LINES OF BEST FIT – FACT AND FICTION

Dave Bock (bock@htva.net)

	L1	L2
	Height (inches)	**Weight** (pounds)
	66	147
	62	124
	71	189
	64	141
	75	172
	70	144
	64	112
	71	165
	69	129
	68	177
Mean	68	150
St. dev.	4	25.05
Correlation	$r = 0.702$	

Typical Exam Question: The table displays the heights and weights of ten students in the Math Club at East Percent High School. Two other club members were absent when these data were collected.

A. One of the absent students is 60 inches tall. Estimate her weight.
B. The other absent student weighs 200 pounds. Estimate his height.

Part A: 3 solutions

(1) USING THE CALCULATOR:

`LinReg L1,L2` gives us the equation $\hat{y} = 4.4x - 149$

Now we predict: $\hat{y} = 4.4(60) - 149 = 115$ pounds

(2) BY REGRESSION:

The absent student is 2 standard deviations below the mean height.
We expect her to be 0.7(2) = 1.4 SD's below the mean weight.
We estimate her weight to be 150 – 1.4(25) = 115 pounds.

(3) USING THE POINT-SLOPE LOBF, where x = Height and y = Weight:

$slope = \frac{rS_y}{S_x} = \frac{0.702(25.0466)}{4} \approx 4.396$, so the equation is $\hat{y} = 4.396x + b$

To pass through the point (68,150): $150 = 4.396(68) + b$, so $b \approx -148.9$

And (again) the equation is: $\hat{Wt} = 4.396Ht - 148.9$, predicting 115 pounds.

Part B: The expected (but WRONG) "solution"

Use the same equation, $\hat{y} = 4.4x - 149$, let y= 200, and solve for x: $200 = 4.4x - 149 \Rightarrow x = 79$ inches.
WHAT? The prediction seems silly – 6'7" is far too tall. In fact, 79 inches is nearly 3 standard deviations above the mean, hardly a cautious or reasonable guess.

Part B: 3 CORRECT solutions

(1) BY REGRESSION:

The absent student is 2 standard deviations above the mean weight.
We expect him to be 0.7(2) = 1.4 SD's above the mean height.
We estimate his height to be 68 + 1.4(4) = 73.6 inches.

(2) USING THE POINT-SLOPE LOBF:

The correct $slope = \frac{rS_x}{S_y} = \frac{0.702(4)}{25.0466} \approx 0.112$, so the equation is $\hat{x} = 0.112y + b$

To pass through the point (68,150): $68 = 0.112(150) + b$, so $b \approx 51.2$

And the equation is: $\hat{Ht} = 0.112Wt + 51.2$, predicting $\hat{Ht} = 0.112(200) + 51.2 = 73.6$ inches.

(3) USING THE CALCULATOR:

`LinReg L2,L1` gives us the correct equation $\hat{x} = 0.112y + 51.2$, predicting a height of 73.6".

Derivation of the Equation for Least Squares Line of Best Fit

(Adapted from *Stats: Modeling the World* and the accompanying *Resource Guide*.)

Step 1: Preparation. Here are some basic facts we'll use in the derivation.

(1) The mean of any set of z-scores is 0, and therefore the sum is also 0. $\bar{z} = 0$ and $\sum z = 0$

(2) The standard deviation of a set of z-scores is 1, and therefore the variance is also 1. $\frac{\sum (z_y - \bar{z}_y)^2}{n-1} = \frac{\sum (z_y - 0)^2}{n-1} = \frac{\sum {z_y}^2}{n-1} = 1$

(3) Correlation is defined in terms of z-scores. $r = \frac{\sum z_x z_y}{n-1}$

Step 2: Find the *y*-intercept. Recall that the LOBF minimizes the sum of squared residuals.

We seek the line $\hat{z}_y = a + mz_x$ that minimizes $\sum \left[z_y - \hat{z}_y \right]^2$

Substitute the equation of the line: $\sum \left[z_y - (a + mz_x) \right]^2$

Rearrange terms: $\sum \left[(z_y - mz_x) - a \right]^2$

Square the binomial: $\sum \left[(z_y - mz_x)^2 - 2a(z_y - mz_x) + a^2 \right]$

Consider the "middle term"; by (1): $\sum \left[2a(z_y - mz_x) \right] = 2a \sum z_y - 2am \sum z_x = 0$

Now we seek to minimize what's left: $\sum \left[(z_y - mz_x)^2 + a^2 \right]$

By choosing $a = 0$ we can be sure that the sum will be minimal. (Adding the square of any other value would make it bigger.) Hence the LOBF must have a *y*-intercept of 0. In the standardized plane the line passes through the origin; in general, it goes through the mean-mean point $(\bar{x}, \bar{y})$.

Step 3: Find the slope. Because the line passes through the origin, its equation will be of the form $\hat{z}_y = mz_x$. We seek the value for m that will minimize the sum of the squared residuals. Actually we'll divide that sum by $n - 1$ and minimize this "mean squared residual", or MSR. Here goes:

Minimize: $MSR = \frac{\sum (z_y - \hat{z}_y)^2}{n-1}$

Since $\hat{z}_y = mz_x$: $MSR = \frac{\sum (z_y - mz_x)^2}{n-1}$

Square the binomial: $MSR = \frac{\sum ({z_y}^2 - 2mz_x z_y + m^2 {z_x}^2)}{n-1}$

Rewrite the summation: $MSR = \frac{\sum {z_y}^2}{n-1} - 2m \frac{\sum z_x z_y}{n-1} + m^2 \frac{\sum {z_x}^2}{n-1}$

Substitute; by (2) and (3): $MSR = 1 - 2mr + m^2$

Wow! That simplified nicely! We see that the *MSR* is a quadratic function of m. Remember that a parabola in the form $y = ax^2 + bx + c$ reaches its minimum at its turning point, occurring when $x = \frac{-b}{2a}$. Hence, we can minimize the mean of squared residuals by choosing $m = \frac{-(-2r)}{2(1)} = r$.

Wow, again! The slope of the line of best fit for z-scores is the correlation, r.