Multiple Imputation for Missing Data Analysis in Proportional Odds Models for Ordinal Response Variables

Xing Liu
Eastern Connecticut State University

Haiyan Bai
University of Central Florida

Hari Koirala
Eastern Connecticut State University

Although multiple imputation (MI) of missing data has been getting popularity in educational research, previous research mainly focuses on normally distributed continuous variables. There is a great need to impute ordinal categorical variables. Further, since there exist various methods for MI, it is unclear which one should be used for specific empirical data. The purpose of this study is to compare and illustrate the implementation of both MI for a single ordinal variable and multiple imputation by chained equations (MICE) for multivariate variables in ordinal logistic regression to predict mathematics proficiency levels. This study helps researchers better understand and implement the methods through comparing various proportional odds (PO) models and the results of these models with different numbers of imputations. For demonstration purposes, the empirical data from the High School Longitudinal Study (2009) are used for the missing data analysis.

Ordinal logistic regression extends binary logistic regression when the ordinal outcome variable has more than two levels. The proportional odds (PO) model (Agresti, 2007, 2010, 2013; Ananth & Kleinbaum, 1997; Armstrong & Sloan, 1989; Hilbe, 2009; Liu, 2009, 2016; Long, 1997; Long & Freese, 2014; McCullagh, 1980; McCullagh & Nelder, 1989; O'Connell, 2000, 2006; O'Connell & Liu, 2011; Powers & Xie, 2000), implemented as the default for ordinal regression analysis in general-purpose statistical software packages, is most commonly used for ordinal response variables. This model estimates the cumulative odds of being at or below a particular level of the ordinal response variable. Thus, it is also called the cumulative odds model.

When conducting ordinal regression analysis, we assume that the data are complete. However, missing data are common in educational research, particularly in large-scale national studies, such as the Early Childhood Longitudinal Study-Kindergarten (ECLS-K), the Educational Longitudinal Study of 2002 (ELS: 2002), and the High School Longitudinal Study of 2009 (HSLS: 2009). By simply ignoring the missing mechanisms and deleting the missing data in the data analysis, we may obtain the biased estimates of parameters and incorrect variance estimates, and thus the misleading results. Therefore, it is critical for researchers to understand and be familiar with techniques for dealing with missing data in ordinal regression analysis.

According to Little and Rubin (2002), the missing data mechanisms normally include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). If missing values on a variable cannot be predicted from any other variables in the data and the unobserved values of that variable itself, it is considered to be MCAR. When the pattern of missing of a variable can be predicted from other variables in the data but is unrelated to that variable itself, the condition is called MAR. The MAR condition is less restrictive than the MCAR and is more commonly treated. When the probability of missing data on a variable depends on that variable itself, this condition is called MNAR. Since the nature and properties of these three types of missingness are different, it is important to use appropriate techniques in missing data analysis.

Several commonly used methods for dealing with missing data when they are MAR included direct deletion or imputation of the missing values, such as the listwise deletion, pairwise deletion, mean substitute, single regression imputation, expectation-maximization (EM) (Aitkin & Rubin, 1985; Lawrence & Reilly, 1990), and multiple imputation (Allison, 2001; Cheema, 2014; Horton & Kleinman, 2007; McKinght, McKnight, Sidami, & Figueredo, 2007; Peng, Harwell, Liou, & Ehman, 2006). Among them, multiple imputation (MI), the only viable method in most situations, is superior over the other methods to handle missing data issues (Enders, 2010; Graham, 2009, 2012; Little & Rubin, 2002; Peugh & Enders, 2004; Rubin, 1987, Schafer, 1999; Schafer & Graham, 2002). Rubin first introduced MI as a process of generating imputed values on the basis of exiting data. Specifically, MI is a simulation-based approach. It replaces missing data with multiple sets of simulated data to create completed data sets, applies standard analysis to each data set, and obtains the unbiased parameter estimates and standard errors (Rubin, 1987). This process includes three steps: the imputation step, estimation step, and pooling step. MI is better than other methods since it is more efficient and flexible. It creates multiple imputations and accounts for the sampling variability. Thus, it provides unbiased parameter estimates and variances.

Although the MI method of dealing with missing data that are MAR has been getting popularity in educational research, this method mainly focuses on normally distributed continuous variables. There is a great need to impute categorical variables, such as binary, ordinal, and nominal variables. However, investigating missing data techniques in ordinal logistic regression analysis is scarce. One major reason is the complexity of the analytic steps researchers need to follow when conducting missing data analysis using MI, which requires advanced analytic skills. Further, since different methods can be performed for MI, such as MI for a single variable and multivariate imputation, it is unclear which one should be recommended for empirical use. Therefore, it is critical to help educational researchers to better understand the methods and the procedures for fitting the ordinal logistic regression model with missing data in practice.

