In press, International Journal of Methods in Psychiatric Research

Multivariate Linear Regression Analysis of Childhood Psychopathology using Multiple Informant Data

MEREDITH A. GOLDWASSER
GARRETT M. FITZMAURICE

Department of Biostatistics, Harvard School of Public Health, Boston MA, USA

Correspondence should be sent to:

Meredith Goldwasser
Department of Biostatistics
Harvard School of Public Health
655 Huntington Avenue
Boston, MA 02115

mgoldwas@hsph.harvard.edu
phone: (617) 432-1056
fax: (617) 739-1781

ABSTRACT *It is common in psychiatric epidemiologic studies of childhood psychopathology to have multiple informant reports of mental health outcomes. The key challenge in analyzing multiple informant data concerns how they should best be represented in statistical models. Here we propose multivariate linear regression as the preferred method when the multiple informant outcome data are continuous. This approach permits the informant-specific information about mental health outcomes to be included in a single regression analysis, at the same time adjusting for the correlation between informant responses. The advantages of using a multivariate model include the ability to: (1) test for informant differences in outcome and assess if the effect of a risk factor on the outcome varies by informant; (2) estimate separate effects for each informant where necessary, or common effects where appropriate; (3) estimate the correlation between informant reports; (4) appropriately handle missing data by including data from all subjects with at least one informant report. To illustrate the application of this approach an example from the Connecticut Child Study is presented, examining risk factors for ``Internalizing'' behavior using parent and teacher informants.*

Key words: bivariate linear regression, multivariate analysis, child behavior disorders, psychometrics, risk factors

Introduction

The use of multiple informant reports of mental health has become standard practice in psychiatric epidemiologic studies. This is especially true in studies of child psychopathology in which parents, teachers, and peers are traditional informants. Psychological constructs are often elusive and, therefore, prone to measurement error. Multiple informant reporting is a common method for reducing the measurement error inherent in assessing psychopathology (Achenbach, McConaughy, Howell, 1987). The need for appropriate methods to analyze multiple informant data is even more pronounced in the study of child psychopathology, where children are assumed to lack the cognitive maturity needed for accurate self-reporting. In addition, psychopathology may be situation-specific or vary greatly by environment, in which case using multiple informant reports can lead to an understanding of the nature of the psychopathology not possible with only one informant response (Achenbach, 1993).

The central issue in analyzing multiple informant data is how to represent multiple outcomes for a subject from different informants in a statistical model. There are several traditional approaches for analyzing multiple informant data in psychiatric research that are not completely satisfactory. One approach is to conduct separate regression analyses for each of the informants and to report the results separately. This approach has a number of distinct drawbacks: (1) separate analyses yield multiple and often differing sets of results for the different informants; (2) separate analyses provide no formal statistical means of evaluating how similar or different the results are across the various informants; (3) in cases where the separate analyses yield results that are sufficiently similar, this approach provides no formal means of summarizing effects in a single set of results;

(4) separate analyses may be based on different subsets of the data if some subjects are missing one informant report and others are missing another.

The other main alternative is to use some sort of ``pooling'' strategy. There are three common pooling strategies which produce a summary measure from informant ratings to be used as the single outcome variable. When the outcome is dichotomous, e.g. presence or absence of psychopathology, the ``And'' rule dictates that the subject has psychopathology if all informants agree. In contrast, the ``Or'' rule indicates psychopathology if at least one informant reports its presence. Depending on which of these two rules is used, under or over-estimation of the prevalence of psychopathology may arise. A third strategy that can be used for both continuous and dichotomous outcomes is consensus pooling. Implemented during the data collection phase, a consensus procedure brings the discordant informants together to arrive at an agreement (Horton, Laird, Zahner, 1999). Practical considerations of the study make the consensus strategy difficult to use. Finally, for the continuous outcome case, another pooling strategy is to take the arithmetic mean of the multiple informant reports. We will return to discuss this strategy in the Discussion section.

