Behavior Genetics, Vol. 15, No. 5, 1985

Multiple Regression Analysis of Twin Data

J. C. DeFries[1] and D. W. Fulker[1]

Received 18 Feb. 1985—Final 10 May 1985

A multiple regression model for the analysis of twin data is described in which a cotwin's score is predicted from a proband's score and the coefficient of relationship (R = 1.0 and 0.5 for identical and fraternal twin pairs, respectively). This model is especially appropriate for the analysis of data on twins in which one member of each pair has been selected because of a deviant score, e.g., low reading performance. When the model is fitted to such data, the partial regression of the cotwin's score on the coefficient of relationship provides a powerful test of the extent to which the difference between the mean for probands and that for the unselected population is heritable, i.e., a test for genetic etiology. By fitting an augmented model containing an interaction term to either selected or unselected data sets, direct estimates of heritability and the proportion of variance due to shared environmental influences can also be obtained (subject, of course, to the usual assumptions underlying twin analyses, e.g., a linear polygenic model, little or no assortative mating, and equal shared environmental influences for identical and fraternal twins).

KEY WORDS: heritability; multiple regression; reading disability; twins.

INTRODUCTION

Twin studies of psychopathology typically ascertain affected index cases (i.e., probands) and then assess the status of their cotwins. For categorical variables, such as the presence or absence of a psychiatric illness, a simple comparison of identical (MZ) and fraternal (DZ) concordance rates provides a sufficient test for genetic etiology. However, if the probands have

This work was supported in part by a program project grant from the NICHD (HD-11681).

[1] Institute for Behavioral Genetics, University of Colorado, Boulder, Colorado 80309.

0001-8244/85/0900-0467$04.50/0

been selected because of deviant scores on a continuous variable, the differential regression of cotwin's scores toward the mean of the unselected population is a more appropriate test. For example, when probands have been selected on the basis of low test scores, cotwins of DZ probands are expected to have higher scores than cotwins of MZ probands if the condition is heritable. Thus, a simple t test of the difference between the means for cotwins of MZ and DZ probands would suffice when the probands have identical mean scores. However, the simultaneous regression of the cotwin's score on the proband's score and coefficient of relationship ($R = 1.0$ and 0.5 for MZ and DZ pairs, respectively) yields a more general test. Furthermore, when the interaction between the proband's score and the coefficient of relationship is added to the regression equation during a second step in the analysis of these data, direct estimates of heritability (h^2) and the proportion of variance due to common or shared environmental influences (c^2) potentially relevant to the unselected population are also obtained. In general, regression analyses of attenuated data are more appropriate than correlation analyses because regression coefficients are less influenced by restriction of range of the dependent variables (Cohen and Cohen, 1975, p. 65; Morton, 1982, p. 60).

The primary objective of the present short communication is to illustrate how multiple regression models may be fitted to twin data to provide tests of the relative importance of genetic and environmental influences. Although multiple regression analysis provides a test for genetic etiology when applied to data on twins in which one member of each pair has been selected because of a deviant score, it can also be used to obtain direct estimates of h^2 and c^2 from both selected and unselected samples. The method that we propose for the analysis of twin data provides a simple and flexible alternative to more conventional model-fitting approaches (e.g., Fulker, 1981). Moreover, it can be applied to data on other genetic relationships and may be generalized to accommodate any linear model of genetic and environmental influences.

MODELS

The basic multiple regression model for the analysis of selected twin data is as follows:

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

where C is a cotwin's predicted score, P is the proband's score, R is the coefficient of relationship, and A is the regression constant. B_1 is the partial regression of cotwin's score on proband's score and is a measure of twin resemblance that is independent of zygosity; B_2 is the partial

regression of cotwin's score on the coefficient of relationship and equals twice the difference between the mean for MZ and that for DZ cotwins after covariance adjustment for any difference that may exist between MZ and DZ probands. Therefore, B_2 provides a test of significance for genetic etiology analogous to that of differential twin concordance. Furthermore, the ratio of B_2 to the difference between the mean for probands and that for the unselected population yields an index of the extent to which the condition is heritable, symbolized h_g^2.

When probands are selected because of deviant scores, the first step in our analysis is to fit the regression model in Eq. (1) to the data and then evaluate the significance of B_2. Subsequently, direct estimates of h^2 and c^2 can be obtained when the following augmented regression model is fitted to the same data set:

$$C = B_3P + B_4R + B_5PR + A, \quad (2)$$

where PR is the product of the proband's score and the coefficient of relationship. Thus, during the second step in the analysis, the cotwin's score is simultaneously regressed on the proband's score, the coefficient of relationship, and their product. B_5, the coefficient corresponding to the interaction term (PR), is equal to twice the difference between the MZ and the DZ regression coefficients; thus, B_5 is a direct estimate of h^2 (assuming an additive model, little or no assortative mating, and equal shared environmental influences for MZ and DZ pairs) equivalent to that obtained using other estimation procedures. Moreover, when the augmented regression model [Eq. (2)] is fitted to twin data, B_3 is a direct estimate of c^2 because it is a measure of twin resemblance that is independent of genetic resemblance as indexed by the interaction term. B_4, on the other hand, does not have the simple expectation that B_2 has when estimated from the basic model.

