Behavior Genetics, Vol. 19, No. 2, 1989

Bias in Correlations from Selected Samples of Relatives: The Effects of Soft Selection

M. C. Neale,[1] L. J. Eaves,[1] K. S. Kendler,[2] and J. K. Hewitt[1]

Received 4 Aug. 1987—Final 1 Feb. 1988

Martin and Wilson (1982) describe two forms of sampling bias in twin studies. One is "hard selection," where individuals above a threshold participate, and those below do not. The second is "soft selection," where the probability of including a pair of relatives varies over the range of the character. We present an alternative model of soft selection which has strikingly different consequences for the resemblance between relatives. In general, the softer the threshold, the more the correlation resembles that in the underlying population. Results are presented where the probability of selection equals the cumulative distribution function of a normal distribution with 10% of the variance of the selected variable. In these circumstances, soft selection usually leads to less severely attenuated correlations than truncate selection.

KEY WORDS: twin studies; sampling bias; statistical selection.

INTRODUCTION

The resemblance of relatives forms the basis of human behavior genetic analysis. In order to draw correct conclusions about the nature of variation, it is important to obtain correlations that are unbiased estimates of the population parameters. One source of systematic and substantial bias is nonrandom sampling. This was noted by Martin and Wilson (1982), who provided tables of the probability of selection and the expected correlation between twins sampled in two nonrandom ways. First, they considered

[1] Department of Human Genetics, Medical College of Virginia, Box 33, Richmond, Virginia 23298.

[2] Departments of Human Genetics and Psychiatry, Medical College of Virginia, Box 710, Richmond, Virginia 23298.

the case of "hard selection" in which the probability of selection is zero for individuals with liabilities below a fixed threshold value and unity for individuals above this value. The effect of this type of selection was to reduce the correlations between relatives. This reduction was found to be proportionately *greater* for smaller than for larger correlations, and hence bias of this type in the classical twin study would lead to reduced estimates of common environmental variation or increased estimates of dominance variation. Second, Martin and Wilson employed a form of "soft selection" in which the probability of selection varied in a sigmoid fashion over the range of the character. This selection function was based on a model described by Curnow and Dunnett (1962). In this model, the observed variate, y, comprises a value x on an underlying distribution of liability, and a random error component z. Truncate (hard) selection is applied to the y variate to effect a sigmoid function of selection on the underlying variate, x. Since the error component z is assumed to be uncorrelated between relatives, quite different results are observed under this type of selection. In particular, as noted by Martin and Wilson, the attenuation of correlation is increased over the case of hard selection. What is not clear is the degree to which this increased attenuation is due to (a) the effect of the shape of the selection function or (b) the addition of variance which is uncorrelated between relatives. In the following account we present a description of a soft selection function which does not involve any attenuation of correlation due to the addition of a random error component to observed scores.

CALCULATION

Following Martin and Wilson, we define hard selection as truncation of a distribution at a fixed value and soft selection as a probability of selection which varies continuously over the character range. We consider only the case of soft selection arising from the sigmoid function $S(x) = \Phi[(x - \mu_s)/\sigma_s]$ obtained from the cumulative normal distribution $\Phi(z) = \int_{-\infty}^{z} \phi(\mathrm{t})\, d\mathrm{t}$, with

$$\phi(t) = (1/\sqrt{2\pi}) \exp\left(-\frac{t^2}{2}\right),$$

where μ_s is the value at which the probability of selection is 0.5 and σ^{-1} is a measure of the sensitivity of the probability of selection to the value of x. If we write $f(x_1,x_2,\rho_T)$ for the bivariate normal joint frequency distribution of x_1 and x_2, then the probability that a pair of relatives will be selected is

$$P_T = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, \rho_T) S(x_1) S(x_2) dx_2 dx_1$$

Variables x_1 and x_2 are normally distributed with mean μ and variance σ^2 and correlate ρ_T. Martin and Wilson computed an approximation to the probability of selection due to Curnow and Dunnett (1962). Owing to advances in computation, it is possible to obtain accurate estimates of multidimensional integrals (NAG, 1984) without excessive computer time. Therefore we calculated the means μ_i, variances σ_i^2, and covariance $\rho\sigma_i\sigma_j$ of the selected samples as follows:

$$\mu_i = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_i f(x_1,x_2,\rho_T)S(x_1)S(x_2)dx_2dx_1}{P_T},$$

$$\sigma_i^2 = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - \mu_i)^2 f(x_1,x_2,\rho_T)S(x_1)S(x_2)dx_2dx_1}{P_T},$$

$$\rho\sigma_1\sigma_2 = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1 - \mu_1)(x_2 - \mu_2)f(x_1,x_2,\rho_T)S(x_1)S(x_2)dx_2dx_1}{P_T}.$$

