DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 27

11-2014

Missing Data and the Statistical Modeling of Adolescent Pregnancy

Dudley L. Poston Dr.
Texas A&M University, d-poston@tamu.edu

Eugenia Conde Dr.
Rutgers University

Recommended Citation

Poston, Dudley L. Dr. and Conde, Eugenia Dr. (2014) "Missing Data and the Statistical Modeling of Adolescent Pregnancy," *Journal of Modern Applied Statistical Methods*: Vol. 13 : Iss. 2 , Article 27.
DOI: 10.22237/jmasm/1414815960
Available at: http://digitalcommons.wayne.edu/jmasm/vol13/iss2/27

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 464-478.

ISSN 1538 − 9472

Missing Data and the Statistical Modeling of Adolescent Pregnancy

Dudley L. Poston
Texas A&M University
College Station, TX

Eugenia Conde
Rutgers University
New Brunswick, NJ

Missing data is a pervasive problem in social science research. Many techniques have been developed to handle the problem. Different ways of handling missing data were shown to lead to different results in statistical models. A demonstration was given based on statistical modeling of the likelihood of a woman reporting having had an adolescent pregnancy by handling missing data with several different approaches. Results indicate that many of the independent variables in the model vary in whether they are, or are not, statistically significant in predicting the log odds of a woman having a teen pregnancy, and in the ranking of the magnitude of their relative effects on the outcome.

Keywords: Missing data, listwise deletion, mean substitution, multiple imputation, proxy variables, adolescent pregnancy, race/ethnicity, logistic regression, logit coefficients, semi-standardized logit coefficients, demography

Introduction

Missing data is a pervasive problem in social science research. “Sooner or later, usually sooner, anyone who does statistical analysis runs into problems with missing data” (Allison, 2001: 1). Many techniques have been developed to handle missing data; often, the results of a statistical model will differ depending on the technique used.

Missing Data Mechanisms

According to Rubin (1976; 1987), there are three missing data mechanisms; the data are either “missing completely at random” (MCAR), “missing at random” (MAR) or “missing not at random” (MNAR). Missing data are said to be missing completely at random (MCAR) when the probability of the missing data for a

Dr. Poston is Professor of Sociology. Email him at d-poston@tamu.edu. Eugenia Conde is a Postdoctoral Fellow.

variable does not depend on the variable itself or on any of the other independent variables in the model. MCAR refers to the "condition in which missing responses to a particular variable are independent of the values of any other variable in the explanatory model and of the true value of the variable in question" Treiman (2009, p. 182). If all the missing data are MCAR, this is usually not a serious problem because the remaining data are considered to be a subsample of the original sample.

Missing data are considered to be missing at random (MAR) if the probability of the missing data does not depend on the values of variables with the missing data, after controlling for other variables in the model. That is, 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).

Missing data are considered to be missing not at random (MNAR) when the MAR assumption is violated. The data are MNAR if the probability that the values were missing depends on the variable itself.

Methods for Handling Missing Data

There are many methods for handling missing data. We discuss several of the more popular approaches and then use each separately in an analysis of adolescent pregnancy.

1. *Listwise Deletion* The method that is the default method in most statistical packages is listwise deletion, also known as case deletion. It drops the missing values from the data set, and the analysis is then conducted using the reduced sample. If the data are MCAR, the resulting smaller sample is considered to be an unbiased subsample of the original dataset (Allison, 2001), and the use of listwise deletion should result in models with unbiased estimates. However, the standard errors will be slightly larger because the sample size is now, obviously, smaller. Statistical power will be reduced and the probability of finding significant results decreased; thus the listwise deletion method is often viewed as conservative provided that the MCAR assumption has been met (Acock, 2005). But if the missing data are MAR and listwise deletion is used, then the estimates will likely be biased (Allison, 2001).

2. *Mean Substitution* Mean substitution is a very simple approach. The missing values for a variable are replaced with the mean value for that variable. Mean substitution is especially problematic when the percentage of missing values is large because this greatly reduces the variance and hence underestimates the correlation between the variable with missing values and any of the other variables in the model (Acock, 2005; Allison, 2001). Mean substitution "is possibly the worst missing data handling method available" Enders (2010, p. 43).

