Approaches for Addressing Missing Data in Statistical Analyses of Female and Male Adolescent Fertility[1]

Eugenia Conde
Texas A&M University

and

Dudley L. Poston, Jr.
Texas A&M University

Introduction

Missing data is a pervasive problem in social science research. Allison (2002: 1) notes that "sooner or later, usually sooner, anyone who does statistical analysis runs into problems with missing data. In a typical 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." Treiman (2009: 182) reminds us that "missing data is a vexing problem in social research. It is both common and difficult to manage."

Missing data almost always occurs and is usually quite difficult to manage properly. Many techniques have been developed to handle missing data, and some are clearly better than others. Also, it is often the case that the results of a statistical model will differ depending on the method used to handle the missing data.

In this chapter we undertake two separate analyses, one for females and the other for males, of the likelihood of the respondent reporting having had a teen birth. We use several independent variables in our analyses that have been shown in prior studies to be important predictors of adolescent fertility. We handle the problem of missing data using several different approaches. We show in our analyses that depending on the method used, many of the independent variables in the sex-specific models vary in whether they are, or are not, statistically significant in predicting the log odds of a person having had a teen birth, and in the ranking of the magnitude of their relative effects on the outcome.

In this chapter, we first discuss the several mechanisms, as set out by Donald Rubin (1976; 1987), for why data may be missing. We then review some of the major methods that researchers have developed to handle missing data, and use eight of them to handle

missing data in our separate models of adolescent fertility . We conduct our analyses using data from the The National Longitudinal Study of Adolescent Health (Add Health) (Harris, 2008).

Mechanisms: Why Are Data Missing?

According to Donald Rubin (1976; 1987), there are three main reasons or mechanisms for why data are missing; 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 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: 182). If all the missing data in the model turn out to be MCAR, the data that are not missing 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 the variables with the missing data, after controlling for the 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: 182). For example, given a data set with the three variables of age, marital status and income, with missing values on the income variable, 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. It is important to point out that

there is no statistical test for determining if the data are MAR because, obviously, one cannot test whether there is a relationship between unobserved and observed data (Allison, 2002).

Missing data are considered to be missing not at random (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.

Methods for Handling Missing Data

Many methods may be used to handle missing data. In this section we discuss several of the more prominent approaches. In a later section of the chapter we use each approach separately in our two analyses of adolescent fertility.

1: Listwise Deletion. This is the method that is the default method in most statistical packages; it is also known as case deletion. The method of listwise deletion drops the missing values from the data set, and the analysis is then conducted using the reduced sample. We noted above that if the missing data are MCAR, the resulting smaller sample may be considered to be an unbiased subsample of the original dataset (Allison, 2002). Consequently, the use of listwise deletion should result in models with unbiased estimates. However, the standard errors may be slightly larger because the sample size is now, obviously, smaller. With larger standard errors, statistical power will necessarily 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 most likely be biased (Allison, 2002). However, Allison (2002: 7) argues that, "… listwise deletion is not a *bad* method for handling missing data. Although it does not use all of the available information, at least it gives valid inferences when

the data are MCAR… Multiple imputation is potentially much better than listwise deletion in many situations, but for regression analysis, listwise deletion is even more robust than these sophisticated methods to violations of the MAR assumption." Enders (2010:55) states that listwise deletion should only be used when the "proportion of missing data is trivially small." However it is not at all clear what he means by "trivially small." For this reason we consider listwise deletion to be a useful albeit somewhat controversial method.

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. For example, if many respondents did not report an answer to a question pertaining to their annual income, the mean value on the annual income question for those giving answers to this question is assigned to those persons not answering the question. One reason why this method is inappropriate is the fact that subjects who do not answer a question on a variable often tend to be at the extreme ends of the distribution and should thus not be assigned the average score of the variable (Acock, 2005; Enders, 2010). Mean substitution has also been shown to be 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, 2002). Enders (2010: 43) writes that mean substitution "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. 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 income value for males, and to the females the average income value for females.

Although this approach will reduce the variance, it is considered to be only slightly better than substituting with the overall mean (Acock, 2005).

4: Proxy Method. When researchers are confronted with lots of missing data on a theoretically important independent variable, they sometimes use the proxy method as a solution. That is, they substitute for the variable with the missing data another variable with little or no missing data, which variable is related substantively and statistically to the variable with the missing data. This method is not usually discussed in the missing data literature, but its use and application are endemic. For example, to address the situation of missing data on a variable such as income, some researchers (Vaquera, 2006; Wahl, 2010) have used educational attainment as a proxy for income (see Francis 2010; and Perreira et al., 2007, for other examples). At best this approach is a substitute solution to the problem and could well lead to model misspecification.

