The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

| U.S. and Canada | | Elsewhere | |
|---|---|---|---|
| **Printed & electronic** | | **Printed & electronic** | |
| 1-year subscription | $115 | 1-year subscription | $145 |
| 2-year subscription | $210 | 2-year subscription | $270 |
| 3-year subscription | $285 | 3-year subscription | $375 |
| 1-year student subscription | $ 85 | 1-year student subscription | $115 |
| 1-year institutional subscription | $345 | 1-year institutional subscription | $375 |
| 2-year institutional subscription | $625 | 2-year institutional subscription | $685 |
| 3-year institutional subscription | $875 | 3-year institutional subscription | $965 |
| **Electronic only** | | **Electronic only** | |
| 1-year subscription | $ 85 | 1-year subscription | $ 85 |
| 2-year subscription | $155 | 2-year subscription | $155 |
| 3-year subscription | $215 | 3-year subscription | $215 |
| 1-year student subscription | $ 55 | 1-year student subscription | $ 55 |

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.




The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

# Analysis of partially observed clustered data using generalized estimating equations and multiple imputation

Kathryn M. Aloisio
Smith College
Northampton, MA
kaloisio@gmail.com

Sonja A. Swanson
Harvard School of Public Health
Boston, MA
sswanson@hsph.harvard.edu

Nadia Micali
University College London
London, UK
n.micali@ucl.ac.uk

Alison Field
Harvard School of Public Health
Boston, MA
alison.field@childrens.harvard.edu

Nicholas J. Horton
Amherst College
Amherst, MA
nhorton@amherst.edu

**Abstract.** Clustered data arise in many settings, particularly within the social and biomedical sciences. For example, multiple-source reports are commonly collected in child and adolescent psychiatric epidemiologic studies where researchers use various informants (for instance, parents and adolescents) to provide a holistic view of a subject's symptoms. Fitzmaurice et al. (1995, *American Journal of Epidemiology* 142: 1194–1203) have described estimation of multiple-source models using a standard generalized estimating equation (GEE) framework. However, these studies often have missing data because additional stages of consent and assent are required. The usual GEE is unbiased when data are missing completely at random in the context of Little and Rubin (2002, *Statistical Analysis with Missing Data* [Wiley]). This is a strong assumption that may not be tenable. Other options, such as the weighted GEE, are computationally challenging when missingness is nonmonotone. Multiple imputation is an attractive method to fit incomplete data models while requiring only the less restrictive missing-at-random assumption. Previously, estimation of partially observed clustered data was computationally challenging. However, recent developments in Stata have facilitated using them in practice. We demonstrate how to use multiple imputation in conjunction with a GEE to investigate the prevalence of eating disorder symptoms in adolescents as reported by parents and adolescents and to determine the factors associated with concordance and prevalence. The methods are motivated by the Avon Longitudinal Study of Parents and their Children, a cohort study that enrolled more than 14,000 pregnant mothers in 1991–92 and has followed the health and development of their children at regular intervals. While point estimates for the missing-at-random model were fairly similar to those for the GEE under missing completely at random, the missing-at-random model had smaller standard errors and required less stringent assumptions regarding missingness.

 

**Keywords:** st0363, ALSPAC study, eating disorders, multiple informants, weighted estimating equations, generalized estimating equations, multiple imputation, missing data, missing at random, missing completely at random

# 1 Introduction

Clustered data arise in many settings, particularly within the social and medical sciences, and they require sophisticated analytical methods. Standard-error estimates that do not account for association within clusters will be inaccurate and inferences will be invalid (Cannon et al. 2001).

For example, multiple-source reports are commonly collected in child and adolescent psychiatric epidemiologic studies where researchers use various informants (for example, parents and adolescents) to provide a holistic view of a subject's symptoms. These clustered reports also arise in other settings, such as geriatric studies, school settings, and health services research (Caria et al. 2011).

Several articles have reviewed methods to integrate reports from multiple sources (Fitzmaurice et al. 1995; Horton and Fitzmaurice 2004; Caria et al. 2011). Fitzmaurice et al. (1995) proposed methodology for simultaneously analyzing information from multiple-source outcomes by applying a generalized estimating equation (GEE) approach (Liang and Zeger 1986). GEEs account for the correlation between reports to model the average response for observations sharing covariates.

A practical difficulty in analyzing multiple-source reports is that there is often a substantial amount of missingness. In multiple-source studies, data may be missing from a single source or multiple sources, because additional stages of consent and assent are required. Analyzing data without appropriately accounting for missingness can induce bias and loss of efficiency.

The usual GEE is unbiased whenever missingness is missing completely at random (MCAR), which means that missingness does not depend on observed or unobserved measurements (Little and Rubin 2002). The GEE permits a report to contribute to one equation and not to the other, but using the available-case method may be biased if the missing mechanism is not MCAR (Liang and Zeger 1986).

Xie and Paik (1997) proposed a weighted GEE that handles missingness when the probability of missingness depends on the outcomes or observed covariates. This method assumes the less restrictive missingness mechanism, which was named missing at random (MAR) by Little and Rubin (2002). To fit the weighted GEE, one must estimate the probability of subjects' being observed, drop all the partially observed subjects, and fit the reweighted model using only the complete cases. Horton et al. (2001) implemented this with multiple-source reports but had to use an ad hoc procedure to account for complex nonmonotone patterns of missingness. The monotone structure is rarely seen in observational studies with many covariates and is absent in the motivating example. Furthermore, accounting for a complex nonmonotone pattern is computationally difficult (Li et al. 2011). Therefore, other approaches are needed.

An alternative approach to this problem implements multiple imputation (MI), a flexible and principled method for fitting incomplete data regression models (Rubin 1987). After specifying an appropriate imputation model, the algorithm "fills in" the missing data with plausible values that account for the uncertainty that comes with using predicted values. The MI method does not require the missingness pattern to be MCAR or monotone.

