Genetic Epidemiology 3:425–433 (1986)

Multiple Regression Analysis of Twin Data Obtained From Selected Samples

Michele C. LaBuda, J.C. DeFries, and D.W. Fulker

Institute for Behavioral Genetics, University of Colorado, Boulder

The multiple regression analysis of twin data in which a cotwin's score is predicted from that of a proband (the member of a twin pair selected because of a deviant score) and the coefficient of relationship provides a powerful test of genetic etiology (DeFries and Fulker: Behav Genet 15:467–473, 1985). Moreover, when an augmented model containing an interaction term is fitted to the same data set, direct estimates of heritability (h^2) and the proportion of variance owing to shared environmental influences (c^2) are also obtained. In the present paper, the expected partial regression coefficients estimated from these models are derived, and the flexibility of the general approach is illustrated. An extended model is formulated for the analysis of data from combined samples of affected and control twin pairs that yields tests for differential h^2 and c^2 in the two groups as well as pooled estimates of these parameters. The application of these models is illustrated by an analysis of data from reading-disabled and control twin pairs. Because of the ease, flexibility, and utility of the multiple regression analysis of twin data, it is an appealing alternative to more traditional model-fitting approaches.

Key words: multiple regression, twins, reading disability

INTRODUCTION

A comparison of identical (MZ) and fraternal (DZ) twin concordance rates for categorical variables (eg, presence or absence of pathology) is often used to infer genetic effects. However, when probands have been selected because of deviant scores on a continuous variable, the differential regression of MZ and DZ cotwins' scores toward the mean of the unselected population is a more suitable test. DeFries and Fulker [1985] recently proposed a multiple regression analysis of selected twin

Received for publication April 21, 1986; revision accepted August 18, 1986.

Address reprint requests to Michele C. LaBuda, Institute for Behavioral Genetics, University of Colorado, Campus Box 447, Boulder, Colorado 80309.

data in which a cotwin's score is predicted from the proband's score and the coefficient of relationship (R = 1.0 for MZ and 0.5 for DZ twin pairs) and showed that the partial regression of cotwin's score on R is a function of the difference between the means of MZ and DZ cotwins after covariance adjustment for any difference that may exist between probands, ie, a direct test of genetic etiology. Moreover, it was shown that when an augmented model that incorporates the interaction between proband's score and coefficient of relationship is fitted to the same data set, estimates of heritability (h^2) and the proportion of variance owing to shared environmental influences (c^2) are also obtained. This multiple regression approach is particularly relevant to the analysis of twin data from selected samples because regression coefficients are less influenced than correlations by restriction of range of the dependent variables [Morton, 1982, p 60].

DeFries and Fulker [1985] illustrated the multiple regression analysis of twin data by fitting their basic and augmented models to simulated data as well as to a small set of data from twin pairs in which one member of each pair had been selected because of low reading performance. Although these applications illustrated that the basic model provided a powerful test of genetic etiology and that the augmented model yielded appropriate estimates of h^2 and c^2, the expected partial regression coefficients for the models were not explicitly derived. Therefore, a primary objective of the present report is to derive the expected partial regression coefficients for both the basic and augmented models. In addition, the flexibility of this approach is demonstrated by formulating an extended model for the combined analysis of data from different groups by fitting this model to data from a sample of reading-disabled and control twin pairs.

METHOD

Basic and Augmented Models

The basic model in which a cotwin's score (C) is predicted from the proband's score (P) and coefficient of relationship (R) is as follows:

$$C = B_1P + B_2R + A, \tag{1}$$

where A is the regression constant, B_1 is the partial regression of cotwin's score on proband's score, ie, a measure of resemblance in the combined sample of MZ and DZ twins, and B_2 is the partial regression of cotwin's score on the coefficient of relationship. As shown below, B_2 is a function of the difference between MZ and DZ cotwins after covariance adjustment for the difference between MZ and DZ probands.

