Behavior Genetics, Vol. 24, No. 3, 1994

A Simple Algebraic Demonstration of the Validity of DeFries–Fulker Analysis in Unselected Samples with Multiple Kinship Levels

Joseph Lee Rodgers[1] and Matt McGue[2]

Received 4 May 1993—Final 5 Mar. 1994

DeFries and Fulker's (*Behav. Genet.* **15**, 467–473, 1985) regression procedure (DF analysis) to estimate c^2 and h^2 was originally applied to selected twin data. Since then, DF analysis has been applied more broadly in unselected data and with multiple (nontwin) kinship levels. Theoretical work based on the matrix algebra of variance–covariance matrices has shown that estimates of c^2 and h^2 are unbiased in selected two-group settings. In this article, a simple proof is presented supporting the validity of DF analysis in broader settings. We use scalar algebra to show that parameter estimates of h^2 and c^2 are unbiased in unselected settings with multiple (more than two) kinship levels. Caveats are offered, and other DF analysis problems are identified.

KEY WORDS: DeFries–Fulker regression procedure; heritability; common-environmental influences; multiple kinship levels; algebra DF analysis; unbiased estimates.

INTRODUCTION

We present a simple algebraic demonstration that DeFries and Fulker's (1985) regression procedure for estimating heritability (h^2) and common environmental influences (c^2) provides unbiased estimates of population parameters in unselected samples with multiple kinship levels. Plomin and Rende (1991) suggested the name "DF analysis." DeFries and Fulker developed the procedure to compare monozygotic (MZ) and dizygotic (DZ) twins in which one of the twins had been selected for an extreme score on the trait of interest. They also suggested that the method could be applied in broader settings.

DeFries and Fulker's (1985) model estimates genetic and common environmental influences on traits under the assumptions of an additive behavioral genetic model. LaBuda *et al.* (1986) used matrix algebra to derive expected values showing that certain regression coefficients from the model are unbiased estimates of h^2 and c^2 in a MZ–DZ twin design in which one of the twins was selected on the basis of an extreme score. They also presented empirical analyses that included unselected twin pairs as well. Cherny *et al.* (1992b) used matrix algebra to derive expected values for parameters of submodels of the augmented model in unselected samples.

A number of empirical DF analyses have been published with unselected data (Cherny *et al.*, 1992a,b; Cyphers *et al.*, 1990; Detterman *et al.*, 1990; LaBuda *et al.*, 1986; Rodgers and Rowe, 1987; Rodgers *et al.*, 1994; Zieleniewski *et al.*, 1987). Several of these (Rodgers and Rowe, 1987; Rodgers *et al.*, 1994; Zieleniewski *et al.*, 1987) incorporated more than two kinship categories. Rodgers *et al.* (1994) adapted derivations from LaBuda *et al.* (1986) to unselected settings for MZ–DZ twin pairs and suggested that theoretical extensions of that algebra to unselected settings and multiple kinship levels should be addressed in future work.

There are several reasons to use DF analysis

[1] Department of Psychology, University of Oklahoma, Norman, Oklahoma 73019.

[2] Department of Psychology, 75 East River Road, Elliott Hall, University of Minnesota, Minneapolis, Minnesota 55455.

0001-8244/94/0500-0259$07.00/0 © 1994 Plenum Publishing Corporation

in broader settings. First, increased power results from using multiple kinship levels compared to the traditional pairwise comparison of correlations from MZ–DZ twins, adopted versus biological siblings, etc. Second, somewhat unusual kinship pairs (e.g., second cousins, offspring of twins, etc.) can be used alongside traditional kinship pairs (e.g., twins, full siblings). Third, several large national datasets have kinship information imbedded within them. DF analysis facilitates applying behavioral genetic models to problems that previously could not be addressed because of data limitations (although maximum-likelihood procedures are attractive in these settings as well). Fourth, DF analysis offers the advantage that it can provide control for genetic and shared environmental influences so that measured indicators of the nonshared environment can be studied without the usual genetic/environmental confounds that have plagued the socialization literature. Rodgers *et al.* (1994) illustrated the last two advantages in an analysis of problem behaviors in preadolescent children from the National Longitudinal Survey of Youth files.

We present the DF analysis model in the next section. Then, we present a simple algebraic derivation for the expectations of the parameters of the model. The derivation both simplifies and extends previous derivations. We conclude with caveats and identification of several other DF analysis problems.