Simulation studies with longitudinal binary data and missing data have been implemented to assess different analytical approaches, including the usual GEE, the weighted GEE, and MI in conjunction with estimating equations (MI–GEE). Beunckens, Sotto, and Molenberghs (2008) found that using the MI–GEE approach was more successful than using the usual GEE and the weighted GEE approach. DeSouza, Legedza, and Sankoh (2009), Yoo (2010), and Birhanu et al. (2011) expanded the simulation study, and each concluded that MI–GEE outperformed the weighted GEE and is a valid analysis tool for nonnormal and repeated binary responses. Frank Liu and Zhan (2011) undertook a similar simulation study and found contrary evidence for MI–GEE but concluded that the null finding may be due to the misspecification of the imputation model. The flexibility of the MI–GEE allows the imputation model to be adjusted. Lloyd et al. (2013) describe how to undertake estimation for longitudinal regression by using the `ice` and `uvis` user-written commands in Stata 11.

Previously, estimation of partially observed clustered data was computationally challenging. However, recent developments in Stata have facilitated their use in practice. This article demonstrates estimation of a GEE model with multiply imputed data by using the `mi` system in Stata 13.

We first describe the motivating study, the Avon Longitudinal Study of Parents and Children (ALSPAC), a long-running cohort study using parent and adolescent questionnaires to research the health and development of the adolescents. Then we describe how GEE models can be used to fit generalized linear models using available case data. Next we introduce MI and simultaneous estimation of GEE models using multiply imputed data within Stata. Then we fit the GEE models to our motivating data by using both available cases and the imputed data. We conclude by discussing the method, possible extensions, and areas for future research.

# 2 Example: Multiple source reports of adolescent eating-disorder behaviors

## 2.1 Study sample

These methods are motivated by data from the ALSPAC, a longitudinal, prospective study of women and pregnancy (Golding et al. 2001; Boyd et al. 2013). All pregnant women living in the area of Avon, UK, who were expected to deliver their babies between 1 April 1991 and 31 December 1992 were invited to take part in the study. Adolescents from 14,541 pregnancies were enrolled. Of these, 12,388 singleton adolescents were alive

at age one and provided the study with complete information on each adolescent's sex and maternal age. Adolescents and their parents were followed to investigate a range of psychological, physical, and social outcomes.

Parents and adolescents who were still enrolled in the study were sent questionnaires when the adolescent was age 14 and again when age 16. The analytic sample consists of 7,986 adolescents that had at least one adolescent or parent report at age 14 and 16, and it includes fully observed family demographics.

Adolescents completed questions on eating disorder symptoms adapted from the purging behavior assessments in the McKnight Risk Factor Survey and the Youth Risk Behavior Surveillance System Questionnaires (Kann et al. 1996). Adolescents were asked whether they had engaged in eating disorder behaviors in the past year, including binge eating (overeating with loss of control; two questions), vomiting, laxative use, and fasting. Parents completed a questionnaire version of the Eating Disorder Developmental and Well-Being Assessment with no skip rules (Goodman et al. 2000; Ford, Goodman, and Meltzer 2003). Parents were asked whether their teenager had engaged in eating disorder behaviors in the past three months, including binge eating (overeating with loss of control; one question), vomiting, laxative use, and fasting.

To demonstrate, this article will focus on predicting reports of vomiting behavior at age 16. Analyses for other eating disorder symptoms at ages 14 and 16 are reported in Swanson et al. (2014).

Ethical approval for the study was obtained from the ALSPAC Laws and Ethics Committee, the Local Research Ethics Committees, and the Smith College and Amherst College Institutional Review Boards.

## 2.2 Variables

The questionnaire sent to participants when the adolescents were age 16 asked the parents, "Over the last 3 months, has your study teenager made herself/himself sick to avoid putting on weight?" It also asked the adolescents, "During the past year, how often did you make yourself throw up (vomit) to lose weight or avoid gaining weight?" Because of inconsistency of possible answer options across informants, these two questions were recoded into two dichotomous variables (`vomit_p16`, `vomit_c16`) as either any or no endorsement.

For the GEE approach, we needed to reshape our dataset from wide form (one row per subject) to long form (two rows per subject). We created a binary-source variable (adolescent report versus parent report, `child`) and combined the outcome variables into `vomit_16`.

The models also included three dichotomous covariates that measured maternal education (A levels or above [college entrance] versus less than A levels, `edua`), adolescent's sex at birth (female versus male, `female`), and maternal parity at birth of the adolescent under study (multiparae [any siblings] versus primiparae, `multiparae`).

Table 1. Prevalence for the covariates

| | Overall ($n = 7968$) | Male ($n = 3834$) | Female ($n = 4134$) |
|---|---|---|---|
| Maternal education *(less than A levels)* | 57.7% (4429/7679) | 57.5% (2135/3715) | 57.9% (2294/3964) |
| Parity *(primiparae)* | 46.6% (3594/7714) | 47.0% (1749/3724) | 46.2% (1845/3990) |

Table 2. Prevalence for report of adolescent vomiting at age 16

| | Overall | Male | Female |
|---|---|---|---|
| Parent report | 0.30% (16/5252) | 0.19% (5/2578) | 0.41% (11/2674) |
| Adolescent report | 4.93% (236/4788) | 0.82% (16/1962) | 7.78% (220/2826) |

# 3 Methods

## 3.1 Notation

Following the notation in Horton and Fitzmaurice (2004), we assume there are $N$ independent subjects, each with an outcome obtained from $J$ sources. Let $Y_{ij}$ represent the dichotomous outcome obtained for the $i$th subject from the $j$th source (with $i = 1, \ldots, N$ and $j = 1, \ldots, J$). The study has two sources ($J = 2$), where $Y_{i1}$ is the first source report (adolescent, `child==1`), and $Y_{i2}$ is the second source report (parent, `child==0`). In addition, let $\mathbf{X}_{ij}$ be a $p \times 1$ vector of covariates associated with the outcome obtained for the $i$th subject from the $j$th source ($\mathbf{X}_{ij}$ contains both source information and subject-specific information). We let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ be the $J \times 1$ outcome vector for the $i$th subject and $\mathbf{X}_i$ be the associated $J \times p$ matrix of covariates.

## 3.2 Analytic approaches

### GEE model for multiple sources

If we had only one source, there would be one observation per subject (no clustering), and we could proceed to fit a logistic regression model for the dichotomous outcome or another model from the generalized linear model family. However, the clustered nature of multiple sources, where two reports from the same adolescent are likely to be positively associated, requires a more sophisticated model.