Estimates of h^2 and c^2 may be obtained by fitting the following augmented model to the same data set:

$$C = B_3P + B_4R + B_5PR + A, \tag{2}$$

where PR is the interaction between a proband's score and coefficient of relationship. As shown below, B_5 is equal to twice the difference between the MZ and DZ regressions and, therefore, provides a direct estimate of h^2. As is the case with the basic model, B_4 is a function of the difference between cotwins; however, it does not

yield a simple test of genetic etiology. As expected, B_3 is a measure of twin resemblance independent of h^2 and, therefore, is a direct estimate of c^2.

To derive the expected partial regression coefficients, the normal equations were formulated in terms of expected variances and covariances as follows:

$$S_X \cdot \boldsymbol{B} = S_{XY}, \tag{3}$$

where S_X is the expected variance-covariance matrix among the independent variables, S_{XY} is a column vector of expected covariances of the independent variables with the dependent variable, and $\boldsymbol{B}$ is a column vector of unstandardized regression coefficients [Cohen and Cohen, 1983, pp 472–473]. The expected regression coefficients were then derived by solution of equation 3:

$$\boldsymbol{B} = S_X^{-1} \cdot S_{XY}. \tag{4}$$

In order to derive the expected variances and covariances, the dependent and independent variables from the basic model were expressed as deviation scores (Table I). As shown in Table I, p_1 and p_2 and c_1 and c_2 are the deviations of an individual's score from its group mean for MZ and DZ probands and cotwins, respectively, and d_p and d_c are the deviations of the proband and cotwin group means from their population means (μ_p and μ_c).

Expected covariances were then formulated from cross-products of the deviation scores, summing across the MZ and DZ groups each weighted by its respective frequency. Allowing for unequal numbers of MZ and DZ twin pairs, the respective frequencies are expressed as n_1/N and n_2/N, where n_1 and n_2 are the number of MZ and DZ pairs, and N is the total number of twin pairs. For example, the expected covariance between the proband's score and the coefficient of relationship is derived as follows:

$$\begin{aligned} E(\mathrm{Cov}_{P,R}) &= E\,(P_i - \mu)\,(R_i - \bar{R}) \\ &= E\,[(n_1/N)\,(p_1 + d_p)(1/4) + (n_2/N)(p_2 - d_p)(-1/4)] \\ &= d_p(n_1 + n_2)/(4N) \\ &= d_p/4 \\ &= (\bar{P}_{MZ} - \bar{P}_{DZ})/8. \end{aligned} \tag{5}$$

TABLE I. Expectations of Variables Under the Basic Model*

Item	Zygosity	P	R	C
Raw score	MZ	$p_1 + d_p + \mu_p$	1	$c_1 + d_c + \mu_c$
	DZ	$p_2 - d_p + \mu_p$	1/2	$c_2 - d_c + \mu_c$
Mean		μ_p	3/4	μ_c
Deviation scores	MZ	$p_1 + d_p$	1/4	$c_1 + d_c$
	DZ	$p_2 - d_p$	−1/4	$c_2 - d_c$

*p_1, p_2 and c_1, c_2 are the deviations of an individual's score from its group mean for MZ and DZ probands and cotwins, respectively. Symbols d_p and d_c represent deviations of the proband and cotwin group means from their population means (μ_p and μ_c).

The remaining covariances and variances for the basic model were derived in a similar manner and are expressed as functions of means, phenotypic variance (V_P), additive genetic variance (V_A), and variance that is due to environmental influences shared by members of twin pairs (V_C):

$$S_X = \begin{array}{c} \\ P \\ \\ R \end{array} \begin{array}{c} \begin{array}{cc} P & R \end{array} \\ \begin{bmatrix} V_P + (\bar{P}_{MZ} - \bar{P}_{DZ})^2/4 & (\bar{P}_{MZ} - \bar{P}_{DZ})/8 \\ & 1/16 \end{bmatrix} \end{array},$$

$$S_{XY} = \begin{array}{c} \\ P \\ \\ R \end{array} \begin{array}{c} \begin{array}{c} C \end{array} \\ \begin{bmatrix} [(n_1 + n_2/2)/N]V_A + V_C + (\bar{P}_{MZ} - \bar{P}_{DZ})\,(\bar{C}_{MZ} - \bar{C}_{DZ})/4 \\ (\bar{C}_{MZ} - \bar{C}_{DZ})/8 \end{bmatrix} \end{array}.$$

