Rose-Hulman Undergraduate Mathematics Journal

Volume 17
Issue 2

Article 11

Filling in the Gaps: Using Multiple Imputation to Improve Statistical Accuracy

Ashley Peterson
Purdue University

Emily Martin
Purdue University

ROSE-
HULMAN
UNDERGRADUATE
MATHEMATICS
JOURNAL

FILLING IN THE GAPS: USING MULTIPLE IMPUTATION TO IMPROVE STATISTICAL ACCURACY

Ashley Peterson [a] Emily Martin [b]

VOLUME 17, NO. 2, FALL 2016

Sponsored by

Rose-Hulman Institute of Technology
Mathematics Department
Terre Haute, IN 47803
mathjournal@rose-hulman.edu
scholar.rose-hulman.edu/rhumj

[a] Purdue University
[b] Purdue University

ROSE-HULMAN UNDERGRADUATE MATHEMATICS JOURNAL
VOLUME 17, NO. 2, FALL 2016

FILLING IN THE GAPS: USING MULTIPLE IMPUTATION TO IMPROVE STATISTICAL ACCURACY

Ashley Peterson Emily Martin

Abstract. Missing data is a problem that many researchers face, particularly when using large surveys. Information is lost when analyzing a dataset with missing data, leading to less precise estimates. Multiple imputation (MI) using chained equations is a way to handle the missing value while using all available information given in the dataset to predict the missing values. In this study, we used data from the Survey of Midlife Development in the United States (MIDUS), a large national study of health and well-being that contains missing data. We created a complete dataset using MI. Following that we performed multiple regression analyses probing the relationships between sociodemographic and psychosocial factors and numbers of chronic conditions. Importantly, we compared the results from analyses using imputed data to those from the original dataset. We found that using multiple imputation substantially increased sample size from 3,204 to 7,108 participants and decreased standard errors by an average of 4.81%. This research supports the use of appropriate methods of multiple imputation to facilitate more accurate estimates of associations between disease risk factors and health outcomes in survey research.

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant No. 1246818. We would like to thank Dr. Friedman, Dr. Ward, and Elizabeth Wehrspann for their guidance and support. We would also like to thank the anonymous reviewer.

1 Introduction

Missing values in datasets are a typical issue facing researchers in many fields such as health research. Before multiple imputation (MI), the most common way to handle missing data was to use list-wise deletion, in which a participant missing even just one variable would be deleted from the set. This method removes valuable non-missing information for the same participant and cuts the sample size, thereby reducing statistical power. One method that preserves statistical power involves replacing the missing values with the mean for the relevant variable. However, mean replacement does not include natural variation that occurs in a dataset and may include bias, limiting the validity of the replaced estimates.

Multiple imputation has become a popular way to more accurately fill in missing data in surveys to have full datasets. It incorporates cases that had missing values, which preserves statistical power and maintains precision in analyses compared to handling the missing data by list-wise deletion or mean replacement. Multiple imputation (MI) is a way to estimate the value of missing data using other information from the dataset. It combines classical statistics, where variables are unknown deterministic quantities, and Bayesian statistics, where variables are treated as random with known distributions, when running multiple iterations of equations [13]. A large joint model was assumed for all variables when multiple imputation procedures were first developed. Using a single imputation would only account for a small amount of the variability in the data. However with larger data sets that have hundreds of differing variables this assumption was not accurate. To solve this problem a multiple imputation software package, multivariate imputation by chained equations, allowed a more flexible approach to joint models. Using this software's procedure, regression models were run in which each variable with missing data is modeled conditional on the other variables in the data. Each variable can be modeled to its prior distribution [1]. Bayesian statistics uses the prior distribution of a variable and one's belief about what probability distribution fits the data to determine what type of distribution to apply (such as normal, uniform, or multivariate distribution) when predicting variables. Once the priors have been given for all the variables, MI uses classical statistics to generate equations based on the given priors. Multiple imputation runs chained equations changing the variables to impute until it creates a nearly complete set including the imputed data from previous equations (this method is further explained in the methods section).

Using MI is preferable to other methods for many reasons beyond having more statistical power [4]. Patrician describes three more advantages of using MI in her 2001 paper about using MI for survey data. Multiple imputation incorporates information already known by the data collector about the variables. Adding priors for the individual variables makes the model more accurate and specific to the dataset. Multiple imputation also adds random error to account for differences in imputations. The adjusted standard error from MI creates better estimates of standard errors as compared to single imputation methods, and it reduces the size of standard errors. Finally, one of the best advantages is that MI can create multiple random equations using different variables and different numbers of variables for predicting a missing value. In other words, it will create different models to predict the missing data for a variable. This is beneficial because it creates unbiased estimates that have more validity compared to other methods for handling missing data. It will also use variables containing missing data in the various predictive equations it builds. Using all the information under specific assumptions, even the variables with missing data, creates the most accurate models for predicting the true values.

These many benefits of MI make it a helpful tool when using national datasets that contain large amounts of missing data. With survey data, there are typically three classes of missing data. Data that are missing completely at random (MCAR) are missing for reasons that are unrelated to both observed and unobserved parameters of interest. Data that are MCAR do not bias analyses, but missing data are rarely MCAR. In contrast, data that are missing at random (MAR) are acknowledged to be missing for non-random reasons, but the missingness can be accounted for by observed variables. Finally, data that are not missing at random (NMAR) are missing for reasons that are linked to unobserved variables. It is challenging to determine if data are missing at random or missing not at random, therefore analysts commonly assume that data are missing at random. Surveys can contain data that are not missing at random. Questions that are not applicable to a participant would then be considered NMAR [10]. Multiple imputation adjusts for the potential systematic bias.

The Survey of Midlife Development in the United States (MIDUS) is a widely used national study of health and well-being with large amounts of missing data, making it a good forum for the use of MI. In this study, we took advantage of diverse types of information related to health to employ MI to better predict the factors that influence the risk of developing chronic conditions.