The purpose of this study is to illustrate the use of both MI for a single ordinal variable and multiple imputation by chained equations (MICE) for multivariate variables in ordinal logistic regression to predict mathematics proficiency levels in educational research. In addition, it compares the results of the PO models with and without imputation and compares the results of the models with different numbers of imputations. For demonstration purposes, the empirical data from the High School Longitudinal Study of 2009 (HSLS: 2009) are employed to conduct the missing data analysis.

Theoretical Framework

The Proportional Odds Model

A binary logistic regression model estimates the odds and the probability of experiencing an event for the dichotomous outcome variable on a set of predictors. The logistic regression model is defined as:

$$\ln(Y') = \text{logit}\,[\pi(\underline{x})] = \ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = \alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_pX_p, \qquad (1)$$

where logit $[\pi(\underline{x})]$ is the log odds of success, and the odds is a ratio between the probability of having an event and the probability of not having that event.

By extending binary logistic regression, the proportional odds (PO) model estimates the odds and the probabilities of being at or below a particular category of an ordinal response variable. This model can be expressed on the logit scale as follows:

$$\ln(Y_j') = \text{logit}\,[\pi_j(x)] = \ln\left(\frac{\pi_j(x)}{1-\pi_j(x)}\right) = \alpha_j + (-\beta_1X_1 - \beta_2X_2 - \ldots - \beta_pX_p), \qquad (2)$$

where $\pi_j(x) = \pi(Y \leq j \mid x_1, x_2, \ldots, x_p)$, which is the cumulative probability of being at or below category j ($j = 1, 2, \ldots, J$-1) , given a set of predictors. The cut points are α_j, and $\beta_1, \beta_2, \ldots, \beta_p$ are the logit coefficients. This PO model estimates different cut points or intercepts, but the effect of any predictor is assumed to be the same across these cut points. To estimate the cumulative odds of being at or below the j^{th} category, this model can be rewritten as:

$$\text{logit}\,[\pi(Y \leq j \mid x_1,x_2,\ldots x_p)] = \ln\left(\frac{\pi(Y \leq j \mid x_1, x_2, \ldots x_p)}{\pi(Y > j \mid x_1, x_2, \ldots x_p)}\right) = \alpha_j + (-\beta_1X_1 - \beta_2X_2 - \ldots - \beta_pX_p \qquad (3)$$

where logit is the log odds of being at or below a particular category relative to being beyond that category. The signs before both logit coefficients on the right side of the equation are negative so that an increase in a predictor variable is associated with the odds of being beyond a particular category.

Multiple Imputation

Single imputation replaces a missing value with a single imputed value. This method ignores the uncertainty existed in the unknown missing values, so it produces biased standard errors of parameter estimates (Allison, 2001). Rather than treating the missing values as known values, multiple imputation replaces each missing value with a set of plausible values that represent a random sample of the missing values, so both the variability among imputed values and the variability due to sampling are considered. Thus, it produces the unbiased parameter estimates and standard errors (Rubin, 1987).

MI involves three steps including the imputation step, estimation step, and pooling step. MI first creates several imputed data sets, each of which contains a plausible value of missing values. Then data analysis is conducted on each imputed data set. Finally, the results from the analyses of all complete data are combined. MI can be conducted on a single variable or a set of variables with missing data. When handling multiple variables with missing data, the multiple imputation by chained equations (MICE) method (Royston & White, 2011) is a better option than the MI method for a single variable; therefore, MICE has been increasingly popular. The major advantage of MICE is to handle various types of variables with missing data, such as continuous, binary, ordinal, nominal, and count variables, and each variable type can have its own imputation model. As introduced by Royston and White, MICE is an iterative process which involves multiple steps to impute a set of variables with missing values. It uses a Gibbs-like algorithm to impute multiple variables in sequence. First, a particular variable, X1, is predicted by the other variables. Missing values in X1 are then filled in by simulated data from the posterior predictive distribution of X1. Next, the second variable X2 is predicted by the other variables including the imputed values of X1. This process is repeated like a chain until the missing data in all other variables are estimated. A number of cycles are repeated to create an imputed dataset and five to 20 cycles are normally conducted. Just like the regular MI, once imputed datasets are created by MICE, the subsequent analysis can be conducted and the results can be pooled.

Methodology

Sample

The High School Longitudinal Study of 2009 (HSLS: 2009), conducted by the NCES, surveyed high school students, parents, teachers, and other school personnel and assessed 9th graders' mathematics skills and reasoning. In the 2009 base-year data, 21,444 high school students from a national sample of 944 schools participated in the study. Students were asked to provide information regarding basic demographics, school and home experience, coursework, and time spent on different activities, mathematics and science attitude, mathematics and science self-efficacy, their feelings about math and science teacher, and future plans.

