Approaches for Addressing Issues of Missing Data in the Statistical Modeling of Adolescent Fertility

Dudley L. Poston, Jr.
Texas A&M University
&
Eugenia Conde
Rutgers University

Missing Data

- Missing data are a pervasive challenge in scientific research.

- Missing data can threaten the validity of the inferences that researchers draw from their findings because it has potential of affecting three key components of scientific research:

    - Construct validity
    - Internal validity
    - External validity

Paul Allison (*Missing Data*, 2002, p. 1)

Allison (2002, p. 1) writes that “sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data. In a typical [secondary] dataset, information is missing for some variables for some cases. ... Missing data are a ubiquitous problem in both the social and health sciences...[Yet] the vast majority of statistical textbooks have nothing whatsoever to say about missing data or how to deal with it.”

Donald Treiman (*Quantitative Data Analysis*, 2009, p. 182)

Treiman writes that "missing data is a vexing problem in social research. It is both common and difficult to manage. Most survey items include nonresponse categories: respondents do not know the answers to some questions or refuse to answer; interviewers inadvertently skip questions or record invalid codes; errors are made in keying data; and so on. Administrative data, hospital records, and other sorts of data have similar problems, namely, invalid or missing responses to particular items."

Sources of Missing Data

Missing cases

- Participants fail to show up for the interview

Missing variables (most common source in demographic analyses)

- Participants do not answer all the questions in the interview

Missing occasions

- In longitudinal studies participants do not complete all the stages of the study

Rubin's Mechanisms of Missing Data

In 1976 Donald Rubin introduced three reasons or mechanisms for why data are missing:

- **Missing completely at Random (MCAR)**
- **Missing at Random (MAR)**
- **Missing not at Random (MNAR)**

Rubin's classification is concerned with the relationship between the variables and the probability of missing data.

Missing Completely at Random

(MCAR)

The probability of the missing data does not depend on the variable with missing data or on any other of the variables in the model.

If this condition is met for all the variables with missing values, the data are considered to be a subsample of the original sample.

Missing at Random

(MAR)

MAR refers to “the condition in which missingness is independent of the true value of the variable in question but not of at least some of the other variables in the explanatory model” (Treiman, 2009, p. 182).

The missing values thus depend on other variables in the model but not on the variable with missing data.

For example, given three independent variables of age, marital status and income; say that income is missing for 15% of the respondents.

The data would be considered MAR if the probability that income is missing is related to age and/or to marital status but not to income; that is, missing data on income would not depend on, say, whether a respondent has low or high income.

Missing Not at Random

(MNAR)

The data are considered MNAR when the MAR assumption is violated.

The data would be MNAR if the probability that the values were missing depended on the variable itself.

In the previous example, the data would be MNAR if the missingness of income depended on whether the respondent had a high or a low income.

Five Traditional Methods of Handling Missing Data

1. Listwise deletion (LWD)

- Drops the cases with missing values.
- Considered a conservative approach if data are MCAR, i.e., standard errors will be larger because the sample size will be smaller.
- If the missing data are MAR and LWD is used, the estimates will likely be biased.
- LWD is the default method in most statistical packages e.g., Stata.

2. Mean Substitution (MS)

MS is a very simple approach. The missing values for a variable are replaced with the mean value for that variable. One reason why MS is inappropriate is because subjects who do not answer a question on a variable often tend to be at the extreme ends of the distribution and not in the middle, and should thus not be assigned the average score of the variable.

MS is problematic when the percentage of missing values is large because this greatly reduces the variance and runs the risk of underestimating the correlation between the variable with missing values and any of the other variables in the model.

Enders (*Applied Missing Data Analysis,* 2010, p. 43) writes that MS "is possibly the worst missing data handling method available. Consequently, in no situation is [it] defensible, and you should absolutely avoid this approach."

3: Mean Substitution for Subgroups (MSS)

MSS is a modification of MS.

It assigns the mean values for the subgroups in the analysis. E.g., one might handle missing data on a variable such as income for the males and females in the sample by assigning to the males with missing data on income the average value for males, and to the females the average value for females.

MSS will usually not reduce the variance in the variable with the missing data as much as MS will, so it is thus considered to be only slightly better than MS.

Should be avoided as much as MS is avoided.

