Regression to the Mean

Garrett Fitzmaurice, ScD

From the Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA

Approximately 4 wk after the birth of our son, I accompanied my wife to our son's 1-mo visit to the pediatrician's office. Our son was prodded, measured, and weighed in the usual manner, and I was delighted to learn that my son's height placed him at the 95th percentile. This seemed to me a very fitting place on the distribution for the son of a statistician. However, on our next visit to the pediatrician's office some months later, we learned that our son's height no longer placed him at the 95th percentile but rather at the 90th percentile. With this news, my wife asked the pediatrician what could be the possible reason for the decline. Before our pediatrician had a chance to respond, I helpfully opined that there was probably a very simple explanation: regression to the mean. "Regression to the what?" my wife exclaimed, while rolling her eyes toward heaven. Our pediatrician was, or at least pretended to be, somewhat more receptive to this explanation and asked what I meant by "regression to the mean." She had recently heard one of her colleagues use this expression and was curious as to what it meant. Ignoring the bewildered expression on my wife's face, I sat back and prepared to give a concise explanation of this common yet fascinating statistical phenomenon. Alas, albeit to the obvious relief of my wife, my minitutorial on regression was interrupted when the receptionist appeared with the next little tyke waiting to be examined by our pediatrician, and I never had the opportunity to complete my explanation. Not to be deterred, I decided to put pen to paper and produce this column.

The origins of the expression *regression to the mean* can be traced back to the late 19th century.[1] Sir Francis Galton, a cousin of Charles Darwin, first coined the term to explain a curious conundrum that he encountered in his studies of human genetics. Galton was interested in the heritability of height, and he studied the relation between the height of parents and the height of their offspring. To be more precise, he related the heights of children to what he referred to as the *mid-parent height*, an average of the height of both parents.* Galton found that parents and their offspring had approximately the same mean or average height. However, he also observed that the offspring of tall parents tended on average to be somewhat shorter than their parents. Similarly, the offspring of parents of short stature tended on average to be somewhat taller than their parents. Galton termed this phenomenon *regression towards mediocrity*, later replacing the word *mediocrity* with *mean* (for a fascinating historical account of regression to the mean, see Stigler[2]).

What Galton had discovered was that if parents were subdivided into groups of equal height and the mean height of their offspring was determined, the means for all of the different subgroups could be plotted along a straight line. This line later became known as the *regression line*. The precise impact of regression to the mean can best be understood by examining the formula for the slope of the regression line. Recall from an previous column[3] that the simple linear regression equation is given by:

$$E(Y|X = x) = \alpha + \beta x$$

where Y denotes the dependent variable (e.g., height of offspring), X denotes the independent variable (e.g., height of parents), and $E(Y \mid X = x)$ denotes the mean of Ys for a given value x.

The regression equation can also be expressed as:

$$E(Y|X = x) = \mu_y + \rho\sigma_y/\sigma_x(x - \mu_x)$$

where μ_y and μ_x are the mean of Y and X, ρ denotes the correlation between Y and X, and σ_y and σ_x are the standard deviations of Y and X, respectively. For the purposes of our discussion, we can simply assume that Y and X have been measured on a common scale and that $\sigma_y = \sigma_x$. The expression for the regression equation then simplifies to:

$$E(Y|X = x) = \mu_y + \rho(x - \mu_x).$$

What is clear from the latter expression is that unless $\rho = 1$ (i.e., all values of Y and X fall along a straight line), the mean of Y for a given value x will on average deviate less from μ_y than does x from μ_x. That is, for any given value of X, that is, say δ units from its mean (i.e., $\delta = x - \mu_x$), the predicted value of Y is only $\rho\delta$ units from μ_y. Note also that this expression implies that the impact of regression to the mean is related to both δ, the distance of x from μ_x, and the strength of the association between Y and X. The larger δ is the greater the regression effect. Also, with $0 < |\rho| < 1$, the weaker the correlation between Y and X, the greater the impact of regression to the mean.

Regression to the mean is ubiquitous in medical research, and its effect can very easily lead the unwary researcher astray. It most commonly occurs in studies in which subjects are selected because they have extreme values on a variable. On reflection, this in not an unusual occurrence in many studies in which patients are enrolled only if they meet certain eligibility criteria. That is, patients are eligible to participate in the study only if they screen high on a marker for disease progression or some other variable that is thought to be related to the disease. By virtue of regression to the mean, we can therefore expect to see a mean reduction from the pretreatment response, regardless of the efficacy of the treatment. For example, subjects with high low-density lipid (LDL) cholesterol levels, say greater than 160 mg/dL, may be enrolled in a study to receive treatment to lower cholesterol. After treatment with an experimental drug, the subjects have their LDL cholesterol levels measured for a second time. However, even if the subjects had not received any treatment, we would expect to see a reduction in LDL cholesterol levels due entirely to regression to the mean. That is, the mean LDL cholesterol levels at the second occasion would be expected to be closer to the overall mean in the general population. It should be clear from this example that failure to acknowledge the impact of regression to the mean in the analysis will lead to a biased estimate of the effect of treatment. In general, the impact of regression to the mean can be eliminated either through the use of randomization with an appropriate control group or by various statistical methods that separate any genuine reduction due

Correspondence to: Garrett Fitzmaurice, ScD, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA. E-mail: fitzmaur@hsph.harvard.edu

*Galton defined the mid-parent height as (father's height + [1.08 × mother's height])/2. Galton multiplied the mother's height by 1.08 instead of the expected 1.0 because this was the ratio of the mean height of men to the mean height of women.

Nutrition 16:81–82, 2000 0899-9007/00/$20.00

PII S0899-9007(99)00218-X

to treatment from the effect of regression to the mean.

In summary, regression to the mean is a fascinating and very common phenomenon. It is also one that is often not well understood. Regression to the mean will necessarily occur when there is a non-perfect correlation among two variables (i.e., almost all of the time). Whenever two variables have a correlation less than 1, cases that have extreme values on one of the variables will, on average, have less extreme values on the other variable. This means that, when the same variable is measured on two occasions, cases that are extreme on the first occasion will be somewhat less extreme on the second occasion. Thus, it was not so surprising to find that my son's height was less extreme on his subsequent visit to the pediatrition. Over time, low scores on the first occasions will, on average, improve, whereas high scores will decline.

Regression to the mean is such a common phenomenon that one need look no further than the very journal you are now reading for an example.[4,5] Before publication, manuscripts submitted to this journal are peer reviewed by experts to determine the overall quality of the papers. However, we all know that referees do not always agree on the merits of a manuscript. Their assessments are subject to measurement error and are certainly not perfectly correlated with the true (but, perhaps, unobservable) quality of the manuscript. Acknowledging that referees' assessments are not entirely perfect but the best yardstick that is available, the editor, even of this journal, is far more likely to be persuaded to publish a manuscript that has received positive reviews from all of the referees involved. However, this necessarily means that, due to regression to the mean, the papers that finally appear in each issue of *Nutrition* are probably not quite as good as our dear editor would like to believe!

REFERENCES

1. Galton F. Regression towards mediocrity in hereditary stature. J Anthropol Inst 1886;15:246
2. Stigler SM. Regression towards the mean, historically considered. Stat Methods Med Res 1997;6:103
3. Pagano M. Simple linear regression and correlation. Nutrition 1995;11:179
4. Rousseeuw PJ. Why the wrong papers get published. Chance 1991;4:41
5. Bland JM, Altman DG. Statistics notes: some examples of regression towards the mean. Br Med J 1994; 309:780