The outcome variable, students' mathematics proficiency levels in high schools, was ordinal with five levels (1 = students can answer questions in algebraic expressions; 2 = students can answer questions and solve problems for multiplicative and proportional situations; 3 = students can understand algebraic equivalents and solve problems; 4 = students can understand systems of linear equations and solve problems; 5 = students can understand linear functions, find and use slopes and intercepts of lines, and can use functional notation). In addition, those students who failed to pass through level 1 were assigned to level 0. Table 1 provides the frequency of six mathematics proficiency levels (i.e., levels 0-5).

Table 1. Proficiency Categories and Frequencies (Proportions)
for the Study Sample, HSLS: 2009 Base Year Data (n = 21,444)

Proficiency Category	Description	Frequency (%)
0	Did not pass level 1	2263 (10.6%)
1	Algebraic expressions	4933 (23.0%)
2	Multiplicative & proportional thinking	5495 (25.6%)
3	Algebraic equivalents	5761 (26.9%)
4	Systems of equations	2396 (11.2%)
5	Linear functions	596 (2.8%)

Data Analysis

Conventional PO model. To compare the results from a conventional PO model with the results from MI model, a conventional PO model with all six predictor variables was fitted first. STATA `ologit` command was used for model fitting. The listwise deletion was used to remove the missing data.

PO model with MI for a Single Ordinal Variable. To implement MI to PO model, the following procedures were followed. First, before multiple imputation was conducted, the STATA `mi set` and `mi register` commands were used to work on the missing data. Second, the `mi impute` command was used for multiple imputations for a single variable. Third, the `mi estimate: ologit` command was used to fit the proportional odds model with missing data.

PO model with MI for Multiple Ordinal Variable. To conduct multiple imputations for multiple variables, the MICE method was used with the `mi impute chained` command. Different variable types were specified and the number of imputations was specified in three different conditions (i.e., 5, 10, and 20).

Results

Proportional Odds (PO) Model without Multiple Imputation

The results of the PO model without multiple imputation are provided in Table 2. The log likelihood ratio chi-square test, $LR\ \chi^2_{(7)} = 2059.56$, $p < 0.001$, which indicated that the model with six predictors provided a better fit than the null model without independent variables.

The logit coefficients of all six predictors except one on the mathematics proficiency level were significantly different from zero. The estimated logit regression coefficient for mathematics identity (*MTHID*), $\beta = 0.603$, $z = 19.57$, $p < 0.001$; the logit coefficient for mathematics self-efficacy (*MTHEFF*), $\beta = .218$, $z = 6.96$, $p < .001$; the logit coefficient for school engagement (*SCHENG*), $\beta = 0.151$, $z = 5.57$, $p < 0.001$; the logit coefficient for mother's education level (Bachelor's degree) (*MEDU:BA*), $\beta = 0.824$, $z = 15.75$, $p < 0.001$; the logit coefficient for mother's education level (Master's degree and beyond) (*MEDU:MA*), $\beta = 1.360$, $z = 17.25$, $p < 0.001$; and finally, for students' individualized education plan (*IEPFLAG*), $\beta = -1.427$, $z = -21.15$, $p < 0.001$. However, the coefficient for student's sense of school belonging (*SCHBEL*) was not significant, $\beta = 0.027$, $z = 1.01$, $p = 0.313$.

Among the six predictors, five of them were positively associated with the logits or log odds of being beyond a proficiency level versus being at or below that level. In terms of odds ratio (OR), the odds of being beyond a proficiency level increased by 1.828 with a one-unit increase in higher level of math identity and increased by 1.243 with one-unit increase in students' mathematics self-efficacy. In addition, students who had higher level of school engagement were associated with the odds of being beyond a mathematics proficiency level. With regard to mother's education level (*MEDU*), mother's Bachelor's degree and Master's degree and beyond were associated with the odds of being beyond a proficiency level. Conversely, students' individualized education plan (*IEPFLAG*) was negatively associated with the logits or log odds of being beyond a proficiency level. Being a student with an individualized education plan decreased the odds of being beyond a proficiency level by a factor of 0.240. In addition, students' sense of school belonging did not influence the odds of being above that proficiency level (OR = 1.028) since they were not significant ($p = 0.313$).