Major disadvantages of these pooling strategies include: (1) the optimal algorithm for combining outcomes depends upon the type of measurement error present; although the measurement error in informant reports of psychopathology is believed to be substantial and also complex, there have been few studies that would support the use of one pooling strategy over another; (2) pooling informant reports does not permit the assessment of potential differences in risk factor effects

across the various informants; (3) many of the pooling algorithms are not clearly defined in the presence of missing informant reports.

For the special case where the outcome variable is dichotomous, Fitzmaurice et al. (Fitzmaurice, Laird, Zahner, Daskalakis, 1995) suggest handling multiple informant data by using multivariate logistic regression. This approach permits the informant-specific information about case status to be included in a single multivariate regression analysis, at the same time adjusting for the correlation between informant responses. The advantages of such an approach include the ability to: (1) test for informant differences in outcome and assess whether the effect of a risk factor on the outcome varies by informant; (2) estimate a common risk factor effect if it does not; (3) obtain measures of prevalence based on each informant or based on combined data from all informants (where appropriate); (4) handle missing data appropriately by including data from all subjects with at least one informant report; (5) assess informant agreement.

The current article extends the approach described in Fitzmaurice et al. to consider the setting where the outcome variable is continuous rather than dichotomous. Multivariate linear regression is proposed in place of multivariate logistic regression. A detailed description of the methodology behind this approach is presented in the next section, followed by an example from the Connecticut Child Study to illustrate how the method can be applied. An analysis of risk factors for ``Internalizing'' behaviors of 6- to 11-year-old children in Connecticut reported by parent and teacher informants is carried out using bivariate linear regression. Note that the use of the term ``bivariate'' here refers to the number of outcomes and not the number of predictors.

Multivariate Linear Regression Model

Similar to the use of multivariate logistic regression in the paper by Fitzmaurice et al., a multivariate linear regression model is proposed for the case of continuous outcome variables. Multivariate linear regression allows us to fit a single model for all informants, taking into consideration the (usually positive) correlation among informant reports on a given subject. Specifically, the multivariate linear regression model is comprised of multiple informant outcomes or responses for each subject, in addition to a set of covariates or predictor variables. Covariates in the model may include risk factors and their interactions, indicator variables for informant status, and informant-risk factor interactions. Informant-risk factor interactions are included in the model to allow the effect of risk factors on the outcome to differ by informant.

To illustrate these ideas consider a simple bivariate example: let Y_1 and Y_2 denote the responses from the two informants. Let X_1 be a risk factor and X_2 be the informant indicator (i.e. $X_2 = 1$ if the response comes from the second informant and $X_2 = 0$ if it comes from the first informant). Then the bivariate model for the mean response from the *j*th informant, as it relates to the risk factor is

$$E(Y_j) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \qquad j = 1,2$$

With the interaction present, this model allows for different intercepts and slopes for the 2 different informants. Therefore, the mean response for each informant follows as

$$E(Y_1) = \beta_0 + \beta_1 X_1$$

$$E(Y_2) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

Formal statistical tests of β_2 and β_3 equal to zero are tests of informant differences, where β_2 represents an informant difference in the intercept and β_3 represents an informant difference in the effect of the risk factor. Therefore, inclusion in the model of a statistically significant informant-risk factor interaction indicates the effect of a risk factor on the outcome depends upon the informant. In the absence of a significant informant-risk factor interaction the $\beta_3 X_1 X_2$ term may be dropped; an overall informant effect can still be retained in the model by including the $\beta_2 X_2$ term. A significant test of β_2 equal to zero then indicates that overall one informant is more or less likely to report symptoms than another. In general, the decision to estimate common or separate effects for each informant should not be based solely on a mechanistic use of significance tests. Rather, where a statistically discernible difference among the effects has been found, it should then be judged on the basis of subject-matter considerations. While small differences among effects may be statistically significant, these differences may be substantively unimportant. This will be especially true with large data sets.