DF ANALYSIS—BACKGROUND AND MODEL SPECIFICATION

DeFries and Fulker (1985) assumed an additive genetic model, no assortative mating and equal environmental influences across kinship categories. If the latter assumption—that c^2 is constant, and not a function of level of relatedness—is violated, c^2 estimates average environmental influence across levels of genetic relatedness.

The "augmented model" from DeFries and Fulker (1985) fits the following regression equation to data:

$$K_1 = b_0 + b_1 K_2 + b_2 R + b_3 (K_2 {*} R) + e \quad (1)$$

where K_1 and K_2 are scores on a given trait from a kinship pair, R is the coefficient of genetic relatedness, e is the residual of the model, and the b's are estimated least-squares regression weights. LaBuda *et al.* (1986) showed that $E(b_1) = c^2$ and $E(b_3) = h^2$. They also demonstrated that in selected twin samples, b_2 tests the equal environments assumption. Rodgers *et al.* (1994) showed that in unselected samples the interpretation of b_2 changes; there, a function of b_2 provides a second indirect estimate of h^2.

In the selected population that DeFries and Fulker (1985) first used, the proband defined the independent variable (K_2) in the model. In unselected settings such as those considered in this paper, the ambiguity as to which member of the kinship pair should be entered as K_1 and which as K_2 is resolved by using double-entry (e.g., Haggard, 1958) where each individual is entered once as K_1 and then again as K_2. This procedure doubles the sample size compared to that from selected settings. The estimates are still unbiased, however. Only statistical tests based on distributional assumptions (e.g., independence of errors) are affected by double-entry. An important feature of the double-entry procedure for our derivation is that K_1 and K_2 sample means and variances are identical (since every individual provides both K_1 and K_2 scores).

A SIMPLE ALGEBRAIC DEMONSTRATION

We first present our proof for the particular case of double-entered unselected MZ and DZ twins, then generalize to multiple kinship levels and general genetic relatedness. Under the additive genetic model (Falconer, 1981) expected values of correlations between scores of MZ and DZ twins are

$$E(r_{\mathrm{MZ}}) = h^2 + c^2 \quad (2)$$

$$E(r_{\mathrm{DZ}}) = .5h^2 + c^2 \quad (3)$$

Further, in the simple regression model

$$K_1 = a_0 + a_1 K_2 + e \quad (4)$$

the coefficient a_1 associated with K_2 can be computed as

$$a_1 = \mathrm{cov}(K_1,K_2)/\mathrm{var}(K_1) \quad (5)$$

But in double-entry settings where $\mathrm{var}(K_1) = \mathrm{var}(K_2)$,

$$a_1 = \mathrm{cov}(K_1,K_2)/\sqrt{\mathrm{var}(K_1)\,\mathrm{var}(K_2)} = r_{K1,K2} \quad (6)$$

Thus, a_1 in this simple model has the same expected value as r_{MZ} if K_1 and K_2 are scores for identical twins and as r_{DZ} if K_1 and K_2 are scores for fraternal twins.

We now expand the augmented DF model from

Eq. (1) for each of these settings. If $R = 1$, then

$$K_1 = b_0 + b_1K_2 + b_2 + b_3K_2 + e \\ = (b_0 + b_2) + (b_1 + b_3)K_2 + e \quad (7)$$

If $R = .5$, then

$$K_1 = b_0 + b_1K_2 + .5b_2 + .5b_3K_2 + e \\ = (b_0 + .5b_2) + (b_1 + .5b_3)K_2 + e \quad (8)$$

In each case, we have reduced the DF model from Eq. (1) to the simple model from Eq. (4) in which the regression coefficients are the MZ and DZ twin correlations. Equating these derived correlations to those defined in Eqs. (2) and (3) gives

$$h^2 + c^2 = E(b_1 + b_3) = E(b_1) + E(b_3) \\ .5h^2 + c^2 = E(b_1 + .5b_3) = E(b_1) + .5E(b_3) \quad (9)$$

These two equations are easily solved to give $E(b_1) = c^2$ and $E(b_3) = h^2$.