In the case of truncated distributions, the formulas are simplified because it is no longer necessary to integrate over the whole distribution, since the probability of selection is 1.0 for values above threshold and 0.0 for values below. Hence we have moments:

$$\mu_i = \frac{\int_{t_1}^{\infty}\int_{t_2}^{\infty} x_i f(x_1,x_2,\rho_T)dx_2dx_1}{P_{T_\mathrm{h}}},$$

$$\sigma_i^2 = \frac{\int_{t_1}^{\infty}\int_{t_2}^{\infty}(x_\mathrm{i} - \mu_i)^2 f(x_1,x_2,\rho_T)dx_2dx_1}{P_{T_\mathrm{h}}},$$

$$\rho\sigma_1\sigma_2 = \frac{\int_{t_1}^{\infty}\int_{t_2}^{\infty}(x_1 - \mu_1)(x_2 - \mu_2)f(x_1,x_2,\rho_T)dx_2dx_1}{P_{T_\mathrm{h}}},$$

where P_{T_h}, the probability of being selected under hard selection, is simply

$$\int_{t_1}^{\infty}\int_{t_2}^{\infty} f(x_1,x_2,\rho_T)dx_2dx_1.$$

Computation was performed using subroutines D01BBF and D01FBF from the NAG library (NAG, 1984). Calculation of integrals with limits $-\infty$ and ∞ was performed using Gauss–Hermite quadrature, while Gauss–Laguerre quadrature was employed for integrals with finite lower limits.

DISCUSSION

Table I shows the probability of accepting a pair into the sample for values of ρ_T, μ_s, and σ_s^2. Tabled values of μ_s correspond to 90, 70, 50,

Table I. The Probability that a Pair of Relatives Is Included in a Sample for Different Values of the Population Correlation, ρ_T, Threshold Value, μ_s, and Variance of Soft Selection Function, σ_s^2

		μ_s				
ρ_T	σ_s^2	−1.282	−0.524	0.0	0.524	1.282
.1	.0	.813	.502	.266	.102	.013
	.1	.794	.489	.264	.107	.016
.3	.0	.822	.528	.298	.128	.021
	.1	.802	.513	.294	.131	.024
.5	.0	.833	.557	.333	.157	.032
	.1	.813	.539	.325	.156	.034
.7	.0	.847	.590	.373	.191	.046
	.1	.826	.569	.360	.186	.048
.9	.0	.873	.638	.429	.238	.068
	.1	.844	.606	.403	.223	.066

30, and 10% selection from one side of the distribution when $\sigma_s^2 = 0.0$. This percentage of selection varies slightly under soft selection, when $\sigma_s^2 = 0.1$. For all values of μ_s and ρ_T, probabilities of selection correspond closely to those reported by Martin and Wilson. We note that for nearly all values of $\mu_s \leq 0$, hard selection yields a higher probability of selection than soft selection. In common with Martin and Wilson, we also conclude that the mean and variance of truncated samples do not differ widely as a function of the correlation ρ_T or the sensitivity σ_s^{-1}, for the tabulated parameters of the selection function. Clearly, for any heritable trait, truncation will cause greater attenuation of the dizygotic (DZ) sample than the monozygotic (MZ) sample. For many volunteer samples, MZ twin pairs are approximately twice as likely to volunteer as DZ twin pairs (Lykken *et al.*, 1978). Unless the heritability of the liability to volunteer is high *and* individual volunteer rates are 10% or less, truncation bias alone will not account for the difference in the volunteer rates between MZ and DZ twin pairs. It seems likely that this difference is caused by a lower threshold of volunteering in MZ twins, associated with greater interest in the phenomenon of twinning. As suggested by Lykken *et al.*, (1987) it is also possible that a selection process in which highly discordant twins are less likely to volunteer is operating.

In Table II we show the effects of sampling bias on the correlation between two relatives. The effects of sampling from a truncated distribution agree with those published by Martin and Wilson. However, quite different results are obtained for the case of soft selection. For the most part, expected correlations derived from a sample under soft selection

Table II. Expected Correlations ρ_T (Calculated by Numerical Integration) for Samples Drawn from Populations with Correlations ρ_T and Selection Functional Parameters μ_s and σ_s^2

		μ_s				
ρ_T	σ_s^2	-1.282	-0.524	0.0	0.524	1.282
.1	.0	.073	.052	.039	.029	.019
	.1	.073	.055	.044	.036	.027
.3	.0	.229	.172	.136	.105	.073
	.1	.230	.180	.149	.122	.094
.5	.0	.407	.326	.269	.218	.159
	.1	.407	.335	.286	.242	.193
.7	.0	.615	.530	.463	.396	.310
	.1	.613	.537	.478	.422	.352
.9	.0	.869	.818	.772	.717	.632
	.1	.859	.814	.775	.731	.666

with $\sigma_s^2 = .1$ are greater than those obtained under hard selection. For high correlation and a low threshold ($\mu_s = -1.282$, with $\rho_T \geq .7$, and $\mu_s = -.524$ with $\rho_T = .9$) soft selection causes slightly greater attenuation of the correlation than does hard selection. This result appears to be due to the nonmonotonic relationship between the mean of the selected sample and the variance of the selection function. The striking difference between "direct" soft selection as applied here and "indirect" soft selection as described by Martin and Wilson is most clearly seen when extremely soft selection is considered. In the limiting case, very soft selection produces a flat normal ogive, so that all members of the population have an equal chance of selection. Under our direct soft selection, the correlation between relatives selected in this (asymptotically random) fashion is not expected to change from the population value. However, under the indirect soft selection of Martin and Wilson, the correlation between relatives will tend to zero, irrespective of the true population correlation.