Note that the methods proposed in this paper can be applied to the multivariate case where there are more than two informants. If, for instance, there were 3 informants then the model would include 2 indicator variables for the informant factor (e.g. representing contrasts of parent vs. peer and teacher vs. peer) and interactions between each informant variable and the risk factors. Interactions between clinically meaningful risk factors and other risk factors could also be included.

In summary, in the multivariate model the multiple informant outcomes are modeled simultaneously, thereby allowing the formal comparison of results across informants. In addition, in the multivariate model the correlation between informants is accounted for in the estimation of

the regression coefficients and their standard errors. In the presence of missing data, use of likelihood-based methods of estimation exploit the correlation among informant reports and allow for information on all subjects for whom there is at least one response to be included in the regression analysis, thereby increasing precision and potentially reducing selection bias.

Unlike the logistic regression scenario presented in Fitzmaurice et al., the regression coefficients here are not log odds ratios, but rather have the standard linear regression interpretation. The regression coefficients, their standard errors, and the within-subject covariance matrix can be estimated using standard software for general linear models for correlated data. Where there are missing data and the missingness mechanism is assumed to be *missing at random* (Rubin, 1976), related only to other observed responses, then likelihood methods lead to consistent and asymptotically normally distributed estimators. Assuming that the multiple informant outcomes have a multivariate normal distribution, likelihood based inference can be conducted using, for example, SAS Proc Mixed. Alternatively, where there are no missing data or the missingness mechanism is assumed to be *missing completely at random*, related only to covariates in the model, then a multivariate normal distribution need not necessarily be assumed, and a Generalized Estimating Equations (GEE) approach could be used instead (e.g., using SAS Proc Genmod). The GEE approach only requires correct specification of the model for the conditional mean of the outcome vector. No additional assumptions or restrictions on the error distribution are required.

Example: Connecticut Child Study

Sample and measures

The Connecticut Child Study consists of data collected on children age 6-11 years from two Connecticut children's mental health surveys. The New Haven Child Survey, conducted in 1986-1987, drew a random sample from 54 public and private schools serving New Haven, Connecticut. Using the same procedures as the New Haven Child Survey, the Eastern Connecticut Child Survey, conducted in 1988-1989, drew a stratified, two-stage cluster sample from 83 public, private and institutional schools nested within strata consisting of small cities, suburban areas, or rural areas (Zahner, Jacobs, Freeman, 1993; Zahner, Pawelkiewicz, DeFrancesco, 1992). Parent questionnaires were distributed through schools with follow-up of non-responders. One parent per household completed the survey for one randomly selected child. 70% and 72% of parents with valid addresses completed surveys in New Haven and Eastern Connecticut, respectively. Teacher surveys were then distributed with parental and schoolboard consent. Missing teacher surveys were the result of either permission refused by parents/schools or failure of the teacher to return the questionnaire. In addition, unscorable questionnaire data were obtained for 1% of parents and 2% of teachers. The analyses reported here are based on responses of 2,501 parents and 1,428 teachers.

The child psychopathology of interest here is ``internalization,'' a continuous outcome variable measuring withdrawn, somatic, and anxious-depressed problems. Internalizing behaviors were measured using the Internalizing Scale with parent ratings of the Child Behavior Checklist (Achenbach, 1991a) and with teacher ratings of the Teacher's Report Form (Achenbach, 1991b). The Child Behavior Checklist and the Teacher's Report Form are parallel versions of the same test. Using national norms of non-referred children, the raw scores were standardized to *t* scores with a mean of 50 and standard deviation of 10. Internalizing scores ranged from 33-93 for parents and 36-90 for teachers. Eight clinically significant categorical variables were included as study factors:

area of residence, social class, single parent status, maternal distress, child's health, grade repetition, child's gender, and family stress. It is important to note that all eight study factors are based on parental report. Table 1 summarizes the coding and distributional information for the outcome and independent variables. Details of missing responses and full descriptions of the study factors can be found elsewhere (Fitzmaurice et al., 1995; Zahner et al. 1992; Zahner et al. 1993). Marginally, the distributions of the dependent variable for parent and teacher informants appear to be very similar. However, this does not imply that their associations with the eight study factors will necessarily be the same.