In a more general setting, consider (say) four levels of genetic relatedness indexed by R coefficients of g_1, g_2, g_3, and g_4 (e.g., $g_1 = 1.0$, $g_2 = .5$, $g_3 = .25$, and $g_4 = .125$ for MZ twins, full siblings, half-siblings, and cousins, respectively). Then the extensions of Eqs. (2) and (3) are the following :

$$E(r_1) = g_1h^2 + c^2 \quad (10)$$
$$E(r_2) = g_2h^2 + c^2 \quad (11)$$
$$E(r_3) = g_3h^2 + c^2 \quad (12)$$
$$E(r_4) = g_4h^2 + c^2 \quad (13)$$

Defining the same basic model as above and letting $R = g_1$, g_2, g_3, and g_4 successively produce DF models that reduce to the simple model in Eq. (4) for each specification:

$$R = g_1 \Rightarrow K_1 = (b_0 + g_1b_2) + (b_1 + g_1b_3)K_2 + e \quad (14)$$
$$R = g_2 \Rightarrow K_1 = (b_0 + g_2b_2) + (b_1 + g_2b_3)K_2 + e \quad (15)$$
$$R = g_3 \Rightarrow K_1 = (b_0 + g_3b_2) + (b_1 + g_3b_3)K_2 + e \quad (16)$$
$$R = g_4 \Rightarrow K_1 = (b_0 + g_4b_2) + (b_1 + g_4b_3)K_2 + e \quad (17)$$

As before, we equate the slope of the basic models in Eqs. (14)–(17) to their equivalent correlations from Eqs. (10)–(13):

$$g_1h^2 + c^2 = E(b_1 + g_1b_3) = E(b_1) + g_1E(b_3) \quad (18)$$
$$g_2h^2 + c^2 = E(b_1 + g_2b_3) = E(b_1) + g_2E(b_3) \quad (19)$$
$$g_3h^2 + c^2 = E(b_1 + g_3b_3) = E(b_1) + g_3E(b_3) \quad (20)$$
$$g_4h^2 + c^2 = E(b_1 + g_4b_3) = E(b_1) + g_4E(b_3) \quad (21)$$

Any pair of these equations can be used to solve for the expected values of b_1 and b_3, which are c^2 and h^2, respectively, as before. Note that this derivation applies to any levels of genetic relatedness and to any number of levels. Since any relation $g_ih^2 + c^2 = E(b_1) + g_iE(b_3)$ is proportional to any other, all such equations are consistent and will provide the solution above.

We now derive the expected value for the b_2 coefficient from Eq. (1). First note that $\overline{K_1} = \overline{K}_2$, because of the double-entry feature. Let this common mean be $\overline{K}$. Then, since the mean of the DV and IV aways fall on the regression line, we can use the DF model from Eq. (1) to assert that when $R = g_1$,

$$\overline{K} = b_0 + b_1\overline{K} + b_2g_1 + b_3g_1\overline{K} \Rightarrow \\ \overline{K}(1 - b_1 - g_1b_3) = b_0 + b_2g_1 \quad (22)$$

Similarly, when $R = g_2$,

$$\overline{K} = b_0 + b_1\overline{K} + b_2g_2 + b_3g_2\overline{K} \Rightarrow \\ \overline{K}(1 - b_1 - g_2b_3) = b_0 + b_2g_2 \quad (23)$$

These two equations can be solved to show that the expected value of b_2 is $-\overline{K}h^2$. This is the same solution found by Rodgers *et al.* (1994) using a different derivation that followed LaBuda *et al.* (1986). Thus, the coefficient b_2 from the DF model in Eq. (1) gives a second estimate of h^2 when it is divided by the negative of the reciprocal of the sample mean of the trait being measured.

DISCUSSION

We have presented a simple algebraic demonstration that DF analysis provides unbiased estimates of c^2 and h^2 in unselected samples and with multiple kinship levels. Of course, caveats must be offered as well. The equal environments assumption has been a major concern in previous DF analysis research. Loehlin (1989) applied maximum-likelihood methods to Bouchard and McGue's (1981) IQ kinship correlations and estimated separate c^2's for twins ($c^2 = .39$) and siblings ($c^2 = .27$). This assumption is even more questionable for cousins or other kinship pairs who do not necessarily live together in the same household. One approach to evaluating this problem with more than two kinship levels is to drop each level out of the model sequentially and refit the equation. While this approach does not provide separate estimates of c^2, it does give a clear indication of how kinship levels differ in contributing to the estimate of average c^2. Contributions to estimating h^2 may be evaluated as well. The ability to do this type of sequential refitting is an attraction of using more than two kinship levels.