4. Proxy Method (PM)

PM involves substituting for the variable with a lot of missing data another variable with little or no missing data, which variable is related substantively and statistically to the variable with the missing data.

E.g., to address the situation of an excessive amount of missing data on a variable such as income, some researchers have dropped the income variable and used a variable such as educational attainment as a proxy for income.

PM is at best a substitute approach.

PM is problematic because it could lead to model misspecification.

5. Dropping the Variable(s) with Missing Data (DRP)

Some research uses DRP, i.e., the variable(s) with excessive amounts of missing data are dropped from the regression equation.
E.g., consider a dataset with, say, 20 percent of the respondents not responding to a question on personal income. If the researcher were to retain the income variable in the equation and use LWD, then the analysis would be conducted with 20 percent fewer cases. If DRP was used, the income variable would not be in the equation, and the analysis would retain those 20 percent of the respondents not reporting incomes.
DRP should be avoided without question because of the obvious problem of model misspecification.

Four More Traditional Methods (I will only mention them)

1. **Pairwise deletion**
    - uses all the available information to compute the summary statistics; not a good strategy; can’t be used in estimating multivariate equations; should be avoided
2. **Dummy variable adjustment**
    - uses all the cases and adjusts for those that have missing values; artificially influences the size and complexity of the model.
3. **Hot deck**
    - missing values are replaced with random values found in the observed data; used with census and ACS data
4. **Cold deck imputation**
    - replacing values with values from another data set

Multiple imputation (MI)

The most popular of the non-traditional methods is multiple imputation (MI), a method first introduced by Donald Rubin in 1987.

Allison (2002, p. 27) argues that MI is the preferred method for handle missing data because "when used correctly, it produces estimates that are consistent, asymptotically efficient and asymptotically normal when the data are MAR."

Treiman (2009, pp. 186-87) states that MI is the current gold-standard approach for dealing with missing data.

Unique about MI is that it does not treat the data as if they are real.

Instead MI estimates the values by taking into account the uncertainty of the missing values component.

Multiple datasets are generated. But MI is not concerned with recovering the missing data.

Concerned with estimating the population variances to make generalizable estimates.

Three stages in MI

The imputation stage creates several data sets (see next slide).

The analysis stage runs the desired analysis in each data set.

The combination stage combines the results (i.e., the estimates and the standard errors from each of the data sets) using rules developed by its creator Donald Rubin; these are known as "Rubin's Rules."

In the imputation stage, auxiliary variables may (or may not) be used to impute the missing values. Auxiliary variables are used that are statistically related to the variables with missing values. They enhance the effectiveness of the imputation stage in the MI process. The auxiliary variables are not used in the regression equation per se, but are used to provide more information about the variances of the independent variables with the missing data. For this reason, some authors (Allison, 2002; Treiman, 2009) hold that the preferred MI equation is the one that uses auxiliary variables.

Two main MI iterative methods

1. The fully conditional specification (FCS) method is also known as imputation by chain equation (ICE); it imputes continuous and categorical variables without assuming a multivariate normal distribution. It is sometimes criticized because it is said to lack theoretical statistical soundness. However, simulation studies have shown that it works reasonably well, and its results are comparable to the Markov chain Monte Carlo method (see below).

2. The Markov chain Monte Carlo (MCMC) method is an iterative procedure that assumes a multivariate normal distribution of all the variables in the model. Hence, it works best when imputing continuous variables. However, it has been shown that it can also be used to impute categorical variables.

We use three MI methods in our analysis of adolescent pregnancy; they are our 6th through 8th missing data methods (the first five are the traditional methods described earlier).

Our three MI methods are:

6. MI using the fully conditional specification (FCS) method; (MI-1)

7. MI using the Markov chain Monte Carlo (MCMC) method with four auxiliary variables (MI-2) (the auxiliary variables are shown on a later slide)

8. MI using the Markov chain Monte Carlo (MCMC) method but only imputing education and income (MI-3)

Of the three MI methods we use in this analysis, MI with auxiliary variables, i.e., MI-2, will be our preferred method.

Data

The National Longitudinal Study of Adolescent Health (Add Health)

Nationally representative sample of adolescents from the 7th to the 12th grades.
Sampled from 80 high schools and 52 middle schools
The survey includes variables on individuals, families, schools and communities.

