

DAVID FREEDMAN  
ROBERT PISANI  
ROGER PURVES

# Statistics

Third Edition

W • W • NORTON & COMPANY

NEW YORK • LONDON



WHARTON REPROGRAPHICS

### 3. DOES THE REGRESSION MAKE SENSE?

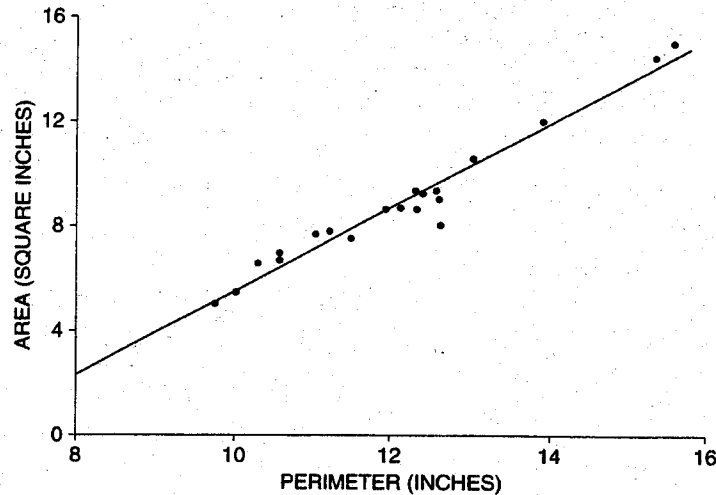
A regression line can be put down on any scatter diagram. However, there are two questions to ask: First, was there a non-linear association between the variables? If so, the regression line may be quite misleading (p. 163). Even if the association looks linear, there is a second question: Did the regression make sense? The second question is harder. Answering it requires some understanding of the mechanism which produced the data. If this mechanism is not understood, fitting a line can be intellectually disastrous.

To make up an example, suppose an investigator does not know the formula for the area of a rectangle. He thinks area ought to depend on perimeter. Taking an empirical approach, he draws 20 typical rectangles, measuring the area and the perimeter for each one. The correlation coefficient turns out to be 0.98—almost as good as Hooke's law. The investigator concludes that he is really on to something. His regression equation is

$$\text{area} = (1.60 \text{ inches}) \times (\text{perimeter}) - 10.51 \text{ square inches}$$

(Area is measured in square inches and perimeter in inches.)

Figure 6. Scatter diagram of area against perimeter for 20 rectangles; the regression line is shown too.



There is a scatter diagram in figure 6, with one dot for each rectangle; the regression line is plotted too. The rectangles themselves are shown in figure 7. The arithmetic is all in order, but the regression is silly. The investigator should have looked at two other variables, length and width. These two variables determine both area and perimeter:

$$\text{area} = \text{length} \times \text{width}, \quad \text{perimeter} = 2(\text{length} + \text{width})$$

Our straw-man investigator would never find this out by doing regressions.

When looking at a regression study, ask yourself whether it is more like Hooke's law, or more like area and perimeter. Of course, the area-perimeter example is hypothetical. But many investigators do fit lines to data without facing the hard issues. That can make a lot of trouble.<sup>11</sup>

*Technical note.* Example 1 in section 1 presented a regression equation for predicting income from education. This is a good way to describe the relationship between income and education. But it may not be legitimate to interpret the slope as the effect on income if you intervene to change education. The problem—the effects of other variables may be confounded with the effects of education.

Many investigators would use multiple regression to control for other variables. For instance, they might develop some measure for the socioeconomic status of parents, and fit a multiple regression equation of the form

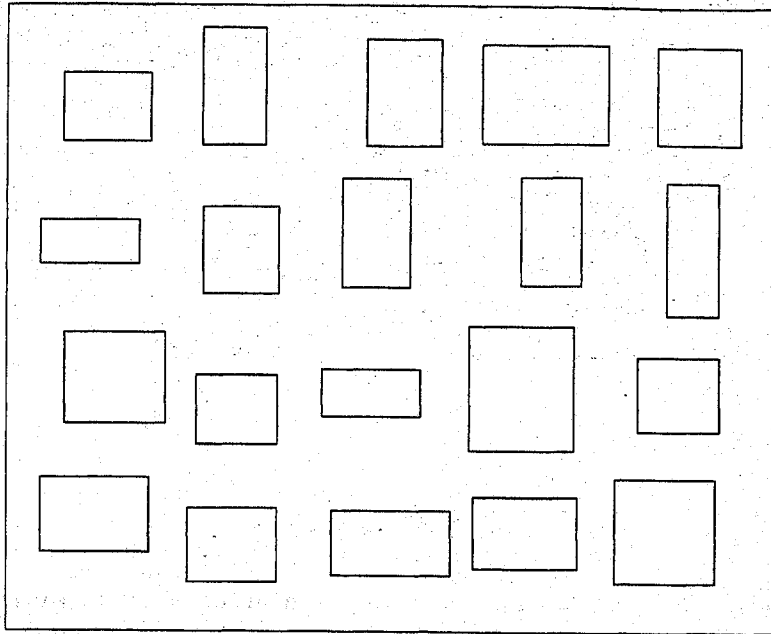
$$y = a + b \times E + c \times S,$$

where

$$\begin{aligned} y &= \text{predicted income, } E = \text{educational level,} \\ S &= \text{measure of parental status.} \end{aligned}$$

The coefficient  $b$  would be interpreted as showing the effect of education, controlling for the effect of parental status.

Figure 7. The 20 rectangles themselves.



SCALE: 0 2 4 6 8 10 INCHES

This might give sensible results. But it can equally well produce nonsense. Take the hypothetical investigator who was working on the area of rectangles. He could decide to control for the shape of the rectangles by multiple regression, using the length of the diagonal to measure shape. (Of course, this isn't a good measure of shape, but nobody knows how to measure status very well either.) The investigator would fit a multiple regression equation of the form

$$\text{predicted area} = a + b \times \text{perimeter} + c \times \text{diagonal}.$$

He might tell himself that  $b$  measures the effect of perimeter, controlling for the effect of shape. As a result, he would be even more confused than before. The perimeter and diagonal do determine the area, but only by a non-linear formula. Multiple regression is a powerful technique, but it is no substitute for understanding.