

Multivariate Regression Models for Analyzing Data from Multiple Informants

Garrett M. Fitzmaurice

Division of General Medicine, Brigham and Women's Hospital
Department of Biostatistics, Harvard School of Public Health

Joint work with Nick Horton (Smith) and Nan Laird (Harvard)

Support provided by NIMH grant R01-MH54693

Outline

- Statement of Problem
- Motivating Examples
- Historical Approaches
- Multivariate Regression Models for Multiple Informant Outcomes
- Multivariate Regression Models for Multiple Informant Predictors
- Concluding Remarks

Background: Multiple Informant Data

In many studies, information on health outcomes and/or risk factors is obtained from multiple sources (or “informants”).

Psychopathology in Children: teacher, parent, and self-reports are commonly used as data sources about underlying psychological state.

Geriatrics: relative, caregiver, and self-reports are commonly used as data sources about cognitive function in the elderly.

- Multiple informants also known as “multiple sources”, proxy informants, co-informants.
- Multiple source data collected to provide better measures of some underlying construct.
- **Note:** Discordance is expected, otherwise there would be no need for multiple source data.

Mean corr. between informants (meta-analysis)

	PAR	TCH	MHW	OBS	Peer
PAR	.59				
TCH	.27	.64			
MHW	.24	.34	.54		
OBS	.27	.42		.57	
Peer		.44 ^a			.73

a: pooled peers

PAR: Parent; TCH: Teacher

MHW: Mental Health Worker; OBS: Observer

Source: Achenbach *et al.*, Psych. Bulletin, 1987

- Term “multiple source” data encompasses data that are simultaneously obtained from multiple informants or raters or via different/parallel instruments or methods.
- Multiple source data are assumed to be *commensurate*: provide multiple measures of same underlying variable and measured on similar scale.
- Multiple sources can provide data on outcomes and/or predictors.
- Key Methodological Challenge: How best to incorporate multiple source data in statistical models?
- In this talk, I will outline a multivariate regression-based method for analyzing multiple source data.

Motivating Example: Simmons Longitudinal Study (SLS)

Researchers: Simmons College School of Social Work

Study: Longitudinal community sample of children (now adults, ~ in their mid-30's).

Unit of observation: Young adult.

Multiple source variables: Family history of substance abuse reported by young adult and mother.

Outcome of interest: Anxiety problems for young adult and their association with familial substance abuse.

Motivating Example: Stirling County Study

Researchers: Harvard School of Public Health

Study: Longitudinal prospective cohort study of psychiatric problems and mortality.

Unit of observation: Individual.

Multiple source variables: Physician and self-reports of psychiatric disorders.

Outcome of interest: Mortality during 16 year follow-up period and its association with psychiatric disorder.

Motivating Example: Eastern Connecticut Child Survey

Researchers: Harvard School of Public Health

Study: Epidemiologic study of psychopathology in children.

Unit of observation: Child.

Multiple source outcome: Parent (CBC) and teacher (TRF) measures of childhood psychopathology (externalizing scale).

Predictors of interest: Single parent status and child's physical health problems.

Notation

Assume outcome and/or predictor(s) obtained from J different “sources”.

Let Y denote the outcome (continuous, binary, ordinal, or count data).

Let X denote the predictor(s)/covariate(s).

Multiple source outcome: Let Y_j represent the outcome obtained from the j^{th} source (with $j = 1, \dots, J$).

Multiple source predictor(s): Let X_j represent the predictor(s) obtained from the j^{th} source (with $j = 1, \dots, J$).

Historical Approaches to Analysis of Multiple Source Data

Historically, the following analytic methods have been used:

Consensus Decision

- “best estimate” diagnosis, where clinicians/experts review data from all sources and arrive at a diagnosis
- by producing single number summary, sidesteps multiple source data analytic issues
- somewhat simpler to implement for binary variables
- may not always be possible
- in general, needs to be implemented at data collection stage

Algorithms for combining sources

For binary data:

- (i) Let $X = 1$ if $X_1 = 1$ or $X_2 = 1$, and $X = 0$ otherwise (“OR” rule).
- (ii) Let $X = 1$ if $X_1 = X_2 = 1$, and $X = 0$ otherwise (“AND” rule).

