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

Aloisio, Kathryn M; Swanson, Sonja A; Micali, Nadia; Field, Alison; Horton, Nicholas J

How to cite

ALOISIO, Kathryn M et al. Analysis of partially observed clustered data using generalized estimating equations and multiple imputation. In: The Stata Journal, 2014, vol. 14, n° 4, p. 863–883.

This publication URL: https://archive-ouverte.unige.ch/unige:151352

Published in final edited form as:
Stata J. 2014 October 1; 14(4): 863–883.

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

Kathryn M. Aloisio,
Smith College, Northampton, MA

Sonja A. Swanson,
Harvard School of Public Health, Boston, MA

Nadia Micali,
University College London, London, UK

Alison Field, and
Harvard School of Public Health, Boston, MA

Nicholas J. Horton
Amherst College, Amherst, MA

Kathryn M. Aloisio: kaloisio@gmail.com; Sonja A. Swanson: sswanson@hsph.harvard.edu; Nadia Micali: n.micali@ucl.ac.uk; Alison Field: alison.field@childrens.harvard.edu; Nicholas J. Horton: nhorton@amherst.edu

Abstract

Clustered data arise in many settings, particularly within the social and biomedical sciences. As an example, multiple–source reports are commonly collected in child and adolescent psychiatric epidemiologic studies where researchers use various informants (e.g. parent and adolescent) to provide a holistic view of a subject's symptomatology. Fitzmaurice et al. (1995) have described estimation of multiple source models using a standard generalized estimating equation (GEE) framework. However, these studies often have missing data due to additional stages of consent and assent required. The usual GEE is unbiased when missingness is Missing Completely at Random (MCAR) in the sense of Little and Rubin (2002). This is a strong assumption that may not be tenable. Other options such as weighted generalized estimating equations (WEEs) are computationally challenging when missingness is non–monotone. Multiple imputation is an attractive method to fit incomplete data models while only requiring the less restrictive Missing at Random (MAR) assumption. Previously estimation of partially observed clustered data was computationally challenging however recent developments in Stata have facilitated their use in practice. We demonstrate how to utilize multiple imputation in conjunction with a GEE to investigate the prevalence of disordered eating symptoms in adolescents reported by parents and adolescents as well as factors associated with concordance and prevalence. The methods are motivated by the Avon Longitudinal Study of Parents and their Children (ALSPAC), 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 were fairly similar to

the GEE under MCAR, the MAR model had smaller standard errors, while requiring less stringent assumptions regarding missingness.

Keywords

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 require sophisticated analytical methods. Standard error estimates that do not account for the association within clusters will be inaccurate and inferences will be invalid (Cannon et al. 2001).

As an example, multiple–source reports are commonly collected in child and adolescent psychiatric epidemiologic studies where researchers use various informants (e.g. parent and adolescent) to provide a holistic view of a subject's symptomatology. These clustered reports also arise in other settings such as geriatric studies, school settings, and health services research (Caria et al. 2011).

A number of papers 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) have proposed methodology for simultaneously analyzing information from multiple–source outcomes 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 due to additional stages of consent and assent required. Analyzing data without appropriately accounting for missingness can induce bias and loss of efficiency.

The usual GEE is unbiased when missingness is "missing completely at random" (MCAR); that is, 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 use of 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 and/or observed covariates. This method assumes the less restrictive missingness mechanism named "missing at random" (MAR) by Little and Rubin (2002). The process to fit the weighted GEE includes estimating the probability of being observed, dropping all of the partially observed subjects, and fitting the re–weighted 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

non–monotone patterns of missingness. The monotone structure is rarely seen in observational studies with many covariates and is not present in the motivating example. Furthermore accounting for a complex non–monotone pattern is computationally difficult (Li et al. 2011). Therefore other approaches are needed.