Upon substitution of these matrices into equation 4, the following expected partial regression coefficients are readily derived:

$$B_1 = [(n_1 + n_2/2)/N]\ V_A/V_P + V_C/V_P \tag{6}$$

$$B_2 = 2[(\bar{C}_{MZ} - \bar{C}_{DZ}) - B_1\,(\bar{P}_{MZ} - \bar{P}_{DZ})]. \tag{7}$$

These results are based upon the assumption of a very simple additive model in which the covariance of MZ twin pairs equals $V_A + V_C$, whereas the covariance of DZ twins equals $V_A/2 + V_C$. Thus, B_1 is a measure of resemblance in the combined sample of MZ and DZ twins. B_2, on the other hand, equals twice the difference between MZ and DZ cotwins after covariance adjustment for the difference between MZ and DZ probands. If the difference between the means of probands and that of unselected population is heritable, DZ cotwins would be expected to exhibit significantly more regression toward the population mean than would MZ cotwins. B_2, therefore, is a direct test of genetic etiology.

In a similar manner, the expected partial regression coefficients may be derived for the augmented model, which includes the interactive product of the proband's score and the coefficient of relationship (PR). Following Table I, the interaction term is presented in mean deviation form in Table II. Using these deviations, as well as those for P, R, and C in Table I, the expected variances and covariances for the augmented model were derived in the manner outlined above. The resulting 3×3 variance-covariance matrix among the predictors (S_X) and the 3×1 covariance matrix of the independent variables with the dependent variable (S_{XY}) specified by the

TABLE II. Mean Deviation Representation of the Interactive Product*

Item	Zygosity	PR
Raw score	MZ	$p_1 + d_p + \mu_p$
	DZ	$p_2/2 - d_p/2 + \mu_p/2$
Mean		$d_p/4 + 3\mu_p/4$
Deviation scores	MZ	$p_1 + 3d_p/4 + \mu_p/4$
	DZ	$p_2/2 - 3d_p/4 - \mu_p/4$

*See footnote to Table I for explanation of the variables.

augmented model are as follows:

$$S_X = \begin{array}{c} \\ P \\ R \\ PR \end{array} \begin{array}{c} \begin{array}{ccc} P & R & PR \end{array} \\ \begin{bmatrix} V_P + (\bar{P}_{MZ} - \bar{P}_{DZ})^2/4 & (\bar{P}_{MZ} - \bar{P}_{DZ})/8 & [(n_1 + n_2/2)/N]V_P \\ & & + (\bar{P}_{MZ} - \bar{P}_{DZ})(\bar{P}_{MZ} - \bar{P}_{DZ}/2)/4 \\ & 1/16 & (\bar{P}_{MZ} - \bar{P}_{DZ}/2)/8 \\ & & [(n_1 + n_2/4)/N]V_P \\ & & + (\bar{P}_{MZ} - \bar{P}_{DZ}/2)^2/4 \end{bmatrix} \end{array},$$

$$S_{XY} = \begin{array}{c} \\ P \\ R \\ PR \end{array} \begin{array}{c} C \\ \begin{bmatrix} [(n_1 + n_2/2)/N]V_A + V_C + (\bar{P}_{MZ} - \bar{P}_{DZ})(\bar{C}_{MZ} - \bar{C}_{DZ})/4 \\ (\bar{C}_{MZ} - \bar{C}_{DZ})/8 \\ [(n_1 + n_2/4)/N]V_A + [(n_1 + n_2/2)/N]V_C + (\bar{P}_{MZ} - \bar{P}_{DZ}/2)(\bar{C}_{MZ} - \bar{C}_{DZ})/4 \end{bmatrix} \end{array}.$$