For continuous variable: Take arithmetic average, $X = \text{mean}(X_1, X_2)$.

- by producing single number summary, X , sidesteps multiple source data analytic issues
- optimal pooling algorithm not always clear (resulting in loss of information)
- can yield overestimate (“OR” rule) or underestimate (“AND” rule) of prevalence
- cannot examine differences in effects across sources
- many algorithms are not clearly defined when there are missing data

Use single source

- analyze data using a single source only, say X_1
- preferred source (discount reports of other sources)
- conceptually simple and easy to implement
- inefficient (but all too common)
- results may be sensitive to choice of source/informant

Separate analyses for each source

- conduct separate analysis for each source
- addresses issue of sensitivity of choice of informant
- multiple (and often differing!) sets of results
- no formal means of evaluating similarity or differences
- analyses may be based on different subsets of the data

Include both sources in analysis

- conduct analysis that simultaneously includes all sources
- model $f(Y|X_1, X_2)$

e.g., $E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

- incorporates all available data in analyses
- may be difficult to interpret, especially when X_1 and X_2 are strongly related?

Multivariate Regression-Based Methods

To overcome some of the shortcomings of existing methods, we propose regression methodology for simultaneously analyzing multiple source data.

These models can be considered special cases of generalized linear models, albeit with potentially correlated outcomes.

For ease of exposition, we describe regression models for multiple source outcomes and predictors separately.

Multivariate Regression Models for Multiple Source Outcomes

Notation: Assume N subjects, each with an outcome, Y , obtained from J different sources.

Outcome can be continuous, binary, ordinal or count data.

Let Y_{ij} represent the outcome obtained for the i^{th} subject from the j^{th} source (with $i = 1, \dots, N$; $j = 1, \dots, J$).

Let X_{ij} be a $p \times 1$ vector of covariates associated with Y_{ij} : each X_{ij} will in general contain both source variables (or indicators for the different sources) and subject-specific covariates.

Consider regression models relating the mean of the outcome measured by each source, $E[Y_{ij} | X_{ij}]$, to the vector of covariates,

$$g(E[Y_{ij} | X_{ij}]) = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp},$$

where $g(\cdot)$ is a known link function.

For example, for a binary outcome a logistic regression model

$$\text{logit}(E[Y_{ij} | X_{ij}]) = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp},$$

might be adopted.

Simple Illustration: *Eastern Connecticut Child Survey*

Epidemiologic study of children that assessed psychopathology using ratings from parents and teachers.

Suppose we are interested in assessing the association of family stressors (binary) with child psychopathology.

The following *bivariate* regression allows a single regression model to be fit to data from both sources,

$$g(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1 SOURCE_{ij} + \beta_2 STRESS_i + \beta_3 (SOURCE_{ij} * STRESS_i) ,$$

where SOURCE (0=parent, 1=teacher) and STRESS (0=no, 1=yes) are indicator variables.

In general, source-related differences in the effect of family stress can be evaluated via test of β_3 equaling zero.

For example, the simplified bivariate regression model,

$$g(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1 SOURCE_{ij} + \beta_2 STRESS_i ,$$

assumes that the effect of family stress on the mean rating does not vary by source:

$$Parent : \quad g(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_2 STRESS_i$$

$$Teacher : \quad g(E[Y_{ij}|X_{ij}]) = (\beta_0 + \beta_1) + \beta_2 STRESS_i$$

Note 1: Regression model can be used to specify hypotheses about the effects of the sources and of covariates on the outcome, as well as possible source by covariate interactions.

Note 2: A test of a source by covariate interaction is equivalent to a test of the differences among the source-specific regression coefficients for the corresponding covariate.

Note 3: Source by covariate interactions represent contrasts of within-subject effects.

Note 4: Because multiple source data are expected to be positively correlated, these regression coefficients are also positively correlated; consequently we have more power than we usually would have for testing interactions.