GEEs were first described by Liang and Zeger (1986) and are an attractive method to fit population-averaged regression models for clustered data. The GEE assumes a "working" correlation matrix and uses an empirical variance estimator (also known as a robust or Huber–White or "sandwich" variance) to obtain estimates for the logistic regression model, which accounts for the clustering within subjects. Liang and Zeger (1986) proved that the GEE yields consistent estimates of the regression parameters and of their variances under mild assumptions about dependence and correct specification of the mean model.

The general form for regression models for the mean of some function of $Y_i$, conditional on both source and risk factors (this setting includes adolescent gender, maternal education, and parity), is given by

$$g\{E(Y_{ij}|\mathbf{X}_{ij})\} = \mathbf{X}'_{ij}\boldsymbol{\beta}$$

where $g(\cdot)$ is a known link function. For our setting with a binary outcome, we can set $g(y) = \log\{y/(1-y)\} = \text{logit}(y)$ (for example, the logit function). The full model applied to the motivating example would be the following:

$$\begin{aligned}\text{logit}\{E(Y_{ij}|\mathbf{X}_{ij})\} = {} & \beta_0 + \beta_1\texttt{multiparae} + \beta_2\texttt{edua} + \beta_3\texttt{female} + \beta_4\texttt{child} \\ & + \beta_5(\texttt{child} \times \texttt{female}) + \beta_6(\texttt{child} \times \texttt{multiparae}) \\ & + \beta_7(\texttt{child} \times \texttt{edua})\end{aligned} \tag{1}$$

The coefficients are log odds-ratios, where $\beta_5$, $\beta_6$, and $\beta_7$ represent the interaction of the source effect with the three covariates.

Model (1) can be simplified if interactions were found to be nonsignificant. For the predicted prevalence of the vomiting behavior model, we dropped the extraneous interactions (those with $p$-values $\geqslant 0.05$) and refit the model to obtain estimates for a parsimonious model, which retained the gender by source interaction:

$$\begin{aligned}\text{logit}\{E(Y_{ij}|\mathbf{X}_{ij})\} = {} & \beta_0 + \beta_1\texttt{multiparae} + \beta_2\texttt{edua} + \beta_3\texttt{female} + \beta_4\texttt{child} \\ & + \beta_5(\texttt{child} \times \texttt{female})\end{aligned} \tag{2}$$

Without the parity by source interaction and maternal education by source interaction in (2), $\exp(\beta_1)$ and $\exp(\beta_2)$ are interpreted as odds ratios for parity and maternal education, respectively, within levels of source and gender. We can interpret the interaction term by extracting the equation for parent reports [presented in (3)]; similarly, we can obtain the equation for adolescent reports (4).

$$\text{logit}\{E(Y_{ij}|\mathbf{X}_{ij}, \texttt{child} == 0)\} = \beta_0 + \beta_1\texttt{multiparae} + \beta_2\texttt{edua} + \beta_3\texttt{female} \tag{3}$$

$$\begin{aligned}\text{logit}\{E(Y_{ij}|\mathbf{X}_{ij}, \texttt{child} == 1)\} = {} & (\beta_0 + \beta_4) + \beta_1\texttt{multiparae} + \beta_2\texttt{edua} \\ & + (\beta_3 + \beta_5)\texttt{female}\end{aligned} \tag{4}$$

Note that for the adolescent report (4), $\beta_4$ is the log odds for additional prevalence for adolescent reports, and $\beta_5$ is the additional log odds for female adolescent reports.

## 3.3 Accounting for missing data

Missing data occur in almost all real-world investigations (Little and Rubin 2002). This was also the case for the ALSPAC study, which is demonstrated using the `miss` option for `tabulate`.

```
. by female: tabulate vomit_c16 vomit_p16, miss
```

-> female = 0

| vomit_c16 | vomit_p16 0 | 1 | . | Total |
|---|---|---|---|---|
| 0 | 1,673 | 2 | 271 | 1,946 |
| 1 | 12 | 1 | 3 | 16 |
| . | 888 | 2 | 982 | 1,872 |
| Total | 2,573 | 5 | 1,256 | 3,834 |

-> female = 1

| vomit_c16 | vomit_p16 0 | 1 | . | Total |
|---|---|---|---|---|
| 0 | 1,986 | 4 | 616 | 2,606 |
| 1 | 135 | 5 | 80 | 220 |
| . | 542 | 2 | 764 | 1,308 |
| Total | 2,663 | 11 | 1,460 | 4,134 |

From this output, we observe that 1,688 male adolescents returned completed questionnaires and that 2,130 female adolescents returned completed questionnaires out of the total 7,968 sample subjects. By adding the `miss` option, we show that 4,150 (52%) of the possible sample are missing adolescent, parent, or both reports, regardless of gender. For instance, for male subjects, 7% $\{(271 + 3)/3834\}$ have adolescent reports but are missing parent reports, 23% $\{(888 + 2)/3834\}$ have missing adolescent reports but have parent reports, and 26% (982/3834) are missing both adolescent and parent reports for the questionnaire from age 16. Accounting for the partially observed responses is crucial for obtaining reliable results for future inferences.

There are three concerns that typically arise with missing data: 1) loss of efficiency; 2) complication in data handling and analysis; and 3) bias due to differences between the observed and unobserved data. Next we introduce a nomenclature for missing data.

### Missing-data nomenclature

For each of the $N$ subjects, the outcome vector, $\mathbf{Y}$, and the vector of predictors, $\mathbf{X}$, are either observed or missing. We denote $\mathbf{Y}^{\text{obs}}$ as the observed component of the outcome and $\mathbf{X}^{\text{obs}}$ as the observed components of the predictors. Similarly, we denote $\mathbf{Y}^{\text{mis}}$ and $\mathbf{X}^{\text{mis}}$ as the unobserved components of the outcome and predictors, respectively. In addition, $\mathbf{Z}^{\text{obs}} = (\mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}})$ and $\mathbf{Z}^{\text{mis}} = (\mathbf{Y}^{\text{mis}}, \mathbf{X}^{\text{mis}})$ denote the vector of observed variables and missing variables, respectively. We also use $\boldsymbol{\gamma}$ to denote the regression

parameters. Lastly, we define a set $\mathbf{R}$ of response indicators (that is, $R_i = 1$ if the $i$th element of $\mathbf{Z}$ is observed, and it equals 0 otherwise).