Upon substitution of these matrices in equation 3 and subsequent solution, the following expected partial regression coefficients were derived for the augmented model:

$$B_3 = V_C/V_P = c^2 \tag{8}$$

$$B_4 = 2\{(\bar{C}_{MZ} - \bar{C}_{DZ}) - [\bar{P}_{MZ}(h^2 + c^2) - \bar{P}_{DZ}(h^2/2 + c^2)]\} \tag{9}$$

$$B_5 = V_A/V_P = h^2. \tag{10}$$

Therefore, the augmented model proposed by DeFries and Fulker [1985] yields unbiased estimates of heritability and the proportion of variance owing to shared environmental influences. The B_4 term is again a function of twice the difference between MZ and DZ cotwins; however, for the augmented model, separate covariance adjustments are included for the MZ and DZ proband means. Whereas proband means in the basic model were weighted by the overall twin resemblance, proband means in the augmented model are weighted by their respective bivariate regressions.

Extended Model

In addition to providing a test of genetic etiology and estimates of h^2 and c^2, the multiple regression analysis of twin data may be extended in a wide variety of ways.

For example, estimates of h^2 and c^2 obtained from a regression analysis of selected twin data are potentially relevant to the unselected population [DeFries and Fulker, 1985]. Therefore, the augmented model can be extended and fitted to data from both selected and unselected samples simultaneously to test hypotheses of differential h^2 and c^2 in the two groups as well as to provide pooled estimates of these parameters.

The extended model incorporates an additional variate indicative of group membership and three interaction terms into the augmented regression equation:

$$C = B_6P + B_7R + B_8D + B_9PR + B_{10}PD + B_{11}RD + B_{12}PRD + A, \quad (11)$$

where D is a dummy variate for "diagnosis" of probands, ie, affected or not affected. Although several choices exist for the representation or coding of D [see Cohen and Cohen, 1983, pp 181–222, for discussion regarding the representation of nominal data in testing specific hypotheses], the desired contrasts can be tested regardless of the coding scheme employed. Nevertheless, as shown below, a straightforward interpretation of the individual regression coefficients from the extended model was facilitated by assigning values of $+0.5$ for affected probands and -0.5 for control probands. Resulting estimates of B_{10} and B_{12} provide direct tests for differential c^2 and h^2, respectively. Moreover, B_6 and B_9 provide pooled estimates of these parameters in the two groups.

Subjects

To illustrate the application of both the basic and augmented models, data from twin pairs in which one member of the pair had been selected owing to low reading performance were analyzed. Recently, as a part of an ongoing study of the genetics of reading disability [DeFries, 1985], an extensive test battery that assesses reading performance and related measures has been administered to 42 MZ and 37 DZ reading disabled twin pairs. To ascertain the twin sample, the school records of twin pairs available for study were examined for evidence of reading problems. Those twins in which at least one member of the pair evidenced reading problems were then administered the psychometric test battery. The member of a twin pair was designated reading disabled only if the following criteria were met: a score of 90 or above was achieved on either the Verbal or Performance subscales of the Wechsler Intelligence Scale for Children-Revised [WISC-R; Weschler, 1974], there was no evidence of neurological problems, and the subject was classified as being affected using a discriminant function analysis. Discriminant function weights were obtained from an analysis of independent reading-related test data (Reading Recognition, Reading Comprehension, and Spelling subtests of the Peabody Individual Achievement Test [Dunn and Markwardt, 1970], WISC-R Coding-B and Digit Span scaled scores [Wechsler, 1974], and the Colorado Perceptual Speed test [DeFries et al, 1981]) from a sample of 140 reading-disabled children and 140 matched control children in which 93.6% of the reading-disabled subjects and 92.9% of the controls were correctly reclassified. Because the members of the twin pairs were selected on the basis of discriminant scores, these data were subjected to multiple regression analyses.

