

## Choices: *The Line? Which Line?*

The term “regression” refers to one particular way of estimating a summary line to fit a cloud of data. There are others. In fact, while persons beginning the study of data analysis sometimes think of it as an intellectual rock: “here is how it’s done”, the truth is that the methods themselves remain the subject of active research — particularly with regard to methods of matching summary lines to clouds of data. There is no way that a person can be a responsible data analyst without realizing that there are alternative methods on the menu and that there are choices to be made — responsibilities that can not be ducked by delegating the choice to the computer, or by adopting the choices of a previous writer, or by referring to a text book,. Choice should also teach the humility to avoid overstating what the data have “told” you: It is difficult to believe results to two or three decimal digits, and perhaps to base policy decisions on fine comparisons among results — if you know full well that different methods would have given somewhat different numbers and if you know full well that there is no clear and obvious argument proving that one is right and the other is wrong.

Among the options available for fitting lines to the data, the first and most obvious option is “none of the above”. Very few real world examples, allow a *routine* application of any line fitting technique. In earlier chapters the relation between fertilizer and crop yield (four data points), the relation between time and soy bean plant length (seven data points), and the relation between time and the size of the population of the United States (twenty-one data points) — were all routine prosaic examples. Yet none of the analyses would have been well served by dumping all of the data into the computer and waiting for regression procedures to come up with a description.

Another option is created by the choice between least squares statistics and minimum absolute deviation statistics. Minimum absolute

deviation statistics are rarely used, as compared to least squares statistics, probably for reasons of custom and mathematical ease.

**OLS versus OLS versus OLS :  
Ordinary Least Squared Error for Y  
Versus Ordinary Least Squared Error for X  
Versus Orthogonal Least Squared Error**

Another option is both more subtle and more drastic in its effect on data analysis. I used least squares to minimize the residuals of one of the variables, always “y”, always represented vertically on the graph. For the moment I acted as if that were the obvious and only thing to do. Now, let me demonstrate alternatives to the fitting of “y”.

To demonstrate let me use these stylized hypothetical data describing combinations of income and education for seven individuals.

Person	Years of Education	Income
1	12	\$20,000
2	12	\$30,000
3	16	\$20,000
4	16	\$30,000
5	16	\$40,000
6	20	\$30,000
7	20	\$40,000

Allright: The one variable distribution is symmetrical in each case — likely I need no re-expression (Real data on income and education would not be so well-behaved — I am simplifying, as before, to make the point transparent.)

So, the means are \_\_\_\_; the standard deviations are \_\_\_\_\_. Then computing the standardized variables, and then computing the mean “cross product” of the standardized variables, I get  $r = .43$ .

Person	Years of Education	Income	Predicted Income	(Observed - Predicted)	Standardized Education	Standardized Income	Product of Standardized Variables
1	12	\$20,000	\$25,000	-\$5,000	-1.3228757	-1.3228757	1.75
2	12	\$30,000	\$25,000	\$5,000	-1.3228757	0	0
3	16	\$20,000	\$30,000	-\$10,000	0	-1.3228757	0
4	16	\$30,000	\$30,000	\$0	0	0	0
5	16	\$40,000	\$30,000	\$10,000	0	1.3228757	0
6	20	\$30,000	\$35,000	-\$5,000	1.3228757	0	0
7	20	\$40,000	\$35,000	\$5,000	1.3228757	1.3228757	1.75
Average	16	\$30,000					0.43
Stand dev.	3.024	\$7,559.289					

Remembering that, in standardized form, the regression equation (best fit, predicting Y from X, best in the sense of least squares) is

$$Y = r X$$

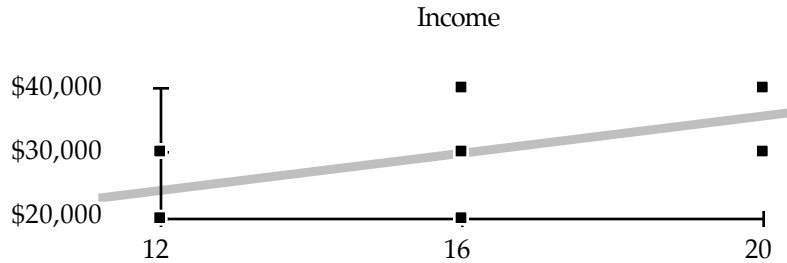
And then substituting to unstandardize:

$$y = \left( r \left( \frac{s_y}{s_x} \right) \right) x + \left( \bar{y} - \left( r \left( \frac{s_y}{s_x} \right) \right) \bar{x} \right)$$

So the regression equation for income as a function of education is:

$$\text{Income} = (\$1,250 / \text{year}) * (\text{Years of Education}) + \$10,000.$$

The regression equation says that a hypothetical person with zero years of formal education would have had an income of \$10,000. Other people would have that \$10,000 plus \$1,250 for each year of formal education. The graph is as shown below.



But now I want to show you a problem: Not so much a problem as it is a requirement that you understand exactly what you are doing when you use a regression line. The analysis I have just completed tells you how much money you can expect (on the average) for another year in school. Answer \$1,250.