Stata offers an MI package to impute data. This MI package combines classical and Bayesian statistical techniques and it relies on specific iterative algorithms to complete imputations. The advantages of MI in Stata are that it uses complete-data methods (using the entire dataset to do data analysis whether a participant has missing values or not) and incorporates the use of priors. The researcher can add realistic priors for the different types of variables in their dataset. Random variation leads to random error during the imputation process. Stata assumes data is missing at random. The standard error results of MI are more plausible because there are repeated estimations. We found that Stata best suited our needs for the MIDUS dataset. Stata allowed us to customize the imputation to our dataset.

The goals of this study were two-fold: 1) to produce a fully imputed dataset using chained imputation and 2) to use the completed dataset to analyze associations between sociodemographic and psychosocial factors and chronic illness. The first goal was accomplished using Stata's MI imputed chained function. For our second goal, we performed correlation and regression analyses comparing the non-imputed dataset to the imputed dataset to assess the usefulness of using MI.

In Section 2 we give background about multiple imputation and how it has been applied to various fields of research. We continue in Section 3 with the data and methods. Next, Section 4, details the cause of missing data in MIDUS. In Section 5 we dive into multiple imputation and discuss our analytic strategy. We finish with Section 6, our results, and Section 7, our conclusion., At the end, Section 8 lists all tables, graphs, and code. Cohen

2 Background

The development of MI began in the 1980s, and has since been improved and become more widely used. Rubin introduced multiple imputation in 1987 based on his work with hot-deck imputation at the US Census Bureau. His method used repeated imputation, which involved computing the mean and variance for each imputed dataset using standard techniques and combining the results. An adjustment for missing-data uncertainty was added, which accounts for the randomness between and within the datasets (this method is discussed in detail in the

methods section). Rubin's method is still the basis for multiple imputations that are performed today. Rubin [15] suggested that multiple imputation be used for large public data files from surveys or censuses because of the benefits of even a small number of imputations. Based on Rubin's recommendation (and considering the many benefits of MI), we used multiple imputation with the MIDUS survey data.

In 1996, Rubin wrote a paper discussing the evolution of multiple imputation and remaining concerns of the technique. Imputation had not changed much from his 1987 paper but it became easier for analysts and data collectors to create complete datasets with the development of statistical software programs and more user-friendly packages [16]. There were still concerns about multiple imputation at this time, including the validity of MI because it uses simulation. Rubin explained that the simulation only involves missing data, while the rest of the data are left unchanged. Therefore, the appropriate number of imputations depends on the number of missing variables, and the pattern of missingness. Variables that are not missing are not imputed, and the number of imputations is less than a complete data inference. Other concerns were whether multiple imputation was too much work for a user and whether running the imputations used up too much computer storage. These were valid concerns at the time because computers and computing techniques were not as developed as they are today.

As computers and statistical packages have improved, multiple imputation has been used more frequently for research on diverse topics, including social relationships [17], academic achievement [8], consumer finances [14], genotypes, [11], and homicide reports [7]. For example, Lavori, Dawson, and Shera [9] used multiple imputation to handle the missing data in a study of drug treatments for psychiatric disorders, and they found that even a one-step imputation yielded results with a more reliable p-value for rejecting the null hypothesis than other approaches to handle the missing data [10]. In a more recent study, Cole, Chu and Greenland [5] examined children with chronic kidney disease to assess the rate of getting end-stage renal disease. They concluded that using the MI was valuable in removing the bias created from missing data or from misclassifying missing data. Researchers who have used MI to handle missing data in their projects have determined that it is a beneficial method for improving the conclusions drawn from data sets with large amounts of missing data.

3 Methods

3.1 Participants and Procedure

Data used for all multiple imputation and regression analyses in the present study were from the Survey of Midlife Development in the United States (MIDUS). The first wave of MIDUS data was collected from 1995 to 1996 (MIDUS 1, N = 7,108). Participants included a nationally representative sample recruited using random digit dial (n = 3,847), five metropolitan area oversamples (n = 757), siblings of those in the nationally representative sample (n = 950) and a twin sample from a national twin registry (n = 1,914). All participants were non-institutionalized, English-speaking adults, aged 25 to 74 (M = 46, SD = 13.0). One adult per household completed both a telephone interview and self-administered questionnaire. The overall response rate was 60.8% [3].

3.2 Measures

3.2.1 Chronic Illness.

Chronic illness was determined using responses to telephone interviews and self-administered questionnaires. Participants indicated if they had experienced or were treated for the following conditions in the past twelve months: hypertension, asthma, arthritis, AIDS/HIV, diabetes, neurological problems, stroke, tuberculosis, and ulcers. Presence of high cholesterol was determined from a questionnaire item asking participants to indicate if they had taken cholesterol medication in the past 30 days. Presence of heart disease was determined from a single item in the telephone interview asking participants if they had ever had heart trouble suspected or confirmed by a doctor. Cancer status was determined from a telephone interview question asking participants if they had ever had cancer. Obesity (body mass index ≥ 30) was calculated from self-reported height and weight. "Yes" responses to these items (and BMI ≥ 30) each received a score of '1' and were summed to create a total chronic conditions score ranging from 0-13.

3.2.2 Demographic Information.

Participants reported their age, sex, and race during the telephone interview. Participants indicated their race as white, black and/or African American, Native American or Aleutian Islander/Eskimo, Asian or Pacific Islander, multiracial, or other. A simple dichotomous race variable was created from these responses (1= nonwhite 0=white). Participants indicated their highest level of schooling using a scale from 1 (no schooling/some grade school) to 12 (professional degree). To condense this variable, we collapsed it to create a variable with three categories: participants with a high school degree or less (1), participants who attended some college (2) and participants who attended college or an advanced degree (3). Data on household income were obtained from the self-administered questionnaire. Total household income information came from participant responses to a series of questions about income from wages, pensions, social security, and government assistance. Totals were adjusted to account for size of the household; total household income was divided by the square root of the number of people in the household and was capped at $300,000. To condense the income variable to the same scale we divided the variable into 5 categories: participants who made $25,000 or less (1), participants who made between $25,001 and $50,0000 (2), Participants that made between $50,001 and $75,000 (3), participants that made between $75,001 and $100,000 (4) and participants that made more than $100,000 (5).

