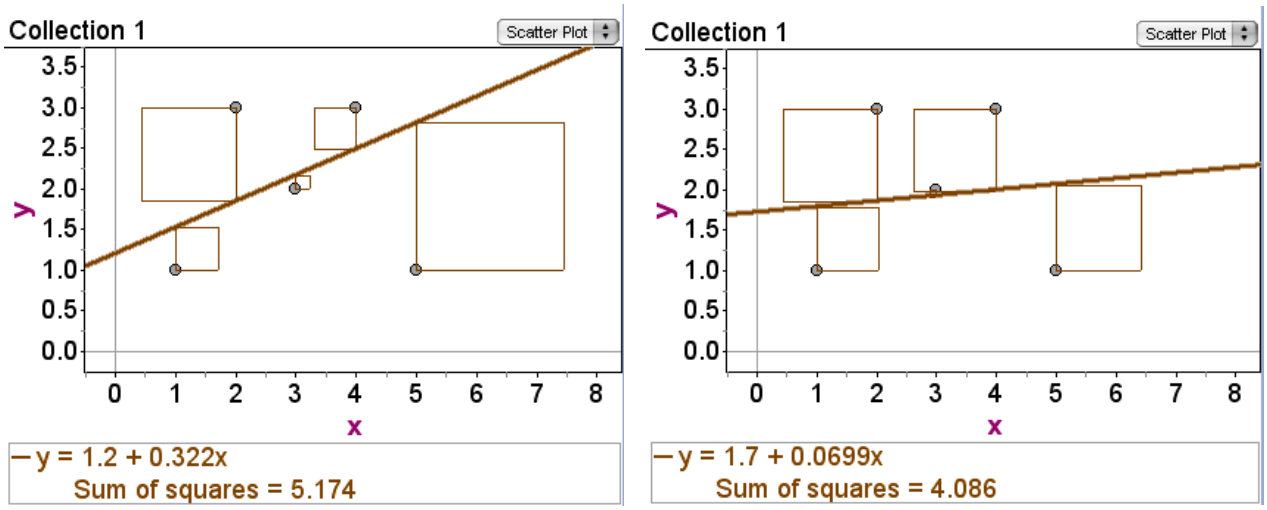
MY FAVORITE LESSON: FINDING THE LEAST SQUARES REGRESSION LINE BY HAND

Linear regression using the method of least squares is a large part of any AP Statistics course, and we often rely on technology to give students a magical solution without having them understand exactly what happened or why *that* line in the calculator is be the best fit for the data. If you're anything like me, you often introduce the topic to students by first explaining residuals and then talking about how we'd like to minimize the sum of their squares. This activity will take students through the mathematics that lie under the hood of the calculator when they press the LINREG button. Engaging with this analysis will strengthen students' analytic skills and give them a taste of higher-level mathematics while reinforcing their understanding of what it means to find a Least Squares Regression Line. Activity sheets are provided at the end of this document.

A variety of software packages will demonstrate the construction of these squared residuals and let students play around with the slope and y-intercept in the hopes of finding a line that generates the *least* amount of squares possible. The figure below shows how Fathom dynamically constructs the squares of the residuals from a given set of points and an arbitrary, moveable line.



The line on the right is a better fit because it has a smaller sum of squared residuals.

Unfortunately, at this point I often direct my students to the **LINREG** command on their calculators and do some hand waving about how the line of best fit is actually found. I say things like, "Trust me--the calculator has found *the* magic combination of slope and y-intercept that *guarantees* the smallest sum of

squares." Students can verify easily enough that, indeed, the calculator's line does seem to give a smaller amount of squared error than any other line students could come up with. But that's not a proof, and my students know it.

So how do we know that the calculator's regression equation is the best we can do? I was never very satisfied with this "let the calculator do the math" approach to regression; and trying to show my students the obscure-looking summation formulas for y-intercept and slope didn't help. But it doesn't have to be this way! Your students can do this math themselves!

Telling students that $b = \dfrac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$ or that $b = r\dfrac{s_y}{s_x}$ don't help!