Insert Table 1 about here

Methods

All analyses were conducted using the Proc Mixed procedure in SAS, exploiting that procedure's capability for analyzing data from repeated measures designs (here, informant represents the repeated measures factor). The Proc Mixed procedure is based on the general linear mixed model, and allows a variety of parametric structures for the covariance matrix. Sample syntax for undertaking the analysis is given in Appendix A. An ``unstructured'' covariance matrix was assumed in the current analyses. Note that, in general, choosing an ``unstructured'' covariance matrix allows the variances to depend on informant and places no restrictions on any of the pairwise correlations between informants. The current analysis uses Restricted Maximum Likelihood (REML) estimation of the covariances. The REML estimates of the covariances were used, in turn, to obtain the Generalized Least Squares (GLS) estimates of the regression coefficients and their standard errors. The nonresponse mechanism was assumed to be *missing at random*.

That is, the analysis presented here is valid provided that teachers' missing responses depend only on parents' evaluations and/or on the covariates in the model, but not on the missing teacher observations themselves.

In choosing a final model the strategy undertaken was to identify potential informant-study factor interactions first. Only after these were chosen were interactions between other study factors considered. The initial model included all 8 study factors, which have been shown to have epidemiologic significance in studies of childhood psychopathology, plus an indicator variable for informant status. After an overall test of all informant-study factor interactions was found to be statistically significant (Likelihood Ratio Test $\chi^2 = 72.56, 11df,\ p < .0001$), stepwise model-selection with Wald test statisics and a significance criterion of $p=.05$ was conducted. The stepwise technique alternates between forward selection and backward elimination steps, ending when none of the variables outside the model and every variable in the model has an F statistic significant at the significance criterion. The resulting model includes 4 interactions of informant with child's health ($p < .0001$), maternal distress ($p \approx .0006$), family stress ($p \approx .0012$), and SES ($p \approx .0543$). The interaction of informant with SES was retained since it was on the borderline of the criterion p-value. Recall that significant informant-study factor interactions indicate significant differences between the regression coefficients for parent and teacher reports.

Because predictors of psychopathology in children are often gender specific, only the interactions between child's gender and the 7 other study factors were considered. Before considering individual interactions of child's gender with other study factors, omnibus Likelihood Ratio tests of three-way, and then two-way interactions with child's gender were performed to control partially

for multiple testing. Because neither of these overall tests was statistically significant ($p > .20$ and $p > .13$, respectively), no further interactions with child's gender were considered. Therefore, the final model contains the 8 study factors, an indicator of informant status, and the 4 interactions of informant with child's health, maternal distress, family stress, and SES.

Results

Histograms (not shown here) of the parent and teacher reported internalization scores indicate that they are approximately normally distributed, lending support to the assumption that the parent and teacher reports have a multivariate normal distribution. In addition, a plot of parent versus teacher residuals (based on the bivariate model) indicates no departures from a linear association. Although there are no formal regression diagnostics available to test the adequacy of our model, separate residual plots for the parent and teacher reports indicate no systematic patterns, and therefore, seem to support the usual regression assumptions.