3.2.3 Psychological Well-Being.

Participants responded to 18 questionnaire items on psychological well-being representing six domains: autonomy, environmental mastery, personal growth, positive relations with others, purpose in life, and self-acceptance. Participants answered using a scale ranging from 1 (strongly agree) to 7 (strongly disagree). A sample item for autonomy was, "I judge myself by what I think is important, not by the values of what others think is important." A sample item for environmental mastery was, "The demands of everyday life often get me down." A sample item for personal growth was, "I think it is important to have new experiences that challenge how I

think about myself and the world." A sample item for positive relations with others was, "Maintaining close relationships has been difficult and frustrating for me." A sample item from purpose in life was "I live life one day at a time and don't really think about the future." A sample item from self-acceptance was, "I like most parts of my personality." This scale comes from a validated 20-item scale used for MIDUS (a=.82). We summed the scores for each subscale and then calculated an overall mean psychological well-being score for which higher scores indicated more psychological well-being.

3.2.4 Depression.

Presence of likely depression was determined using the short forms of the Composite International Diagnostic Interview (CIDI-SF) that was part of the telephone interview. To be categorized as depressed the participants had to have experienced either a depressed mood or anhedonia for at least two weeks, for most of the day, nearly every day and also have at least 4 of 7 symptoms that often coincide with depression (for example loss of appetite). Likely depression was coded as 1, and an absence of depression was coded as 0.

3.2.5 Health behaviors.

Data on smoking habits, drinking habits, and levels of physical activity were included as health behaviors. To assess smoking habits, participants were asked the age they started smoking, if they had ever smoked cigarettes, and if they were smoking cigarettes at the time of the survey. Using these three items, participants were classified as non-smokers, ex-smokers, and smokers. Regarding drinking habits, participants indicated the age at which they first had an alcoholic drink as well as if they ever drank greater than three times per week during one year. Participants indicated that they drank more than three times per year were categorized as drinkers, and all others were categorized as non-drinkers. Participants indicated the frequency with which they participated in moderate and vigorous exercise; a mean score was used in the present analyses for which higher scores indicated a higher average level of activity.

4 Missing Data

There were four missing codes in the MIDUS dataset: "inappropriate", "refused/missing", "invalid", and "don't know". "Don't know" indicated that a participant did not know the answer to a question. Questions participants refused to answer or were missing an answer were coded as "refused/missing". An "inappropriate" code meant that the question was not applicable to a participant (such as part of a skip pattern). If a question was left blank it was considered missing at random. The missing data from the MIDUS survey could not be considered missing completely at random because there are too many factors that can cause missing data. Therefore we assumed missing at random for most variables. Table 1 shows the number and reasons for missing for each variable.

5 Analytic Strategy

5.1 Goal 1: Multiple Imputation using Chained Equations

In order to produce a fully imputed dataset, we used the multiple imputation package in Stata. We used the chained function which imputes the missing values with chained equations (see description below). The benefit of using chained equations is that this process uses different models for each variable based on its distribution, and it can use incomplete and imputed variables in the equations used for imputing other variables [2]. Originally we tried using 60 imputations, but using such a large number of imputations yielded similar results and took considerable time (nearly an hour). To determine the number of imputations, we used the "rule of thumb" given in the Stata help manual. It states "M ≥ 100× (Fraction of Missing) provides an adequate level of reproducibility of MI analysis" with M being the number of imputations [19]. Our FMI = 0.18 therefore we used 20 imputations. We also checked the Monte Carlo errors for the coefficient and p-value. Most of the variables met the requirements, and since we have a large sample we assumed normality. [20] Each imputation had 100 iterations.

Once imputations were complete, we diagnosed the quality of the imputations by checking for trends among the imputed data, running descriptive statistics, and running correlations to make sure imputed data aligned with the original data. Trends in the imputation values would suggest a relation between imputations. Related imputations values would suggest that the values are no longer random. Randomness is evidence that each imputed dataset is independent of the other imputed sets. Having random samples helps to eliminate the unobserved factors, thereby making the model more reliable. To visually assess randomness across imputations, we created a line graph with the averages of each iteration for each variable. We expected the graphs to have variability (a horizontal line with lots of noise around it versus a cone shape). Given the density of the graphs when all twenty imputations were included, we used an overall average to graph as a single line on top of the collection of imputation lines.

5.2 Multiple Imputation using Chained Equations

The following description summarizes the imputation sequence used with "mi impute chained" function, the Stata command we used for our analyses.

Let *X1, X2, …, Xp* be the set of variables to be imputed. Stata orders the variables from the least missing to the most missing. The user gives Stata the prior information to create the model. They are as follows: "regress" for continuous variables, "logit" for dichotomous, "ologit" for ordinal, "mlogit" for nominal, and "poisson" for a count variable (see Table 2 for the linking functions used in our analyses).

First, mi impute chained uses the given variables types to create univariate imputation models. Let Z be the set of independent variables which are complete predictors and let X be variables to be imputed. The Stata manual stated, "By default, fully conditional specification of prediction equations is used". [19]

Then, at iteration t=0 it simulates the missing values using conditional densities in the form

$$f_i(X_i|{X_1}^{(0)}, {X_2}^{(0)}, \dots, {X_{i-1}}^{(0)}, Z, \theta_i) \quad (1)$$

where the conditional density is determined by the variable categorization and θ_i is the set of parameters with uniform prior.