Three waves

First wave collected in 1994-1995 (20,000 adolescents)
Second wave collected in 1996 (15,000 students)
Third wave collected in 2001-2002 (15,197 adolescents)

Our Models

Eight logistic regression models were estimated predicting the log odds of a female having an adolescent birth, each differing on which missing data method was used. The methods were 1. listwise deletion (LWD), 2. mean substitution (MS), 3. mean substitution for subgroups (MSS), 4. the proxy method (PM), 5. the dropping variables method (DRP), and three varieties of multiple imputation (MI-1, MI-2, MI-3).

The complex survey design of the Add Health data was taken into consideration via our use of Stata's "svy" suite of commands for logistic regression .

We estimated eight logistic regression equations, using eight different Methods of Handling Missing Data

1. **Listwise Deletion (LWD)**
2. **Mean Substitution (MS)**
3. **Mean Substitution for Subgroups (MSS); we assigned the means for the missing data for each of the race/ethnic groups**
4. **Proxy Method (PM) using mother's education as a proxy for income**
5. **Dropping Variables with Excessive Missing Data (DRP); we dropped parental education and income.**
6. **Multiple Imputation (fully conditional specification); we imputed all variables (M-1)**
7. **Multiple Imputation (Markov chain Monte Carlo) we imputed all variables and we used four auxiliary variables (M-2)**
8. **Multiple Imputation (Markov chain Monte Carlo); we imputed only the two variables with the most missing data, namely household income and parent education (M-3)**

We used four auxiliary variables. Two questions were asked of the parents, namely, “How important is religion to you?” and “Do you have enough money to pay your bills.” And two questions were asked of the students, namely, “Since school started this year, how often do you have trouble getting along with your teachers?” and “How much do you want to go to college?” All four auxiliary questions were answered on a 1-4 or a 1-5 point scale from low to high. These are related to the two variables in our model with the most missing data (family income and parental education).

Dependent variable:

- 1 = had a teen birth
- 0 = did not have a teen birth

Independent Variables:

1. Virginity pledge (yes = 1)
2. Race and ethnicity (dummy variables):
   African American, Mexican-origin, Other Latina, Other race, and non-Hispanic whites (reference group)
3. Religion (dummy variables):
   None, Protestant, Evangelical, Black Protestant, Jewish, Other Religion, and Catholic (reference group)
4. Household income (continuous, in ’000s of dollars)
5. Parent’s education (in years)
6. Importance of Religion (4 categories; 1= not imp to 4= very imp).
7. Likelihood of attending college
   (1 to 5 scale, 1= not likely to 5=most likely)

Descriptive Data: 6,719 Adolescent Females, The National Longitudinal Study of Adolescent Health, Waves 1 and 3

Variable	Cases	Percent missing	Mean	SD
Dependent Variable				
Teen pregnancy	6,710	0.24	0.18	0.38
Seven Independent Variables				
1. Virginity pledge	6,644	1.22	0.15	0.36
2. Race / Ethnicity	6,719	0.10		
White	3,568		0.67	0.47
African American	1,510		0.17	0.37
Mexican	539		0.06	0.24
Other Latinas	538		0.05	0.23
Other	564		0.05	0.21
3. Religion	6,620	1.60		
Catholic	1,757		0.24	0.43
None	744		0.12	0.32
Protestant	1,447		0.22	0.42
Evangelical	1,056		0.20	0.40
Black Protestant	884		0.11	0.31
Other	682		0.11	0.31
Jewish	50		0.01	0.09
4. Household Income (in thousands)	4,983	26.00	$42.7	$27.0
5. Parental Education (in years)	5,708	15.14	13.27	2.45
6. Religious importance	6,717	0.13	3.12	0.93
7. Likelihood, college	6,681	0.67	4.25	1.13

Statistical Significance of the Variables Predicting Adolescent Pregnancy: Eight Equations Using Different Methods to Handle Missing Data