Proportional Odds Model with Multiple Imputation for a Single Ordinal Variable

Before multiple imputation was conducted, the STATA `mi set` and `mi register` commands needed to be used to work on the missing data. First, the `mi set` command set a MI dataset so that the imputed data could be saved in different styles, such as the wide and mlong styles. Second, the `mi register` command was used to specify the variable(s) for imputation once the data style was defined. Figure 1 displays the Stata syntax for creating the imputed data and specifying the variable. Third, since *MEDU* was an ordinal variable with three categories, the `mi impute ologit` command was used as the univariate method for multiple imputation. Figure 2 displays the Stata syntax for imputing the ordinal variable. Finally, the `mi estimate: ologit` command was employed to fit the PO model to the imputed data ($m = 5$). The results of the PO model with multiple imputation are provided in Table 2.

The results of the PO model with multiple imputation are presented in Table 2. The *F* statistic for the overall model, $F_{(7,\ 3688.3)} = 287.70$, $p < .001$, which indicated that the model with all six predictors was significant in predicting mathematics proficiency. The logit coefficients of all the predictors were significantly different from zero. Five predictors including mathematics identity (*MTHID*), mathematics self-efficacy (*MTHEFF*), school belonging (*SCHBEL*), school engagement (*SCHENG*), and mother's education level (*MEDU*) were positively related to the logits or log odds of being beyond a proficiency level. However, the predictor, individualized education plan (*IEPFLAG*), had a negative logit coefficient.

```
. mi set mlong
. mi register imputed MOMEDU
(6145 m=0 obs. now marked as incomplete)
```