Little and Rubin (2002) defined classifications for the probability distribution generating the missing data. MCAR is characterized as

$$P(\mathbf{R}|\mathbf{Z}, \boldsymbol{\gamma}) = P\left(\mathbf{R}|\mathbf{Z}^{\text{obs}}, \mathbf{Z}^{\text{mis}}, \boldsymbol{\gamma}\right) = P(\mathbf{R}|\boldsymbol{\gamma})$$

That is, the probability of being missing is the same for all cases. Heuristically, the reasons for missingness are unrelated to the observed or unobserved data. MCAR is simple but is unlikely to happen in practice.

The mechanism MAR assumes

$$P(\mathbf{R}|\mathbf{Z}, \boldsymbol{\gamma}) = P\left(\mathbf{R}|\mathbf{Z}^{\text{obs}}, \boldsymbol{\gamma}\right)$$

That is, the probability of being missing is the same after conditioning on the observed data. Heuristically, this states that missingness depends on only observed quantities, including outcomes, predictors, and auxiliary variables. Most analyses start with this assumption because it is more likely to happen than MCAR, particularly within datasets containing many variables (Collins, Schafer, and Kam 2001). It is possible to test the MCAR assumption against the alternative hypothesis that missingness is MAR (Diggle and Kenward 1994).

Missing not at random (MNAR) concerns researchers and analysts the most because MNAR means that the probability of data being missing varies for reasons that are unknown to the researcher (missingness is related to the unobserved quantities). Symbolically, $P(\mathbf{R}|\mathbf{Z})$ cannot be simplified, and it must be modeled as part of the likelihood. Little and Rubin (2002) call this "nonignorable". While MNAR missingness is important when undertaking sensitivity analyses, we will not consider it further.

The pattern of missingness, monotone versus nonmonotone, can also influence how we address missing data. A dataset is said to have a monotone-missing pattern when the variables in the dataset can be arranged in a stair-step pattern (that is, nonincreasing or nondecreasing) when missingness on one implies missingness on the other (Little and Rubin 2002). The monotone pattern is generally uncommon with observational studies, as with the motivating study, where we have some subjects missing a parent report and others missing an adolescent report.

We can use `misschk` (Long and Freese 2014) to display the missingness pattern for a subset of the variables used in our motivating example.

```
. misschk female edua multiparae vomit_c16 vomit_p16
Variables examined for missing values
   #  Variable        # Missing   % Missing
-------------------------------------------
   1  female                  0         0.0
   2  edua                  289         3.6
   3  multiparae            254         3.2
   4  vomit_c16            3180        39.9
   5  vomit_p16            2716        34.1
```

| Missing for which variables? | Freq. | Percent | Cum. |
|---|---|---|---|
| _2345 | 32 | 0.40 | 0.40 |
| _234_ | 12 | 0.15 | 0.55 |
| _23_5 | 18 | 0.23 | 0.78 |
| _23__ | 17 | 0.21 | 0.99 |
| _2_45 | 79 | 0.99 | 1.98 |
| _2_4_ | 40 | 0.50 | 2.48 |
| _2__5 | 50 | 0.63 | 3.11 |
| _2___ | 41 | 0.51 | 3.63 |
| __345 | 57 | 0.72 | 4.34 |
| __34_ | 27 | 0.34 | 4.68 |
| __3_5 | 23 | 0.29 | 4.97 |
| __3__ | 68 | 0.85 | 5.82 |
| ___45 | 1,578 | 19.80 | 25.63 |
| ___4_ | 1,355 | 17.01 | 42.63 |
| ____5 | 879 | 11.03 | 53.66 |
| _____ | 3,692 | 46.34 | 100.00 |
| Total | 7,968 | 100.00 | |

| Missing for how many variables? | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 3,692 | 46.34 | 46.34 |
| 1 | 2,343 | 29.41 | 75.74 |
| 2 | 1,735 | 21.77 | 97.52 |
| 3 | 166 | 2.08 | 99.60 |
| 4 | 32 | 0.40 | 100.00 |
| Total | 7,968 | 100.00 | |

Note that the most common missing patterns include subjects missing both adolescent and parent reports (20%, $n = 1578$), subjects missing just the adolescent report (17%, $n = 1355$), and subjects missing just the parent report (11%, $n = 879$). However, we do not have a monotone-missingness pattern, because for each of the covariates (parity and maternal education), there are cases missing adolescent, parent, or both reports.

### Available-case method

The available-case method includes all cases where the variable of interest is present. This method is more efficient than complete-case analyses, where any case with a missing value is removed. In the motivating study, there were $n = 7968$ available cases versus

$n = 3692$ complete cases, as shown in the missing-patterns table above. The available-case method is also unbiased when missingness is MCAR. However, complications can arise because the analytic sample base changes from model to model and may lead to problems of comparability.

### Weighted estimating equations

Weighted estimating equations (Xie and Paik 1997; Horton et al. 2001; Li et al. 2011) are an attractive approach if missingness is monotone. However, using them is not feasible in this setting, even for modeling parent and adolescent reports for one age, because some subjects have missing adolescent reports while others have missing parent reports. However, weighted estimating equations are supported in Stata 13. For more details, see `help weight`.

### MI

MI is a principled method used to account for missing data (Rubin 1976). It involves a three-step approach for fitting incomplete data regression models. First, it creates plausible values for missing observations that reflect uncertainty about the nonresponse model. These values are used to "fill in" or impute the missing values (generally under a MAR assumption). This process is repeated, which results in the creation of several "completed" datasets. Second, each of these datasets is analyzed using complete-data methods. Finally, the results are combined, which allows the uncertainty regarding the imputation to be considered (Little and Rubin 2002). Because increasing the number of imputed datasets minimizes variability introduced into the results because of the imputation process (Horton and Lipsitz 2001; White, Royston, and Wood 2011; van Buuren 2012), we recommend a set of 25 imputations, though more are computationally possible.