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

Various 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 multiple imputation in conjunction with estimating equations (MI–GEE). Beunckens et al. (2008) found that using the MI–GEE approach was more successful in comparison with the usual GEE and the weighted GEE. DeSouza et al. (2009), Yoo (2010), and Birhanu et al. (2011) expanded the simulation study each concluding that for non–normal and repeated binary responses MI–GEE outperformed the weighted GEE and is a valid analysis tool. Liu and Zhan (2011) undertook a similar simulation study, 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 for adjustments to the imputation model. Lloyd et al. (2013) describe how to undertake estimation for longitudinal regression using the ICE/UVIS add-on routines in Stata 11.

Previously estimation of partially observed clustered data was computationally challenging however recent developments in Stata have facilitated their use in practice. The goal of this paper is to demonstrate estimation of a GEE model from multiply imputed data using the mi system in Stata 13.

We begin by describing the motivating study, the Avon Longitudinal Study of Parents and Children, a long–running cohort study, utilizing parent and adolescent questionnaires to research the health and development of the adolescent. This is followed by a description of how GEE models can be used to fit generalized linear models using available case data. Next, we introduce multiple imputation and simultaneous estimation of GEE models using multiply–imputed data within Stata. Then we fit the GEE models to our motivating data using both available cases and the imputed data. We conclude with a discussion of 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 Avon Longitudinal Study of Parents and their Children (ALSPAC), a longitudinal, prospective study of women and pregnancy (Golding et al. 2001; Boyd et al. 2013). All pregnant women living in the geographical area of Avon, UK, who were expected to deliver their baby between 1st April 1991 and 31st December

1992 were invited to take part in the study. The adolescents from 14,541 pregnancies were enrolled. At age one, there were 12,388 singleton adolescents alive with complete information on adolescent's sex and maternal age. Adolescents and their parents have been 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 at age 16. The analytic sample consists of 7,986 adolescents that had at least one adolescent or parent reports at age 14 and 16, as well as 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 complete a questionnaire version of the Eating Disorder– Developmental and Well–Being Assessment (DAWBA) with no skip rules (Goodman et al. 2000; Ford et al. 2003). Parents were asked whether their study 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.

For the purpose of demonstration, the current 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 and the Local Research Ethics Committees as well as the Smith College and Amherst College Institutional Review Boards.

2.2 Variables

The age 16 questionnaire asked the parent "Over the last 3 months, has your study teenager made herself/himself sick to avoid putting on weight" and the adolescent "During the past year, how often did you make yourself throw up (vomit) to lose weight or avoid gaining weight?" Due to 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 need to reshape our dataset from wide form (one row per subject) into 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).

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 ith subject from the jth source (with $i = 1, \ldots, N$; $j = 1, \ldots, J$). 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 X_{ij} be a $p \times 1$ vector of covariates, associated with the outcome obtained for the ith subject from the jth source (X_{ij} contains both source information and subject–specific information). We let $Y_i = (Y_{i1}, \ldots, Y_{iJ})'$ be the $J \times 1$ outcome vector for the ith subject, and X_i 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 just a single 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 (GLM) family. However, the clustered nature of multiple sources, where two reports from the same adolescent are likely to be positively associated, require a more sophisticated model.

Generalized estimating equations (GEEs) first described by Liang and Zeger (1986), 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 (a.k.a. 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 (in this setting include adolescent gender, maternal education and parity), is given by:

$$g(E[Y_{ij}|X_{ij}]) = X_{ij}'\beta$$

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

$$\begin{aligned}\text{logit}(E[Y_{ij}|X_{ij}]) = {} & \beta_0 + \beta_1 \text{multipare} + \beta_2 \text{edua} + \beta_3 \text{female} + \beta_4 \text{child} \\ & + \beta_5(\text{child} \times \text{female}) + \beta_6(\text{child} \times \text{multipare}) + \beta_7(\text{child} \times \text{edua}).\end{aligned} \quad (1)$$