3. ***Mean Substitution for Subgroups*** A modification of mean substitution assigns the mean values for subgroups of the analysis. For example, a researcher might handle missing data on a variable such as income for the males and females in the sample by assigning to the males the average value of income for males, and to the females the average value of income for females. Although this modification reduces the variance, it is considered to be only slightly better than substituting with the overall mean (Acock, 2005).

4. ***Proxy Method*** When confronted with an excessive amount of missing data on an independent variable, some have used the proxy method as a solution. That is, they have substituted for the variable with the missing data another variable with little or no missing data that is related substantively and statistically to the variable with the missing data. For example, to address the situation of an excessive amount of missing data on a variable such as income, one could use educational attainment as a proxy for income.

5. ***Dropping the Variable(s) with Missing Data*** This approach simply drops from the analysis the variable (or variables) with excessive amounts of missing. It should be avoided without question because of the obvious problem of model misspecification.

The above are five of the "traditional" methods used for handling missing data. With the exception of listwise deletion when the data are MCAR, all five are problematic. For one thing, they will often produce biased estimates and inefficient standard errors. And when listwise deletion is used with MAR data, the estimates will be biased and the standard errors inefficient.

(Other traditional methods not used in this paper include dummy variable adjustment and hot and cold deck imputation. Dummy variable adjustment uses all the cases and adjusts for those that have missing values by adding a dummy variable scored 1 if the value for the variable is missing, and 0 if not missing. Hot deck imputation also uses all the cases but replaces the missing values with random values found in the observed data. Cold deck imputation is similar but replaces the missing values with those from another data set. These methods may seem to be appealing because they use all the cases, but they have been shown to produce biased estimates irrespective of whether or not the data are MCAR, MAR or MNAR (Acock, 2005; Allison, 2001).

6-8. ***Multiple Imputation (MI) - three versions*** The most popular of the non-traditional methods is multiple imputation (MI), a method first introduced by Rubin in 1987. There are several variations of MI.

It has been argued 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" (e.g., Allison, 2001, p. 27). MI has become the gold-standard approach for dealing with missing data (Treiman, 2009, p. 186-186).

Multiple imputation is not concerned with recovering the missing data like the traditional methods mentioned above. Instead, it is concerned with estimating the population variances so as to produce generalizable estimates (Acock, 2005; Allison, 2001; Enders, 2010; Rubin, 1987). Unique about this method is that it does not treat the data as if "they were real" (Allison, 2001). Instead MI estimates the values by taking into account the uncertainty of the missing values. MI recognizes that even if the missing values are imputed, there is still uncertainty in those values, so it adjusts the variances to take this into account.

MI has three steps: imputation, analysis, and the combination of datasets. The imputation stage creates several data sets; the analysis stage runs the desired analysis in each data set; and the combination stage combines the results from the imputations using rules developed by its creator, Donald Rubin.

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, so to enhance the effectiveness of the imputation stage. The auxiliary variables are not used as independent variables in the regression equation per se, but are used to provide more information about the variances of the independent variables with the missing data. A preferred MI equation is usually one that uses auxiliary variables (Allison, 2001; Treiman, 2009).

The two main MI iterative methods for handling missing data are the fully conditional specification (FCS) method, and the Markov chain Monte Carlo (MCMC) method. The fully conditional specification (FCS) method is sometimes known as imputation by chain equation (ICE); it imputes continuous and categorical variables without assuming a multivariate normal distribution. Simulation studies have shown that it works reasonably well, and the results are comparable to the MCMC method (Lee & Carlin, 2010).

The Markov chain Monte Carlo (MCMC) method is an iterative procedure that assumes a multivariate normal distribution of all the variables in the model. It works best when imputing continuous variables (Schafer, 1997), but it can also be used to impute categorical variables (Allison, 2001; Lee and Carlin, 2010).

Following the above discussion, we will use three MI methods in our analysis of adolescent pregnancy, as follows: 6. MI using the fully conditional specification (FCS) method; 7. MI using the Markov chain Monte Carlo (MCMC) method with auxiliary variables; and 8. MI using the Markov chain Monte Carlo (MCMC) method but only imputing education and income.

Thus, eight models of adolescent pregnancy will be estimated, with missing data handled differently in each of the eight models.