The estimated variances of the parent and teacher reports are 93.12 and 101.27 respectively, with an estimated correlation of .1632. Table 2 shows the regression coefficients and standard errors for the final model. A common estimate for parent and teacher indicates that there are no statistically discernible differences among the effects of the study factor on the parent and teacher reports of internalization. Separate estimates are reported when there are statistically discernible differences among the parent and teacher regression coefficients (at $\alpha = .05$), that is, when there is a significant informant-study factor interaction. As can be seen in Table 2, the effects of SES, maternal distress, child's health, and family stress on internalizing behaviors all differ by informant. Area of residence, single parent status, grade repetition, and child's gender do not depend on

informant. These results are quite similar to those reported in Fitzmaurice et al. based on the same data but with internalizing scores dichotomized at the cut-point for clinical psychopathology (*t* score > 60); the only exception being that SES was not found to differ by informant in the results presented by Fitzmaurice et al., while here the informant by SES interaction is marginally significant ($p = .0543$). In addition, Fitzmaurice et al. reported a significant interaction of child's gender with family stress. Along with different model selection criteria, the main differences in the two sets of analyses appear to be the result of a gain in statistical power from allowing the outcome variable to remain continuous, rather than dichotomizing it at a somewhat arbitrary cut-point. However, a more direct comparison of the logistic and linear regression coefficients is not meaningful since they have different interpretations.

Insert Table 2 about here

Study factor effects are generally in the expected direction. For instance, we would expect a decrease in SES to be associated with an increase in Internalizing score, whereas the unexpected negative association observed for single parent status (yes vs. no) is small and not statistically significant. The effects of maternal distress, child's health, and family stress are reported as separate effects, found to be significantly greater for parents than teachers. In particular, child's health and family stress are only significant predictors for the parent reports; maternal distress is significantly related to both parent and teacher reports, though it is more pronounced for the parents. One explanation for these parent-teacher differences is that these 3 study factors may be more salient for parents than teachers, keeping in mind that all factors are based on parental report.

The effects of maternal distress and child's health for parents represent a 4.68 and 4.21 respective change in Internalizing score, nearly half a standard deviation on the *t* score scale.

The effect of socioeconomic status was also found to differ significantly by informant, and separate effects are reported. Low socioeconomic status is significantly related to both the parent and teacher reports, though the effect is more pronounced for teachers. Middle socioeconomic status appears to have a significant effect only for parents. Of the 4 common effects for parents and teachers reported in Table 2, only grade repetition is significant, associated with a 2.13 increase in Internalizing score. Area of residence, single parent status, and child's gender do not appear to be significantly related to Internalizing score.

A caveat of the results reported here is that they have neglected to take into account selection probabilities or to make variance adjustments for stratified, multi-stage cluster sampling. Ignoring the complex sample design will generally result in overly optimistic estimates of precision. However, in a previous paper that examined the validity of results that did not take into account the complex sample design (Fitzmaurice et al., 1995), the average design effects obtained for separate regressions of the parent and teacher reports of internalization were 1.4 and 1.6, respectively. As a result, we conjecture that neglecting to account for the complex sample design has led to only a modest underestimation of the sampling variance of the regression estimates. That is, because the design effects are relatively small, accounting for the complex sample design would most likely lead to an inflation of the reported standard errors by a factor of 1.18 to 1.26 (the square-root of the design effects). This would not have any substantial impact on the overall substantive conclusions of the analyses reported here. Finally, we note that, in principle, it is possible to incorporate

sampling design adjustments into likelihood-based multivariate regression models. However, the statistical software to conduct such adjusted analyses does not currently exist. Extending the proposed method to incorporate sampling design adjustments is a topic that will be explored in future research.

Discussion

In this paper we propose an extension of the approach outlined by Fitzmaurice et al. to analyze multiple informant data that are continuous. Often it is more desirable both clinically and statistically to work with a continuous outcome variable, rather than imposing an arbitrary dichotomy. By using a continuous outcome variable there is more information available, thus producing a gain in both precision and power. The application of this new approach to dealing with multiple informant data follows directly from the dichotomous to the continuous case. Multivariate linear regression, accounting for the correlation among repeated measures, replaces multivariate logistic regression, but the structure of the model for the mean remains the same; in both cases there is a ``linear predictor'' that includes indicator variables representing informant status and their interactions with risk factors. The association between the repeated responses is modeled in multivariate logistic regression using the odds ratio; in multivariate linear regression the covariance is a natural metric for association. In either case, the estimated regression coefficients and their standard errors are adjusted for the effect of the correlations among informant reports.