The coefficients are log odds ratio (log(OR)), where β_5, β_6, and β_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 vomiting behavior model we dropped the extraneous interactions (those with *p*–values ≥ 0.05) and refit the model to obtain estimates for a parsimonious model, which retained the gender by source interaction:

$$\text{logit}(E[Y_{ij}|X_{ij}])=\beta_0+\beta_1\text{multipare}+\beta_2\text{edua}+\beta_3\text{female}+\beta_4\text{child}+\beta_5(\text{child}\times\text{female}) \quad (2)$$

Without the parity by source and maternal education by source interaction in model (2) exp(β_1) and exp(β_2) are interpreted as odds ratios for parity and maternal education, respectively, within levels of source and gender. To interpret the interaction term we can extract the equation for parent reports (presented in model (3)) and similarly we obtain the equation for adolescent reports (model (4)).

$$\text{logit}(E[Y_{ij}|X_{ij},\text{child}==0])=\beta_0+\beta_1\text{multipare}+\beta_2\text{edua}+\beta_3\text{female} \quad (3)$$

$$\text{logit}(E[Y_{ij}|X_{ij},\text{child}==1])=(\beta_0+\beta_4)+\beta_1\text{multipare}+\beta_2\text{edua}+(\beta_3+\beta_5)\text{female} \quad (4)$$

Note that for the adolescent report model (4), β_4 is the log odds for additional prevalence for adolescent reports and β_5 is the additional log odds for female adolescent reports.