A second problem—one that is especially likely in using large national data sets—is that an individual may be a part of more than one kinship pair (e.g., one child may be sibling to two others and cousin to yet a third). In this case, a correlated error structure can result that violates the assumption of independence of errors on which statistical tests are based. A possible solution involves coding the data within the context of a design that actually estimates the variance–covariance structure across members of a family (e.g., treating the different family members as repeated treatments in a repeated-measures design and using multivariate analysis). Alternatively, maximum-likelihood methods may be used that explicitly model correlated errors.

In general, maximum-likelihood procedures provide an attractive alternative to the regression approach on which DF analysis is based. Their advantages include increased power and fewer restrictive assumptions. On the other hand, regression procedures are more conceptually straightforward in most settings, and the estimation of maximum-likelihood models may be impractical with extremely large data sets. It is important to note that estimates obtained from least-squares and maximum-likelihood procedures can differ, and researchers must trade off the advantages offered by the two different approaches in deciding on the appropriate estimation procedure to use.

We conclude by noting that the development of DF analysis has an historical counterpart that occurred in the context of the development of regression analysis in the late 19th century. Sir Francis Galton (1885) developed his "reversion" technique (shortly thereafter renamed "regression") because he was interested in pairs of scores from a selected population: Given that a father is tall, will a son be tall as well? Quickly, regression was adapted to apply to unselected settings as well. Similarly, DF analysis was motivated by considering differential regression toward the mean of the cotwins of selected individuals who fell in the tail of the distribution of a trait. Like Galton's setting, however, DF analysis applies equally well to unselected pairs, and is in fact a useful and powerful method in such settings. A whole analytic arena is opened up by the ability to apply behavioral genetic modeling based on least-squares approaches (e.g., DF analysis) or maximum-likelihood methods to national probability samples.

REFERENCES

Bouchard, T., and McGue, M. (1981). Familial studies of intelligence: A review. *Science* **250:**223–238.

Cherny, S. S., Cardon, L. R., Fulker, D. W., and DeFries, J. C. (1992a). Differential heritability across levels of cognitive ability. *Behav. Genet.* **22:**153–162.

Cherny, S. S., DeFries, J. C., and Fulker, D. W. (1992b). Multiple regression of twin data: A model-fitting approach. *Behav. Genet.* **22:**489–497.

Cyphers, L. H., Phillips, K., Fulker, D. W., and Mrazek, D. A. (1990). Twin temperament during the transition from infancy to early childhood. *J. Am. Acad. Child Adolesc. Psychiat.* **29**(3):392–397.

DeFries, J. C., and Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behav. Genet.* **15:**467–473.

Detterman, D. K., Thompson, L. A., and Plomin, R. (1990). Differences in heritability across groups differing in ability. *Behav. Genet.* **20:**369–384.

Falconer, D. S. (1981). *Introduction to Quantitative Genetics,* Longman, New York.

Galton, F. (1885). Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst.* **15:**246–263.

Haggard, E. A. (1958). *Intraclass Correlation and the Analysis of Variance,* Dryden Press, New York.

LaBuda, M. C., DeFries, J. C., and Fulker, D. W. (1986). Multiple regression analysis of twin data obtained from selected samples. *Genet. Epidemiol.* **3:**425–433.

Loehlin, J. C. (1989). Partitioning environmental and genetic contributions to behavioral development. *Am. Psychol.* **44:**1285–1292.

Plomin, R., and Rende, R. (1991). Human behavioral genetics. In Rosenzweig, M. R., and Porter, L. W. (eds.) *Annual Review of Psychology,* Annual Reviews, Palo Alto, CA.

Rodgers, J. L., and Rowe, D. C. (1987). IQ similarity in twins, siblings, half-siblings, cousins, and random pairs. *Intelligence* **11:**199–206.

Rodgers, J. L., Rowe, D. C., and Li, C. (1994). Beyond nature vs. nurture: DF Analysis of nonshared influences on problem behaviors. *Dev. Psychol.* **30:**374–384.

Zieleniewski, A. M., Fulker, D. W., DeFries, J. C., and LaBuda, M. C. (1987). Multiple regression analysis of twin and sibling data. *Personal. Indiv. Diff.* **8:**787–791.

Edited by J. Meyer