ANNALS OF HUMAN BIOLOGY, 1978, VOL. 5, NO. 3, 277–280

SHORT REPORTS

Regression to the mean

M. J. R. HEALY
London School of Hygiene and Tropical Medicine
and H. GOLDSTEIN
Institute of Education, University of London

[Received 25 October 1977; revised 1 February 1978]

Summary. The notion of "regression to the mean" is widely misunderstood. This paper explains the concept in simple terms and shows how it arises in studies of mental and physical development.

The notion of regression to the mean, as used for example by Galton (1886), though one of the oldest in modern statistics, is still regarded as somewhat mysterious. In a paper entitled "Regression to the mean—a confused concept", Clarke, Clarke and Brown (1960) claim that only one article, by McNemar (1940), "approaches a satisfactory description of regression". The difficulties are partly associated with what actually occurs and partly with its explanation on statistical or other grounds.

The phenomenon that Galton christened "regression" is connected with the occurrence and measurement of change, originally the change in a particular measurement from one generation to the next. Galton noted that tall fathers tended to have tall sons; but the sons, while taller than average, tended to be less extreme in this respect than their fathers were. Galton described this situation by saying that the sons "regressed towards the mean". In more recent times, the idea of regression has occurred frequently in the psychological literature in connection with repeated measurements of intelligence, attainment, etc. in children. Again it is found, using standardized scores, that children who have high scores on an initial test tend to have scores on a subsequent test which are higher than average but lower than the initial ones. The questions arising are:

(*a*) does regression always occur?
(*b*) is it real or just a statistical artefact?
(*c*) if it is real, what causes it?

The discussion is bedevilled by a change that has taken place in the terminology. The word "regression" has acquired a technical meaning in statistics (to be explained further below) which differs quite radically from the sense in which Galton originally used it. If x is an initial measurement and y a subsequent one, suppose that $y=a+bx$, in a sense to be made precise later. Then Galton (and, following him, Clarke *et al.*) would say that there was "no regression" if $b=1$, so that y increases and decreases by the same amount as x. The modern statistician, however, would apply the same phrase to the case $b=0$, in which y is unaffected by the value of x. This modern statistical usage, though etymologically doubtful, is now universal and we shall assume it unless otherwise stated.

Suppose then that we have a population of items each of which carries the values

of two variates, x and y. In Galton's example the items were father–son pairs with x and y the two heights; often x and y are the same measurement made on a single individual on two different occasions. Select out from this population all those items for which x takes a particular value. These items constitute a sub-population which defines a probability distribution of y-values. Denote the mean of this distribution by Y. This mean can be defined for any particular value of x and so we can define a mathematical function $Y(x)$ which is the *regression function of y on x*. Note that the regression relationship is asymmetric in that the regression function of x on y will usually be different. It is not in fact necessary for x to be a variate, i.e. a variable with an associated probability distribution, but for the present discussion we assume that this is the case.

Suppose now that the regression function $Y(x)$ is linear. Then the regression equation of y on x can be written in the form

$$Y=\eta+\frac{\rho\sigma_y}{\sigma_x}(x-\xi), \tag{1}$$

where (ξ, η) is the mean of the bivariate distribution of x and y, σ_x and σ_y are the marginal standard deviations and ρ the correlation coefficient, which for convenience we suppose to be positive. It follows that, if x exceeds its mean ξ by $k\sigma_x$, the expected value of *y for this x* will exceed its own mean η by $\rho k\sigma_y$, and this will be less than $k\sigma_y$ except in the limiting case of perfect correlation. This limiting case aside, it can be seen that y, or rather Y, *always* "regresses" *if x and y are measured in units of their own marginal standard deviation*, and that this regression is towards the overall mean of y. If we standardize the distributions of x and y so that the standard deviations are made equal and variable values are expressed in terms of deviations from their respective means, equation (1) in terms of the new variables Y', x' becomes simply

$$Y'=\rho x' \tag{2}$$

Of course, in terms of the original units, "regression to the mean" may or may not occur when the standard deviations are unequal. For example, the model of a linear regression function is a fair first approximation to the relationship in IQ between offspring and their parents, and accordingly offspring IQ will regress towards the offspring mean in a group of families in the sense of the term that we have used. Clarke and Clarke (1974) state that "regression to the mean" is not to be expected when the mid-parent IQ is used in the comparison; this is correct since they are working in unstandardized units. In fact, on the simplest model of multifactorial inheritance, $\rho=1/\sqrt{2}$, but $\sigma_y/\sigma_x=\sqrt{2}$ (since x is the average of the two parental values) so that the regression coefficient is $1\cdot0$. If x is 10 units of IQ above its mean, the expected value of y will be 10 units above its mean; but in standard deviation units, the excess of y is only $0\cdot71$ times that of x. Note that standardizing the individual parental scores to standard deviations of $1\cdot0$ does not produce a unit standard deviation for mid-parent values.

It is often stated that regression takes place towards the mean of an individual's own population; but any particular individual may belong to several different populations, for example he may be a member of one social group or a geographical region etc. However, the regression phenomenon relates to averages of expected values and our expectations will vary according to our information about the individuals concerned. Using Galton's example, suppose that a father's height is 180 cm and we know only that he belongs to a general population whose average height is

160 cm; then (with $\rho = 0 \cdot 5$ and equal means and standard deviations) we expect his son's height to be 170 cm on average, regressing by 10 cm. If we now learn that the father belongs to an upper social class whose mean height is 166 cm, he is less extreme and in the light of the extra information we now expect his son's height to be 173 cm. This difference is basically no more mysterious than that between our expectations for the father himself (160 cm and 166 cm) when we regard him as a random member of the two populations.

