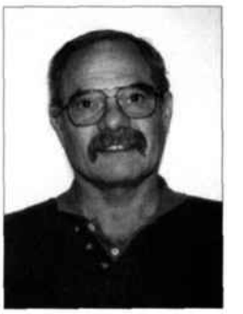
VISUAL REVELATIONS



Howard Wainer, Column Editor

Kelley's Paradox

When we use regression to try to predict one event from another we always find that the variation in the prediction is smaller than that found in the predictor. In 1889 Francis Galton pointed out that this always occurred whenever measurements were taken with imperfect precision and was what he called "regression toward the mean."

Although regression has been well understood by mathematical statisticians for more than a century, the terminology among appliers of statistical methods suggests that they either thought of it as a description of a statistical method or as only applying to biological processes. In 1924, Frederick C. Mills, the economic statistician, wrote that "the original meaning has no significance in most of its applications," (p. 394).

Stephen Stigler (1997, p. 112) pointed out that this was "a trap waiting for the unwary, who were legion." The trap has been sprung many times. One spectacular instance of a statistician getting caught was "in 1933, when a Northwestern University professor named Horace Secrist unwittingly wrote a whole book on the subject, *The Triumph of Mediocrity in Business*. In over 200 charts and tables, Secrist 'demonstrated' what he took to be an important economic phenomenon, one that likely lay at the root of the great depression: a tendency for firms to grow more mediocre over time." Secrist showed that the firms with the highest earnings a decade earlier were currently performing only a little better than average; moreover, a collection of the more poorly performing firms had improved to only slightly below average. These results formed the evidence supporting the title of the book. Harold Hotelling, in a devastating review published the same year, pointed out that the seeming convergence Secrist obtained was a "statistical fallacy, resulting from the method of grouping." He concluded that Secrist's results "prove nothing more than that the ratios in question have a tendency to wander about."

It is remarkable, especially considering how old and well-known regression effects are, how often these effects are mistaken for something substantive. Although Secrist himself was a professor of statistics, Willford I. King, who, in 1934 wrote a glowing review of Secrist's book, was president of the American Statistical Association! This error was repeated in 1985 by W.F. Sharpe, a Nobel laureate in economics (p. 430) who ascribed the same regression effect Secrist described to economic forces. His explanation of the convergence, between 1966 and 1980, of the most profitable and least profitable companies was that "ultimately economic forces will force the convergence of profitability and growth rates of different firms." The explanation is statistical, not economic. Apparently this led Milton Friedman (in 1992), yet another Nobel laureate in economics, to try to set his colleagues straight.

In 1927, Truman Kelley described a specific instance of a regression formula of great importance in many fields, although it was proposed for use in educational testing. It shows how you can estimate an examinee's true score from his/her observed score on a test. "True score" is psychometric shorthand for the mean of the distribution of observed scores that someone would get if parallel forms of the same test were repeated infinitely. Kelley's equation relates the estimated true score (τ) to the observed score (x). It tells us that the best estimate is obtained by regressing the observed score in the direction of the mean score (μ) of the group that the examinee came from. The amount of the regression is determined by the reliability (ρ) of the test. Kelley's equation is

$$\hat{\tau} = \rho x + (1 - \rho)\mu \qquad (1)$$

Note how Kelley's equation works. If a test is completely unreliable ($\rho = 0$), as would be the case if each examinee's score was just a random number, the observed score would not count at all and the estimated true score is merely the group mean. If the test scores were perfectly reliable ($\rho = 1$), there would be no regression effect at all and the true score would be the same as the observed score. The reliability of virtually all tests lies between these two extremes, so the esti-

Column Editor: Howard Wainer, Principal Research Scientist, Measurement-Statistics Data Research, Educational Testing Service (15-T), Rosedale Road, Princeton, NJ, 08541-0001, USA; E-mail *hwainer@rosedale.org*.