5: Dropping the Variable(s) with Missing Data. Occasionally one finds research analyses in which the variable(s) with excessive amounts of missing data is (are) simply dropped from the regression equations. For instance, consider the situation of a dataset with, say, 20 percent of the respondents not responding to a question asking about their personal income. If the researcher were to retain the income variable in the equation and use listwise deletion as the method for handling the missing data, then the analysis would be conducted with 20 percent fewer cases. An alternative would be to drop income entirely from the analysis and hence be able to retain those 20 percent of the respondents who failed to report information about their incomes. Like the proxy method just reviewed, this method is not usually discussed in the missing data literature. But it should be avoided without question because of the obvious problem of model misspecification.

The above are five “traditional” methods used for handling missing data. We will use each of them in our analyses of the adolescent fertility of females and males. With the exception

of listwise deletion when the data are MCAR, all 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.[2]

6-8: Multiple Imputation (MI) - three versions. In the analyses we report in this chapter we will use three different versions of multiple imputation, a method first introduced by Donald Rubin in 1987. Recently has it become a popular method owing to its availability in statistical packages. For instance, the Stata statistical software package did not include a multiple imputation routine until its release of version 11 in 2008. MI is a more complex and sophisticated method than the ones reviewed above. And there are several variations of MI.

Many hold 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" (Allison, 2002: 27). Some hold that MI is the current gold-standard approach for dealing with missing data (Treiman, 2009: 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, 2002; Enders, 2010; Rubin, 1987). Unique about this method is that it does not treat the data as if "they were real" (Allison, 2002). Instead, MI estimates the values by taking into account the uncertainty of the missing values component. 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: 1) imputation, 2) analysis, and 3) the combination of the 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. Auxiliary variables may be used that are statistically

related to the variables with missing values. They are thought to 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 argue that a preferred MI equation is one that uses auxiliary variables (Allison, 2002; Treiman, 2009). We too subscribe to this assessment.

In addition, the imputation stage needs to have the same structure and variables of the analysis. In other words, it needs to include all the variables in the model. MI calculates the variances within and between the datasets and uses these to adjust the parameter estimates and to produce more accurate estimates than if the data were treated as if they were "real" (Acock, 2005; Allison, 2002).

Multiple imputation is especially attractive because it can be used with most statistical models. 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. It is sometimes criticized because it is said to lack theoretical soundness; however, simulation studies have shown that it works reasonably well, and the results are comparable to the Markov chain Monte Carlo method (Lee and 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; hence, it works best when imputing continuous variables (Schafer, 1997). However, it has been shown that this method can also be used to impute categorical variables (Allison, 2006; Lee and Carlin, 2010).

Nonetheless, multiple imputation does have some problems. For example, there are no set standards with regard to the number of imputations that should be used, the maximum or minimum amount of data that should be imputed, and how many, if any, auxiliary variables should be used. This lack of specific guidelines may be problematic because different decisions by researchers regarding the above issues could well change the results.

Following the above discussion, we will use three MI methods in our analysis of adolescent fertility, 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 the variables with the most missing data, namely education and income**.

We will hence estimate eight models of adolescent fertility, handling missing data separately with each of the above eight methods. We will show that the regression results do indeed vary, and significantly so, depending on which method one uses to handle the missing data.

Data and Method

The data we use in this chapter are taken from the National Longitudinal Study of Adolescent Health (Add Health) (Harris, 2008). This dataset is a nationally representative stratified sample of adolescents in the 7^{th} through the 12^{th} 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, including Hawaii and Alaska. 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 and male students and their parents.

