The effects of age grouping on the estimate of a correlation coefficient

H. GOLDSTEIN

Institute of Education, University of London

Received 18 December 1980

Summary. When a correlation coefficient is calculated between two growth measurements over an age range, a biased estimate of the 'instantaneous' correlation is obtained. A correction formula is given.

Healy (1962) showed that the variance of a growth measurement was inflated when calculated from a sample spanning an age range during which the average value of the measurement changed. Using a reasonable set of assumptions, he showed how a correction factor could be calculated in order to obtain an unbiased estimate of the 'instantaneous' variance. The present note extends Healy's results to the calculation of covariances and correlations.

Healy used the characteristic function to obtain his results and although this is a simple way of obtaining estimates of the skewness and kurtosis, the basic results on the variance and covariance can be obtained using simpler expectation methods and these are used here.

Consider two measurements x, y taken over a unit (say one year) interval, whose mean and variance change with age over the interval. Following Healy, for age t, let

$$E(x|t)=a_1+b_1t, \qquad \mathrm{Var}(x|t)=c+dt,$$
$$E(y|t)=a_2+b_2t.$$

If the assumption is made that the ratio of the variances of y and x remains constant over the age range, then

$$\mathrm{Var}(y|t)=k^2\,\mathrm{Var}(x|t).$$

The further assumption is made that the correlation ρ between x and y is constant over the age range, which gives

$$\mathrm{Cov}(xy|t)=\rho\,\{\mathrm{Var}(x|t)\,\mathrm{Var}(y|t)\}^{1/2}.$$

Following Healy, assume that the population is distributed uniformly over the interval so that the distribution function can be written

$$F(t)=t, \quad -\tfrac{1}{2}\leqslant t\leqslant\tfrac{1}{2}.$$

Hence

$$E(x)=\int_{-1/2}^{1/2}(a_1+b_1t)\,\mathrm{d}t=a_1,$$
$$E(x^2)=\int_{-1/2}^{1/2}\{(a_1+b_1t)^2+c+dt\}\,\mathrm{d}t,$$
$$=c+a_1{}^2+b_1{}^2/12,$$

and $\mathrm{Var}(x)=c+b_1{}^2/12.$

0301-4460/81/0802 0181 $02·00 © 1981 Taylor & Francis Ltd

With analogous results for y. These are the results given by Healy. Thus the 'instantaneous' variance at the centre of the age interval is estimated simply by subtracting $b_1{}^2/12$ from the overall variance.

$$\begin{aligned} E(xy) &= \int_{-1/2}^{1/2} \{k\rho(c+dt)+(a_1+b_1t)(a_2+b_2t)\}\,\mathrm{d}t \\ &= k\rho c + a_1a_2 + b_1b_2/12 \end{aligned}$$

and $\operatorname{Cov}(xy) = k\rho c + b_1b_2/12.$

It follows that

$$c = \operatorname{Var}(x) - b_1{}^2/12,$$

$$k = \{[\operatorname{Var}(y) - b_2{}^2/12]/[\operatorname{Var}(x) - b_1{}^2/12]\}^{1/2},$$

$$\text{so that } \rho = \frac{\operatorname{Cov}(xy) - b_1b_2/12}{\{[\operatorname{Var}(y) - b_2{}^2/12][\operatorname{Var}(x) - b_1{}^2/12]\}^{1/2}}.$$

The following data on height (x) and weight (y) from the 1966 London County Council Survey were kindly supplied by Dr Noel Cameron (Cameron 1979). The sample consists of 261 girls uniformly distributed over the age range 6·0–7·0 years.

We have

$$\begin{aligned} \operatorname{Cov}(xy) &= 7{\cdot}93, \\ \operatorname{Var}(x) &= 19{\cdot}06, \\ \operatorname{Var}(y) &= 7{\cdot}45, \end{aligned}$$

and estimates of slopes are

$$\begin{aligned} b_1 &= 5{\cdot}2, \\ b_2 &= 2{\cdot}2, \end{aligned}$$

whence $\hat{\rho} = 0{\cdot}64$.

The unadjusted correlation is 0·67.

With extensive enough data, it would be possible to test some of the assumptions used in the derivation, in particular that the ratio of variances is constant and that the correlation is constant. If, failing this, it is suspected that these assumptions are not justified, an alternative method can be used, which is based on calculating residuals from the regression lines for each variable on age. Details of this procedure for a single variable are given in Goldstein (1979) and the extension of two variables is straightforward.

References

Cameron, N., 1979, The growth of London schoolchildren 1904–1966. *Annals of Human Biology*, **6**, 505–525.

Goldstein, H., 1979, *The Design Analysis of Longitudinal Studies* (London, New York: Academic Press).

Healy, M. J. R., 1962, The effect of age-grouping on the distribution of a measurement subject to growth. *American Journal of Anthropology*, **20**, 49–50.

Address correspondence to: Professor H. Goldstein, Department of Statistics and Computing, Institute of Education, Bedford Way, London WC1.

Zusammenfassung. Wenn ein Korrelationskoeffizient zwischen zwei Wachstumsmaßen über eine Altersspanne berechnet wird, ergibt sich eine fehlerhafte Schätzung der “Augenblicks”-Korrelation. Es wird eine Korrekturformel gegeben.

Résumé. Quand un coefficient de corrélation est calculé entre deux mensurations de croissance couvrant une certaine amplitude d’âge, on obtient une estimation biaisée de la corrélation ”instantanée”. Une formule de correction est donnée.