But now, let me ask a different question of the same data. Now you are a consultant to an advertising agency. Your agency is working on different products, targeted to people with different levels of income. And so they ask you, if we are targeting consumers whose income is in the neighborhood of \$20,000 range, what level of education should we expect of these consumers? And if, by contrast, we are targeting consumers with incomes in the neighborhood of \$40,000 group how many years of formal education should we expect, on the average. Here it is

$$X = r Y$$

And then substituting to unstandardize:

$$x = \left( r \left( \frac{s_x}{s_y} \right) \right) y + \left( \bar{x} - \left( r \left( \frac{s_x}{s_y} \right) \right) \bar{y} \right) \text{So:}$$

$$\text{Years of Education} = (.0002 \text{ years / dollar}) * (\text{Income}) + 10.$$

Or, making a concession to my intuition, which has trouble with values like .0002 years per dollar, I'll write the equation in thousands of dollars:

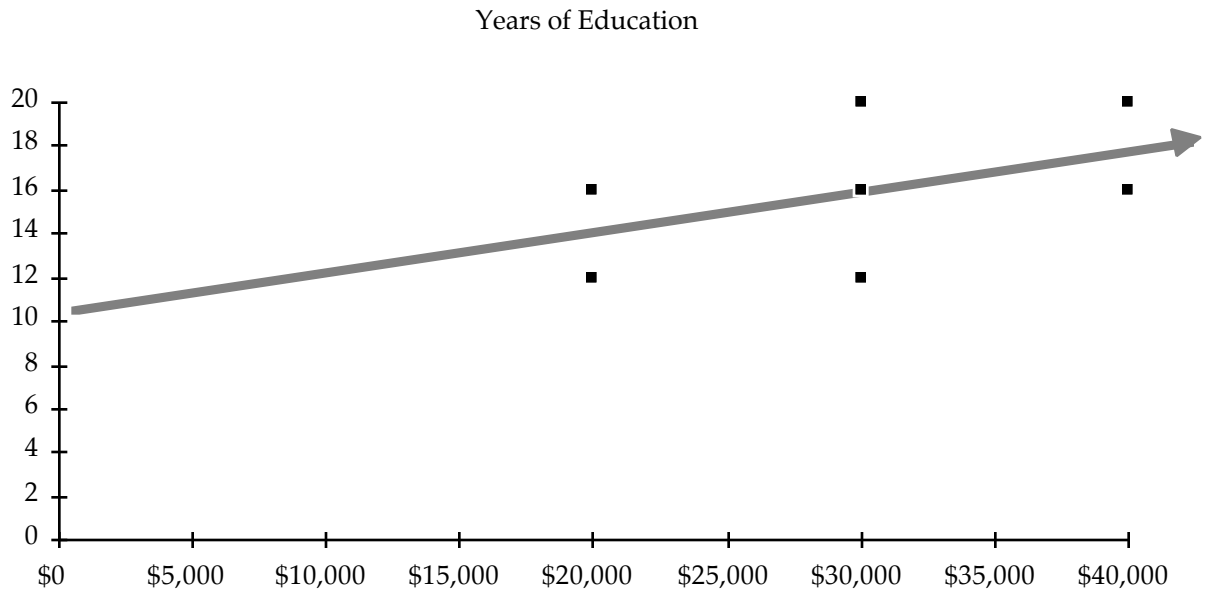
Years of education

$$= (.2 \text{ years of education per thousand dollars}) * \text{income (in thousands)} + 10.85 \text{ years.}$$

or in tens of thousands of dollars,

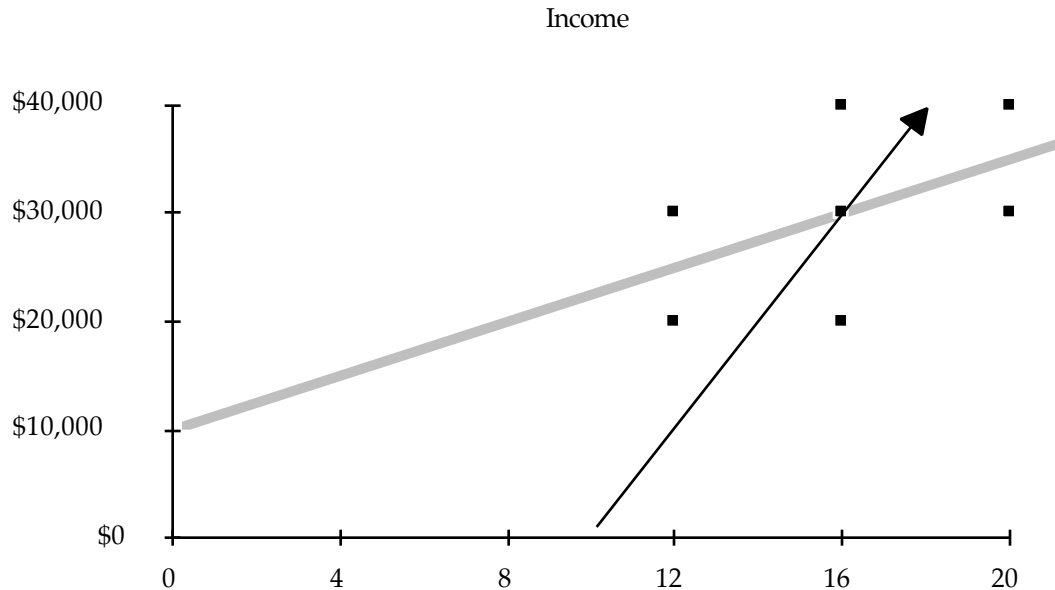
Years of education

$$= (2 \text{ years of education per ten thousand dollars}) * \text{income (in tens of thousands)} + 10.85 \text{ years.}$$