We use logistic regression to estimate the log odds of females or males to have had a live birth when they were adolescents, i.e., when they were between the ages of 15-19. Therefore, our dataset only includes participants who were 20 years old or older at the time the wave III data were collected. We constructed the dependent variable, whether the respondent had a live birth when she or he was an adolescent, using several questions from the Add Health Survey. The females and males were first asked in the Add Health Survey to assemble a table of pregnancies in which they were involved. For each pregnancy they were asked to "Please indicate the outcome of this pregnancy by selecting the appropriate response: 1. miscarriage, 2. abortion, 3. single stillbirth, 4. live birth, 5. pregnancy not yet ended, 6. multiple, no live birth, 7. multiple, involving both a live birth and another outcome." Then they were asked for information on the month and year of each pregnancy. We calculated their age when each pregnancy ended to determine if they were teens when the births occurred. Subjects who responded that one or more of the pregnancies resulted in a live birth, or in a multiple involving both a live birth and another outcome, prior to their reaching their 20^{th} birthday, were designated by us as having at least one adolescent birth and were scored 1 on the dummy variable of having a live birth while an adolescent; subjects not having an adolescent birth were scored 0.

We selected six theoretically relevant independent variables to predict the log odds of having a teen birth, as follows: 1. the adolescent's race/ethnicity; we measured one's race/ethnicity by first separating Latinos from non-Latinos; the Latinos were divided into Latinos of Mexican Origin and Latinos of other origins (referred to as other Latinos); the non-Latinos were then separated into whites, African Americans and other non-Latinos; in our regression equations we use a series of dummy variables for race/ethnicity (non-Latino African American, Non-Latino White, other Non-Latino, Mexican Origin, and other Latino; non-Latino white was used as the reference); 2. the adolescent's religion was measured with six dummy variables (no

religion, Protestant, Evangelical Protestant, Black Protestant, other religion, Jewish, and Catholic; the Catholic dummy was used as the reference group); 3. household income as reported by the parent in wave I (measured in thousands) with $100,000 as the ceiling; 4. parental education as reported by the parent in wave I and measured as number of years of school completed; 5. the importance of religion to the adolescent ("How important is religion to you?"), ranging from a value of 1 if the young man or woman reported no religious affiliation or responded "not important at all" to a value of 4 if he or she reported "very important"; and 6. the respondent'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 an adolescent had a live birth (cf., Bean and Swicegood 1985; Klepinger et al., 1995).

Results

We have data for 6,726 females and 6,143 males. In Table 1 we present descriptive data for the females on the dependent variable and the six independent variables.

We show in the first data column of Table 1 the number of women for whom we have data for that variable. The maximum number of female cases is 6,726. In column 2 of the table we show for our female subjects the percentage of the cases with data missing for each respective variable. Of the seven variables we use in our logit regression equations (the dependent variable and six 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.

In Table 2 we present the descriptive data for the 6,143 males. Similarly, for our male subjects, only three have missing data percentages of more than one percent: household income, 24.1 percent; parental education, 14.5 percent; and religion 1.8 percent.

With around one quarter of the females and the males both 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.

Of the females in the analysis (Table 1), we show in the third data column that 14 percent report having had a teen birth. Almost 70 percent are white, and their mean household income is just under $43 thousand. Religion is fairly to very important for most of the female subjects, and most believe it is very likely they will attend college.

The data for the males (Table 2) are remarkably similar to those for the females, except for the percentage reporting having had an adolescent birth. Only five percent of the males report having had an adolescent birth, compared to 14 percent of the females.

The fact that the percentage of males reporting an adolescent birth is quite a bit lower than that of the females is likely due to the males not having as direct a knowledge of a pregnancy as the females. This finding of differential male and female fertility rates is consistent with empirical research examining male and female fertility (Greene and Biddlecom, 2000; Zhang, Poston and Chang, 2014).

More than two-thirds (67 %) of the males are white, and their average household income is just over $42 thousand. Religion is important to fairly important for them, and most state it is very likely they will attend college (see Table 2).

The fertility data for the females and males were analyzed using the eight different approaches discussed above for handling missing data, namely, 1. listwise deletion, 2. overall mean substitution, 3. mean substitution where the mean values were substituted on the basis of

the race/ethnicity of the respondents, 4. the proxy method where mother's education was used as a proxy for income, 5. dropping the two variables with excessive amounts of missing data, namely, parental education and household income, 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 three auxiliary variables,[3] and 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 parental education. In each of the three MI applications, a total of 100 imputations were undertaken. For the females, the 16 cases (only 0.2 percent of all the female respondents) that were missing on the adolescent fertility dependent variable were imputed in the imputation stage, but they were dropped from the analysis (von Hippel, 2007). We followed the same strategy for the 21 males (only 0.3 percent of all the male respondents) with missing data on the fertility question.