Multiple regression analysis of twin data using the augmented model is similar to that previously reported by Ho *et al.* (1980) and Rose and Ditto (1983). However, they indexed zygosity using a categorical dummy variable instead of the coefficient of relationship and employed a hierarchical regression approach that focused largely on the significance of interaction terms. Although their method yields a significance test for genetic influence, it does not provide direct estimates of either h^2 or c^2. Thus, a very simple modification of their approach facilitates estimates of genetic and environmental parameters and tests of their significance from data on both selected and unselected samples.

APPLICATIONS

To illustrate our method, the basic and augmented regression models were fitted to two different sets of data: a simulated data set and reading

Table I. Simulated Twin Data to Illustrate Multiple Regression Analysis

Data		
Cotwin	Proband	Coefficient of relationship
62	54	1.0
62	52	1.0
64	50	1.0
58	48	1.0
54	46	1.0
72	54	0.5
74	52	0.5
64	50	0.5
74	48	0.5
66	46	0.5

	Summary statistics[a]					
Zygosity	$\overline{P}$	$\overline{C}$	s_p^2	s_c^2	r_{cp}	b_{cp}
MZ	50	60	10	16	0.79	1.0
DZ	50	70	10	22	0.40	0.6

[a] $\overline{P}$, proband mean; $\overline{C}$, cotwin mean; s^2, variance; r_{cp}, correlation; b_{cp}, regression of cotwin's score on proband's score.

performance data from reading-disabled probands and their cotwins. First, consider the simulated MZ and DZ twin data presented in Table I. Note that the means for MZ and DZ probands are equal, whereas the mean for MZ cotwins is 10 units lower than that for DZ cotwins. As expected for a heritable disorder, a greater regression toward the mean of the unselected population has occurred for DZ cotwins than for MZ cotwins. Thus, when the basic model of Eq. (1) is fitted to these data, $B_2 = -20.0 \pm 4.8$ ($P = 0.004$). B_2 is exactly equal to twice the difference between the mean for MZ and that for DZ cotwins because the means for the MZ and DZ probands are equal in this example.

If we assume that the mean of the unselected population is 100, then $h_g^2 = B_2/(50 - 100) = 0.4$ is an estimate of the extent to which the depressed scores of probands are due to heritable causes. It is important to note that this index may differ from h^2, a measure of the extent to which individual differences in the unselected population are heritable. Clearly, the etiology of the difference between the mean for probands and that for the unselected population may differ from that of differences among individuals within the normal range. For example, deviant scores may be due to a major-gene effect, a chromosomal anomaly, a special environmental insult, etc., whereas variation among individuals within

the normal range may be multifactorial in origin. If, on the other hand, probands merely represent the lower end of a normal distribution of individual differences, h_g^2 would be expected to equal h^2 as estimated during the second step in our analysis.

From Table I it may also be seen that the regression of the cotwin's score on the proband's score for MZ twins is 1.0, whereas that for DZ pairs is 0.6. Doubling the difference between these two regression coefficients yields an estimate for h^2 of 0.8, and subtracting this estimate from the MZ regression coefficient results in an estimate for c^2 of 0.2. As expected, when the augmented regression model [Eq. (2)] is fitted to these twin data, exactly the same results are obtained: $B_5 = 0.8$ and $B_3 = 0.2$. Additionally, when the analysis is performed using any available package of statistical computer programs, standard errors for these parameter estimates are automatically provided: 1.80 and 1.42 for h^2 and c^2, respectively. As with any other analysis of these data from five pairs of MZ and five pairs of DZ twins, the parameter estimates for h^2 and c^2 do not approach statistical significance. Given the summary statistics shown in Table I, a sample 20 times larger would be required to a yield a significant h^2. However, it is of some special interest to note that B_2 was significant when the basic model of Eq. (1) was fitted to this small data set. This finding suggests that our method may be powerful for detecting a genetic etiology of a disorder even with a relatively small sample of MZ and DZ twin pairs.

As expected with attentuated data, the correlations between cotwins' and probands' scores reported in Table I are lower than the corresponding regression coefficients. If we had doubled the difference between the MZ and the DZ correlations to estimate h^2, the resulting value (0.78) would have been similar to that obtained from our regression analysis. However, the estimate for c^2 of 0.01 would have been considerably different. This comparison demonstrates the advantage of regression methods over conventional correlation methods for the analysis of attenuated data sets.