3.3 Accounting for missing data

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

```
. by female: tabulate vomit_c16 vomit_p16, miss
------------------------------------------------------------------------
-> female = 0
| vomit_p16
vomit_c16 | 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_p16
vomit_c16 | 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 1,688 male adolescents returned completed questionnaires and 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 either adolescent, parent, or both reports, regardless of gender. For instance for male subjects, 7% ([271+3]/3,834) have observed adolescent reports but missing parent reports, 23% ([888+2]/3,834) have missing adolescent reports but observed parent reports, and 26% (982/3,834) have both adolescent and parent reports missing at age 16. Accounting for the partially observed responses is crucial for obtaining reliable results for future inferences.

There are three types of 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 $\mathbf{Y}^{obs}$ as the observed component of the outcome and $\mathbf{X}^{obs}$ as the observed components of the predictors. Similarly, we denote $\mathbf{Y}^{mis}$ and $\mathbf{X}^{mis}$ as the unobserved components of the outcome and predictors, respectively. In addition, $\mathbf{Z}^{obs} = (\mathbf{Y}^{obs}, \mathbf{X}^{obs})$ and $\mathbf{Z}^{mis} = (\mathbf{Y}^{mis}, \mathbf{X}^{mis})$ denote the vector of observed variables and missing variables, respectively. We will also use γ to denote the regression parameters. Lastly, we define a set $\mathbf{R}$ of response indicators (i.e. $R_i = 1$ if the *i*th element of $\mathbf{Z}$ is observed, and equals 0 otherwise).

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

$$P(\mathbf{R}|\boldsymbol{Z},\gamma)=P(\mathbf{R}|\boldsymbol{Z}^{\text{obs}},\boldsymbol{Z}^{\text{mis}},\gamma)=P(\mathbf{R}|\gamma) \quad (5)$$

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

The mechanism missing at random (MAR) assumes

$$P(\mathbf{R}|\boldsymbol{Z},\gamma)=P(\mathbf{R}|\boldsymbol{Z}^{\text{obs}},\gamma) \quad (6)$$

(i.e., the probability of being missing is the same after conditioning on the observed data). Heuristically, this states that missingness depends only on observed quantities, including outcomes, predictors and/or auxillary variables. Most analyses start with this assumption as

it is more believable to happen than MCAR, particularly within datasets containing many variables (Collins et al. 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) provides the most concern for researchers and analysts since MNAR means that the probability of being missing varies for reasons that are unknown to the researcher (missingness is related to the unobserved quantities). Symbolically, $P(\mathbf{R}|\mathbf{Z})$ can not be simplified and it must be modeled as part of the likelihood, Little and Rubin (2002) call this “nonignorable”. While important for undertaking sensitivity analyses, we will not further consider MNAR missingness.

The pattern of missingness, monotone versus non–monotone, 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 stairstep–like pattern (i.e. non–increasing or non–decreasing) 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 parent report and others missing adolescent report.

Using misschk we can 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%, $n = 1, 578$), missing just the adolescent report (17%, $n = 1, 355$), and missing just the parent report (11%, $n = 876$). However, we do not have a monotone missingness pattern because for each of the covariates (parity and maternal education) there are cases missing either 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 $n = 7, 968$ available cases versus $n = 3, 692$ 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 since 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, this is not feasible in this setting even for just modeling parent and adolescent reports for one age, since some subjects have missing adolescent reports while others have missing parent reports. It should be noted that there is support for weighted estimating equations in Stata 13, see Stata help for weights for more details.

Multiple imputation—Multiple imputation is a principled method due to Rubin (1976) to account for missing data. It involves a three–step approach for estimation of incomplete data

regression models. First, plausible values for missing observations are created that reflect uncertainty about the nonresponse model. These values are used to "fill–in" or impute the missing values (generally under an assumption of MAR). This process is repeated, resulting in the creation of a number of "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 taken into account (Little and Rubin 2002). Since increasing the number of imputed datasets minimizes variability introduced into the results due to the imputation process (Horton and Lipsitz 2001; White et al. 2011; van Buuren 2012), we recommend a set of 25 imputations, though more are computationally possible.

Specifying imputation model: Multiple imputation requires the analyst to provide an appropriate specification of the imputation model. If this model is misspecified, there is the potential for bias (White et al. 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 reporting 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. Additionally, to preserve the source by gender interaction we have to account for the interaction term. We chose to do 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 mi impute chained).

In addition to all the variables that may be used in the analysis model, any auxiliary variables that may contain information about missing data should also be included. For our model, a measure for self–reported Body Mass Index, the mother's age at delivery, and the adolescent's age at the time of reporting were included. Furthermore, the outcome variable should always be present in the imputation model to obtain valid results (Moons et al. 2006). By including all of the variables necessary for the model as well as 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 et al. 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 has 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 using chained equations with a variable–by–variable approach. The imputation model is specified separately for each variable, involving the other variables as predictors. At each stage of the algorithm, an imputation is generated for all the missing values in a given variable, then this imputed variable is used in the imputation of the next variable. This process repeats imputing missing values 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 "Ru-bin's rules" to combine the repeated–imputation results, where the total variance stems from three sources (Little and Rubin 2002):

1. The variance caused by the fact that we are taking a sample rather than observing the entire population. This is the conventional statistical measure of variability.
2. The extra variance caused by the fact that there are missing values in the sample.
3. The extra simulation variance caused by the fact that the estimate itself is estimated for finite number of imputations.

4 Application in Stata

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

To use the mi system, we begin by reading in the dataset and creating the analytic set. Additional variables from the cohort study are included in the imputation model to make the MAR assumption more plausible (Collins et al. 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 as it uses slightly less memory compared with the wide (wide) structure. Though the wide format is slightly faster.

```
. mi set mlong
```

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

The variables that require imputation 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 how many 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
... (23 lines removed)
26 | 7 0.09 99.94
27 | 5 0.06 100.00
-------------+-----------------------------------
Total | 7,968 100.00
```

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

4.2 Imputation model specification

We then create the imputed datasets in Stata using mi impute. This requires the user to specify the imputation model. The first selection is the imputation method. For univariate imputation (where the pattern of missingness is by definition monotone) the user 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 (i.e. a mixture of continuous and discrete) the situation is more complicated. If the pattern of missingness is monotone mi impute monotone can be used and each variable is assigned an imputation method. If there is an arbitrary missing patten (as in the present analysis) mi imputed mvn for multivariate normal variables or mi impute chained for the chained equation method are used. Table 3 lists these and other options for selection of the imputation method.