**Specifying imputation model.** MI requires the analyst to provide an appropriate specification of the imputation model. If this model is misspecified, there is potential for bias (White, Royston, and Wood 2011). In general, the imputation model must be compatible with the model used for the analysis, with all potential covariates and important higher-order associations included (Little and Rubin 2002). For example, in model (2), we want to assess the source by gender interaction with reported instances of vomiting. Even though gender is fully observed, we need to include gender in the imputation model because we include gender effects in the analysis model. Also, to preserve the source by gender interaction, we have to account for the interaction term. We did this by stratifying the imputation model by gender (Royston 2005), though this could also have been accomplished by including the interaction when specifying custom prediction equations (see `help mi impute chained`).

In addition to all the variables that can be used in the analysis model, any auxiliary variables that may contain information about missing data should be included. For our model, we included a measure for self-reported body mass index, the mother's age at delivery, and the adolescent's age at the time of reporting. Furthermore, the

outcome variable should always be present in the imputation model to obtain valid results (Moons et al. 2006). By including all the variables necessary for the model and any auxiliary variables that may contain information about missing data, the MAR assumption becomes more plausible, and the quality of the imputed values improves (Collins, Schafer, and Kam 2001).

**Specifying imputation method.** The choice of imputation method depends on the pattern of missing values. As opposed to having a monotone-missing pattern, our data have an arbitrary missing pattern. When a pattern of missing values is arbitrary, iterative methods are used to fill in missing values. To accommodate our arbitrary missing-value patterns, we imputed the data using chained equations with a variable-by-variable approach. The imputation model is specified separately for each variable and involves the other variables as predictors. At each stage of the algorithm, an imputation is generated for all the missing values in a given variable, and this imputed variable is used to impute the next variable. This process repeatedly imputes missing values by using a Gibbs sampling procedure until the process reaches convergence. For this example, we used 25 iterations.

**Combining complete-case results.** The last step of the imputation method uses "Rubin's rules" to combine the repeated-imputation results, where the total variance stems from the following three sources (Little and Rubin 2002):

1. The variance is a result of taking a sample rather than observing the entire population. This is the conventional statistical measure of variability.

2. The extra variance is caused by missing values in the sample.

3. The extra simulation variance is a result of the estimate being estimated for a finite number of imputations.

# 4 Application in Stata

MI can be used in combination with the estimation of a wide variety of models, including the GEE model, using the `mi` system in Stata 13.

To use the `mi` system, we begin by reading in the dataset and creating the analytic set. We include additional variables from the cohort study in the imputation model to make the MAR assumption more plausible (Collins, Schafer, and Kam 2001).

```
. use alspac_informant, clear
. keep vomit_c14 vomit_p14 vomit_c16 vomit_p16
>     lax_c14 lax_p14 lax_c16 lax_p16
>     fast_c14 fast_p14 fast_c16 fast_p16
>     binge_c14 binge_p14 binge_c16 binge_p16
>     anyedsx_c14 anyedsx_p14 anyedsx_c16 anyedsx_p16
>     thin_c14 thin_p14
>     edua multiparae m_age_at_delivery female weightkg heightm c_age_at_report
>     bmi cid_153a
```

## 4.1 Registering variables

Next we need to set how Stata should add additional imputations. We chose to use the marginal long (`mlong`) data structure because it uses slightly less memory than the wide (`wide`) data structure. However, the wide format is slightly faster.

```
. mi set mlong
```

Then we register each of the variables within the dataset as either variables to impute or variables to not impute.

The variables that must be imputed require registration:

```
. mi register imputed vomit_c14 vomit_p14 vomit_c16 vomit_p16 lax_c14 lax_p14
>                     lax_c16 lax_p16 fast_c14 fast_p14 fast_c16 fast_p16
>                     thin_c14 thin_p14 binge_c14 binge_p14 binge_c16 binge_p16
>                     anyedsx_c14 anyedsx_p14 anyedsx_c16 anyedsx_p16
>                     edua multiparae weightkg heightm c_age_at_report bmi
(6009 m=0 obs. now marked as incomplete)
```

The variables that do not require imputation but will be used in the imputation model are registered as regular variables.

```
. mi register regular m_age_at_delivery female
```

With the added covariates, we redisplay the table from `misschk` listing missingness for different numbers of variables.

| Missing for how many variables? | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 1,959 | 24.59 | 24.59 |
| 1 | 345 | 4.33 | 28.92 |
| 2 | 106 | 1.33 | 30.25 |
| (*output omitted*) | | | |
| 26 | 7 | 0.09 | 99.94 |
| 27 | 5 | 0.06 | 100.00 |
| Total | 7,968 | 100.00 | |

Note that with these 31 variables, there are only 1,959 complete cases. Three variables are completely observed: the ID, the gender, and the age of mother at the time of delivery.

## 4.2 Imputation model specification

We then create the imputed datasets in Stata by using `mi impute`. This requires us to specify the imputation model. We must first select the imputation method. For univariate imputation, where the pattern of missingness is monotone, we can choose from a variety of imputation models based on the type of variable. For example, `mi`

`impute regress` will fit a linear regression model for a continuous variable or `mi impute poisson` for a count variable.

For multivariate imputation with different types of variables (that is, a mixture of continuous and discrete), the situation is more complicated. If the pattern of missingness is monotone, we can use `mi impute monotone` to assign an imputation method to each variable. If there is an arbitrary missing pattern (as in the present analysis), we can use `mi imputed mvn` for multivariate normal variables or `mi impute chained` for the chained-equation method. Table 3 lists these options and other options that can be selected as the imputation method.

Table 3. MI methods available within Stata 13

| Method | Description |
|---|---|
| Univariate | |
| `regress` | Linear regression |
| `pmm` | Predictive mean matching |
| `truncreg` | Truncated regression |
| `intreg` | Interval regression |
| `logit` | Logistic regression |
| `ologit` | Ordered logistic regression |
| `mlogit` | Multinomial logistic regression |
| `poisson` | Poisson regression |
| `nbreg` | Negative binomial regression |
| Multivariate | |
| `monotone` | Sequential imputation using a monotone missing pattern |
| `chained` | Sequential imputation using chained equations |
| `mvn` | Multivariate normal regression |

Because our data setting did not feature monotone missingness and the study variables were not normally distributed, we adopted the chained-equation approach (Raghunathan et al. 2001; White, Royston, and Wood 2011; van Buuren 2012). Using 25 chains for 25 iterations, we fit a linear regression model `regress` for the incomplete continuous variables and used predicted mean matching (`pmm`) for the binary variables. Predicted mean matching is similar to the regression method except that for each missing value, it imputes a value randomly drawn from a set of observed values whose predicted values are closest to the predicted value for the missing value from the simulated regression model. Generally, predicted mean matching is used for continuous variables. However, predictive mean matching proves to be unbiased for dichotomous variables, it ensures that imputed values are plausible, and it may be more appropriate if the normality assumption is violated (Horton, Lipsitz, and Parzen 2003).