Since the Add Health Survey is based on multistage probability sampling, one should not make inferences with these data to the larger population of U.S. women and men from which the sample was drawn without first taking into account the sampling design. Otherwise, the data will be treated by the statistical software as based on a simple random sample. Thus we used the "svy" suite of statistical sample adjustment methods available in the Stata 13 statistical package (StataCorp, 2013) that introduce survey adjustment estimators.

In Table 3 we present the results from the eight logistic regressions modelling the log likelihood of a female having a live birth while a teenager. Each regression equation handles missing data in a different way, as discussed earlier. Our preferred method for handling missing data is "multiple imputation using auxiliary variables," shown as model 7

(M7) in the table. In Table 4 we present the regression results from these eight logistic regressions for males.

The values in the first line for each variable in Tables 3 and 4 are the logistic regression coefficients predicting the log odds of a female (Table 3) or a male (Table 4) having an adolescent birth; if the coefficient is statistically significant, it is asterisked (see legend at the bottom of the tables). 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, i.e., the logit coefficient has been multiplied by its standard deviation (Long and Freese, 2006: 96-98). Alongside each of the statistically significant semi-standardized coefficients, in parentheses, is shown the ranking in that equation of its relative effect on the outcome of having a teen birth.

The regression results in Tables 3 and 4 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. For the female respondents (Table 3), three of the religion variables (Evangelical, Black Protestant and Jewish) are statistically significant in predicting the likelihood of the woman having an adolescent birth in all eight equations, as are the household income variable and the likelihood to attend college variable. With respect to the male subjects (Table 4), only the parental education variable is statistically significant in all eight equations predicting the log odds of the male reporting an adolescent birth.

Some of the variables are not statistically significant in any of the eight regression equations. For the female subjects (Table 3), the “other” race/ethnicity variable and the “other” religion variable are never statistically significant. Among the male respondents (Table 4), eight of the variables are not statistically significant in any of the eight equations,

namely, African American and the Other Latino race/ethnicity, five of the six religion variables, and the religion importance variable.

But the statistical significance of all the other variables depends on which missing data method is used in the equation. In all three multiple imputation methods, among the females (Table 3), being an African American has no significant effect on the likelihood of having an adolescent birth; but this variable does have an effect on adolescent fertility in four of the other equations, including the equation using listwise deletion (Model 1) the default method for handling missing data in most statistical packages. Being a Mexican-origin woman or being an "other" Latina are not significant predictors in any of the multiple imputation methods but are significant when mean substitution and dropping the variables are used. The importance of religion has a similar pattern. Parental education is only significant when education is used as a proxy for income.

For the male respondents (Table 4), being a member of an "other" race/ethnicity does not have a significant effect on the log odds of having an adolescent birth in four of the equations, but does have a significant effect in the other four equations. Household income has a statistically significant effect on the likelihood of males having an adolescent birth in each of the multiple imputation models, but it does not have a statistically significant effect in Model 1, the listwise deletion model. The variable designating a person being a Jew was automatically dropped, from all the models except the preferred model with auxiliary variables, because it was a perfect prediction of failure.

It is worth noting that, ideally, multiple imputation is best implemented with auxiliary variables (Model 7); however, since auxiliary variables are not always available in one's dataset, it is acceptable to use MI without them. It turns out that in the analyses we conduct in this

chapter, our regression results are fairly, but not entirely, consistent across the three MI methods we use.

Clearly, for many of the variables, for both females and males, 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 fertility.

Another way to evaluate the logit regression results in Tables 3 and 4 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 have been multiplied by their standard deviations (Long and Freese, 2006: 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) (Poston, 2002: 342), their values nonetheless indicate the relative effects of each of the independent variables on the log odds of the woman or the man having a teen birth. In the second row for each variable in each of the eight columns of Tables 3 and 4 we show the rankings of the magnitude of the semi-standardized coefficient in predicting the outcome.

Among the female subjects (Table 3), 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 having an adolescent birth; but in two of the equations, those using mean substitution (M2 and M3), it has the second greatest relative effect.

The degree to which being an Evangelical is influential in predicting the outcome varies tremendously according to the method used to handle missing data. If mean substitution with race (M3) is used, this variable has the 6^{th} most influential effect, but if the method using education as a proxy (M4) is employed, it has the 2^{nd} most influential effect on the outcome. The

importance of the effect on the outcome of a woman being a Jew varies from the 1st most important effect in several of the equations (M1, M2, M3 and M4) to the 4th most important effect in the equation (M7) that we have designated as the preferred model.