One of the key features of the multivariate regression approach to handling multiple informant mental health data is that it allows us to test for informant differences in outcome and for differences in risk factor effects. When the effects of risk factors depend on informant, separate

estimates can be reported; when informant differences are not statistically significant, common estimates of risk factor effects can be reported. An advantage of the multivariate approach over traditional methods is that it takes into consideration the correlated nature of the data. By modeling the covariance structure between informants, estimated standard errors of informant-specific regression coefficients in the multivariate regression analysis will, in general, be smaller than their counterparts in univariate regressions for each informant. Furthermore, when the agreement among informant reports is also of scientific interest, the correlation between informant reports can be estimated. The multivariate approach is especially important in the handling of missing data where use of all available information on both informants may produce a gain in precision and a reduction in bias (when missing reports are *missing at random*).

In the earlier discussion of the disadvantages of traditional approaches for handling multiple informant data, it was briefly mentioned that one pooling strategy is to take the arithmetic mean of the multiple informant reports. It should be noted that this pooling strategy is encompassed within the proposed approach. Under a set of strong assumptions, the univariate regression coefficients estimated using the arithmetic mean of the multiple informant reports as the outcome will be identical to those estimated under the multivariate regression model proposed here (a proof is outlined in Appendix B). The two approaches will be equivalent when there are no missing data and under the assumptions of compound symmetry (a covariance matrix structure indicating equal variances among informants and equal correlations between informants) and no informant effects. The proposed approach allows these assumptions to be tested; informant effects can be tested as previously described, and the assumption of compound symmetry can be assessed by constructing a likelihood ratio test that compares compound symmetry with the unstructured covariance. The

arithmetic mean pooling approach, on the other hand, makes these assumptions a priori. In addition, when there are missing informant reports the arithmetic mean pooling approach uses a form of ``mean imputation,'' and thus can yield biased estimates of the regression coefficients and standard errors when data are *missing at random*. As a result, the arithmetic mean pooling approach should be implemented, when appropriate, using the multivariate regression method proposed here.

In conclusion, the ability to compare and formally test informant differences is a major advantage of the multivariate regression approach with potentially important consequences. Differences between informants may reflect reporting biases or may point to situation-specific psychopathology that would otherwise be overlooked. Thus, in the absence of a gold standard, using multivariate linear regression to identify risk factor effects for multiple informants may lead to advances in our understanding of the nature of childhood psychopathology.

Acknowledgements

We are grateful to Dr. Gwendolyn Zahner for the use of her data set. The Connecticut child surveys were conducted under contract to the Connecticut Department of Children and Youth Services while Dr. Zahner was on the faculty of the Yale University Child Study Center.

This research was supported by grants MH17119 and GM29745 from the National Institutes of Health.

References

Achenbach TM. Implications of multiaxial empirically based assessment for behavior therapy with children. Behav Ther 1993;24:91-116.

Achenbach TM. Manual for the child behavior checklist/4-18 and 1991 profile. Burlington, VT: University of Vermont Department of Psychiatry, 1991.

Achenbach TM. Manual for the teacher's report form and 1991 profile. Burlington, VT: University of Vermont Department of Psychiatry, 1991.

Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioural and emotional problems: implication of cross-informant correlations for situational specificity. Psychological Bulletin 1987; 101:213-32.

Fitzmaurice, GM, Laird, NM, Zahner, GEP, Daskalakis, C. Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. American Journal of Epidemiology 1995;142(11):1194-1203.

Horton NJ, Laird, NM, Zahner, GEP. Use of multiple informant data as a predictor in psychiatric epidemiology. International Journal of Methods in Psychiatric Research 1999;8(1):6-18.

Littell, RC, Milliken, GA, Stroup, WW, Wolfinger, RD. *SAS System for Mixed Models*. Cary, NC: SAS Institute, Inc., 1996.