Since 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 et al. 2011; van Buuren 2012). Using 25 chains run 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 has been shown to be unbiased for dichotomous variables, ensures that imputed values are plausible, and may be more appropriate if the normality assumption is violated (Horton et al. 2003).

Both sets of models included all symptoms from both sources at each age as well as 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(R) Core(TM)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, 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 as well as survey regression models. The present study uses mi estimate xtgee to fit the GEE models due to 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 into long form (two rows per subject). This is straightforward to accomplish using the post mi data manipulation commands. To help 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 where $j = 1$ indicates adolescent report and $j = 0$ indicates 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 with the long format we have doubled the number of rows (15,936). 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 model (2) includes the gender by source interaction. The recoding of a new variable using imputed data is implemented with mi passive, we named the interaction variable femchild.

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

Prior to fitting the model we must declare the type of complex data (e.g. 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 model (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 controlling for other factors the odds for exhibiting vomiting behavior for a adolescent with siblings is 1.31 (95% CI 1.02–1.70) times the odds for an only adolescent exhibiting vomiting behavior. Maternal education was found not to be significantly associated with vomiting behavior (OR 1.14; 95% CI 0.89–1.46), after controlling for the other factors.

To interpret the gender by source interaction we calculated the four predicted probabilities using model (3) and (4) with the other covariates set to 0 and the inverse logit function (invlogit $(\beta)=\frac{\exp(\beta)}{1+\exp(\beta)}$). The predicted probability for male adolescent report (invlogit(β_0 + β_4), 0.8%, [95% CI 0.4%–1.1%]), male's parent report (invlogit(β_0), 0.3%, [95 % CI 0.08%–0.5%]), female adolescent report (invlogit(β_0 + β_3 + β_4 + β_5), 6.8%, [95% CI 5.3%–8.3%]), and female's parent report (invlogit(β_0 + β_3), 0.4%, [95% CI 0.1%–0.6%]).