Figure 1. Create a MI data and specify a single imputed variable

```
. mi impute ologit MOMEDU MTHID MTHEFF SCHBEL SCHENG IEPFLAG, add(5)
rseed(1234) force

Univariate imputation                     Imputations =        5
Ordered logistic regression                     added =        5
Imputed: m=1 through m=5                      updated =        0

------------------------------------------------------------------
                   |              Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
            MOMEDU |      15299         6145      2004 |     21444
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Figure 2. Impute the single ordinal variable using `mi impute ologit.`

Table 2. Results of PO Models with Multiple Imputation for a Single Ordinal Variable, MEDU (Number of Imputations: 0, 5, 10, and 20)

	No Imputations: (n=0)		Imputations: (n = 5)		Imputations: (n = 10)		Imputations: (n = 20)	
Variables	b(se(b))	OR	b(se(b))	OR	b(se(b))	OR	b(se(b))	OR
α_1	-2.595		-2.477		-2.479		-2.478	
α_2	-0.810		-0.732		-0.733		-.733	
α_3	0.542		0.571		0.570		.570	
α_4	2.292		2.275		2.273		2.275	
α_5	4.203		4.155		4.153		4.154	
MTHID	0.603**	1.828	0.568**	1.765	0.567**	1.763	0.567**	1.763
	(0.031)		(0.026)		(0.026)		(0.026)	
MTHEFF	0.218**	1.243	0.197**	1.210	0.191**	1.211	0.190**	1.209
	(0.031)		(0.026)		(0.026)		(0.027)	
SCHBEL	0.027	1.028	0.048*	1.049	0.048*	1.049	0.048*	1.048
	(0.027)		(0.023)		(0.023)		(0.023)	
SCHENG	0.151**	1.163	0.185**	1.203	0.185**	1.203	0.184**	1.202
	(0.027)		(0.023)		(0.023)		(0.023)	
MEDU(BA)	0.824**	2.281	0.602**	1.825	0.600**	1.821	0.602**	1.825
	(0.052)		(0.054)		(0.051)		(0.051)	
MEDU(MA)	1.360**	3.898	1.001**	2.720	0.995**	2.705	0.999**	2.714
	(0.079)		(0.073)		(0.073)		(0.073)	
IEPFLAG	-1.427**	0.240	-1.422**	0.241	-1.424**	.240	-1.423**	.241
	(0.067)		(0.059)		(0.059)		(.059)	
Model Fit	χ^2_7= 2,059.56**		$F_{(7, 3688.3)}$ = 287.70**		$F_{(7, 8827.5)}$ = 286.06**		$F_{(7, 18210.2)}$ = 285.15**	

Note: Significant at * $p < 0.05$; ** $p < 0.01$. Imputed Variable: MEDU (ordinal).

The odds ratios in this model could be interpreted as the same way as those in the conventional PO model without multiple imputation. The odds of being above a proficiency level increased by 1.765 with a one unit increase in mathematics identity and increased by 1.210 with a one unit increase in mathematics self-efficacy. The odds also increased by 1.049 and 1.203 with a one unit increase in school belonging and school engagement, respectively. In addition, being a mother with a Bachelor's degree and being a mother with a Master's degree and beyond rather than with a high school diploma or below increased the odds of being in higher mathematics proficiency levels (OR = 1.825 for BA and OR = 2.720 for MA, respectively). However, being a student with an individualized education plan decreased the odds of being above a mathematics proficiency level (OR = 0.241).

Comparison of the Parameter Estimates and Standard Errors from the PO Models with and without Multiple Imputation

Table 2 presents a comparison of the results of the PO models with and without multiple imputation. The parameter estimates and standard errors were different between these two models. After the PO model was fitted to the MI data, the estimated logit coefficients and their standard errors were different from those in the PO model with the listwise deletion of the missing data. Compared to those in the PO model without MI, the logit coefficients for mathematics identity (*MTHID*), mathematics self-efficacy (*MTHEFF*), and mother's education level (*MEDU*) in the PO model with MI decreased. However, the logit coefficients for the other three predictors, school belonging (*SCHBEL*), school engagement (*SCHENG*), and individualized education plan (*IEPFLAG*) increased. With regard to the changes of standard errors, all predictors except the dummy-coded variable mother's education level (*MEDU: BA*) decreased. Notably, the logit coefficient of school belonging (*SCHBEL*) changed from non-significance in the PO model without MI to statistical significance in the model with MI due to the change in parameter estimates and standard errors. In the PO model without MI, the logit coefficient of school belonging (*SCHBEL*), $\beta = 0.027$, $z = 1.01$, p=0.313; whereas in the PO model with MI, $\beta = 0.048$, $z = 2.10$, $p < 0.05$.