Next, it preforms the given number of iterations generating values for each iteration, t, following

$$g_i\left(X_i \vee {X_1}^{(t)}, {X_2}^{(t)}, \ldots, {X_{i-1}}^{(t)}, {X_{t+1}}^{(t-1)}, \ldots {X_p}^{(t-1)}, Z, \varphi_i\right) \quad (2)$$

with the same conditions for g_i and φ_i as in equation (1).

It repeats equations (1) and (2) with all variables to obtain multiple imputation.

5.3 Goal 2: Regression Analyses

After imputing the data, in order to examine the associations between sociodemographic and psychosocial factors and chronic illness, we used multiple regression analyses. First, we ran a regression with sociodemographic and psychosocial factor to predict chronic conditions using the non-imputed dataset (N = 3,204). Then, we ran the same regression with the imputed dataset (N =7,108). We then compared standard errors between the two datasets in order to assess the benefits of using an imputed dataset. A smaller standard deviation would give evidence of a smaller standard error. Comparing the two regression estimates was beneficial for checking that the imputed dataset did not produce extreme variations from the non-imputed regression, and that the predicted estimates were more accurate due to smaller standard errors.

5.4 Regression Analyses using Imputed Data

The following description summarizes the equations used with “mi estimate: regress”, the Stata syntax used for regression. Stata's “mi estimate” command runs the regression with the imputed data sets that were created, and it adjusts the coefficients and standard errors to account for the variability between imputations following Rubin's rules.

Let Q be a regression coefficient from the data set j (j=1, 2, 3...m) and let U be the standard error of each Q. Let B be the variation between imputations and T be the total variance. First, the overall average of the coefficients is found by:

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m} Q_j$$

Then, the overall standard error is found by taking the average of the standard errors.

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U_j$$

Next, the between imputation variation is found.

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(Q_j - \overline{Q}\right)^2$$

The total variance found from the mi estimate regression is

$$T = U + \left(1 + \frac{1}{m}\right) B.$$

We checked how Stata was computing the regression coefficient and the variance by running the regression and comparing the results with the results from the above equations [12].

6 Results

6.1 Goal 1: Multiple Imputation using Chained Equations

To ensure the most accurate imputation, we graphed the imputations from their iteration values to check for variability. The graphs for each variable looked similar (see examples with Graphs 1 and 2). All graphs had a flat line with variation around it like a cloud with no apparent pattern, which passed the visual inspection for variability.

6.2 Goal 2: Regression Analyses using Imputed Data

6.2.1 Descriptive Statistics.

Table 3 and Table 4 include correlations and descriptive statistics for the non-imputed and imputed variables respectively. Chronic conditions and its predictors were significantly correlated in the expected direction. The descriptive statistics showed that with a completed dataset our means and standard deviations decreased or stayed the same.

6.2.2 Regression Results: Non-Imputed Data.

The complete variable depression, and imputed variables sex, age, income, smoking, exercise, drinking, race, education, chronic conditions, and psychological well-being, predicted sum of chronic conditions, $F(10, 5860) = 117.57$, $p < .001$ (see Table 5). Eight variables were significantly associated with chronic conditions. These variables were depression, income, age, exercise, drinking, race, education, and psychological well-being. The estimated coefficients, standard errors, t-value, and p-values are listed in Table 5. For example, while holding all other variables constant, with one year increase in age, chronic conditions are predicted to increase by 0.03. The other variables can be interpreted in a similar way.

6.2.3 Regression Results: Imputed Data.

The results of the imputed data more accurately predicted an individual's chronic conditions than the non-imputed dataset, $F(10, 4599.5) = 122.73$ (see Table 6). Seven variables were statistically significantly related to chronic conditions, and compared to the non-imputed data, the associations were stronger as indicated by larger beta values. The seven variables were age, depression, income, exercise, education, race, and psychological well-being. Drinking was no

longer significant in the imputed regression. For example, holding all other variables constant, with an increase in education level chronic conditions are predicted to decrease by 0.08. Running an ANOVA table for the imputed dataset does not work because the degrees of freedom changes depending on the predictor in each imputation.

6.2.4 Standard Error Comparison.

As shown in Tables 5 and 6, the standard errors for the variables stayed the same or decreased. This indicated more precise estimates in the imputed dataset compared to the non-imputed set. Percent change was found before rounding. Education had the largest percent change of 7.11%. Race had the smallest percent change of 2.10%. Most variables changed between 3-5%. The changes in standard errors inform us that the imputed data set gives us better results for analysis than the non-imputed set.

7 Conclusion

Overall, our goal in this project was to make a complete dataset by using multiple imputation and to use the completed dataset to analyze associations between sociodemographic and psychosocial factors and chronic illness. We found that using MI was the most beneficial strategy to fill in the missing data. Other methods, such as maximum likelihood, have a minimal use of auxiliary variables, where multiple imputation can be used to include the auxiliary variables [6]. We used Stata's MI because it had more user options regarding data management and variable priors. After the imputations were completed, a visual inspection confirmed that the individual imputations were independent of one another ensuring randomness in the imputed dataset. Then, in our regression analyses, we compared standard errors between the imputed and non-imputed dataset. We saw that the imputed dataset gave lower standard errors, which were evidence of more accurate estimates.

The variables in our data set had varying amounts of missingness. This accounts for the change in mean and standard error from imputed to non-imputed datasets. On average, the standard error decreased by 4.81% with the range being 2-7%. Overall, all of the standard errors decreased which is another indication that the imputed dataset would be better to use for future analyses.