Rubin, DB. Inference and Missing Data. Biometrika 1976;63:581-590.

Zahner GEP, Jacobs JH, Freeman DH, et al. Rural-urban child psychopathology in a northeastern U.S. state: 1986-1989. J Am Acad Child Adolesc Psychiatry 1993;32:378-87.

Zahner GEP, Pawelkiewicz W, DeFrancesco JJ, et al. Children's mental health service needs and utilization patterns in an urban community: an epidemiological assessment. J Am Acad Child Adolesc Psychiatry 1992;31:951-60.

Table 1: Coding and distribution of the dependent and independent variables

Internalizing (dependent variable)		
Parent informant	n=2501	
mean	50.72	
std. dev.	10.28	
median	51	
75th%	58	
25th%	43	
Teacher informant	n=1428	
mean	49.77	
std. dev.	10.21	
median	49	
75th%	57	
25th%	43	
Independent Variables	#	%
Area		
1=rural	874	35
2=suburban	428	17
3=small city	386	15
4=large city	813	33
Social Class		
1=high	1240	50
2=middle	949	38
3=low	312	12
Single Parent		
1=yes	519	21
0=no	1982	79
Maternal Distress		
1=yes	391	16
0=no	2110	84
Child's Health		
1=poor health	1172	47
0=good health	1329	53
Grade Repetition		
1=yes	466	19
0=no	2035	81
Child's Sex		
1=boy	1207	48
2=girl	1294	52
Family Stress		
1=yes	1596	64
0=no	905	36

Table 2: Estimated regression coefficients ± standard errors for the Internalization Model

	Bivariate Linear Regression*	
	Teacher	Parent
Intercept	47.6896 ± 0.5954	45.4168 ± 0.4651
Area		
large city		-0.7133 ± 0.4529
small city		-0.1529 ± 0.5116
suburban		-0.9233 ± 0.4873
Social Class		
low	3.524 ± 0.8965**	1.3664 ± 0.6861**
middle	0.9156 ± 0.5887	1.1190 ± 0.4305**
Single Parent (yes)		-0.2901 ± 0.4790
Maternal Distress (yes)	1.7604 ± 0.7349**	4.6795 ± 0.5450**
Child's Health (poor)	0.6951 ± 0.5372	4.2069 ± 0.3912**
Grade Repetition (yes)		2.1273 ± 0.4381**
Child's Sex (boy)		0.6064 ± 0.3310
Family Stress (yes)	0.6924 ± 0.5617	2.8081 ± 0.4094**

* Single estimates indicate common effects for teachers and parents; separate estimates indicate effects that are significantly different between informants at $\alpha = .05$.
** Statistically significant ($\alpha = .05$) risk factors for Internalizing score.

Appendix A: Sample SAS Commands

Sample Data

ID	PARENT	TEACHER	GENDER	SES
1	51	53	1	1
2	43	.	0	1
3	70	57	0	2

Above is an example of 3 subjects from the data analysis presented in the paper, with only Gender and SES included as covariates. Subject 2 has a missing teacher response, represented by a period. Note that Proc Mixed requires the data to be in a univariate form, with as many records as there are informant reports. Below is the SAS code for carrying out the data transformation, followed by the resulting SAS output. The ``intern'' and ``inf'' variables, representing internalizing score and the corresponding informant, are created in the SAS code. It can be seen in the output that there are two entries per subject, one row or record for parent report and one for teacher report.