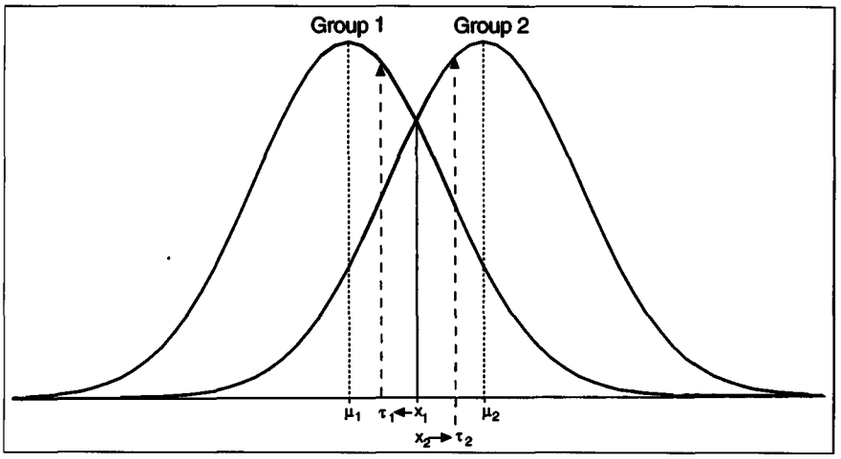


Figure 1. A graphical depiction of Kelley's equation for two groups. The two distributions and their means are shown. How the true scores are regressed when two identical observed scores come from each of the two different score distributions is also indicated.

mated true score will be somewhere between the observed score and the mean.

Intuition about how Kelley's equation works when there are multiple groups is aided by a diagram. Shown in Fig. 1 are the distributions of scores for two groups of individuals, here called Group 1 (lower scoring group) and Group 2 (higher scoring group). If we observed a score x midway between the means of the two groups, the best estimate of the true score of the individual who generated that score depends on which group that person belonged to. If that person came from Group 1, we should regress the score downward; if from Group 2 we should regress it upward. The regression effect is because we know that there is some error in the score. The average error is considered to be 0, so some errors will be positive and some negative. Thus, if someone from a low-scoring group has a high score we can believe that to some extent that person is the recipient of some positive error that is not likely to reappear on retesting, so we regress their score downward. Similarly, if someone from a high-scoring group has an unusually low score, we regress that score upward.

So far this is merely an equation. What is the paradox? Webster defines a paradox as a statement that is opposed to common sense and yet is true. So long as Kelley's equation deals solely with abstract groups named 1 and 2, no paradox emerges. But suppose we call Group 1 the "Low SES Group" and Group 2 the "High SES Group." Now when we see someone from Group 1 with a high score, despite their coming from an environment of intellectual and material deprivation, we suspect that they must be very talented indeed and their true ability ought to be considered somewhat higher. Similarly, someone who comes from a more privileged background but who scores low leads us to suspect a lack of talent and hence ought to be rated lower still.

Do people truly make this sort of mistake? In the August 31, 1999, issue of the *Wall Street Journal* an article appeared about a research project done under the auspices of the Educational Testing Service called "Strivers." The goal of "Strivers" was to aid colleges in identifying applicants (usually minority applicants) who have a better chance of succeeding in college than their test scores and high school grades might otherwise suggest. The basic idea was to predict a student's SAT score from a set of background variables (e.g., ethnicity, SES, mother's education, etc.) and characterize some of those students who do much better than their predicted value as "Strivers." These students might then become special targets for college admission's officers. In the newspaper interview the project's director, Anthony Carnevale, said, "When you look at a Striver who gets a score of 1000, you're looking at someone who really performs at 1200." Harvard emeritus professor Nathan Glazer, in an article on Strivers in the September 27, 1999, *New Republic*, indicated that he shares this point of view when he said (p. 28), "It stands to reason that a student from a materially and educationally impoverished environment who does fairly well on the SAT and better than other students who come from a similar environment is probably stronger than the unadjusted score indicates."

Unfortunately for the goals of affirmative action, neither Mr. Carnevale nor Mr. Glazer are correct. When you look at a Striver who gets a score of 1000, you're probably looking at someone who really performs at 950. And, alas, a Striver is probably *weaker* than the unadjusted score indicates.

References and Further Reading

Friedman, M. (1992), "Do Old Fallacies Ever Die?" *Journal of Economic Literature*, 30, 2129–2132.

Galton, F. (1889), *Natural Inheritance*, London: Macmillan.

Hotelling, H. (1933), "Review of the Triumph of Mediocrity in Business by Secrist H," *Journal of the American Statistical Association*, 28, 463–465.

Kelley, T.L. (1927), *The Interpretation of Educational Measurements*, New York: World Book.

King, W.I. (1934), "Review of the Triumph of Mediocrity in Business by Secrist H.," *Journal of Political Economy*, 42, 398–400.

Mills, F.C. (1924), *Statistical Methods. Applied to Economics and Business*, New York: Henry Holt.

Secrist, H. (1933), *The Triumph of Mediocrity in Business*, Evanston, IL: Bureau of Business Research, Northwestern University.

Sharpe, W.F. (1985), *Investments* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Stigler, S. (1997), "Regression Toward the Mean, Historically Considered," *Statistical Methods in Medical Research*, 6, 103–114.