Findings from the multiple regression analyses revealed that age, exercise, education, psychological well-being and depression were associated with chronic conditions. When an individual was older, he or she, on average, had more chronic conditions, but when a person had a higher exercise level or did not have psychological well-being, the number of chronic conditions was lower. The associations with the imputed variables showed similar results as the non-imputed. The same variables were significantly predictive other than drinking. Five variables had strong associations when predicting chronic conditions: age, exercise, education, psychological well-being, and depression. These results were consistent with previous studies. For example, Cole, Shu and Greenland [5] showed that multiple imputation reduced standard error, making an imputed dataset more accurate.

7.1 Limitations and Future Directions

Multiple Imputation and multiple regression analyses were used to examine the links between psychosocial factors and sociodemographic with chronic conditions. Because we were interested in psychosocial and sociodemographic factors we chose specific variables from the larger MIDUS dataset for our imputation. However, a potential limitation of our project is not using all the available data in MIDUS. One reason to use a dataset for which all variables are used in the imputation is that the imputed dataset can then be used for any conceivable analysis. That said there are no hard and fast rules for choosing the number of variables to use in an imputation; it is project specific. There may also be a cost for increasing the number of variables included in an imputation process. Performing an imputation using a joint model would require considerable amount of memory due to the large number of nuisance parameters and be time consuming. For our project, we were interested in the psychosocial factors that play a role in having chronic conditions. Therefore, we chose the variables that were theoretically and empirically related to that question. Using all the variables would not guarantee any more precision than using the most likely predictors. When performing a specific analysis, such as our analysis of chronic conditions, a smaller imputed dataset is often better because small changes caused by unrelated variables are eliminated.

Another limitation, one about which scholars have voiced concern, is that MI creates data that do not actually exist. Creating the data can cause extreme estimates, but with the improvement of computation the researcher can constrain the range for an imputation. Importantly, multiple imputation does not "make up" data, but mean substitution does. When a missing value is replaced with the mean there is no noise, no margin of error, and no variability around the estimate of the missing value. There is an assumption that with absolute certainty the mean would be the observed value if that point were missing. However, multiple imputation adds noise which gives variability to the predicted values that is seen with the observed data. Adding the variability to the missing values makes each estimate plausible [10]. Multiple imputation should not be used for general cases; the results found with MI are specific to the dataset. In general, the benefits of MI, mostly a more complete dataset, outweigh the drawbacks.

Additionally, large survey datasets, such as MIDUS, can contain bias due to the large amount of missing data, and using multiple imputation removes this bias. With the missing data, analyses cannot be performed until the missing data is handled. Creating an adjustment to account for the missing data is difficult to derive, but MI gives a way to handle the missing without removing any data from the dataset. The uncertainty about how data is missing is adjusted for in multiple imputation's MAR assumption. If the MAR assumption is true, the bias is reduced with multiple imputation from the chained equations that can contain missing data.. Multiple imputation allows for a more accurate survey dataset with less bias, the use of prior distributions, and adjustment for missing patterns.

MI is a helpful tool that is more frequently used, but there are still improvements that could be made. Future goals should be to expand the user base of MI. Another goal is to make the programs more user-friendly by creating easier syntax to impute data, combining steps for imputations and graphs, and having more options for analysis after the imputation. Overall, multiple imputation is a helpful tool when handling missing data that could improve the results of future studies.

8 Appendix

8.1 Tables

Table 1
Number of Missing by Variable

Variable Name	**Reasons for Missing**	***N***
Age	Refused	59
Sex	Refused	81
Income	Not Calculated	998
Smoking	Don't know, Inappropriate	5
Exercise	Refused/Missing	801
Drinking	Don't know, Inappropriate	10
Race	Refused/Missing	932
Education	Don't Know	13
Psychological Well-being	Invalid	841
Chronic Conditions	Don't know, Refused/Missing	872

Note: Depression had no missing data.
This table summarizes the classifications and number of missing data for each of the variables used in the study.

Table 2
Linking functions for Variables

Linking Function	**Variable(s)**
Regress	Age, Exercise, Psychological Well-being
Logit	Income, Race, Drinking, Depression
Ordinal	Education, Sex
Nominal (> 2 categories)	Smoking
Poisson	Chronic Conditions

The variables are categorized based on the linking functions used in imputation

Table 3
Correlation and Descriptive Statistics without Imputation

Variables	1	2	3	4	5	6	7	8	9	10
1. Age	1.00									
2. Sex	0.02	1.00								
3. Income	-0.11***	-0.14***	1.00							
4. Exercise	-0.22	-0.14	0.14	1.00						
5. Drinking	-0.05***	-0.29***	0.05***	0.06***	1.00					
6. Depression	-0.11***	0.08***	-0.08***	-0.05***	0.05***	1.00				
7. Race	0.10***	-0.03*	0.09***	0.09***	0.04**	-0.02	1.00			
8. Education	-0.11***	-0.10***	0.31***	0.16***	0.06***	-0.05***	0.02	1.00		
9. Psychological Well-being	-0.00	-0.02	0.18***	0.22***	-0.05***	-0.23***	0.01	0.19***	1.00	
10.Chronic Conditions	0.33***	0.04***	-0.12***	-0.23***	0.00	0.07***	-0.02	-0.13***	-0.15***	1.00
M	46.38	1.52	2.89	7.61	0.42	0.13	0.91	1.92	16.63	1.17
SD	13.00	0.50	1.45	4.35	0.49	0.34	0.29	0.83	2.36	1.50

Note: (*$p<0.05$ **$p<0.01$ ***$p<0.001$) Sex: 1=Male 2=Female Drinking: 0=don't drink 1=heavy drinkers Depression: 0=negative 1=positive Race: 0=White 1=Nonwhite BMI: 0=Not Obese 1=Obese
The table summarizes the correlations between the variables in the non-imputed dataset.