Both sets of models included all symptoms from both sources at each age and other covariates stratified by gender to account for the interaction (StataCorp 2013).

```
mi impute chained (regress) weightkg heightm c_age_at_report bmi          ///
                  (pmm) vomit_c14 vomit_p14 vomit_c16 vomit_p16           ///
                        lax_c14 lax_p14 lax_c16 lax_p16                   ///
                        fast_c14 fast_p14 fast_c16 fast_p16               ///
                        thin_c14 thin_p14                                 ///
                        binge_c14 binge_p14 binge_c16 binge_p16           ///
                        anyedsx_c14 anyedsx_p14 anyedsx_c16 anyedsx_p16   ///
                        edua multiparae = m_age_at_delivery,              ///
                        dots noisily add(25) by(female) augment
```

This model was implemented on an Intel® Core™2 Duo Processor and took approximately two hours.

## 4.3 GEE with imputed datasets

With the imputed datasets, complete-case methods can be used to estimate models using `mi estimate`. Stata supports estimation of many regression models with imputed data, including linear regression models, binary-response regression models, count-response regression models, ordinal-response regression models, categorical-response regression models, quantile regression models, survival regression models, panel-data models, and survey regression models. The present study uses `mi estimate xtgee` to fit the GEE models because of the clustering within subjects.

To preserve associations between the parent and adolescent reports, we imputed the data in wide form (one row per subject). However, to fit the model, we need to reshape our datasets from wide form to long form (two rows per subject). This is easy to do using the post `mi` data manipulation commands. To clarify the process, we will display the data for the first five subjects. We can select the original dataset with `mi xeq 0`.

```
. mi xeq 0: list id female edua multiparae vomit_c16 vomit_p16 if id < 6
 m=0 data:
-> list id female edua multiparae vomit_c16 vomit_p16 if id < 6
```

| | id | female | edua | multip~e | vomi~c16 | vomi~p16 |
|---|---|---|---|---|---|---|
| 1. | 1 | 0 | 0 | . | . | . |
| 2. | 2 | 0 | 0 | 1 | . | 0 |
| 3. | 3 | 0 | 0 | 0 | . | 0 |
| 4. | 4 | 0 | 1 | 1 | 0 | 0 |
| 5. | 5 | 0 | 0 | 1 | 0 | 0 |

Then we rename the outcome variables for the `reshape` command so that $j = 1$ indicates an adolescent report and $j = 0$ indicates a parent report.

```
. mi rename vomit_c16 vomit_161
. mi rename vomit_p16 vomit_160
```

```
. mi reshape long vomit_16, i(cid_153a) j(child)
reshaping m=0 data ...
(note: j = 0 1)
Data                                wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                      7968   ->   15936
Number of variables                   35   ->      35
j variable (2 values)                      ->   child
xij variables:
                     vomit_160 vomit_161   ->   vomit_16
-----------------------------------------------------------------------------
```

Now we have doubled the number of rows (15,936) in the long format. We then display the same first five subjects in the long format.

```
. mi xeq 0: sort id; list id female edua multiparae vomit_16 if id <6
 m=0 data:
-> sort id
-> list id female edua multiparae child vomit_16 if id <6
```

| | id | female | edua | multip~e | child | vomit_16 |
|---|---|---|---|---|---|---|
| 1. | 1 | 0 | 0 | . | 1 | . |
| 2. | 1 | 0 | 0 | . | 0 | . |
| 3. | 2 | 0 | 0 | 1 | 0 | 0 |
| 4. | 2 | 0 | 0 | 1 | 1 | . |
| 5. | 3 | 0 | 0 | 0 | 1 | . |
| 6. | 3 | 0 | 0 | 0 | 0 | 0 |
| 7. | 4 | 0 | 1 | 1 | 1 | 0 |
| 8. | 4 | 0 | 1 | 1 | 0 | 0 |
| 9. | 5 | 0 | 0 | 1 | 0 | 0 |
| 10. | 5 | 0 | 0 | 1 | 1 | 0 |

Recall that (2) includes the gender by source interaction. We recode a new variable using imputed data by implementing `mi passive`. We named the interaction variable `femchild`.

```
. mi passive: generate femchild = female*child
```

Before fitting the model, we must declare the type of complex data (for example, `mi stset` for survival data or `mi svyset` for survey data). For panel data, we use the `xtset` command. Note that we are paneling on the adolescent ID variable.

```
. mi xtset cid_153a
       panel variable:  cid_153a (balanced)
```

Then we proceed to fit (2).

```
. mi estimate: xtgee vomit_16 multiparae edua female child femchild,
> family(binomial) link(logit) corr(inde)
Multiple-imputation estimates                   Imputations        =         25
GEE population-averaged model                   Number of obs      =      15936
Group variable:                     cid_153a    Number of groups   =       7968
Link:                                  logit    Obs per group: min =          2
Family:                             binomial                   avg =        2.0
Correlation:                     independent                   max =          2
Scale parameter:                           1
                                                Average RVI        =     0.6721
                                                Largest FMI        =     0.4821
DF adjustment:                  Large sample    DF:     min        =     107.46
                                                        avg        =     143.06
                                                        max        =     208.34
Model F test:                      Equal FMI    F(   5,  642.9)    =      46.50
Within VCE type:                Conventional    Prob > F           =     0.0000
```

| vomit_16 | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| multiparae | .2721927 | .1302315 | 2.09 | 0.038 | .0153082 | .5290773 |
| edua | .1300673 | .1272539 | 1.02 | 0.308 | -.120803 | .3809375 |
| female | .2884962 | .5023234 | 0.57 | 0.567 | -.7072522 | 1.284245 |
| child | .9982947 | .4376222 | 2.28 | 0.024 | .1322118 | 1.864378 |
| femchild | 1.975446 | .5530279 | 3.57 | 0.001 | .8798404 | 3.071051 |
| _cons | -5.883525 | .3970033 | -14.82 | 0.000 | -6.67001 | -5.09704 |

The model indicates that when we control for other factors, the odds for exhibiting vomiting behavior for an adolescent with siblings is 1.31 [95% confidence interval (CI) 1.02–1.70] times the odds for an only-child adolescent exhibiting vomiting behavior. Maternal education was found to not be significantly associated with vomiting behavior (odds ratio 1.14; [95% CI 0.89–1.46]) after we controlled for other factors.

