# LSRL

**QUESTION:**

Why does the AP Stats curriculum have students determine a LSRL in the form y = a + bx rather than y = ax + b (which is more parallel to their y = mx + b Algebra background)?

**ANSWERS:**

1) Statisticians prefer y=a+bx, or better still, y=a0+a1x (a-subscript-0 and a-subscript-1). The reason is that a model gets richer as terms are added, and it makes more sense to add terms to the right into empty space than to squeeze them in the front and push everything else out. For example, a quadratic model might be y=a0 + a1x + a2x^2. Or a multi-variate model with three predictive variables x1, x2, and x3 might be y=a0 + a1x1 + a2x2 + a3x3. The a0 is a kind of "base prediction" of y, what you would predict y to be in the absence of any other predictive variables at all. (Given a set of data, the mean of the observed y's would be a reasonable estimate of a0.) Then you add terms as you add explanatory variables. And note that with subscript notation, you don't have to worry about running out of letters of the alphabet!

2) Because in statistics we are eventually going to worry about the multivariable case where we have y dependent on a number of variables rather than just one as in: y = a + b-sub1 x-sub1 + b-sub2 x-sub2 + b-sub3 x-sub3 + . . . Note that really 'a'should be b-sub0. Each of the x-subi's actually will then have a second subscript within that variable for each instance of the variable.

AP stats does not cover the multi-variable situation other than to teach the students to think about potentially confounding or lurking variables when setting up an experiment or analyzing a study. But we don't actually teach them how to attempt to tease out the different influences of the different variables. For example, age and gender and weight wrt height all have an influence on resting pulse -- no one of them completely explains the variation in pulse rate (for that matter, all three of them don't completely explain it, but they come a lot closer than any one alone). So you would have a (or b-sub0) as a y-intercept which is then adjusted based on the person's age (x-sub1), the person's gender (which is categorical rather than quantitative -- so I guess it would be used to adjust the b-sub1 and the b-sub2 -- and possibly the b-sub0), and the person's weight wrt their height (x-sub2).

Even though we don't cover it, the students need to know that it comes up a lot -- few real-world problems are so kind as to be explicable by a single relationship. The software output always uses the a + bx form since it is easily extensible to the more complicated case. Since students are expected to be able to read real output from real statistics packages, we use the stat version rather than the algebra version (which is limited to a 2-dimensional coordinate plane rather than an n-dimensional vector space)

3) I think using y = a +bx makes more sense once you think of the simple linear model ( one explanatory variable) as a special case of the regression model, which has many explanatory variables. In that case we have y= a + B_1 X_1 + B_2X_2 + B_3X_3+ ... It would be kind of wierd to write it y = ax + b as then the constant term would come in between the first and second explanatory variables. The qustion remains, though, why do we use y = mx + b so much in other classes?

4) Because eventually (although only *after* the AP exam) we'll be looking at multiple regression models of the form:

y = b0 + b1x1 + b2x2 + ...

That's why (BTW), I prefer b0 for the intercept, and b1 for the slope right from the start.

(And why *do* we write "m" for the slope in Algebra, anyway?)

5) The y = a + bx form is most likely used b/c it is parallel to the later statistical _multiple_ regression models that statistics students will see. In multiple regression the constant in the model is labeled alpha and the coefficients are labeled as beta's. The subsequent estimates of those quantities are naturally then an a and however many b's.

6) The y = a + bx form is preferred because it is closer to the general form of the regression equation y = b(0) + b(1)x(1) + b(2)x(2) + ... + b(n)x(n) [linear regression on several variables x(1), x(2), ..., x(n)]. That is, there's room on the right to keep adding more variables. Logically, this also has the advantage that it shows a "starting position" (intercept) followed by change [would you guess I'm teaching an applied calculus course this semester?].

The y = mx + b form carries over better into the algebraically popular form for polynomials (higher powers first) P(x) = a(n) x^n + a(n-1) x^(n-1) + . . . + a(1) x + a(0), so it's more popular in computational algebra (and thus in Algegra I & II). However, in some [mathematical] theoretical work [and always, as far as I can see, when involved with linear regression] polynomials also get written in the other order to allow the coefficients to proceed in "indexed" order P(x) = a(0) + a(1) x^1 + . . . + a(n) x^n.

This is just one more of the situations in which the "standard form" depends on context ("Whose standard?").

Charlie Peltier

7) I think a better question would be, "Why do algebra texts insist on using y = mx + b when it is much more common in advanced courses (in mathematics as well as statistics) to see it the other way around?"

y = 'starting value' + 'rate' * x seems much more intuitive for an introductory course anyway.