Two Facets of the Linear Regression Process

David R. Duncan and Bonnie H. Litwiller

Mathematics teachers are always looking for ways to incorporate statistical concepts into their classrooms. Since calculators can readily produce the equation of the "least-squares" regression line for a data sample, this topic is sometimes introduced in secondary school mathematics.

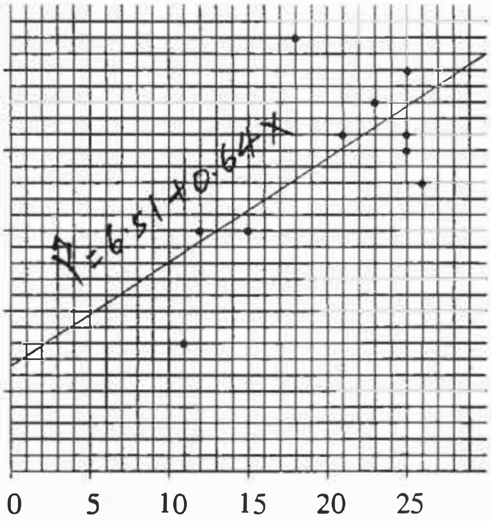
We will show two examples that illustrate facets of the regression process that are appropriate for the secondary school level. Since regression involves graphing, these examples also connect to algebra concepts.

Example 1

Consider the following set of 10 pairs of scores. Each pair represents a different student in a class. For a given student, the x-value is the score on test 1 and the y-value is the score on test 2. Set of scores = {(11,8), (12,15), (15,15), (18,27), (21,21), (23,23), (25,20), (25,21), (25,25), (26,18)}. The built-in statistical capability of the TI-85 (or some other calculator) reports that the regression line has the equation: $\hat{y} = 6.51 + 0.64x$. This line provides the "best" linear fit to the observed data.

Graph 1 displays the original 10 points and the regression line.

Graph 1



Have your students produce this graph with the given points and the equation cited above.

In what sense does this regression line provide the best line of fit for the data? The criterion used in statistics is that the best line of fit should minimize the sum of the squares of the deviations of each true y value from the $\hat{y}$ value predicted by the regression line.

Table 1 reports the 10 original (x,y) pairs, the predicted y-value $(\hat{y})$ for each point, the difference between the true y and the predicted y-value $(\hat{y})$ for each point, the difference between the true y and the predicted $\hat{y}$ $(y - \hat{y})$, and the square of these differences $(y - \hat{y})^2$.

The sum of the entries of the $(y - \hat{y})^2$ column is approximately 159. The theory of linear regression asserts that the sum of $(y - \hat{y})^2$ column is minimized when the regression line $\hat{y} = 6.51 + 0.64x$ is used. Any other regression line, even though it might appear to the eye to better approximate the data, will yield a larger $(y - \hat{y})^2$ sum and hence be less effective overall.

Table 1

x	y	$\hat{y} = 6.51 + 0.64x$	$(Y - \hat{y})$	$(Y - \hat{y})^2$
11	8	13.55	–5.55	30.8025
12	15	14.19	0.81	0.6561
15	15	16.11	–1.11	1.2321
18	27	18.03	8.97	80.4609
21	21	19.95	1.05	1.1025
23	23	21.23	1.77	3.1329
25	20	22.51	–2.51	6.3001
25	21	22.51	–1.51	2.2801
25	25	22.51	2.49	6.2001
26	18	23.15	–5.15	26.5225

Table 2
Using the Regression Line $\hat{Y} = X$:

X	Y	$\hat{Y} = X$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
11	8	11	–3	9
12	15	12	3	9
15	15	15	0	0
18	27	18	9	81
21	21	21	0	0
23	23	23	0	0
25	20	25	–5	25
25	21	25	–4	16
25	25	25	0	0
26	18	26	–8	64
				204

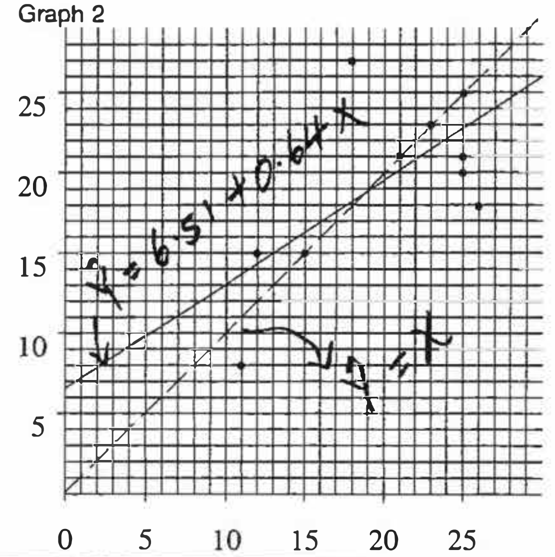
To test this assertion we note that 4 of the 10 points satisfy $y = x$. Consequently, one might suppose that the regression line $\hat{y} = x$ might better approximate the data than the line yielded by the calculator. Is this true?

Table 2 reports calculations similar to those of Table 1 for the alternative regression line.

Note that the $(y - \hat{y})^2$ column sums to 204; this is greater than the total of 159 of Table 1. Even though the alternative regression line has the advantage of predicting four y values exactly, in total, it does a worse job of fitting the data linearly than does the original regression line. Graph 2 displays both regression lines and the original data on the same axes.

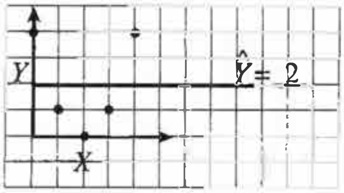
Is it obvious to your students that the solid graph fits the data better than the broken one? Opinions based on the appearance of Graph 2 must be tested by calculations as we have done in Tables 1 and 2.

Graph 2



Example 2

Consider the following small sample of paired scores: {(0,4), (1,1), (2,0), (3,1), (4,4)}. The TI-85 reports the linear regression equation to be $\hat{y} = 2 + 0x$ or $\hat{y} = 2$. These data and the regression line are depicted in Graph 3. The theoretical meaning of this horizontal regression line is that no linear pattern could be detected; the regression equation then predicts the mean value of y for each of the x values. Although one might say that this regression line does not predict the y values well, it does a better job by minimizing the sum of $(y - \hat{y})^2$ entries than would any other straight line. Graph 3 displays the data points and the regression line.



The striking facet of these data points is that they lie exactly on the parabola $y - (x - 2)^2$. The use of a parabola to predict the ys would have predicted perfect results. However, the linear regression process built into the TI-85 looks only for a straight line to fit the data; it is oblivious to any other possibility. This demonstrates the inappropriateness of rigidly using a procedure that may not fit a given situation.

Find other regression models in textbooks or in your calculator. On what assumptions are they built?