Comparison of the Parameter Estimates and Standard Errors from the PO Models with Different Numbers of Imputations (m = 5, 10, and 20)

When MI was implemented, the results obtained from different numbers of imputations (m = 5, 10, and 20) were compared. Table 2 displays the results across different numbers of imputations. The results of the parameter estimates and corresponding standard errors were consistent across these three numbers of imputations.

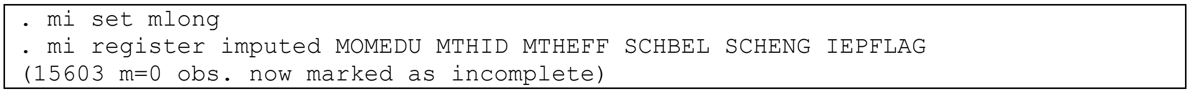
Proportional Odds (PO) Model with MICE for Multiple Variables with Different Types

The preceding PO model with MI focused on a single variable with missing data. To deal with multiple variables with missing data, the `mi impute chained` command was used for MI. By following the same steps as those for the MI for a single variable, we created the MI data and specified the six variables with missing data for imputation. Figure 3 displays the Stata syntax for creating the imputed data and specifying the six variables.

These variables included four continuous variables (i.e., *MTHID*, *MTHEFF*, *SCHBEL*, and *SCHENG*), an ordinal variable with three categories (i.e., *MEDU*), and one binary variable, *IEPFLAG*. Different imputation methods including regression, ordinal logistic regression, and binary logistic regression were specified in the multiple imputation for chained equations for continuous, ordinal, and binary variables, respectively. Figure 4 displays the STATA syntax for the `mi impute chained` command, the chained equations for different types of imputed variables, and the imputed number of observations for these six variables.

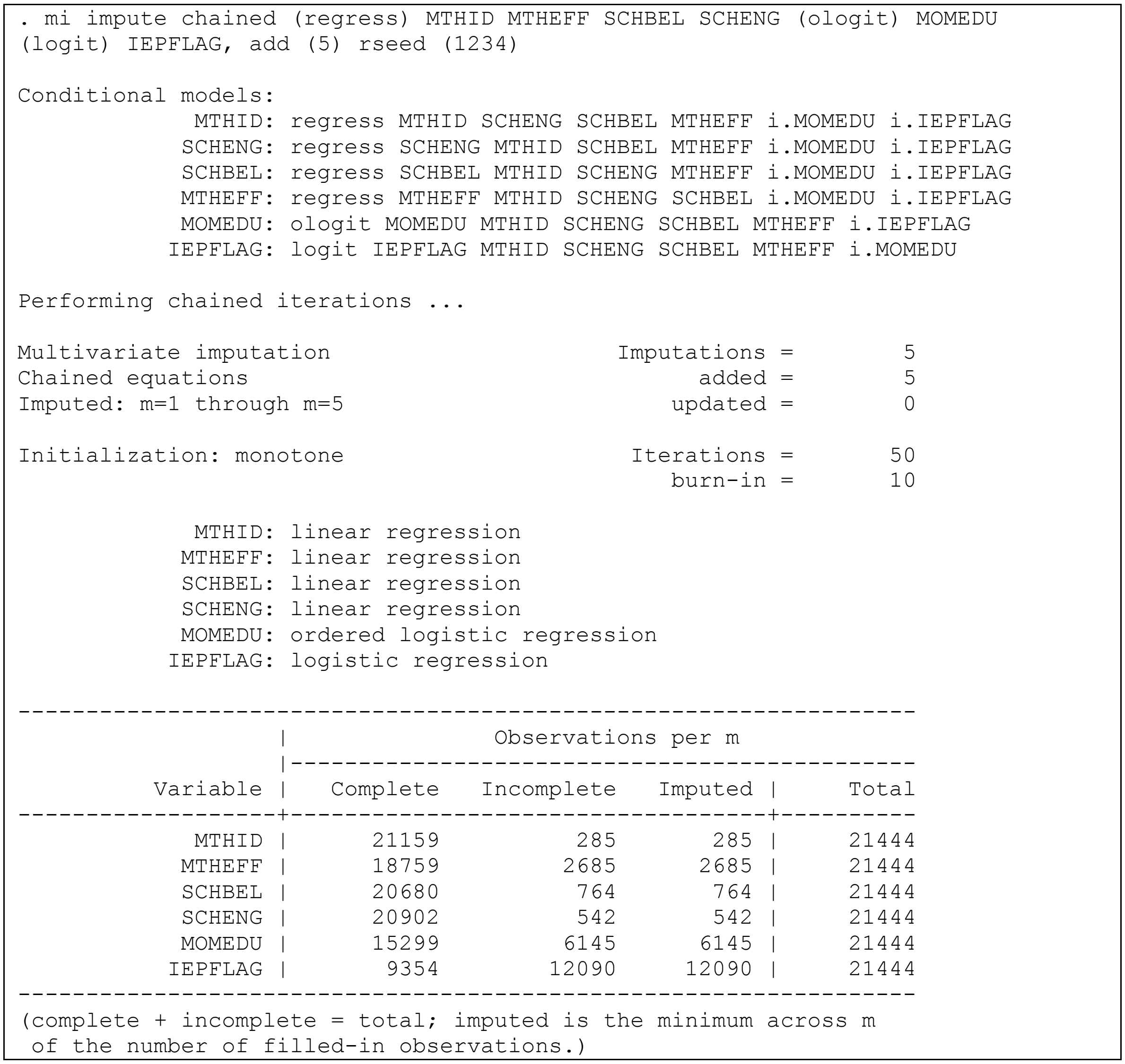
Table 3 presents the results of the PO model with MICE for multiple variables. The F statistic for the overall model, $F_{(7,\ 393.7)} = 517.62$, $p < 0.001$, which indicated that the model with all six predictor variables was significant in predicting the ordinal outcome variable, mathematics proficiency. The logit coefficients of all the predictors were significantly different from zero. All six predictors except *IEPFLAG* (individualized education plan) were positively associated with the logits or log odds of being in higher proficiency levels. However, the *IEPFLAG* predictor had a negative logit coefficient, β = -0.603, t = -11.95, $p < 0.001$.

Five predictors were associated with the odds of being above a proficiency level rather than being at or below that level. The odds of being above a proficiency level increased by a factor of 1.799 with a one-unit increase in mathematics identity and increased by a factor of 1.157 with a one-unit increase in mathematics self-efficacy. In addition, the odds increased by 1.067 and 1.197 with a one-unit increase in school belonging and school engagement, respectively. Furthermore, being a mother with a Bachelor's degree (OR = 1.696 for BA) and being a mother with a Master's degree and beyond (OR = 2.669 for MA) rather than with a high school diploma or below increased the odds of being in higher mathematics proficiency levels. However, being a student with an individualized education plan decreased the odds of being above a mathematics proficiency level (OR = 0.547). In other words, having an individualized education plan increased the odds of being in lower proficiency levels.