BUILD THE EQUATION FOR THE SUM OF SQUARES

Let's take a very simple set of data, say (1,1), (2,3), (3,2), (4,3), and (5,1), and work out the sum of squares. We don't know what the LSRL will be, so let's just go ahead and call it $\hat{y} = a + bx$ and leave the values of *a* and *b* as unknowns. You've already talked to your students about how a residual is the difference between the observed y and the predicted $\hat{y}$, so applying that to the five data points isn't that big of a stretch for them.

$$\begin{aligned} RESID &= y - \hat{y} \\ &= y - (a + bx) \end{aligned}$$

Our first data point is (1, 1), so the residual would become $y - (a + bx) = 1 - (a + b \cdot 1)$. Squaring this difference results in the area of the square associated with any line with *y*-intercept *a* and slope *b*.

$$\begin{aligned} RESID^2 &= \left(y - \hat{y}\right)^2 \\ &= \left(1 - (a + b \cdot 1)\right)^2 = \left(1 - a - b\right)^2 \\ &= a^2 + b^2 - 2a - 2b + 2ab + 1 \end{aligned}$$

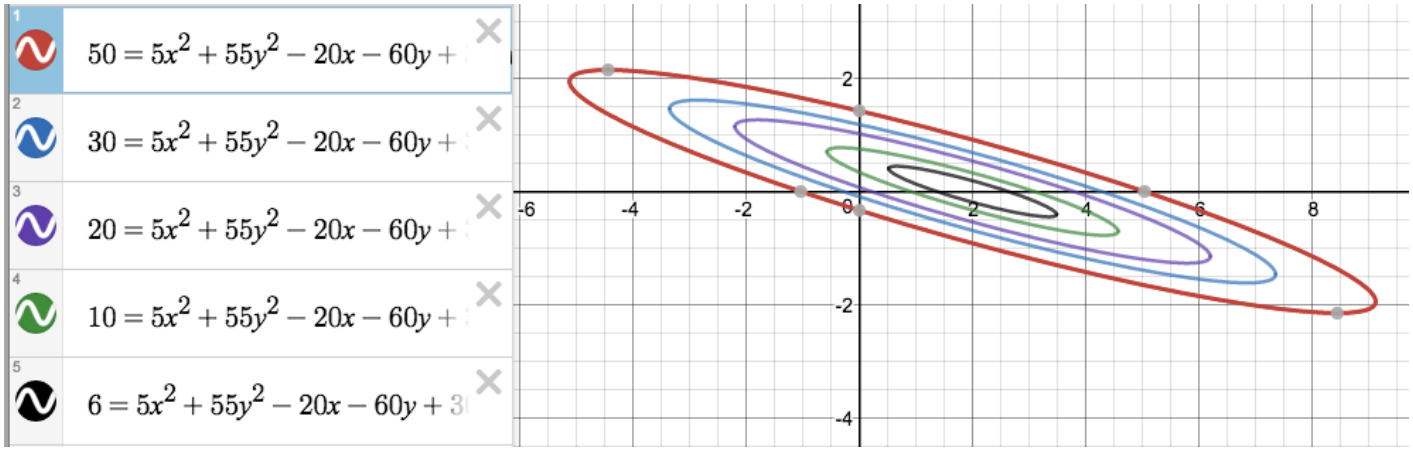
We repeat this calculation for each of the five points, and write the equation for the total sum of squares of the model. This equation is a function of the unknown *y*-intercept and slope values, *a* and *b*.

$$\text{SUM}_{a,\,b} = 5a^2 + 55b^2 - 20a - 60b + 30ab + 24$$

ANALYZE THE EQUATION

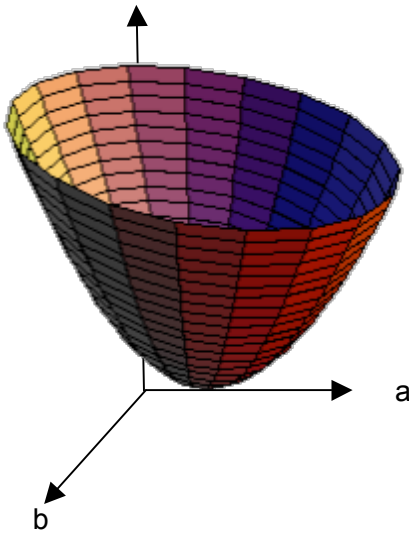
We now have a measure for the sum of squares associated with any line $\hat{y} = a + bx$ we choose. It's helpful here to let students experiment with different values of *a* and *b* to see how low they can get the sum. For instance, the line $\hat{y} = 3 + 1x$ (letting *a* = 3 and *b* = 1) would yield 94 units^2 for the sum of squares. Different choices for *a* and *b* will yield different values for the sum of squares. My students have a fun time competing with each other, trying to get a lower sum of squares than their peers. Some may realize that they can win this competition by having their calculator find the LSRL for them. Those values will yield the lowest possible sum, and they're the values that we calculate by hand in just a minute!