To interpret the gender by source interaction, we calculated the four predicted probabilities using (3) and (4) with the other covariates set to 0 and the inverse logit function ($\texttt{invlogit}(\beta) = \exp(\beta)/\{1 + \exp(\beta)\}$). We calculated the predicted probability for the male adolescent report ($\texttt{invlogit}(\beta_0 + \beta_4)$, 0.8%, [95% CI 0.4%–1.1%]); the male's parent report ($\texttt{invlogit}(\beta_0)$, 0.3%, [95 % CI 0.08%–0.5%]); the female adolescent report ($\texttt{invlogit}(\beta_0+\beta_3+\beta_4+\beta_5)$, 6.8%, [95% CI 5.3%–8.3%]); and the female's parent report ($\texttt{invlogit}(\beta_0 + \beta_3)$, 0.4%, [95% CI 0.1%–0.6%]).

From these, we can determine important distinct patterns. First, estimates for vomiting are higher when vomiting is reported by the adolescent instead of his or her parent for both male and female adolescents. In addition, for adolescent reporting, there is a significant difference between females and males reporting endorsement of vomiting behaviors. However, there is not a significant gender difference for parent reporting. This result has implications for our understanding of the diagnosis and the prevalence of reported symptoms, as discussed in more detail by Swanson et al. (2014).

To compare these results with an available case model, we can fit the model to only the original dataset by using `mi xeq 0`.

```
. mi xeq 0: xtgee vomit_16 multiparae edua female child femchild,
> family(binomial) link(logit) corr(inde)

 m=0 data:
-> xtgee vomit_16 multiparae edua female child  femchild,
   family(binomial) link(logit) corr(inde)

Iteration 1: tolerance = 7.568e-07

GEE population-averaged model                   Number of obs      =      9618
Group variable:                    cid_153a     Number of groups   =      5926
Link:                                 logit     Obs per group: min =         1
Family:                            binomial                    avg =       1.6
Correlation:                    independent                    max =         2
                                               Wald chi2(5)       =    224.44
Scale parameter:                          1    Prob > chi2        =    0.0000

Pearson chi2(9618):                9668.38     Deviance           =   1835.64
Dispersion (Pearson):              1.005238    Dispersion         =  .1908548
```

| vomit_16 | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| multiparae | .3196914 | .1369466 | 2.33 | 0.020 | .051281 | .5881019 |
| edua | .3216119 | .1349857 | 2.38 | 0.017 | .0570449 | .5861789 |
| female | .7626511 | .5401381 | 1.41 | 0.158 | -.2960001 | 1.821302 |
| child | 1.375299 | .517371 | 2.66 | 0.008 | .3612708 | 2.389328 |
| femchild | 1.59392 | .6035489 | 2.64 | 0.008 | .410986 | 2.776854 |
| _cons | -6.550324 | .4617182 | -14.19 | 0.000 | -7.455275 | -5.645373 |

When the two methods are compared, the estimates are similar, and the standard errors from the MI model assuming MAR are consistently smaller than the standard errors for the MCAR model.

# 5 Discussion

Clustered data with partially observed responses and predictors arise in many situations. In this article, we have detailed how to account for clustering when MI is used to account for missingness.

Multiple-source data often occur when one analyzes studies with complex survey designs. Along with clustering, stratification and sampling weights must be considered in the analysis. This can be done in Stata by using the survey design tools (Horton and Fitzmaurice 2004).

Many analytic approaches rely on the accuracy of the assumptions associated with the proposed method. Negligence or inaccurate analysis of the collected data can introduce bias. Assumptions that missingness is MAR are inherently unverifiable without auxiliary information. When one uses methods that incorporate other variables associated with missingness and responses, the possibility of bias is reduced, and the data are represented more accurately.

The GEE model is attractive because it can account for clustering or repeated measures induced by longitudinal data. However, the assumption of MCAR is very restric-

tive because reasons for missingness are generally more complex than just being due to chance.

The weighted GEE loosens the often implausible MCAR missingness assumption. If a weighted model were feasible, it could be incorporated using survey weights, as described by Horton and Fitzmaurice (2004). However, the requirement that the patterns of missing be monotone is a major limitation. The use of MI is attractive because it can incorporate auxiliary variables to make MAR more tenable, and it does not require monotone missingness. One disadvantage of using MI is that it requires additional work to specify the imputation model. Further research could make it easier for users to specify imputation models.

While our estimates from the multiply imputed data were similar to those found using the GEE under MCAR, the MAR model had smaller standard errors and less restrictive assumptions regarding missingness. The ability to fit clustered data models within MI provides great flexibility for analysts. This principled analytic method was once limited by computational access, but, as we demonstrated, it is now readily available within general-purpose statistical software.

# 6 Acknowledgments

We are extremely grateful to all the families who took part in this study, to all the midwives who helped recruit participants, and to the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. The UK Medical Research Council (Grant ref: 74882), the Wellcome Trust (Grant ref: 076467), and the University of Bristol provide core support for ALSPAC. This research was specifically funded by grant R01-MH087786-04 from the National Institutes of Health and the Smith College Borie and Tomlinson Funds. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, or the Department of Health. The funders had no involvement in any aspect of the study.

# 7 References

Beunckens, C., C. Sotto, and G. Molenberghs. 2008. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis* 52: 1533–1548.

Birhanu, T., G. Molenberghs, C. Sotto, and M. G. Kenward. 2011. Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics* 21: 202–225.