In addition to the reading-disabled pairs, a group of 79 pairs of control twins were also tested ($N = 45$ MZ pairs and 33 DZ pairs). To ensure comparability of the samples, control twin pairs were matched to the reading-disabled pairs when possible on the basis of age, sex, and zygosity. Whereas the reading-disabled subjects were

selected owing to low reading performance, the controls were selected to be of normal or above normal reading ability. Hence, both groups are selected at least to some degree. The combined sample of reading-disabled and control twins was used to illustrate the extended multiple regression model.

APPLICATIONS

The average discriminant function scores of MZ and DZ probands and controls are presented in Table III. Inspection of the observed means for the reading-disabled sample suggests a greater regression toward the mean for DZ than MZ cotwins.

To test the hypothesis of genetic etiology of reading disability, the basic model represented by equation 1 was fitted to data from the reading-disabled sample. Resulting estimates for B_1 and B_2 are .86 ± .09 and −.90 ± .26, respectively. The significance of overall twin resemblance is manifested in the B_1 estimate. B_2, which is equal to twice the difference between MZ and DZ cotwins after covariance adjustment for the difference between probands, is also highly significant ($P < .001$). Thus, results of the multiple regression analysis of selected discriminant function twin data provide compelling evidence for genetic etiology of reading disability.

Application of the augmented model separately to data from the reading-disabled and control groups resulted in the partial regression estimates reported in Table IV. As expected, the estimate of heritability (B_5) is exactly equal to twice the difference between the MZ and DZ regressions (1.04 and .70 for the reading-disabled group and .82 and .39 for the control group, respectively), and the estimate of c^2 (B_3) equals the MZ regression minus h^2; ie, 1.04 − .68 = .36 in the reading-disabled sample, and .82 − .86 = −.04 in the control sample.

In order to test the hypothesis of differential h^2 and c^2 in the two groups, the extended model was fitted to the data on both groups simultaneously. Resulting estimates of B_6, B_9, B_{10}, and B_{12} are presented in Table V. By comparing the results of the separate and combined analyses, it may be seen that B_6 is the exact average of the c^2 estimates generated separately for each group. In a similar manner, B_9 is the average h^2. B_{10} is precisely the difference between the c^2 estimates, and B_{12} is the

TABLE III. Mean Discriminant Function Scores for MZ and DZ Probands and Cotwins*

	Reading-disabled		Control	
Zygosity	Proband	Cotwin	Proband	Cotwin
MZ	−.96	−.38	1.31	1.24
DZ	−1.08	−.03	1.51	1.18

*See text for explanation of discriminant function employed.

TABLE IV. Unstandardized Partial Regression Coefficients (± SE) Resulting From Application of the Augmented Model to Discriminant Scores of Reading-disabled and Control Twins

Coefficient	Reading-disabled	Control
B_3	.36 ± .28	−.04 ± .36
B_4	−.21 ± .45	−.85 ± .74
B_5	.68 ± .36	.86 ± .45

TABLE V. Coefficients Resulting From the Simultaneous Analysis of Data From Reading-disabled and Control Twins

Coefficient	Interpretation	Estimate (± SE)
B_6	Average c^2	.16 ± .23
B_9	Average h^2	.77 ± .29
B_{10}	Test of differential c^2	.40 ± .46
B_{12}	Test of differential h^2	−.18 ± .59

difference between the corresponding h^2 estimates. Neither B_{10} nor B_{12} is significant; thus, there is no evidence for differential c^2 or h^2 in the two groups. B_9, the pooled estimate of h^2, is highly significant ($P < .01$), whereas $B_6 = c^2$ is not.

DISCUSSION

The utility of multiple regression analyses of twin data make them appealing alternatives to conventional model-fitting approaches. They may be applied to either selected or unselected data sets, but are especially applicable to data from twins in which one member of each pair is selected because of a deviant score. When the basic model is fitted to these data, the partial regression of the cotwin's score on the coefficient of relationship provides a test of the extent to which the difference between probands and the unselected population is heritable. When the augmented model is fitted to the same data set, unbiased estimates of h^2 and c^2 are obtained.