Table 4

Correlation and Descriptive Statistics with Imputation

Variables	1	2	3	4	5	6	7	8	9	10
1. Age	1.00									
2. Sex	0.02	1.00								
3. Income	-0.12***	-0.15***	1.00							
4. Exercise	-0.23	-0.15	0.16	1.00						
5. Drinking	-0.05***	-0.29***	0.05***	0.06***	1.00					
6. Depression	-0.11***	0.09***	-0.09***	-0.06***	0.05***	1.00				
7. Race	0.10***	-0.03***	0.09***	0.09***	0.04***	-0.02	1.00			
8. Education	-0.11***	-0.10***	0.33***	0.17***	0.06***	-0.05***	-0.03*	1.00		
9. Psychological Well-being	0.00	-0.02	0.19***	0.22***	-0.05***	-0.25***	0.01***	0.20***	1.00	
10.Chronic Conditions	0.35***	0.05***	-0.13***	-0.24***	0.00	0.07***	-0.02	-0.14***	-0.15***	1.00
M	46.38	1.52	2.87	7.62	0.42	0.13	0.91	1.92	16.60	1.17
SD	12.96	0.50	1.36	4.13	0.49	0.34	0.27	0.83	2.24	1.42

Note: Sex: 1=Male 2=Female Drinking: 0=don't drink 1=heavy drinkers Depression: 0=negative 1=positive Race: 0=White 1=Nonwhite BMI: 0=Not Obese 1=Obese
The table summarizes the correlations between the variables in the imputed dataset.

Table 5

Regression Table of Non-imputed Variables

Summary of Simple Regression Analyses for Variables Predicting Sum of Chronic Conditions

	Sum of Chronic Conditions				
Variable	*Estimated Coefficient*	*Standard Error*	*B*	*t-value*	*p-value*
Age	0.03	0.00	0.30	23.76	0.00
Sex	0.04	0.04	0.01	1.01	0.31
Income	-0.03	0.01	-0.03	-2.41	0.02
Smoking	-0.01	0.02	-0.00	-0.30	0.77
Exercise	-0.04	0.00	-0.13	-9.83	0.00
Drinking	0.08	0.04	0.03	2.18	0.03
Depression	0.27	0.06	0.06	4.89	0.00
Race	-0.17	0.06	-0.03	-2.72	0.01
Education	-0.09	0.02	-0.05	-3.73	0.00
Psychological Well-being	-0.06	0.01	-0.09	-6.80	0.00
R^2		0.15			
F		105.97			

The results from the regression predicting chronic conditions with the non-imputed dataset are listed.

Table 6

Regression Table of Imputed Variables

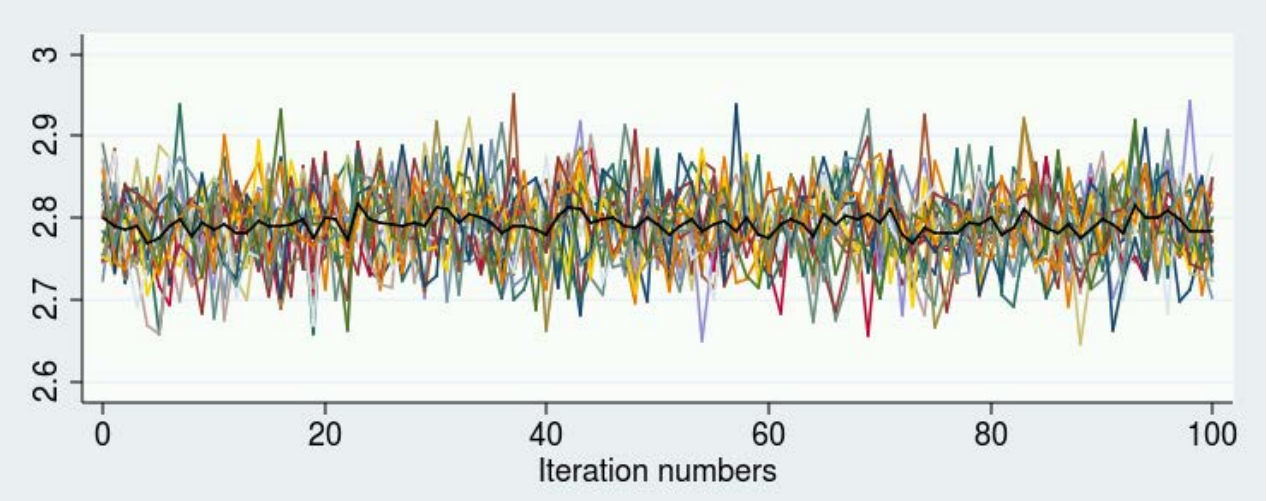
Summary of Simple Regression Analyses for Variables Predicting Sum of Chronic Conditions

Variable	Sum of Chronic Conditions						
	Estimated Coefficient	*Standard Error*	*B*	*t-value*	*p-value*	*MCE for Coefficient*	*MCE for p-value*
Age	0.04	0.00	0.31	25.40	0.00	0.00	0.00
Sex	0.04	0.04	0.01	1.08	0.28	0.01	0.08
Income	-0.03	0.01	-0.03	-2.54	0.01	0.01	0.01
Smoking	-0.01	0.02	0.00	-0.43	0.67	0.00	0.04
Exercise	-0.04	0.00	-0.13	-10.49	0.00	0.00	0.17
Drinking	0.07	0.04	0.02	1.84	0.07	0.03	0.08
Depression	0.30	0.05	0.07	5.52	0.00	0.02	0.01
Race	-0.17	0.06	-0.04	-2.79	0.01	0.04	0.18
Education	-0.08	0.02	-0.05	-3.82	0.00	0.01	0.12
Psychological Well-being	-0.06	0.01	-0.09	-7.25	0.00	0.00	0.07
R^2		0.16					
F		122.73					