We have also recently used the regression models of Eqs. (1) and (2) to conduct some preliminary analyses of data on 29 MZ and 20 DZ twin pairs in which at least one member of each pair is reading disabled. As part of a study of the etiology of reading disability (DeFries, 1985), a newly developed test battery that includes measures of cognitive abilities, reading and language processes, and patterns of electrophysiological activity has been administered to this sample. [See Decker and Vandenberg (1985) for an overview of the twin component of this program project and alternative analyses of data from a smaller sample.] With regard to the Reading Recognition subtest of the Peabody Individual Achievement Test (Dunn and Markwardt, 1970), mean z scores of MZ and DZ probands are

-1.64 and -1.68, respectively, whereas those for cotwins of the MZ and DZ probands are -1.34 and -0.93. When the basic regression model of Eq. (1) was fitted to these twin data, $B_2 = -0.88 \pm 0.20$. This result suggests that the difference between disabled and normal readers is due at least in part to heritable influences. If it is assumed that the mean for control subjects tested in this study (0.0) is equal to that for the unselected population, then $h_g^2 = -0.88/(0.0 - 1.66) = 0.53$.

The regression of the MZ cotwin's Reading Recognition score on the proband's score is 0.83, whereas that for DZ pairs is 0.37. Thus, application of the augmented regression model [Eq. (2)] to the twin data yields estimates for h^2 and c^2 of 0.92 ± 0.44 and -0.09 ± 0.38, respectively. Although h_g^2 is somewhat lower than h^2 in this example, they are not significantly different. In fact, the estimate for h_g^2 may be too low because the control mean in this study probably exceeds that of the unselected population.

DISCUSSION

The multiple regression analysis of twin data outlined above has several distinct advantages over conventional model-fitting procedures. First, our method is easy to understand and apply. Multiple regression computer programs are readily available and are familiar to researchers in the behavioral sciences. Small data sets are even amenable to analysis on personal computers using currently available statistical packages such as SYSTAT (Wilkinson, 1984). Second, our models may also be fitted to data on other genetic relationships. For example, when data from nonadoptive and adoptive sibling pairs (coefficients of relationship of 0.5 and 0.0, respectively) are analyzed, B_5 and B_3 again provide direct estimates of h^2 and c^2. Third, it is possible to analyze data from more than two relationships simultaneously (e.g., MZ and DZ twins, siblings, parent–offspring pairs, etc.). However, it would be necessary to include additional coefficients as dummy variables producted by probands' scores to model shared environmental influences that vary as a function of relationship. Fourth, as originally proposed by Ho *et al.* (1980), the method can be easily extended to facilitate age adjustment and/or developmental analyses by including age as another independent variable. Finally, other independent variables, such as ethnic group, gender, socioeconomic status, various environmental indices, etc., can also be readily incorporated in a comprehensive multiple regression analysis.

Our method is general for a variety of linear models and may be applied to either selected or unselected data sets but is especially applicable to data on twins in which one member of each pair is selected

because of a deviant score, e.g., low reading performance. When the basic model [Eq. (1)] is fitted to such data, the partial regression of the cotwin's score on the coefficient of relationship (B_2) provides a test of significance for genetic etiology. When the augmented model [Eq. (2)] is fitted to the same data set, B_5 and B_3 yield direct estimates of h^2 and c^2, respectively. This estimate of h^2 should be similar to our index of the extent to which the difference between probands and the unselected population is heritable (h_g^2) if affected individuals represent the lower end of a normal distribution of individual differences.

ACKNOWLEDGMENTS

We thank Sadie N. Decker and Steven G. Vandenberg for recruiting the twin sample, George P. Vogler for analyzing the reading data, Robert Plomin for making many helpful suggestions, and Rebecca G. Miles for providing expert editorial assistance.

REFERENCES

Cohen, J., and Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, N.J.

Decker, S. N., and Vandenberg, S. G. (1985). Colorado Twin Study of Reading Disability. In Gray, D., and Kavanagh, J. (eds.), *Biobehavioral Measures of Dyslexia*, York Press, Parkton, Md. (in press).

DeFries, J. C. (1985). Colorado Reading Project. In Gray, D., and Kavanagh, J. (eds.), *Biobehavioral Measures of Dyslexia*, York Press, Parkton, Md. (in press).

Dunn, L. M., and Markwardt, F. C. (1970). *Examiner's Manual: Peabody Individual Achievement Test*, American Guidance Service, Circle Pines, Minn.

Fulker, D. W. (1981). The genetic and environmental architecture of psychoticism, extraversion, and neuroticism. In Eysenck, H. J. (ed.), *A Model for Personality*, Springer-Verlag, New York.

Ho, H.- Z., Foch, T. T., and Plomin, R. (1980). Developmental stability of the relative influence of genes and environment on specific cognitive abilities during childhood. *Dev. Psychol.* **16**:340–346.

Morton, N. E. (1982). *Outline of Genetic Epidemiology*, Karger, New York.

Rose, R. J., and Ditto, W. B. (1983). A developmental-genetic analysis of common fears from early adolescence to early adulthood. *Child Dev.* **54**:361–368.

Wilkinson, L. (1984). *SYSTAT: The System for Statistics*, SYSTAT, Evanston, Ill

Edited by C. Robert Cloninger