```
. mi set mlong
. mi register imputed MOMEDU MTHID MTHEFF SCHBEL SCHENG IEPFLAG
(15603 m=0 obs. now marked as incomplete)
```

Figure 3. Create a MI data and specify six imputed variables.



```
. mi impute chained (regress) MTHID MTHEFF SCHBEL SCHENG (ologit) MOMEDU
(logit) IEPFLAG, add (5) rseed (1234)

Conditional models:
             MTHID: regress MTHID SCHENG SCHBEL MTHEFF i.MOMEDU i.IEPFLAG
            SCHENG: regress SCHENG MTHID SCHBEL MTHEFF i.MOMEDU i.IEPFLAG
            SCHBEL: regress SCHBEL MTHID SCHENG MTHEFF i.MOMEDU i.IEPFLAG
            MTHEFF: regress MTHEFF MTHID SCHENG SCHBEL i.MOMEDU i.IEPFLAG
            MOMEDU: ologit MOMEDU MTHID SCHENG SCHBEL MTHEFF i.IEPFLAG
           IEPFLAG: logit IEPFLAG MTHID SCHENG SCHBEL MTHEFF i.MOMEDU

Performing chained iterations ...

Multivariate imputation                     Imputations =        5
Chained equations                                 added =        5
Imputed: m=1 through m=5                        updated =        0

Initialization: monotone                     Iterations =       50
                                                burn-in =       10

             MTHID: linear regression
            MTHEFF: linear regression
            SCHBEL: linear regression
            SCHENG: linear regression
            MOMEDU: ordered logistic regression
           IEPFLAG: logistic regression

------------------------------------------------------------------
                  |               Observations per m
                  |-----------------------------------------------
         Variable |   Complete   Incomplete   Imputed |     Total
------------------+-----------------------------------+----------
            MTHID |      21159          285       285 |     21444
           MTHEFF |      18759         2685      2685 |     21444
           SCHBEL |      20680          764       764 |     21444
           SCHENG |      20902          542       542 |     21444
           MOMEDU |      15299         6145      6145 |     21444
          IEPFLAG |       9354        12090     12090 |     21444
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Figure 4. Impute the six different types of variables using `mi impute chained`.

Comparison of the Results of the PO Models with and without MICE for multivariate variables

Table 3 provides the parameter estimates and the odds ratios (ORs) obtained from the PO models with and without MICE for multivariate variables. The logit coefficients and their standard errors in the PO model with MICE for missing data were different from those in the PO model with the listwise deletion of the missing data. Compared to those in the PO model without MICE, the logit coefficients of three predictors in the PO model with MICE decreased while those of the other three predictors increased. In addition, the standard errors of all six predictors decreased. Specifically, the logit coefficient for mathematics identity (*MTHID*) decreased by 2.7% and its standard error decreased by 48%; the logit coefficient for mathematics self-efficacy (*MTHEFF*) decreased by 33% and its standard error decreased by 48%; the logit coefficients for mother's education level (*MEDU: BA*) and mother's education level (*MEDU: MA*) decreased by 36% and 28%, respectively, and their corresponding standard errors decreased by 44% and 46%, respectively. However, the logit coefficients for the other three predictors, school belonging (*SCHBEL*), school engagement (*SCHENG*), and individualized education plan (*IEPFLAG*) increased. The logit coefficient for school belonging (*SCHBEL*) increased by 117%, but its

Table 3. Results of PO Models with MICE for Multiple Variables (Number of Imputations: 0, 5, 10, & 20)

	No Imputations: (n=0)		Imputations: (n = 5)		Imputations: (n = 10)		Imputations: (n = 20)	
Variables	b(se(b))	OR	b(se(b))	OR	b(se(b))	OR	b(se(b))	OR
α_1	-2.595		-2.231		-2.232		-2.227	
α_2	-0.810		-0.587		-0.586		-0.581	
α_3	0.542		0.659		0.660		0.665	
α_4	2.292		2.324		2.326		2.332	
α_5	4.203		4.177		4.179		4.185	
MTHID	0.603**	1.828	0.587**	1.799	0.586**	1.796	0.587**	1.798
	(0.031)		(0.016)		(0.016)		(0.016)	
MTHEFF	0.218**	1.243	0.146**	1.157	0.146**	1.158	0.146**	1.567
	(0.031)		(0.016)		(0.016)		(0.017)	
SCHBEL	0.027	1.028	0.065**	1.067	0.066**	1.068	0.067**	1.070
	(0.027)		(0.014)		(0.014)		(0.014)	
SCHENG	0.151**	1.163	0.180**	1.197	0.180**	1.198	0.182**	1.199
	(0.027)		(0.015)		(0.014)		(0.014)	
MEDU(BA)	0.824**	2.281	0.528**	1.696	0.533**	1.704	0.539**	1.714
	(0.052)		(.029)		(0.029)		(0.030)	
MEDU(MA)	1.360**	3.898	0.982**	2.669	0.983**	2.674	0.989**	2.689
	(0.079)		(0.043)		(0.049)		(0.049)	
IEPFLAG	-1.427**	0.240	-0.603**	.547	-0.610**	0.543	-0.601**	0.548
	(0.067)		(0.050)		(0.042)		(0.040)	
Model Fit	χ^2_7= 2,059.56**		$F_{(7,\ 393.7)}$ = 517.62**		$F_{(7,\ 1626)}$ = 552.73**		$F_{(7,\ 4525.2)}$ = 566.60**	

Note: Significant at * $p < 0.05$; ** $p < 0.01$. Imputed Variable: MTHID, MTHEFF, SCHBEL, SCHENG (continuous); MEDU (ordinal); IEPFLAG (binary).

standard error decreased by 48%; the logit coefficient for school engagement (*SCHENG*) increased by 19%, but its standard error decreased by 44%; the logit coefficient for individualized education plan (*IEPFLAG* increased by 58%, but its standard error decreased by 25%. It was also found that the logit coefficient of school belonging (*SCHBEL*) was not significant in the PO model with listwise deletion for missing data but was significant in the model with MICE. This change was impacted by the tremendous change in the parameter estimate and standard error between these two models.