Among the male respondents (Table 4), the rankings of the significant predictors do not vary as much as they do among the females because there are fewer statistically significant effects in the male equations. Nevertheless, among the males, the perceived likelihood of college varies from being the most important predictor of having a teen birth in three of the equations (M2, M3 and M5), to being the 2nd most important predictor in three more equations (M4, M6 and M8), to being the 3rd most influential predictor in M7 (the preferred equation), to not having a statistically significant effect in M1, listwise deletion, the default method in most statistical packages.

Clearly the importance of the relative effects of the independent variables on the likelihood of a woman or a man having an adolescent birth vary considerably depending on how missing data are handled in the regression equation. We turn now to a discussion of our results.

Discussion

In this chapter we discussed the importance of missing data and many of the different methods that have been developed to address the problem. We used data on over six thousand women and over six thousand men from the National Longitudinal Study of Adolescent Health to predict the likelihood of the woman or the man having had a birth when she or he was a teenager. We handled the problem of missing data using eight different approaches. Depending on the method used, our results indicate that many of the independent variables in our models vary in whether they are, or are not, statistically significant in predicting the log odds of a person 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 variable. In summary, 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.

Understanding differences between minority group members and whites, and the differential influences of minority membership on an outcome such as adolescent fertility, is a very important sociological question with substantial political and social implications. But we showed in our chapter that the issue of missing data and how a researcher chooses to handle the missing data can have an impact on how we understand this social issue. To illustrate, we showed that if a researcher used listwise deletion, mean substitution, a proxy, or dropped the variables to handle the problem of missing data in equations modelling whether or not a woman had an adolescent birth, the results would lead the researcher to conclude that after controlling for all the other variables in the model, African American women were more likely than White women to have had an adolescent birth. But if the researcher used multiple imputation with or without auxiliary variables as the method to handle the missing data, the results would indicate no statistically significant difference between African American women compared to White women regarding the log odds of having had a teen birth. In other words, listwise deletion, the default method in most statistical packages, and multiple imputation with auxiliary variables, the so-called “gold standard,” give exactly opposite results regarding the log odds of an African American woman compared to a White woman having an adolescent birth.

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

Another interesting research question concerns the extent to which the socioeconomic characteristics of the respondents influence their likelihood of having teen births. We consider here the case of the males. Our results show that when we use the so-called "gold standard" method (M7) for handling missing data, the respondent's household income and his perceived likelihood of attending college have negative effects on his likelihood of having a teen birth. But if we use listwise deletion (M1) the default method for handling missing data, these two variables do not have statistically significant effects on the log odd of having a teen birth. Using different methods in this case results in very different conclusions with regard to the effect of socioeconomic status on fertility.

Some researchers have handled missing data using proxy variables. The use of proxies also has important implications for scientific research. We showed that among our female respondents when parental education is used as a proxy for household income, it has a statistically significant effect in modelling teen fertility, but when one uses household income in the equation the effect of parental education disappears. In the male equations, the parental education has by far the most influential effect on teen fertility when it is used as a proxy for household income; but this variable is much less influential in the other equations when it is used as a predictor along with household income.

These findings are very important for at least 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 often very different. Thus, knowing that the two variables have different effects predicting the likelihood of an adolescent birth 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 usually 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.

The analyses we conducted in this chapter have shown clearly and conclusively 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 do we handle missing data 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 by Paul Allison: "The only good solution to missing data is not to have any" (Allison, 2002: 2). But since 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. As we stated earlier, listwise deletion is still controversial, and MI is considered to be the gold standard. In addition, the researcher should try different methods of multiple imputation, i.e., with auxiliary variable and without them, and with several different numbers of imputations (say, 50, 75, and 100), so 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 the missing data are handled, and also when there is no strong theory, then the results should be rendered as inconclusive.

Finally, an important recommendation of our chapter is that the effect of missing data on scientific research requires more scrutiny. Journal editors should require their 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 the missing data (Sterne et al., 2009). We specifically recommend that researchers 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 serious deliberation and forethought, and the findings, if inconsistent across method, should be interpreted with caution.

Endnotes

1: This research uses data from the National Longitudinal Study of Adolescent Health (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.