Data and Method

Data were taken from the National Longitudinal Study of Adolescent Health (Add Health) (Harris, 2008), a nationally representative stratified sample of adolescents in the 7th through the 12th grades who were followed across four waves between 1994 and 2008. The sample was collected from 80 high schools and 52 middle schools and junior high schools across the United States. The first wave of data was collected in 1994-1995, the second in 1996, the third in 2001-2002, and the fourth in 2007-2008. Data on the parents of the school children were collected in the first wave. We use data from wave I and wave III for the female students and their parents.

Logistic regression is used to estimate the log odds of females who had a pregnancy when they were between the ages of 15-19. Seven theoretically relevant independent variables were selected, as follows: (1) a dummy variable from wave 1 regarding whether or not the adolescent ever made a pledge to remain a virgin until marriage, scored 1 if yes and 0 if no; (2) the adolescent’s race/ethnicity measured with a series of dummy variables (African American, non-Hispanic white, Mexican-origin, other Latina; other race; and non-Hispanic white, which was used as the reference); (3) the adolescent’s religion measured with six dummy variables (no religion, Protestant, Evangelical Protestant, Black Protestant, other religion, and Catholic; the Catholic dummy was used as the reference group); (4) household income as reported by the parent in wave 1 (measured in thousands) with $100,000 as the ceiling; (5) parental education as reported by the parent in wave 1 and measured as number of years of school completed; (6) the importance of religion to the adolescent (“How important is religion to you?”), ranging from a value of 1 if the woman reported no religious affiliation or responded “not important at all” to a value of 4 if she reported “very important”; and (7) the adolescent’s perceived likelihood to attend college, with 1 as the lowest category and 5 as the highest. All these independent variables have been previously shown to be influential in models predicting whether or not a woman had a teen pregnancy (see, e.g., Bean and Swicegood 1985; Klepinger et al., 1995; Rosenbaum, 2006).

Results

Table 1. Descriptive Data: 6,719 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 Latina	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	42.70	27.00
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 of college	6,681	0.67	4.25	1.13

Table 1 presents descriptive data on the dependent variable and the independent variables for the 6,719 females of age 20 years or higher in our sample. We show in the first data column the number of women for whom we have data for each variable. The maximum number of cases is 6,719. In column 2 we show the percentage of the cases with data missing for each variable. Of the nine variables we use in the logit regression equations (the dependent variable and eight independent variables), only three have missing data percentages of more than one percent: household income, 26.0 percent; parental education, 15.1 percent; and religion 1.6 percent. With more than one quarter of the cases having missing data on income, this means we would lose at least this percentage of respondents from the analysis were we to rely on listwise deletion as the method for handling missing data.

In the third data column of Table 1, note that 18 percent of the women in the sample reported having had a teen pregnancy, 15 percent reported having made a pledge while a teenager to remain a virgin until marriage. Almost 67 percent of the respondents were white, and their mean household income was over $42.7 thousand. Religion was fairly to very important for most of the respondents, and most of them believed it is very likely that they will attend college.

These data were analyzed using the eight different approaches discussed above for handling missing data:

1. Listwise deletion

2. Overall mean substitution

3. Mean substitution where the mean values were substituted on the basis of the race and ethnic groups of the women

4. The proxy method where mother's education was used as a proxy for income

5. Dropping the variables with excessive amounts of missing data; parental education and household income, the two variables with the most missing data, were excluded from the equation

6. Multiple imputation in which we imputed all the variables with missing data using the fully conditional specification iterative method

7. Multiple imputation using the Markov chain Monte Carlo iterative method with four auxiliary variables (via 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)

8. Multiple imputation using the Markov chain Monte Carlo iterative method to impute only the two variables with the most missing data, namely household income and parent education. In each of these three MI applications, a total of 100 imputations were undertaken. The 16 cases (only 0.2 percent of all the respondents) that were missing in the teen pregnancy dependent variable were imputed in the imputation stage, but they were dropped from the analysis (von Hippel, 2007).

Because the Add Health Survey is based on multistage probability sampling, one cannot make inferences with these data to the larger population of U.S. women from which the sample was drawn without first taking into account the sampling design. Thus, the "svy" suite of statistical sample adjustment methods available in the Stata 12 statistical package (StataCorp, 2011) was used to introduce survey adjustment estimators.

The results from eight logistic regressions modelling the log likelihood of a woman becoming pregnant while a teenager are compiled in Table 2. Each regression equation handles missing data in a different way, as discussed earlier. The preferred method for handling missing data is multiple imputation using auxiliary variables, shown as model 7 (M7) in the table.