Independent Variable	LWD	MS	MSS	PM	DRP	MI-1	MI-2	MI-3
Virginity Pledge	**	*	*	**	*	*	*	*
Race/ethnicity (White is reference)								
African-American	†	*	*	***	***	ns	ns	*
Mexican-Origin	*	*	†	*	**	ns	ns	†
Other Latina	ns	†	*	*	**	ns	ns	ns
Other	ns	ns	ns	ns	ns	ns	ns	ns
Religion (Catholic is reference)								
None	ns	ns	ns	ns	ns	ns	ns	ns
Protestant	ns	ns	ns	ns	ns	ns	ns	ns
Evangelical	ns	*	*	**	**	*	*	*
Black Protestant	***	***	***	***	***	***	***	***
Jewish	ns	ns	ns	ns	ns	ns	ns	ns
Other Religion	ns	ns	ns	ns	ns	ns	ns	ns
Household Income	***	***	***	---	---	***	***	***
Parental Education	ns	ns	ns	**	---	ns	ns	ns
Religious Importance	†	*	*	†	*	*	†	*
Likelihood to attend college	***	***	***	***	***	***	***	***

†p<0.05 (one tail);*p<0.05 (two tail); **p<0.01 (two tail);***p<.001 (two tail); ns = not significant

We then calculated semi-standardized logit coefficients for all the X variables that were statistically significant in each of the eight models. We then rank ordered them.

Logit coefficients that have been standardized in terms of the variances of their independent variables are simply the logit coefficients multiplied by their standard deviations.

The semi-standardized logit coefficient for the i^{th} X variable is $\mathbf{b^*(x)_i}$

$$\mathbf{b^*(x)_i = b_i * s_i}$$

Ranks of the Statistically Significant Semi-standardized Logit Coefficients Predicting Adolescent Pregnancy: Eight Equations Using Different Methods to Handle Missing Data

Independent Variable	LWD	MS	MSS	PM	DRP	MI-1	**MI-2**	MI-3
Virginity Pledge	4	8	7	5	8	5	**5**	7
Race/ethnicity (White is reference)								
African-American	6	6	6	3	3	--	**--**	6
Mexican-Origin	5	7	8	7	5	--	**--**	8
Other Latina	--	9	--	9	6	--	**--**	--
Other	--	--	--	--	--	--	**--**	--
Religion (Catholic is reference)								
None	--	--	--	--	--	--	**--**	--
Protestant	--	--	--	--	--	--	**--**	--
Evangelical	--	4	4	4	4	4	**4**	5
Black Protestant	2	1	1	1	2	2	**3**	2
Jewish	--	--	--	--	--	--	**--**	--
Other Religion	--	--	--	--	--	--	**--**	--
Household Income	1	2	2			1	**1**	1
Parental Education	--	--	--	6		--	**--**	--
Religious Importance	7	5	5	8	7	6	**6**	4
Likelihood to attend college	3	3	3	2	1	3	**2**	3

Conclusions

Depending on the method used, our results indicate that many of the independent variables in our model vary in whether they are, or are not, statistically significant in predicting the log odds of a woman having a teen birth; and many of the independent variables that are statistically significant vary in the ranking of the magnitude of their relative effects on the outcome. Our results show that the levels of significance of the effects, the size of the effects, and their relative importance vary considerably depending on the method used to handle the missing data.

Missing data is a critical component of scientific research. We have shown that different techniques will lead to different statistical results.

What’s the best solution if you have missing data?

Paul Allison (2002, p. 2) states that “the only good solution to missing data is not to have any.”

But we almost always have missing data.

So, what should we do?

We propose that it is reasonable to ask researchers who are conducting analyses with lots of missing data to report the results of both LWD and MI; try different methods of MI, i.e., with auxiliary variables and without them, to determine the level of consistency of the findings.

Analyses with strong theories and consistent results across different methods of handling missing data should not be problematic.

But when the findings are inconsistent, that is, they vary depending on how missing data are handled, and also when there is no strong theory, then the results should be rendered as inconclusive.

Finally, we note that the effect of missing data on scientific research requires more scrutiny.

We suggest that journal editors should require their authors to report precisely the amount of missing data in each of their variables, as well as to specify and justify the method they used to handle missing data.

We specifically recommend that researchers with more than small amounts of missing data should estimate their models with both LWD and with MI (with and without auxiliary variables) and report if there are any differences that would lead to different theoretical or empirical conclusions. Research conducted with large amounts of missing data should be scrutinized with great deliberation and forethought, and the findings, if inconsistent across method, should be interpreted with caution.

END of PRESENTATION

TEXAS A&M UNIVERSITY