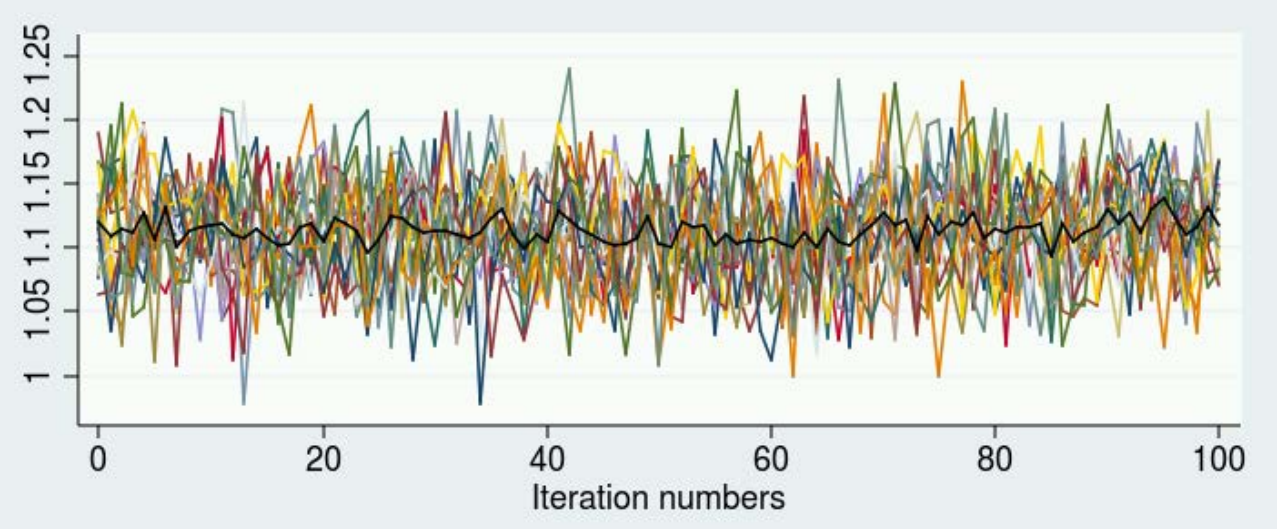
The results from the regression predicting chronic conditions with the imputed dataset are listed.
Note: MCE is the Monte Carlo error.