The values in the first line for each variable in Table 2 are the logistic regression coefficients predicting the log odds of a woman having an adolescent pregnancy; if the coefficient is statistically significant, it is asterisked (see legend at the bottom of the table). Immediately below the logit coefficient is its semi-standardized coefficient; this is the logit coefficient that has been standardized in terms of the variance of the independent variable, that is, the logit coefficient has been multiplied by its standard deviation (Long & Freese, 2006, p. 96-98). Alongside each of the semi-standardized coefficients that is statistically significant, in parentheses, is shown the ranking in that equation of its relative effect on the outcome of teen pregnancy.

Table 2. Eight Logistic Regression Models of Teen Pregnancy According to the Method Used to Handle Missing Data: Females Surveyed in The National Longitudinal Study of Adolescent Health, Waves 1 and 3

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
1. Virg-Pledge	.455**	-.323*	-.322*	-.420**	-.307*	-.331*	-.328*	-.327*
	-.164(4)	-.117(8)	-.117(7)	.152(5)	-.112(8)	-.119(5)	-.118(5)	-.118(7)
2. Race/ethnicity								
White	Ref	ref	ref	ref	ref	ref	ref	Ref
African American	.351†	.369*	.340*	.507***	.485***	.232	.249	.343*
	.125(6)	.137(6)	.126(6)	.184(3)	.180(3)	.082	.088	.122(6)
Mexican-origin	.602*	.535*	.493†	.591*	.691**	.394	.401	.482†
	.135(5)	.127(7)	.116(8)	.137(7)	.163(5)	.088	.090	.108(8)
Other Latina	.245	.325†	.296	.360*	.462**	.247	.253	.295
	.056	.073(9)	.066	.083(9)	.104(6)	.057	.058	.068
Other	-.035	-.145	-.141	.072	-.081	-.174	-.170	-.157
	-.007	-.031	-.030	.015	-.017	-.035	-.034	-.031
3. Religion								
Catholic	Ref	ref	ref	ref	ref	ref	ref	Ref
None	.148	.038	.038	.176	.089	-.008	.035	.026
	.048	.012	.012	.057	.029	-.003	.011	.008
Protestant	.254	.183	.183	.228	.175	.189	.191	.185
	.108	.076	.076	.095	.073	.080	.081	.079
Evangelical	.306	.365*	368*	.449**	.459**	.344*	.343*	.351*
	.121	.145(4)	.146(4)	.178(4)	.182(4)	.136(4)	.135(4)	.138(5)
Black Protestant	.757***	.726***	.722***	.763***	.766***	.748***	.711***	.699***
	.220(2)	.224(1)	.223(1)	.230(1)	.236(2)	.217(2)	.206(3)	.203(2)
Other	.162	.148	.149	.262	.190	.133	.133	.143
	.050	.046	.046	.081	.059	.042	.042	.045
Jewish	-.258	-.831	-828	-.771	-1.021	-.761	-.745	-.757
	-.023	-.079	-.078	-.076	-.097	-.069	-.068	-.069
4. Hh Income	-.010***	-.009***	-.009***			-.010***	-.009***	-.009***
	-.281(1)	-.217(2)	-.220(2)			-.259(1)	-.253(1)	-.254(1)
5. Par-Educ	-.016	-.023	-.021	-.057**		-.024	-.023	-.021
	-.039	-.052	-.048	-.139(6)		-.059	-.056	-.050
6. Relig-imp	-.127†	-.158*	-.158*	-.119†	-.157*	-.130*	-.115†	-.162*
	-.113(7)	-.143(5)	-.144(5)	-.106(8)	-.142(7)	-.116(6)	-.102(6)	-.144(4)
7. College Lik	-.172***	-.190***	-.190***	-.200***	-.229***	-.191***	-.191***	-.175***
	-.188(3)	-.211(3)	-.211(3)	-.218(2)	-.254(1)	-.208(3)	-.209(2)	-.191(3)
Intercept	-.129	.142	.129	.021	-.449	.119	.044	.075
F	6.82	8.06	7.84	8.01	8.55	8.59	8.06	9.27
N	4,822	6,530	6,530	5,557	6,530	6,710	6,710	6,530

†p<0.05 (one tail);*p<0.05 (two tail); **p<0.01 (two tail);***p<.001 (two tail)

Table 2 (contd.)