Multivariate Regression Models for Multiple Source Predictors

Multiple source reports are also commonly used as predictors of an outcome.

Notation: Assume N subjects, each with a predictor, say X , obtained from J different sources.

Outcome, denoted Y , can be continuous, binary, ordinal or count data.

Let X_{ij} represent the predictor obtained for the i^{th} subject from the j^{th} source (with $i = 1, \dots, N$; $j = 1, \dots, J$).

Let Z_i denote other covariates of interest (discrete or continuous), included in model.

For simplicity of exposition, we consider two binary source reports.

Consider the following joint models:

$$g(E[Y_i|X_{i1}, Z_i]) = \beta_0 + \beta_1 X_{i1} + \beta_2 Z_i$$

$$g(E[Y_i|X_{i2}, Z_i]) = (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) X_{i2} + (\beta_2 + \alpha_2) Z_i,$$

where the same outcome appears in each equation, but with different source-dependent predictors.

The parameter α_1 denotes the source-related differences in the effect of X on Y (conditional on Z).

Simple Illustration: *Stirling County Study*

Suppose we are interested in assessing the association of mortality, Y , and psychiatric diagnosis, X :

$$g(E[Y_i|X_{i1}]) = \beta_0 + \beta_1 \text{SELFDIAG}_i$$

$$g(E[Y_i|X_{i2}]) = (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) \text{GPDIAG}_i,$$

where the same outcome appears in each equation, but with different source-dependent predictors.

When there are no significant source differences, a simplified model may be fit that pools all of the information:

$$g(E[Y_i|X_{i1}]) = \beta_0 + \beta_1 \textit{SELFDIAG}_i$$

$$g(E[Y_i|X_{i2}]) = \beta_0 + \beta_1 \textit{GPDIAG}_i.$$

In this model β_1 is interpreted as the association between X and Y, for any source report.

Conceptually quite distinct from model that uses pooled or averaged X, $X = \text{mean}(X_{i1}, X_{i2})$, as the covariate or that conditions on both reports.

The latter would yield an attenuated estimate of effect, with the degree of attenuation depending on correlation between X_1 and X_2 .

Estimation

Multiple source data are usually positively correlated.

This correlation must be accounted for when analyzing either multiple source outcomes or predictors.

There are two broad approaches:

- (i) completely specify joint distribution of $(Y_{i1}, \dots, Y_{iJ}) \implies$ ML estimation;
- (ii) specify model for the correlation among the Y_{ij} 's only \implies GEE methods.

When the Y_{ij} are continuous, the former approach can be implemented using existing software; more difficult in the discrete data setting.

The latter approach can be implemented using existing software (e.g., PROC GENMOD in SAS, xtgee in Stata).

Application: *Eastern Connecticut Child Survey*

Data from Connecticut Child Surveys of mental health.

A standardized measure of childhood psychopathology was used both by parents (CBC) and teachers (TRF).

Focus on externalizing scale (assesses delinquent/aggressive behavior).

Scale dichotomized at the cut point for borderline/clinical psychopathology.

Note: Because of the multiple levels of permissions and reporting, a substantial number of children were missing the TRF.

Our analysis is based on 1428 children who had both parent and teacher responses, and an additional 1073 children with only a parent response.

In this example the two sources are the children's parents and teachers.

The objective of the analysis is to study the influence of single parent status (coded 1: single, 0: otherwise) and child's physical health problems (coded 1: fair to bad health, 0: good health) on the prevalence of externalizing behavior.

In addition, interested in determining whether the effects of the two covariates depend on informant.

Basic approach: Use two separate logistic regression models, one with the CBC as an outcome and one with the TRF as an outcome.

Both models have the same set of covariates, but the coefficients *may* differ for the different sources.

Let μ^P and μ^T denote the probability of a positive response on externalizing behavior as measured by parents and teachers, respectively.