Comparison of the Results of the PO Models with MICE for Multiple Variables with Different Numbers of Imputations (m= 5, 10, and 20)

Table 3 displays the results of the PO models with MICE across different numbers of imputations. When the number of imputations for MICE increased from 5 to 10 and 20, the results of the parameter estimates and corresponding standard errors were identical or similar across these three numbers of imputations.

Conclusions

This article illustrated the use of Stata to fit the PO models with missing data which are MAR and how the MI and MICE approaches can be applied to deal with missing values of a single variable and multiple variables with different types, respectively. Models were fitted from the PO with listwise deletion for missing data to the PO model with MI for a single ordinal variable and finally to the PO model with MICE for multiple variables. The results of all fitted models were interpreted. In addition, the results from the PO model with MI for a single variable and the PO model with MICE for multiple variables were compared with those from the PO model with listwise deletion. Further, the results of the PO models with different numbers of imputations in MI and MICE were compared.

The current study demonstrated that the logit coefficients in the PO models with MI and MICE can be interpreted in the same way as that in the PO model with listwise deletion. However, researchers should be cautious with the change in parameter estimates, standard errors, and statistical significance in these models. To obtain unbiased parameter estimates and standard errors, the study results suggested that appropriate imputation techniques should be used for unbiased estimations and standard errors than simply deleting the missing data with listwise deletion.

It is worth noting that although the results of the PO models with different numbers of imputations in MI and MICE were consistent in this study, further research needs to be conducted to confirm if this finding holds when there are high percentage of missing data.

Although Markov Chain Monte Carlo (MCMC) is a popular method for multiple imputation, it assumes that data are multivariate normal. It is commonly used to impute continuous variables with normally distribution. This study focused on ordinal categorical missing data and the data with mixed types of continuous, binary, and ordinal variables, so it was more appropriate to use the MICE approach. Future research will be conducted on how the MCMC approach can be extended to categorical and non-normal data.

Missing data are pervasive in educational research. This study provides empirical evidence for researchers to conduct ordinal regression analysis with missing data that are MAR. It contributes to the fields of both ordinal logistic regression and missing data analysis by making close connections between the two research topics. With the available software package for missing data analysis, this study will be useful to help educational researchers apply the method to their research.

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: John Wiley & Sons.

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: John Wiley & Sons.

Aitkin, M., & Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67-75.

Allison, P. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology, 26*, 1323-1333.

Armstrong, B. B., & Sloan, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology, 129*(1), 191-204.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 84*(4), 487–508.

Enders, C. (2010). *Applied missing data analysis*. New York: Gilford Press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Graham, J. W. (2012). *Missing data: Analysis and design*. New York: Springer.

Hilbe, J. M. (2009). *Logistic regression models*. Boca Raton, FL: Chapman & Hall/CRC.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician*, *61*, 79–90.

Lawrence, C. E., & Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Bioinformatics*, *7*(1), 41-51.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Liu, X. (2009). Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS. *Journal of Modern Applied Statistical Methods, 8(2), 632-645.*

Liu, X. (2016). *Applied ordinal logistic regression using Stata: From single-level to multilevel modeling*. Thousand Oaks, CA: Sage.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed.). College Station, TX: Stata Press.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Ser. B*, 42, 109-142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

McKnight, P., McKnight, K., Sidani, S., & Figueredo, A. (2007). *Missing data: A gentle introduction*. New York: Gilford Press.

O'Connell, A. A., (2000). Methods for modeling ordinal outcome variables. *Measurement and Evaluation in Counseling and Development*, *33*(3), 170-193.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage.

O'Connell, A.A., & Liu, X. (2011). Model diagnostics for proportional and partial proportional odds models. *Journal of Modern Applied Statistical Methods*, *10*(1), 139-175.

Peng, C., Harwell, M., Liou, S., & Ehman, L. (2006). Advanced in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real data analysis* (pp. 31-78). Charlotte, NC: New Information Age.

Peugh, J., & Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74,* 525-556.

Powers D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. San Diego: Academic Press.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software, 45*(4), 1-20.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3-15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Send correspondence to:

Xing Liu
Eastern Connecticut State University
Email: liux@easternct.edu