From these we can determine important distinct patterns. First, estimates for vomiting are higher when reported by the adolescent relative to their 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 understandings of the diagnosis and 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 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
------------------------------------------------------------------------------
-------------------------
```

Comparing the two methods, the estimates are similar and the standard errors from the multiple imputation 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 paper, we have detailed how to account for clustering when multiple imputation is used to account for missingness.

Multiple source data often arise in the analysis of studies with complex survey designs. Along with clustering, stratification and sampling weights must be taken into account in the analysis. This can be undertaken in Stata using the survey design tools (Horton and Fitzmaurice 2004).

Many analytic approaches rely on the accuracy of the assumptions associated with the proposed method. Biases can be introduced from negligence or inaccurate analysis of the collected data. Assumptions that missingness is “missing at random” is inherently unverifiable without auxiliary information. By utilizing methods that incorporate other variables associated with missingness and/or responses, the possibility of biases will be reduced and the data will be represented more accurately.

The GEE model is attractive because of its ability to account for clustering or repeated measures induced by longitudinal data. However, the assumption of MCAR is very restrictive in a world where 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, this 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. Use of multiple imputation is attractive because it can incorporate auxiliary variables (to make MAR more tenable) and does not require monotone missingness. One disadvantage is that additional work is needed to specify the imputation model. Additional research could help facilitate the process of specifying imputation models.

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

Acknowledgments

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and 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 Smith College Borie and Tomlinson Funds. The views expressed in this publication are those of the author(s) 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.

References

Beunckens C, Sotto C, Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. Computational Statistics & Data Analysis. 2008; 52(3):1533–1548.

Birhanu T, Molenberghs G, Sotto C, Kenward MG. Doubly robust and multiple–imputation–based generalized estimating equations. Journal of Biopharmaceutical Statistics. 2011; 21(2):202– 225. [PubMed: 21390997]

Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Smith GD. Cohort Profile: The 'Children of the 90s'–the index offspring of the Avon Longitudinal Study of Parents and Children. International Journal of Epidemiology. 2013; 42(1):111–127. [PubMed: 22507743]

Cannon MJ, Warner L, Taddei JA, Kleinbaum DG. 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. 2001; 20:1461–1467. [PubMed: 11343366]

Caria MP, Bellocco R, Galanti MR, Horton NJ. The impact of different sources of body mass index assessment on smoking onset: an application of multiple-source information models. Stata Journal. 2011; 11(3):386–402. [PubMed: 22065944]

Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001; 6(4):330–351. [PubMed: 11778676]

DeSouza CM, Legedza ATR, Sankoh AJ. An overview of practical approaches for handling missing data in clinical trials. Journal of Biopharmaceutical Statistics. 2009; 19(6):1055–1073. [PubMed: 20183464]

Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. Applied Statistics. 1994; 43:49–73.

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. [PubMed: 7485066]

Ford T, Goodman R, Meltzer H. 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. 2003; 42:1203–1211. [PubMed: 14560170]

Golding J, Penbrey M, Jones R. the ALSPAC Study Team. ALSPAC—The Avon Longitudinal Study of Parents and Children. Paediatric and Perinatal Epidemiology. 2001; 15:74–87. [PubMed: 11237119]

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

Horton NJ, Fitzmaurice GM. Regression analysis of multiple source data from complex survey samples. Statistics in Medicine. 2004; 23(18):2911–2933.10.1002/sim.1879 [PubMed: 15344194]

Horton NJ, Laird NM, Murphy JM, Monson RR, Sobol AM, Leighton AH. Multiple informants: mortality associated with psychiatric disorders in the Stirling County Study. American Journal of Epidemiology. 2001; 154(7):649–656. [PubMed: 11581099]

Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician. 2001; 55(3):244–254.

Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. The American Statistician. 2003; 57(4):229–232.

Kann L, Warren CW, Harris WA, Eckenrode J, Zielinski D, Smith E, Marcynyszyn LA, Henderson J, CR, Kitzman H, Cole R, Powers J, Olds DL. Youth Risk Behavior Surveillance. Development and Psychopathology. 1996; 13:873– 890.

Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. Statistical Methods in Medical Research. 2011; 22(1):14–30. [PubMed: 21705435]

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

Little, RJA.; Rubin, DB. Statistical analysis with missing data, 2nd edition. John Wiley & Sons; New York: 2002.

Liu GF, Zhan X. Comparisons of methods for analysis of repeated binary responses with missing data. Journal of Biopharmaceutical Statistics. 2011; 21(3):371–392. [PubMed: 21442514]

Lloyd JEV, Obradović J, Carpiano RM, Motti-Stefanidi F. JMASM 32: Multiple imputation of missing multilevel, longitudinal data: a case when practical considerations trump best practices? Journal of Modern Applied Statistical Methods. 2013; 12(1):261–275.

Moons KGM, Donders RA, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006; 59(10):1092–1101. [PubMed: 16980150]

Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 2001; 27(1):85–95.

Royston P. Multiple imputation of missing values. Stata Technical Journal. 2005; 5(4):527–536.

Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–590.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. Wiley; 1987.

StataCorp. Stata Multiple-Imputation Reference Manual: Release 13.0. Stata Corporation; College Station, TX: 2013.

Swanson SA, Aloisio KM, Horton NJ, Sonneville K, Crosby R, Eddy K, Field AE, Micali N. Assessing eating disorder symptoms in adolescence: Is there a role for multiple informants? International Journal of Eating Disorders. 201410.1002/eat.22250

van Buuren, S. Flexible Imputation of Missing Data. Chapman and Hall/CRC; 2012.

White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011; 30(4):377–399. [PubMed: 21225900]

Xie F, Paik MC. Generalized estimating equation model for binary outcomes with missing covariates. Biometrics. 1997; 53:1458–1466. [PubMed: 9423260]

Yoo B. The impact of dichotomization in longitudinal data analysis: a simulation study. Pharmaceutical Statistics. 2010; 9(4):298–312. [PubMed: 19904810]

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)

Table 3

Multiple imputation 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