Boyd, A., J. Golding, J. Macleod, D. A. Lawlor, A. Fraser, J. Henderson, L. Molloy, A. Ness, S. Ring, and G. Davey Smith. 2013. Cohort profile: The 'Children of the

90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 42: 111–127.

Cannon, M. J., L. Warner, J. A. Taddei, and D. G. Kleinbaum. 2001. What can go wrong when you assume that correlated data are independent: An illustration from the evaluation of a childhood health intervention in Brazil. *Statistics in Medicine* 20: 1461–1467.

Caria, M. P., R. Bellocco, M. R. Galanti, and N. J. Horton. 2011. The impact of different sources of body mass index assessment on smoking onset: An application of multiple-source information models. *Stata Journal* 11: 386–402.

Collins, L. M., J. L. Schafer, and C.-M. Kam. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 6: 330–351.

DeSouza, C. M., A. T. Legedza, and A. J. Sankoh. 2009. An overview of practical approaches for handling missing data in clinical trials. *Journal of Biopharmaceutical Statistics* 19: 1055–1073.

Diggle, P., and M. G. Kenward. 1994. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C* 43: 49–93.

Fitzmaurice, G. M., N. M. Laird, G. E. P. Zahner, and C. Daskalakis. 1995. Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. *American Journal of Epidemiology* 142: 1194–1203.

Ford, T., R. Goodman, and H. Meltzer. 2003. The British Child and Adolescent Mental Health Survey 1999: The prevalence of DSM-IV disorders. *Journal of the American Academy of Child and Adolescent Psychiatry* 42: 1203–1211.

Frank Liu, G., and X. Zhan. 2011. Comparisons of methods for analysis of repeated binary responses with missing data. *Journal of Biopharmaceutical Statistics* 21: 371–392.

Golding, J., M. Pembrey, R. Jones, and the ALSPAC Study Team. 2001. ALSPAC—the Avon longitudinal study of parents and children. I. Study methodology. *Paediatric and Perinatal Epidemiology* 15: 74–87.

Goodman, R., T. Ford, H. Richards, R. Gatward, and H. Meltzer. 2000. The Development and Well-Being Assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry* 41: 645–655.

Horton, N. J., and G. M. Fitzmaurice. 2004. Regression analysis of multiple source and multiple informant data from complex survey samples. *Statistics in Medicine* 23: 2911–2933.

Horton, N. J., N. M. Laird, J. M. Murphy, R. R. Monson, A. M. Sobol, and A. H. Leighton. 2001. Multiple informants: Mortality associated with psychiatric disorders in the Stirling County Study. *American Journal of Epidemiology* 154: 649–656.

Horton, N. J., and S. R. Lipsitz. 2001. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician* 55: 244–254.

Horton, N. J., S. R. Lipsitz, and M. Parzen. 2003. A potential for bias when rounding in multiple imputation. *American Statistician* 57: 229–232.

Kann, L., C. W. Warren, W. A. Harris, J. L. Collins, B. I. Williams, J. G. Ross, and L. J. Kolbe. 1996. Youth risk behavior surveillance—United States, 1995. *Journal of School Health* 66: 365–377.

Li, L., C. Shen, X. Li, and J. M. Robins. 2011. On weighting approaches for missing data. *Statistical Methods in Medical Research* 22: 14–30.

Liang, K.-Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Lloyd, J. E. V., J. Obradović, R. M. Carpiano, and F. Motti-Stefanidi. 2013. JMASM 32: Multiple imputation of missing multilevel, longitudinal data: A case when practical considerations trump best practices? *Journal of Modern Applied Statistical Methods* 12: 261–275.

Long, J. S., and J. Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd ed. College Station, TX: Stata Press.

Moons, K. G., R. A. Donders, T. Stijnen, and F. E. Harrell, Jr. 2006. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* 59: 1092–1101.

Raghunathan, T. E., J. M. Lepkowski, J. V. Hoewyk, and P. Solenberger. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27: 85–95.

Royston, P. 2005. Multiple imputation of missing values: Update of ice. *Stata Journal* 5: 527–536.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63: 581–592.

———. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

StataCorp. 2013. *Stata 13 Multiple-Imputation Reference Manual*. College Station, TX: Stata Press.

Swanson, S. A., K. M. Aloisio, N. J. Horton, K. R. Sonneville, R. D. Crosby, K. T. Eddy, A. E. Field, and N. Micali. 2014. Assessing eating disorder symptoms in adolescence: Is there a role for multiple informants? *International Journal of Eating Disorders* 47: 475–482.

van Buuren, S. 2012. *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC.

White, I. R., P. Royston, and A. M. Wood. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 30: 377–399.

Xie, F., and M. C. Paik. 1997. Generalized estimating equation model for binary outcomes with missing covariates. *Biometrics* 53: 1458–1466.

Yoo, B. 2010. The impact of dichotomization in longitudinal data analysis: a simulation study. *Pharmaceutical Statistics* 9: 298–312.

### About the authors

Kathryn Aloisio graduated from Smith College, Northampton, MA, with a BA in mathematics and statistics and is currently a master's student in the Department of Mathematics and Statistics at the University of Massachusetts, Amherst, MA. Her research involves statistical computing for the analysis of clustered and incomplete data.

Sonja Swanson is a doctoral candidate in the Department of Epidemiology at the Harvard School of Public Health, Boston, MA. Her current research focuses on psychiatric epidemiology with a particular interest in improving methods for measurement and classification of psychiatric disorders.

Nadia Micali is a senior lecturer in child and adolescent psychiatry at the Institute of Child Health, University College London and a psychiatrist at Great Ormond Street Hospital, London, UK. She is a trained epidemiologist, and her research focuses on eating disorder epidemiology, risk factors for eating disorders, and developmental risk for psychiatric disorders.

Alison E. Field is an associate professor of pediatrics at Harvard Medical School and an associate professor in the Department of Epidemiology at the Harvard School of Public Health. Her research focuses on classification, risk factors, and the course of obesity and eating disorders.

Nicholas Horton is a professor of statistics in the Department of Mathematics and Statistics at Amherst College, Amherst, MA. His research interests involve the development and dissemination of methods for the analysis of clustered and incomplete data.