In addition to the ease of the regression procedures, they are highly flexible. For example, tests of differential h^2 and c^2 in different groups can be accomplished by a straightforward extension of the augmented model. Many other possible applications can be readily envisioned. For example, an issue of considerable interest to the study of reading disability is the identification and characterization of subtypes. Analogous to the previous extension, differential h^2 and c^2 in these possible subtypes could be assessed by the addition of a dummy variate indicative of subtype classification. Results of such analyses could provide evidence for the external validity of alternative typologies. In general, a variety of independent variables such as ethnic group, gender, socioeconomic status, age, etc could be incorporated into the multiple regression analysis of twin data to assess differential h^2 and c^2.

The multiple regression analysis of twin data may be applied to other genetic relationships (eg, adoptive and nonadoptive siblings), as well as to the analysis of more than two relationships simultaneously. For example, reading ability data from selected twin and sibling pairs have recently been analyzed using an extension of the augmented model to estimate h^2 and differential c^2 for twins and siblings [Zieleniewski and Fulker, 1986]. Furthermore, with the incorporation of additional genetic relationships, multiple regression analysis may allow for testing of more complex genetic models than the simple additive system assumed in the traditional twin method. Just as the coefficient of relationship is used to reflect the degree of genetic resemblance that is due to additive genetic effects, additional dummy variates can be included to indicate resemblance that is due to nonadditive genetic effects. Although analysis of twin data alone would not yield separate estimates of additive and nonadditive genetic variances, the inclusion of data from other genetic relationships (eg, twins, sibs, and adoptive relations) would facilitate their resolution.

The multiple regression analysis of twin data is particularly advantageous when applied to data from selected samples; however, it requires the usual assumptions of more traditional twin analyses, eg, a linear polygenic model in which the coefficient of relationship for MZ and DZ twins is 1.0 and 0.5, respectively, and equal shared-environmental influences. In addition, the analysis is based upon data from pairs of relatives and assumes that each genetic relationship included in the analysis is independent; thus, the methodology is not optimal for the analysis of data from unbalanced pedigrees. Furthermore, the estimation procedure is unconstrained. However, the ease of application makes it an appealing alternative to more traditional model-fitting approaches.

ACKNOWLEDGMENTS

This work was supported in part by a program project grant from the NICHD (HD-11681), and the report was prepared while M.C. LaBuda was supported by NICHD training grant HD-07289. We wish to acknowledge the invaluable contributions of staff members of the Boulder Valley and St. Vrain Valley school districts and of the families who participated in the study. The authors would also like to thank an anonymous referee for critical review and careful editing of an earlier draft of the manuscript.

REFERENCES

Cohen J, Cohen P (1983): "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences," (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

DeFries JC (1985): Colorado reading projects. In Gray D, Kavanagh J (eds): "Biobehavioral Measures of Dyslexia." Parkton, MD: York Press.

DeFries JC, Fulker DW (1985): Multiple regression analysis of twin data. Behav Genet 15:467–473.

DeFries JC, Plomin R, Vandenberg SG, Kuse AR (1981): Parent-offspring resemblance for cognitive abilities in the Colorado Adoption Project: Biological, adoptive, and control parents and one-year-old children. Intelligence 5:245–277.

Dunn LM, Markwardt FC (1970): "Examiner's Manual: Peabody Individual Achievement Test." Circle Pines, MN: American Guidance Service.

Morton NE (1982): "Outline of Genetic Epidemiology." New York: Karger.

Wechsler DI (1974): "Examiner's Manual: Wechsler Intelligence Scale for Children, Revised." New York: The Psychological Corporation.

Zieleniewski A, Fulker DW (1986): Multiple regression analysis of twin and sibling data. Behav Genet 16:640 (abstract).

Edited by D.C. Rao