2: There are other "traditional" methods that researchers have used for handling missing data. Among them are dummy variable adjustment and hot and cold deck imputation. Although we will not use any of these methods in the analyses we undertake in this chapter, we mention each of them here, as follows. **Dummy variable adjustment** is an approach widely used in the social sciences; it is also known as the missing indicator method. According to Treiman (2009: 184), "for each independent variable with substantial missing data, the mean (or some other constant) is substituted, and a dummy variable, scored 1 if a value has been substituted and

scored 0 if otherwise, is added to the regression equation." Some prefer this method because it is also a test of the MCAR assumption. "If any of the dummy variables has a (significant) nonzero coefficient, the data are not MCAR" (Treiman, 2009: 184). Although some argue that this approach corrects the missing data for nonrandomness, it has been has shown that it produces biased estimates (Treiman, 2009). And Acock adds that "it gives a false sense of statistical power" (Acock 2005:101-7). Also, if this method is applied to multiple independent variables, one may well have problems with multicollinearity if many respondents fail to provide data on two or more of the same variables (Acock 2005). In sum, this method might seem appealing since it uses all the cases, but it has been shown to produce biased estimates regardless of whether or not the data are MCAR, MAR or MNAR (Acock 2005; Allison 2002). **Hotdeck imputation** is a method used by the U.S. Census Bureau to construct complete data public use samples. According to Treiman (2009: 185) the "sample is divided into strata… Then each missing value within a stratum is replaced with a value randomly drawn (with replacement) from the observed cases within the stratum. As a result, within each stratum the distribution of values for the imputed cases is (within the limits of sampling error) identical to the distribution of values for the observed cases. When the imputation model is correctly specified (that is, when all variables correlated with the missingness of values on a given variable are used to impute the missing values), this method produces unbiased coefficients but biased standard errors. It also tends to perform poorly when a substantial fraction of cases have at least one missing value ..." **Cold deck imputation** follows the same approach as hotdeck imputation, but it replaces the missing values with those from another data set rather than from the same data set. The hotdeck and cold deck methods may seem to be appealing because they use all the cases, but they have been shown to produce biased estimates irrespective of the reason why the data are missing.

3: We used three 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 one question was asked of the students, namely, "How much do you want to go to college?" All three auxiliary questions were answered on a 1-4 or a 1-5 point scale from low to high.

References

Acock, Alan. 2005. "Working with Missing Values." *Journal of Marriage and Family* 67: 1012-1028.

Allison, Paul David. 2002. *Missing Data*. Thousand Oaks, CA: Sage Publications.

Bean, Frank D. and Gray Swicegood. 1985. *Mexican American Fertility Patterns*. Austin, TX: Univeristy of Texas Press.

Enders, Craig K. 2010. *Applied Missing Data Analysis*. New York, NY: Guilford Press.

Francis, S. A. 2010. "Using the Primary Socialization Theory to Predict Substance Use and Sexual Risk Behaviors between Black and White Adolescents." *Substance Use & Misuse* 45: 2113-29.

Greene, M.E. and A.E. Biddlecom. 2000. "Absent and Problematic Men: Demographic Accounts of Male Reproductive Roles." *Population and Development Review* 26: 81-115.

Harris, Kathleen Mullan. 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, Daniel H., Shelly Lundberg, and Robert D. Plotnick. 1995. "Adolescent Fertility and the Educational Attainment of Young Women." *Family Planning Perspectives* 27:23-28.

Lee, K. J., and J. B. Carlin. 2010. “Multiple Imputation for Missing Data: Fully Conditional Specification versus Multivariate Normal Imputation.” *American Journal of Epidemiology* 171: 624–632.

Long, J. Scott and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata. Second Edition*. College Station, TX: Stata Press.

Mirowsky, John and Catherine E. Ross. 2003. *Education, Social Status, and Health*. New York, NY: A. de Gruyter.

Perreira, Krista, Kathleen Harris, and Dohoon Lee. 2007. "Immigrant Youth in the Labor Market." *Work and Occupations* 34 :5-34.

Poston, Dudley L., Jr. 2002. "Son Preference and Fertility in China." *Journal of Biosocial Science* 34: 333-347.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63: 581-590.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.

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

Sen, Amartya. 1999. *Development As Freedom*. New York, NY: Knopf.

StataCorp. 2013. *Stata Survey Data Reference Manual, Release 13*. College Station, TX: StataCorp.

Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. "Multiple Imputation for Missing Data in Epidemiological and

Clinical Research: Potential and Pitfalls." *British Medical Journal* 338, b2393: http://www.bmj.com/content/338/bmj.b2393 (accessed 12-19-2013)

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

Vaquera, Elizabeth. 2006. "The Implications of Choosing 'No Race' on the Salience of Hispanic Identity: How Racial and Ethnic Backgrounds Intersect among Hispanic Adolescents." *Sociological Quarterly* 47: 375-396.

von Hippel, Paul T. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiple Imputed Data." *Sociological Methodology* 37: 83-117.

Wahl, A. M. 2010. "Gender, Acculturation and Alcohol Use among Latina/o Adolescents: A Multi-ethnic Comparison." *Journal of Immigrant & Minority Health* 12: 153-65.

Zhang, Li, Dudley L. Poston, Jr., and Chiung-Fang Chang. 2014. "Male and Female Fertility in Taiwan." Chapter 9 (pp. 151-161) in Dudley L. Poston, Jr., Wen Shan Yang, and Demetrea Nicole Farris (editors), *The Family and Social Change in Chinese Societies*. New York: Springer.

Table 1
Descriptive Data: 6,726 Females,
The National Longitudinal Study of Adolescent Health, Waves 1 and 3

Variable	Cases Without Missing Data	Percent missing	Mean	SD
Dependent Variable				
Teen birth	6,710	0.24	0.14	0.35
Six Independent Variables				
1. 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
2. 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
3. Household Income (in thousands)	4,983	26.00	$42.7	$27.0
4. Parental Education (in years)	5,708	15.14	13.27	2.45
5. Religious importance	6,717	0.13	3.12	0.93
6. Likelihood of college	6,681	0.67	4.25	1.13

Table 2
Descriptive Data: 6,143 Males,
The National Longitudinal Study of Adolescent Health, Waves 1 and 3

Variable	Cases Without Missing Data	Percent missing	Mean	SD
Dependent Variable				
Teen birth	6,122	0.34	0.05	0.22
Six Independent Variables				
1. Race / Ethnicity	6,140	0.05		
White	3,287		0.67	0.47
African American	1,182		0.16	0.36
Mexican	542		0.07	0.25
Other Latina	512		0.05	0.22
Other	617		0.06	0.23
2. Religion	6,035	1.76		
Catholic	1,600		0.25	0.43
None	794		0.14	0.34
Protestant	1,335		0.22	0.41
Evangelical	898		0.18	0.38
Black Protestant	652		0.10	0.29
Other	717		0.12	0.32
Jewish	39		0.01	0.09
3. Household Income (in thousands of dollars)	4,660	24.14	$42.2	$26.0
4. Parental Education (in years)	5,250	14.54	13.33	2.41
5. Religious importance	6,137	0.10	2.99	0.95
6. Likelihood of college	6,101	0.68	3.98	1.24

Table 3: Eight Logistic Regression Models Predicting Having a Teen Birth, 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. Race/ethnicity								
White	Ref	ref	ref	ref	ref	ref	ref	Ref
African American	.376†	.316†	.340	.475*	.474**	.181	.198	.272
	.133(6)	.117(7)	.103	.172(6)	.175(5)	.064	.070	.096
Mexican-origin	.496	.469†	.416	.507	.673**	.333	.336	.377
	.111	.112(8)	.100	.118	.161(6)	.075	.076	.085
Other Latina	.325	.320	.284	.459*	.507*	.258	.260	.264
	.075	.072	.064	.106(9)	.114(7)	.059	.060	.061
Other	-.490	-.357	-.351	.216	-.272	-.391	-.386	-.385
	-.098	-.077	-.075	.044	-.058	-.078	-.077	-.077
2. Religion								
Catholic	Ref	ref	ref	ref	ref	ref	ref	Ref
None	.337	.217	.217	.407†	.277	.188	.222	.197
	.109	.071	.070	.131(8)	.090	.061	.072	.064
Protestant	.304†	.234	.232	.357*	.222	.254	.256	.240
	.129(7)	.097	.097	.150(7)	.092	.108	.108	.102