Now, I want to point out one not too minor detail: As you can see, I have switched axes: In one graph I have education left to right, used to

predict income, bottom to top. In the last graph I have income left to right, used to predict education, bottom to top. Let me combine the two of these in one graph, superimposing one upon the other:



These are two different lines: Same data, different lines. When I posed two different questions of the one set of data I got two different answers.

That is very important because it means that neither of these regression lines is an objective description of the data. Neither regression line can claim to be “the facts and nothing but the facts — don’t argue with the numbers.” No, depending on the question that I, the analyst, chose to ask one of these answers, or the other of these answers (or neither of these answers) is correct.

Does it matter? Yes, a great deal. Expressing these two slopes in a common unit, one slope says \$1,250 per year. The other says \$5,000 per year of education — separated by a factor of four.

There is nothing wrong with either of these equations. Each is the right answer to the question that was asked. Discuss: If you want to predict Income from education, then you should minimize errors with respect to predictions of income. You will get the line with slope \$1,250 per year. If you want to predict Education from Income, then you should minimize errors with respect to predictions of Education. You will get the line with slope 2 years per \$10,000 (corresponds to \$5,000 per year). But you have to be clear that prediction is not the same as objective description. (Again, neither is an objective description of the data.)

To make the point by reducing it to a possible absurdity, consider that we have data reporting the physical heights for pairs of brothers. Randomly, I will assign the height of one of the two brothers as "x" and assign the height of the other brother as y. Now, I ask you to use the height of one brother, which I will tell you, to predict the height of the other brother. Here is the pattern. If I tell you that one brother is 5'10", about average for these hypothetical data, you can reasonable expect his brother to be about average, about 5'10". If I tell you that one brother is 4'10" considerably below average, you can reasonably expect his brother to be considerably below average, but not so extreme. You might expect something like 5". And if I tell you that one brother is 6'10", considerably above average, you can reasonably expect his brother to be considerably above average, but not so extreme. You might expect something like 6'8".

This is the logic of "regression": If you have somebody who is above average, you expect a result that is above average — but a little closer to the mean: "regression toward the mean of the population".

Now, let me try to confuse you while you try to resist: Suppose I report to you that one brother is 6'8". You can reasonably expect this tall man to have a tall brother — but not quite so extreme: Perhaps 6'6".

That means that men who are 6'10" would be predicted to have brothers who are, on the average 6'8". And men who are 6'8" would be expected to have brother who are, on the average, 6'6".

That sounds strange: It sounds like I am saying that a man who is 6'10" will have a brother who is 6'8". But a man who is 6'8" will have a brother who is 6'6". That is impossible, but that is not what I am saying: I am saying that the *average* in one case is 6'8" and the average in the other case is 6'6". That is possible and that is regression analysis.

There is a third demand I can make of these data and it leads to a third answer. Suppose that these data are part of a larger data set, pairs of persons different by birth order, sex, nutrition, whatever. Then in this one case where the persons being compared are, in fact, the same except for random assignment to group "x" or group "y", I want the data analysis to give me the numerical equivalent of the English statement that they are the same except for random variation.

The answer to this question come by thinking about "the line that is closest to the data". This line is descriptive. It contains no built in assumption about predicting y from x or x from y. It is the line that is closest to the data. As in ordinary Euclidean geometry, the distance between a point and a line is the length of the perpendicular from the point to the line. This is the Orthogonal Least Squares Line and it has the equation

$$Y = X \quad (\text{the } r \text{ has vanished}).$$

That looks too simple in standardized form, but restoring reality by unstandardizing, the descriptive (orthogonal least squares best line is)

$$y = \frac{s_y}{s_x} x + \left( \bar{y} - \frac{s_y}{s_x} \bar{x} \right)$$

When you do not, yourself, wish to impose order on the variables, when you are not trying to predict one from the other. choose the orthogonal least squares line to treat the variables symmetrically. It is

entirely possible to use several procedures on the same data — you protect yourself and your reader by saying what you have done and providing sufficient data and numbers to allow the reader both to reproduce what you have done and to complete alternative analysis. This is not simply a matter of courtesy. It is a necessity because in perfectly ordinary data different estimates of slope, for example, will differ from each other by factors of three or ten or more — depending on the choice of method. You maintain integrity, first, by making it absolutely clear what you have done and, second, by equipping your reader to explore the paths you have not taken.