SAS Code and Output - transformation to univariate form

```
data internal;
   input id parent teacher gender ses;
   intern=parent;inf=0; output;
   intern=teacher;inf=1; output;
   drop parent teacher;
run;
```

ID	GENDER	SES	intern	inf
1	1	1	51	0
1	1	1	53	1
2	0	1	43	0
2	0	1	.	1
3	0	2	70	0
3	0	2	57	1

SAS Code - Proc Mixed

```
proc mixed data=internal noclprint;
    class id inf ses;
    model intern=inf gender ses inf*gender inf*ses/s;
    repeated inf/type=un subject=id;
run;
```

The multivariate linear regression analysis is carried out using the SAS Proc Mixed code presented above. The Model statement specifies the regression of internalizing score onto an indicator of informant status, the Gender and SES risk factors, and their interactions with informant. The Repeated statement in Proc Mixed is used to identify observations that are correlated and to model the covariance structure. Informant (``inf'') represents the repeated measures factor, and in this example ensures that missing informant reports are handled correctly. An unstructured covariance matrix is specified with the ``type=un'' option.

Appendix B: Proof of Equivalence of Mean Pooling Approach and Multivariate Linear Regression

Arithmetic Mean Pooling Approach - Ordinary Least Squares (OLS) Regression

Let $Z_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$, where Y_{ij} is the *j*th informant report on the *i*th subject $(i = 1,\ldots,N,\ \ j = 1,\ldots,n)$.

We assume $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix} = X\beta + \varepsilon$

where $\varepsilon \sim N(0, \sigma^2 I_N)$, $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{bmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, and I_N denotes an $N\,x\,N$ identity matrix.

Then the OLS estimate of β is $\hat{\beta} = (X'X)^{-1}X'Z$.

Multivariate Linear Regression Approach

Assume no missing data, compound symmetry covariance matrix, no informant effects, and

$E(Y_i) = X_i\beta$, where $Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in} \end{pmatrix}$, $X_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{ip} \\ 1 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & \cdots & x_{ip} \end{bmatrix}$.

Note, as seen in X_i above, a subject's covariates do not change across informant since we have assumed no informant effects.

Using Generalized Least Squares (GLS) estimation, the estimate of β is

$\hat{\beta} = (\sum_{i=1}^{N} X_i' \Sigma^{-1} X_i)^{-1} \sum_{i=1}^{N} X_i' \Sigma^{-1} Y_i$.

Under the assumption of compound symmetry, $\Sigma = \sigma^2(1-\rho) I_n + \sigma^2 \rho J_n$, where I_n is the identity matrix and J_n is a matrix of 1's, both of size *n x n*.

Then $\Sigma^{-1} = \frac{1}{\sigma^2(1-\rho)}(I_n - cJ_n)$, where $c = \frac{\sigma^2 \rho}{\sigma^2(1-\rho) + n\sigma^2\rho}$ is a constant.

Substitution of the above yields:

$X_i' \Sigma^{-1} X_i = \frac{1-cn}{\sigma^2(1-\rho)} X_i' X_i$;

$\sum_{i=1}^{N} X_i' \Sigma^{-1} X_i = \frac{1-cn}{\sigma^2(1-\rho)} nX'X$, where X is defined under the mean pooling approach.

$X_i' \Sigma^{-1} Y_i = \frac{1-cn}{\sigma^2(1-\rho)} X_i' Y_i$;

$$\sum_{i=1}^{N} X_i' \Sigma^{-1} Y_i = \frac{1-cn}{\sigma^2(1-\rho)} \sum_{i=1}^{N} \begin{pmatrix} \sum_{j=1}^{n} Y_{ij} \\ x_{i1} \sum_{j=1}^{n} Y_{ij} \\ \vdots \\ x_{ip} \sum_{j=1}^{n} Y_{ij} \end{pmatrix}.$$

Finally, combining the results above, the GLS estimate can be rewritten as

$$\hat{\beta} = (X'X)^{-1} \sum_{i=1}^{N} \begin{pmatrix} \frac{1}{n} \sum_{j=1}^{n} Y_{ij} \\ x_{i1} \frac{1}{n} \sum_{j=1}^{n} Y_{ij} \\ \vdots \\ x_{ip} \frac{1}{n} \sum_{j=1}^{n} Y_{ij} \end{pmatrix} = (X'X)^{-1} X'Z .$$

This is identical to the OLS estimate, $\hat{\beta}$, from the arithmetic mean pooling approach.