No general statement can be made when the regression function is not a straight line—indeed in pathological cases, such as that in which $Y(x)$ is a parabola with the marginal distribution of x-values spanning the vertex, it is not possible to describe the regression phenomenon in simple terms. With near-linear regression, the simple situation is likely to give a reasonable qualitative description of what actually occurs.

If we seek a cause of regression to the mean, we must ascribe it to all those factors or combinations of factors that result in ρ being less than 1. Among these will usually be errors of measurement, though in many instances (especially when x and y are physical measurements) these will be of relatively minor importance.

The nature of "errors of measurement" is another topic which has been much debated. The underlying notion is that of a "true" or "stable" value belonging to a subject, from which observed values differ by amounts which may be treated as random quantities. In the physical sciences, these quantities may be principally those associated with the measuring instruments, in that their average size may be reducible by modifications and improvements to these instruments. In biology and psychology, this type of error will also occur and may often be large, but genuine short term fluctuations in the quantity measured will also often be of major importance and the two may only be distinguishable with difficulty. If the combined "error" variances can be estimated (usually by repeating the measurements), the observed correlation can be adjusted upwards, using what amounts to the well-known "correction for unreliability". It is also possible to investigate the hypothesis that the "true" values of x and y are perfectly correlated (Healy, 1958), so that for them no regression in Galton's sense takes place. This is what must be envisaged by those who assert that regression to the mean is merely a statistical artefact; in practice it rather rarely describes what actually takes place.

Errors of measurement aside, factors of all kinds which affect different subjects to different extents can lead to a value of ρ which is less than 1; they thus lead to the occurrence of regression. Clarke *et al.* (1960) refer to individuals who tend to follow a common growth curve but depart from it in a smooth but irregular manner—they refer to this as non-linearity of the growth processes. Departures of this kind may be ascribed to physiological or psychological factors which vary on a time scale which is short relative to the interval over which growth is being considered. When the time-scale is very short, factors of this kind become indistinguishable from errors of measurement.

Even with no irregular fluctuations, there will usually be individual differences in the average rate of change of the quantity measured over the period concerned. If in this situation we write x and y for the initial and final values as before, we can write $y = x + z$ and it is then easy to show that the correlation between x and y is given by

$$\rho^2 = 1 - \mathrm{Var}\,(z|x)/\mathrm{Var}\,(y)$$

when $\mathrm{Var}\,(z|x)$ is the variance of z given the value of x. This expression is less than

1·0 whenever the increment z is less than perfectly correlated with the initial value x. This result does not depend in any way upon the shape of the individual or average growth curves. It will occur, for example, when the growth curves are all perfectly linear with slopes which differ from one individual to another.

Regression to the mean greatly complicates the analysis of changes in biological and psychological attributes. Some of the problems are discussed by Oldham (1968, section 6.6), Fletcher, Peto, Tinker and Speizer (1976), Kendall and Stuart (1961, chapter 29), Davis (1976) and James (1973). It is essential, especially when attempting to relate changes to other factors, to incorporate all the sources of variability and their inter-correlations into a suitable statistical model, and to be clear whether the questions at issue relate to observed values or to theoretical "error-free" values that are supposed to underlie them.

Acknowledgments

We would like to thank Dr. Ann Clarke for her helpful comments.

References

Clarke, A. D. B., Clarke, A. M., and Brown, R. I. (1960). Regression to the mean—a confused concept. *British Journal of Psychology*, **51**, 105–117.

Clarke, A. M., and Clarke, A. D. B. (1974). Genetic–environmental interactions in cognitive development. In *Mental Deficiency: the Changing Outlook* (3rd edition), ed. Clarke, A. M. and Clarke, A. D. B., London: Methuen; New York: Full Press.

Davis, C. E. (1976). The effect of regression to the mean in epidemiologic and clinical studies. *American Journal of Epidemiology*, **104**, 493–498.

Fletcher, C., Peto, R., Tinker, C., and Speizer, S. (1976). *Natural History of Chronic Bronchitis and Emphysema*. London: Oxford University Press.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, **15**, 246–263.

Healy, M. J. R. (1958). Variation within individuals in human biology. *Human Biology*, **30**, 210–218.

James, K. E. (1973). Regression towards the mean in uncontrolled clinical studies. *Biometrics*, **29**, 121–130.

McNemar, Q. (1940). A critical examination of the University of Iowa studies of environmental influence upon the I.Q. *Psychological Bulletin*, **37**, 63–92.

Oldham, P. D. (1968). *Measurement in medicine*. London: English Universities Press.

Address correspondence to: M. J. R. Healy, Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT.

Zusammenfassung. Der Begriff "Regression auf den Mittelwert" wird häufig mißverstanden. Diese Arbeit erklärt das Konzept in einfacher Weise und zeigt, wie es in Studien der geistigen und körperlichen Entwicklung entsteht.

Résumé. La notion de "régression à la moyenne" est largement mal comprise. Ce travail explique le concept en termes simples et montre comment il survient dans des études de développement mental et physique.