Then the two regression models are:

$$\text{logit}(\mu_i^P) = \beta_0^P + \beta_1^P \text{ Single Parent} + \beta_2^P \text{ Child Health}$$

and

$$\text{logit}(\mu_i^T) = \beta_0^T + \beta_1^T \text{ Single Parent} + \beta_2^T \text{ Child Health}$$

Can fit these two regression models simultaneously using bivariate methods that take the association between the two outcomes into account.

Need to specify this association; here we use the odds ratio rather than correlation.

Table 1: Results of fitting two regression models simultaneously to externalizing behavior data on 2501 children using GEE method.

Informant	Intercept [†]	Single Parent [†]	Child Health [†]
Parent	-2.154 ± 0.091	0.616 ± 0.124	0.598 ± 0.113
Teacher	-1.683 ± 0.104	0.602 ± 0.155	0.146 ± 0.135

[†]Estimated coefficient \pm empirical standard error.

The model given above is very general model; its advantages over fitting two separate regressions are that:

- (i) We can test whether $\beta_k^P = \beta_k^T$ for the k^{th} covariate (or for the whole vector, test $\beta^P = \beta^T$, using $Cov(\hat{\beta})$ provided by the GEE analysis);
- (ii) We can use all available data; and
- (iii) It provides a measure of association (odds ratio) between the two informants.

However, with a large number of covariates, we will usually want to fit simpler models.

To formulate simpler models, we need to create a dichotomous indicator variable of informant.

To illustrate, we introduce a dichotomous variable (X_1) which is 1 if the informant is the parent, and 0 if the informant is the teacher.

Denoting single parent status and child health problems by X_2 and X_3 , consider a model of the form

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3}.$$

This model specifies that effects of single parent status (β_2) and child's health (β_3) are same regardless of source, but log odds may be higher/lower (β_1) depending on source.

Forcing the coefficients of single parent status to be equal seems reasonable in view of the results presented in Table 1, but not for child health problems.

Can allow effect of child's health to depend on source by adding interaction:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4},$$

where $X_{ij4} = X_{ij1} * X_{ij2}$.

Table 2: Results of fitting regression models, with common or shared parameters, simultaneously to data on externalizing behavior using GEE method.

Variable	Estimate	SE	Z
Intercept	-1.685	0.100	-16.85
Informant	-0.467	0.118	-3.96
Single Parent	0.611	0.108	5.68
Child Health	0.146	0.135	1.08
Informant \times Child Health	0.452	0.157	2.87

†Estimated coefficient \pm empirical standard error.

Results become more transparent if the model is written separately for the parent and teacher informant:

$$\text{logit}(\mu_{i1}) = \text{logit}(\mu_i^P) = (\beta_0 + \beta_1) + \beta_2 X_{ij2} + (\beta_3 + \beta_4) X_{ij3};$$

$$\text{logit}(\mu_{i2}) = \text{logit}(\mu_i^T) = \beta_0 + \beta_2 X_{ij2} + \beta_3 X_{ij3}.$$

For single parent status, $\hat{\beta}_2$ is the common coefficient for both informants; notice that its standard error is considerably smaller than the two corresponding standard errors reported in Table 1.

Finally, $\hat{\beta}_4$ estimates the difference in effects of child health problems as estimated by parent and teacher evaluations.

As a general rule, if source interactions are included for all the covariates, then the model is basically equivalent to fitting separate regressions.

When simpler models are fit (i.e., not all interactions with source are present) we can expect to gain efficiency in the analysis for the common coefficients.

This point is illustrated by comparing the standard errors of the coefficient for single parent status in Table 1 with Table 2.

Concluding Remarks

Methods reviewed provide a regression modelling approach for incorporating (often discordant) multiple source reports.

Methods have advantages over more *ad hoc* approaches that combine reports.

Methods allow formal assessment of whether covariate effects vary by source.

Methods allow for “pooling” of data from different sources *when appropriate*.

In contrast to more *ad hoc* “pooling” techniques, proposed methods optimally weight the contribution of data from each source.

Methods can be implemented using existing, general purpose statistical software (e.g., SAS, Stata, SPSS).

Methods can be extended to handle missing source reports, complex survey designs...