Model 1: Listwise deletion
Model 2: Full Mean substitution
Model 3: Mean substitution by race and ethnicity
Model 4: Education as a proxy for income
Model 5: Income and education variables dropped
Model 6: Multiple imputation using the fully conditional specification method
Model 7: Multiple imputation using Markov chain Monte Carlo method with auxiliary variables
Model 8: Multiple imputation using Markov chain Monte Carlo method (imputed only education and income)

The regression results in Table 2 indicate that for some independent variables, whether they are or are not statistically significant does not depend at all on which missing data method is used. The virginity pledge variable is statistically significant in predicting the likelihood of a woman having an adolescent pregnancy in all eight equations, as are the Black Protestant variable, the household income variable, the importance of religion variable, and the likelihood to attend college variable. Five variables are not statistically significant in any of the eight equations, namely, Other race/ethnicity, No religion, Protestant religion, Other religion, and Jewish religion.

However, the statistical significance of all the other variables depends on which missing data method is used in the equation. In the preferred equation, Model 7 (see above), being an African American has no significant effect on the likelihood of having an adolescent pregnancy; but is does have an effect on adolescent pregnancy in six of the other equations. The same pattern holds for the Mexican origin variable and for the Other Latina variable.

A woman being an Evangelical does not have a statistically significant effect on the likelihood of her having an adolescent pregnancy if listwise deletion (M1) is used as the method for handling missing data. But being an Evangelical does have a significant effect on the outcome in all seven of the other equations. Similarly parental education has a significant effect on the outcome in the equation where it is used as a proxy for income (M4), but it does not have a significant effect in any of the other equations.

Clearly, for many of the variables, the method used to handle missing data has an important influence on whether or not the independent variables have significant effects in models of adolescent pregnancy. The statistical significance of most of the race/ethnicity variables (African American, Mexican-origin, Other Latina) depends on the method used for handling missing data; if certain methods are used, e.g., mean imputation, these variables are significant in predicting the outcome; if other methods are used, e.g., two of the three multiple imputation methods, including the preferred method (M7), these variables are not significant. A similar statement may be made regarding one of the religion variables (Evangelical) and the parental education variable.

Another way to evaluate the logit regression results in Table 2 is via the rankings of the statistically significant semi-standardized coefficients. As noted above, these are the logit coefficients that have been standardized in terms of the variances of their independent variables, that is, the logit coefficients are multiplied by their standard deviations (Long & Freese, 2006, p. 96-98). Although there is a problem in the interpretation of the meaning of a semi-standardized coefficient when the independent variable is a dummy variable (there are many dummy variables in the equations, Long, 1997; Poston, 2002, p. 342), their values nonetheless indicate the relative effects of each of the independent variables on the log odds of the woman having a teen pregnancy. In the second row for each variable in each of the eight columns of Table 2 we show the rankings of the magnitude of the semi-standardized coefficient in predicting the outcome. In four of the equations, household income is ranked first, that is, in four equations it has the greatest relative effect on the outcome of adolescent pregnancy; but in two of the equations, those using mean substitution (M2 and M3), it has the second greatest relative effect.

The degree the virginity pledge is influential in predicting the outcome varies according to the method used to handle missing data. If listwise deletion (M1) is used, this variable has the 4th most influential effect, but if mean substitution (M2) is used it has the 8th most influential effect on the outcome. The importance of the effect on the outcome of a woman being an African American varies from the 3rd most important effect in two of the equations (M4 and M5) to the 6th most important effect in four of the equations (M1, M2, M3 and M8). The relative effect on the outcome of the importance of religion variable varies from the 4th most important effect in one equation (M8) to the 8th most important effect in another equation (M4). Clearly the importance of the relative effects of the independent variables on the likelihood of a woman having an adolescent pregnancy vary considerably depending on how missing data are handled in the regression equation.

Discussion

The 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. Understanding differences between minority group members and whites, and the differential influences of minority membership on an outcome such as adolescent pregnancy is a very important sociological question with substantial political and social implications. But the issue of how a researcher chooses to handle the missing data can have an impact on how this social issue is understood. If a researcher used listwise deletion or mean substitution to handle the problem of missing data in equations modelling whether or not a woman had an adolescent pregnancy, the conclusion would be after controlling for all the other variables in the model, Mexican origin women and African American women were more likely than White women to have had an adolescent pregnancy. But if multiple imputation with auxiliary variables as the method to handle the missing data, the results would indicate no statistically significant difference between Mexican origin women and African American women compared to White women with regard to the odds of having had a teen pregnancy. In other words, listwise deletion, the default method in most statistical packages, and multiple imputation with auxiliary variables, the so-called "gold standard," gave the opposite results regarding the odds of a minority woman as compared with a White woman having an adolescent pregnancy.