Our equation for the sum of squares is a lot of fun to experiment with on a graphing platform like Desmos. Either individually or as a whole-group discussion, students can ask, "Can the sum of squares be equal to 50?" Interesting, there is a whole set of *a* and *b* values that can generate 50 as the sum of squares: for instance, the equations $\hat{y} = -5 + 1.746x$ as well as $\hat{y} = 1.683 + 1x$ both yield 50 for their sum of squares. Can we get the sum to equal 40? How low can we go? The solution sets for these equations are ellipses on a diagonal axis—something that's appealing to students who have only seen conics oriented along the major axes. Students may realize that these ellipses are all centered at the same point, (2,0), which happens to be the solution to this particular set of data. Namely, the LSRL is when *a* = 2 and *b* = 0, or $\hat{y} = 2 + 0x$. Students may also realize that no solutions occur if you try to set the sum of squares below the value 4.



Solutions to $S_{a,\,b} = 5a^2 + 55b^2 - 20a - 60b + 30ab + 24$ for several values

MINIMIZE THE SUM OF SQUARES



If students can begin to view these ellipses as slices of a three-dimensional surface, the analysis can open up into a 3-dimensional plot. The equation for sum of squares is an elliptic paraboloid, and we'll locate the (*a*, *b*) coordinates of the bottom point of the surface. Depending on the background of your students, we can differentiate finding this solution for students who have knowledge of derivatives as well as for those who don't.

The Derivative Approach:

To minimize $S_{a,b} = 5a^2 + 55b^2 - 20a - 60b + 30ab + 24$ we can take the partial derivatives of the function with respect to both *a* and *b*.

With respect to *a*: $\frac{\partial}{\partial a} 5a^2 + 55b^2 - 20a - 60b + 30ab + 24 = 10a - 20 + 30b$

With respect to *b*: $\frac{\partial}{\partial a} 5a^2 + 55b^2 - 20a - 60b + 30ab + 24 = 110b - 60 + 30a$

We need the minimum in both dimensions, so we set both derivatives to zero, and simultaneously solve them for *a* and *b*. The solution to that system, *a* = 2 and *b* = 0, tells us that the least squares regression line is $\hat{y} = 2 + 0x$.

The Algebra 2 Approach:

For younger students who have not yet learned derivatives, they can apply their knowledge of quadratics to arrive at the same conclusion. They will rewrite the function as a quadratic in each dimension, then use the fact that the vertex of any quadratic $Ax^2 + Bx + C$ is located at $x = -B/(2A)$.

Written as a quadratic in *a*: $S_{a,b} = (5)a^2 + (-20 + 30b)a + (55b^2 - 60b + 24)$

which has a vertex at $a = \frac{-(-20 + 30b)}{2 \cdot 5}$

Written as a quadratic in *b*: $S_{a,b} = (55)b^2 + (-60 + 30a)b + (5a^2 - 20a + 24)$

which has a vertex at $b = \frac{-(-60 + 30a)}{2 \cdot 55}$

Again, solving this system of equations yields *a* = 2 and *b* = 0, which tells us that the least squares regression line is $\hat{y} = 2 + 0x$.

Overall, this is an excellent and engaging trip into higher-level mathematics that are accessible to your AP Statistics students, regardless of their previous math experience. Activity sheets are attached, and if you choose to do this with your students, I'd love to hear how it went. Send me an email if you have ways it could be improved or other thoughts! Thanks for reading :)

David Custer

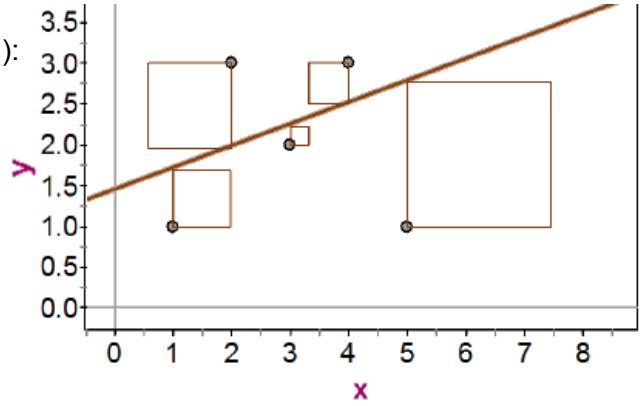
Mathematics Department Chair and ITL
NCTM Department Editor for *Mathematics Teacher*
Decatur High School
310 N. McDonough St.
Decatur, GA 30030
dcuster@csdecatur.net
404.370.4170