Evangelical	.475*	.546***	.551***	.675***	.663***	.518***	.519***	.516***
	.187(5)	.217(5)	.219(6)	.268(2)	.264(4)	.204(5)	.205(5)	.204(5)
Black Protestant	.794***	.853***	.846***	.974***	.897***	.885***	.843***	.810***
	.230(4)	.262(4)	.260(4)	.293(3)	.276(3)	.265(3)	.244(2)	.235(3)
Other	.127	.198	.199	.344†	.250	.208	.209	.189
	.039	.061	.062	.106(10)	.078	.065	.065	.059
Jewish	-2.558***	-3.219***	-3.217***	-3.123***	-3.474***	-3.104***	-2.537***	-3.103***
	-.233(3)	-.304(1)	-.304(1)	-.308(1)	-.328(1)	-.282(2)	-.230(4)	-.282(2)
3. Hh Income	-.014***	-.013***	-.013***			-.013***	-.013***	-.013***
	-.392(1)	-.301(2)	-.300(2)			-.357(1)	-.360(1)	-.357(1)
4. Par-Educ	-.023	-.031	-.027	-.080***		-.034	-.033	-.034
	-.056	-.070	-.061	-.195(5)		-.083	-.082	-.084
5. Relig-imp	-.094	-.133†	-.134†	-.108	-.133*	-.109	-.095	-.139†
	-.083	-.121(6)	-.122(5)	-.096	-.121(8)	-.096	-.084	-.144(6)
6. College Lik	-.220***	-.238***	-.239***	-.236***	-.285***	-.213***	-.211***	-.212***
	-.241(2)	-.266(3)	-.266(3)	-.259(4)	-.319(2)	-.233(4)	-.231(3)	-.232(4)
Intercept	-.284	-.044	-.003	-.106	-.759	-.084	-.138	-.022
F	12.05	13.94	13.72	11.66	13.58	13.80	11.56	15.21
N	4,847	6,583	6,583	5,586	6,583	6,710	6,710	6,583

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

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)

Table 4: Eight Logistic Regression Models Predicting Having a Teen Birth, According to the Method Used to Handle Missing Data: Males 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. Race/ethnicity								
White	Ref	ref	ref	ref	ref	ref	ref	Ref
African American	.303	.363	.339	.390	.462	.359	.377	.313
	.106	.130	.122	.137	.166	.125	131	.109
Mexican-origin	.587	.498	.451	.510	.714*	.348	.342	.392
	.137	.125	.113	.122	.179(2)	.081	.080	.092
Other Latino	.011	.281	.250	.269	.472	.135	.142	.209
	.003	.063	.056	.060	.106	.030	.031	.046
Other	.645	.563†	.575†	.737†	.637†	.485	.490	.536
	.138	.132(4)	.135(4)	.159(3)	.149(3)	.104	.105	.115
2. Religion								
Catholic	Ref	ref	ref	ref	ref	ref	ref	Ref
None	.573	.340	.345	.503	.397	.337	.425	.359
	.199	.117	.119	.173	.137	.117	.148	.125
Protestant	.066	-.239	-.236	-.083	-.236	-.236	-.193	-.213
	.028	-.099	-.097	-.035	-.097	-.099	-.081	-.090

Evangelical	.353	.098	.108	.215	.175	.103	.112	.097
	.136	.038	.042	.083	.068	.040	.043	.037
Black Protestant	.362	.049	.058	.212	.107	-.091	-.123	.067
	.104	.014	.017	.061	.031	-.026	-.035	.019
Other	-.027	-.040	-.036	-.131	-.045	-.066	-.060	-.023
	-.008	-.013	-.012	-.042	-.015	-.020	-.019	-.007
Jewish							-2.972**	
							-.238(2)	
3. Hh Income	-.009	-.009†	-.009†			-.011*	-.011*	-.009*
	-.227	-.203(2)	-.205(2)			-.275(1)	-.274(1)	-.234(1)
4. Par-Educ	-.114**	-.073†	-.065†	-.114***		-.072†	-.075†	-.081*
	-.269(1)	-.164(3)	-.148(3)	-.275(1)		-.170(3)	-.178(4)	-.191(3)
5. Relig-imp	.025	.049	.047	.013	.051	.050	.070	.048
	.022	.046	.044	.012	.047	.045	.063	.044
6. College Lik	-.116	-.190**	-.191**	-.150*	-.246***	-.162*	-.160*	-.167*
	-.143	-.234(1)	-.235(1)	-.184(2)	-.304(1)	-.199(2)	-.196(3)	-.205(2)
Intercept	-1.156	-1.239	-1.321	-1.264	-2.403	-1.268	-1.325	-1.224
F	5.29	4.23	4.08	4.71	2.54	4.11	3.81	4.33
N	4,504	5,954	5,954	5,094	5,954	6,083	6,122	5,954

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

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)