In order to provide some confirmation of our results, we simulated data from a correlated bivariate normal distribution using the SAS statistical package (SAS, 1985). Pairs of scores were then retained or rejected as a function of their joint probability of selection. One hundred thousand pairs of selected scores were simulated in this fashion for each of the 25 combinations of correlation and mean of the cumulative normal selection function. The proportion of pairs of scores retained during simulation differed by a maximum of .003 from the proportions reported in Table I. Pearson product–moment correlations were calculated between the pairs of scores and are shown in Table III. The correlations have a larger stan-

Table III. Product–Moment Correlations ρ_T for Simulated Data of Populations with Correlations ρ_T and Selection Functions with Threshold μ_s and Variance σ_s^2

		μ_s				
ρ_T	σ_s^2	−1.282	−0.524	0.0	0.524	1.282
.1	.1	.073	.053	.041	.034	.022
.3	.1	.229	.182	.149	.122	.097
.5	.1	.409	.339	.280	.236	.188
.7	.1	.611	.539	.482	.423	.352
.9	.1	.857	.815	.775	.733	.667

dard error than the proportions, but there is still very close agreement between these Monte Carlo results and those found by numerical integration shown in Table II. For 100,000 pairs, the z-transformed correlations have an expected standard error of approximately $1/\sqrt{n-3}$ = .0032, and 68% of the differences between the z-transformed correlations obtained by the two methods lie within one standard error of zero. We are therefore confident that our methods provide an accurate account of the effects of direct soft selection.

There is clear and occasionally substantial disagreement between the estimates of correlations given here and those given by Martin and Wilson. Under the sigmoid selection function employed by Martin and Wilson, most of the attenuation of correlation occurs as a result of the addition of a random error component to each individual's true score. In our view, this treatment involves unnecessary assumptions about the distribution of and covariation between relatives of a second source of variation on the trait. The alternative model described in this paper does not involve these assumptions, but clearly other assumptions are made. Effectively, soft selection is performed on all sources of variation in the trait, which may be inappropriate in some cases. However, it may be quite inappropriate to assume that sources of variation not relevant to selection are uncorrelated between relatives, as is the case for indirect soft selection. The decision as to which type of soft selection is the more appropriate will depend on the circumstances of selection. It is a simple matter to apply our existing software to explore cases which are a combination of direct and indirect soft selection. In addition, both methods assume independent selection of the members of a twin pair, and the same degree of selection for MZ and for DZ twins. These assumptions may also be relaxed.

When considering any particular variable, the investigator normally has an estimate of the prevalence of the disorder of interest, or of the proportion of the population included in the sample. This single statistic is clearly insufficient to estimate appropriate values for both the sensitivity and mean of the selection function. The situation is improved when the investigator has information about the mean and variance of the general population. In these circumstances, maximum-likelihood estimates of both statistics describing the selection function may be obtained. In the case of bivariate data (e.g., collected from twins), the correlation in the general population may also be estimated.

Despite the difference in direction of the change in predicted correlation under the two models of soft selection, similar general conclusions may still be drawn. Of particular note for human behavior genetic studies is that small correlations are reduced by nearly the same *amount* as large correlations. Hence the effects of factors such as the common environment shared by twins, or assortative mating, will be obscured in samples in which the variable under study affects (or is correlated with) the probability of selection. This effect may be marked for certain personality variables. For example, the twin study of self-report altruism conducted on a volunteer population of twins by Rushton *et al.* (1984) may well be drawn from an unrepresentative distribution and therefore be giving erroneous estimates of genetic and environmental variance.

REFERENCES

Curnow, R. N., and Dunnett, C. W. (1962). The numerical evaluation of certain multivariate normal integrals. *Ann. Math. Stat.* **33**:571–579.

Lykken, D. T., Tellegen, A., and De Rubeis, R. (1978). Volunteer bias in twin research; the rule of two-thirds. *Soc. Biol.* **25**:1–9.

Lykken, D. T., McGue, M., and Tellegen, A. (1987). Recruitment bias in twin research: the rule of two-thirds reconsidered. *Behav. Genet.* **17**:343–362.

Martin, N. G., and Wilson, S. R. (1982). Bias in the estimation of heritability from truncated samples of twins. *Behav. Gent.* **12**:467–472.

NAG (1984). *Numerical Algorithms Group FORTRAN Library Manual, Mark 11*, NAG, Oxford.

Rushton, J. P., Fulker, D. W., Neale, M. C., Blizard, R. A., and Eysenck, H. J. (1984). Altruism and genetics. *Acta Genet. Med. Gemellol.* **33**:265–271.

SAS (1985). *SAS User's Guide: Basics, Version 5 Edition*, SAS Institute Inc., Cary, N.C.

Edited by N. G. Martin