After controlling for other relevant variables, are minority women more likely than white women to have had an adolescent pregnancy? If listwise deletion or mean substitution was used to handle missing data, the answer is yes. If multiple imputation with auxiliary variables to handle the problem of missing data, the answer is no.

Missing data can also be handled using proxy variables. The use of proxies also has important implications for scientific research. It was showed that when parental education is used as a proxy for household income, it has a statistically significant effect in modelling teen pregnancy, but when household income was used in the equation the effect of parental education disappears.

This finding is very important for two reasons. First from a social policy perspective, the mechanisms and policies that can have an impact on income versus those that can have an effect on education are very different. Thus, knowing that the two variables have different effects on predicting the likelihood of an adolescent pregnancy depending on how one handles the problem of missing data is critical for conducting sociological research. Second, from a theoretical perspective, the use of proxies can have important implications because they might be measuring completely different constructs. For example, the health literature has shown that the effect of education on health is not the same as the effect of income on health (Mirowsky and Ross, 2003). Education taps human capital while income is restricted to financial resources (Sen, 1999). Therefore the effect of education versus that of income can potentially have very different effects on other models related to health outcomes.

This analysis has shown that missing data is indeed a critical component of scientific research, and that different techniques will often lead to different statistical and theoretical conclusions. The next logical question is, how are missing data to be handled when there are potential problems, even with the gold standard of multiple imputation. One of the best and most interesting responses to this question is: "The only good solution to missing data is not to have any" (Allison, 2001, p. 2). Becaise this is an unrealistic option, we propose that it is reasonable to ask researchers who are conducting analyses with missing data to report the results of both listwise deletion and multiple imputation. In addition, the researcher should try different methods of multiple imputation, i.e., with auxiliary variable 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 is handled, and also when there is no strong theory, then the results should be rendered as inconclusive.

Finally, an important recommendation of our paper is that the effect of missing data on scientific research requires more scrutiny. The editors of peer reviewed journals should require the authors to report precisely the amount of data that is missing in their variables, as well as to specify and justify the method they used to handle missing data (Sterne et al., 2009). We specifically recommend that researchers should estimate their models with both listwise deletion and with multiple imputation 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.

Acknowledgment

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

References

Acock, A. (2005). Working with missing values. *Journal of Marriage and Family 67*, 1012-1028.

Allison, P. D. (2001). *Missing Data*. Thousand Oaks, CA: Sage Publications.

Bean, F. D., & Swicegood, G. (1985). *Mexican American Fertility Patterns*. Austin, TX: University of Texas Press.

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.

Harris, K. M. (2008). *The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002 [machine-readable data file and documentation]*. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Klepinger, D. H., Lundberg, S., & Plotnick, R. D. (1995). Adolescent fertility and the educational attainment of young women. *Family Planning Perspectives 27*, 23-28.

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, *171*, 624–632.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Long, J. S., & Freese, J. (2006). *Regression Models for Categorical Dependent Variables Using Stata. Second Edition*. College Station, TX: Stata Press.

Mirowsky, J., & Ross, C. E. (2003). *Education, Social Status, and Health*. New York: A. de Gruyter.

Poston, D. L., Jr. (2002). Son preference and fertility in China. *Journal of Biosocial Science, 34*, 333-347.

Rosenbaum, J. E. (2006). Reborn a virgin: Adolescents' retracting of virginity pledges and sexual histories. *American Journal of Public Health, 96*, 1098-1103.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-590.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.

Sen, A. (1999). *Development As Freedom*. New York: Knopf.

StataCorp. (2011). *Stata Survey Data Reference Manual, Release 12*. College Station, TX: StataCorp.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, *338*(b2393). Retrieved from http://www.bmj.com/content/338/bmj.b2393

Treiman, D. J. (2009). *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco: Jossey-Bass.

von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiple imputed data. *Sociological Methodology*, *37*, 83-117.