Getting the LSRL by hand:

I. Take some points: (1,1), (2,3), (3,2), (4,3), and (5,1):

II. Put an arbitrary line on the points $\hat{y} = a + bx$

we will calculate the specific *a,b* combination that will give us the *least* amount of squarage



III. Recalling that Residuals = $y - \hat{y}$

Let's write an equation for the Sum of Squares in terms of our unknowns, *a* and *b.*

$$\textbf{SUM} = (1-(a+b\cdot 1))^2 + (3-(a+b\cdot 2))^2 + (2-(a+b\cdot 3))^2 + (3-(a+b\cdot 4))^2 + (1-(a+b\cdot 5))^2$$

$$= (1-a-b)^2 + (3-a-2b)^2 + (2-a-3b)^2 + (3-a-4b)^2 + (1-a-5b)^2$$

$$\textbf{SUM}_{a,b} = \boxed{5a^2 + 55b^2 - 20a - 60b + 30ab + 24}$$

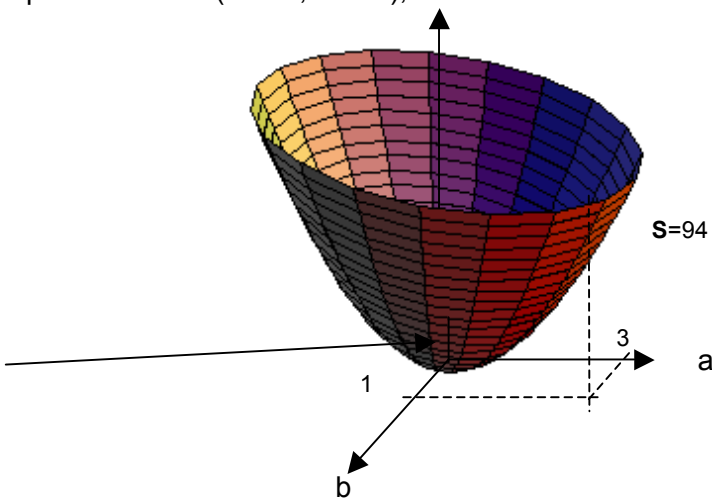
IV. So, for any choice of *a* and *b*, you'll get a different **SUM OF SQUARES**, as defined by that crazy equation above. The equation makes an elliptic paraboloid -- a 3D parabola ☺

i.e. if you used $\hat{y} = 3 + 1x$ as your prediction line (*a*=3, *b*=1), then **S** would be 94.

i.e. if you used $\hat{y} = 1.5 + 0.5x$ as your prediction line (*a*=1.5, *b*=0.5), then **S** would be 11.5.

Can we find an (*a,b*) that does better?

what are the (a,b) coordinates of the minimum?



V. If you know about derivatives, finding minimums is a pretty easy task:

- take the derivative with respect to a;
- take the derivative with respect to b;
- set both derivatives equal to zero and solve the system… you just found the minimum.

if you don't know about derivatives… flip this sheet over and we'll talk about how to get the same result with Algebra

SUM $_{a,b}$ = $\boxed{5a^2 + 55b^2 - 20a - 60b + 30ab + 24}$

VI. Fortunately, since we all passed Alg2, we know how to find the minimum of a parabola in 2D. The minimum of a parabola $y = ax^2 + bx + c$ is at $x = \frac{-b}{2a}$. So let's write that SUM OF SQUARES equation as a quadratic.

SUM$_a$ = _____a^2 + _____a + _____

So the minimum *a* =

SUM$_b$ = _____b^2 + _____b + _____

So the minimum *b* =

VII. Solve that system of equations to see that (a,b) = (2,0). The Line of Best Fit is $\hat{y} = 2 + 0x$. The *least amount of squares* possible is **S = 4**.

PRACTICE:

Repeat the math and find the LSRL by hand for your own set of points. Check it with your calculator to see if you're right

Let x = {the first 3 letters of your first name turned into numbers}
Let y = {the first 3 letters of your last name tuned into numbers}

i.e. **DAV-**id **CUS-**ter
x = {4,1,22) y={3,21,19)