8.2 Graphs

Graph 1: Income



Graph 2: Sum of Chronic Conditions



8.3 Stata Code

```
***Coding Variables ***
recode A1PA11 A1PA36 A1SA10C A1SA9A A1SA9B A1SA9D A1SA9O A1SA9S A1SA9X
A1SA9Y A1SA9Z A1SA9AA (2=0)
mvdecode A1SA9A A1SA9B A1SA9D A1SA9O A1SA9X A1SA9Y A1SA9Z A1SA9AA,
mv(8)
mvdecode A1SHHTOT, mv(98)
mvdecode A1SBMI A1SPWBA A1SPWBE A1SPWBG A1SPWBR A1SPWBU A1SPWBS,
mv(99)
mvdecode A1PAGE_M2, mv(98)
mvdecode A1PRSEX, mv(8)
mvdecode A1PB1, mv(97)
mvdecode A1SHHTOT, mv(999999)
mvdecode A1SS7, mv(8)
mvdecode A1SS7, mv(9)


***Creating New Variables***

*RACE DICHOT white=1 nonwhite=0
gen raceclean = A1SS7
replace raceclean=0 if (A1SS7=2)
replace raceclean=0 if (A1SS7=3)
replace raceclean=0 if (A1SS7=4)
replace raceclean=0 if (A1SS7=5)
replace raceclean=0 if (A1SS7=6)

*educlean highschool or less = 1, some college = 2, college or more =
3
gen educlean= A1PB1
replace educlean =1 if (A1PB1 =1)
replace educlean =1 if (A1PB1 =2)
replace educlean =1 if (A1PB1 =3)
replace educlean =1 if (A1PB1 =4)
replace educlean =1 if (A1PB1 =5)
replace educlean=2 if (A1PB1 ==6)
replace educlean=2 if (A1PB1 ==7)
replace educlean=2 if (A1PB1 ==8)
replace educlean=3 if (A1PB1=9)
replace educlean=3 if (A1PB1=10)
replace educlean=3 if (A1PB1=11)
replace educlean=3 if (A1PB1=12)

*new BMI (above 30 obese)
gen newBMI = A1SBMI
replace newBMI =1 if (A1SBMI >=30)
replace newBMI = 0 if (A1SBMI<30)
```

```
*created variable that combined all of the chronic conditions
gen sumchronic = A1PA11 + A1PA36 + A1SA10C + A1SA9A + A1SA9B + A1SA9D
+ A1SA9O + A1SA9S + A1SA9X + A1SA9Y + A1SA9Z + A1SA9AA + newBMI

*created variable that combined all of the psychological variables
gen avgPWB = (A1SPWBA + A1SPWBE + A1SPWBG + A1SPWBR + A1SPWBU +
A1SPWBS)/6

* new income
gen income=A1SHHTOT/10000
replace income=1 if (A1SHHTOT<=25000)
replace income=2 if (A1SHHTOT>25000)
replace income=3 if (A1SHHTOT>50000)
replace income=4 if (A1SHHTOT>75000)
replace income=5 if (A1SHHTOT>100000)
replace income=. if (A1SHHTOT==.)

***Imputation***

mi set wide
mi register imputed A1PAGE_M2 A1SS7 A1PB1 A1PRSEX income NewSmokeM1
Exercise Drinking A1PA11 A1PA36 A1SA10C A1SA9A A1SA9B A1SA9D A1SA9O
A1SA9S A1SA9X A1SA9Y A1SA9Z A1SA9AA A1SBMI A1SPWBA A1SPWBE A1SPWBG
A1SPWBR A1SPWBU A1SPWBS A1PDEPDX raceclean educlean newBMI sumchronic
avgPWB
mi impute chained (regress) A1PAGE_M2 (logit) raceclean (ologit)
educlean (ologit) A1PRSEX (ologit) income (mlogit) NewSmokeM1
(regress) Exercise (logit) Drinking (poisson) sumchronic (regress)
avgPWB (logit) A1PDEPDX, add(20) rseed(367485) dots
mi estimate, mcerror: regress sumchronic A1PAGE_M2 A1PRSEX income
NewSmokeM1 Exercise Drinking raceclean educlean avgPWB A1PDEPDX
mibeta sumchronic A1PAGE_M2 A1PRSEX income NewSmokeM1 Exercise
Drinking raceclean educlean avgPWB A1PDEPDX
*mibeta use for chained beta values, install the mibeta package*
*New Mean Variables for Correlation Table using Imputations*
gen meanage=(_1_A1PAGE_M2+ _2_A1PAGE_M2+ _3_A1PAGE_M2+ _4_A1PAGE_M2+
_5_A1PAGE_M2+ _6_A1PAGE_M2 +_7_A1PAGE_M2+ _8_A1PAGE_M2 +_9_A1PAGE_M2
+_10_A1PAGE_M2+ _11_A1PAGE_M2+ _12_A1PAGE_M2+ _13_A1PAGE_M2+
_14_A1PAGE_M2 +_15_A1PAGE_M2+ _16_A1PAGE_M2 +_17_A1PAGE_M2
+_18_A1PAGE_M2+ _19_A1PAGE_M2+ _20_A1PAGE_M2)/20
gen
meansex=(_1_A1PRSEX+_2_A1PRSEX+_3_A1PRSEX+_4_A1PRSEX+_5_A1PRSEX+_6_A1P
RSEX+_7_A1PRSEX+_8_A1PRSEX+_9_A1PRSEX+_10_A1PRSEX+_11_A1PRSEX+_12_A1PR
SEX+_13_A1PRSEX+_14_A1PRSEX+_15_A1PRSEX+_16_A1PRSEX+_17_A1PRSEX+_18_A1
PRSEX+_19_A1PRSEX+_20_A1PRSEX)/20
gen meanincome=(_1_income+ _2_income +_3_income +_4_income+ _5_income+
_6_income+ _7_income +_8_income +_9_income+ _10_income+ _11_income
```

```
+_12_income+ _13_income +_14_income +_15_income +_16_income+
_17_income +_18_income +_19_income +_20_income)/20
gen
meanexercise=(_1_ExerciseFull+_2_ExerciseFull+_3_ExerciseFull+_4_Exerc
iseFull+_5_ExerciseFull+_6_ExerciseFull+_7_ExerciseFull+_8_ExerciseFul
l+_9_ExerciseFull+_10_ExerciseFull+_11_ExerciseFull+_12_ExerciseFull+_
13_ExerciseFull+_14_ExerciseFull+_15_ExerciseFull+_16_ExerciseFull+_17
_ExerciseFull+_18_ExerciseFull+_19_ExerciseFull+_20_ExerciseFull)/20
gen
meandrink=(_1_DrinkingFull+_2_DrinkingFull+_3_DrinkingFull+_4_Drinking
Full+_5_DrinkingFull+_6_DrinkingFull+_7_DrinkingFull+_8_DrinkingFull+_
9_DrinkingFull+_10_DrinkingFull+_11_DrinkingFull+_12_DrinkingFull+_13_
DrinkingFull+_14_DrinkingFull+_15_DrinkingFull+_16_DrinkingFull+_17_Dr
inkingFull+_18_DrinkingFull+_19_DrinkingFull+_20_DrinkingFull)/20
gen
meandepress=(_1_A1PDEPDX+_2_A1PDEPDX+_3_A1PDEPDX+_4_A1PDEPDX+_5_A1PDEP
DX+_6_A1PDEPDX+_7_A1PDEPDX+_8_A1PDEPDX+_9_A1PDEPDX+_10_A1PDEPDX+_11_A1
PDEPDX+_12_A1PDEPDX+_13_A1PDEPDX+_14_A1PDEPDX+_15_A1PDEPDX+_16_A1PDEPD
X+_17_A1PDEPDX+_18_A1PDEPDX+_19_A1PDEPDX+_20_A1PDEPDX)/20
gen
meanrace=(_1_raceclean+_2_raceclean+_3_raceclean+_4_raceclean+_5_racec
lean+_6_raceclean+_7_raceclean+_8_raceclean+_9_raceclean+_10_raceclean
+_11_raceclean+_12_raceclean+_13_raceclean+_14_raceclean+_15_raceclean
+_16_raceclean+_17_raceclean+_18_raceclean+_19_raceclean+_20_raceclean
)/20
gen
meanedu=(_1_educlean+_2_educlean+_3_educlean+_4_educlean+_5_educlean+_
6_educlean+_7_educlean+_8_educlean+_9_educlean+_10_educlean+_11_educle
an+_12_educlean+_13_educlean+_14_educlean+_15_educlean+_16_educlean+_1
7_educlean+_18_educlean+_19_educlean+_20_educlean)/20
gen
meanpsych=(_1_avgPWB+_2_avgPWB+_3_avgPWB+_4_avgPWB+_5_avgPWB+_6_avgPWB
+_7_avgPWB+_8_avgPWB+_9_avgPWB+_10_avgPWB+_11_avgPWB+_12_avgPWB+_13_av
gPWB+_14_avgPWB+_15_avgPWB+_16_avgPWB+_17_avgPWB+_18_avgPWB+
_19_avgPWB+_20_avgPWB)/20
gen
meanchron=(_1_sumchronic+_2_sumchronic+_3_sumchronic+_4_sumchronic+_5_
sumchronic+_6_sumchronic+_7_sumchronic+_8_sumchronic+_9_sumchronic+_10
_sumchronic+_11_sumchronic+_12_sumchronic+_13_sumchronic+_14_sumchroni
c+_15_sumchronic+_16_sumchronic+_17_sumchronic+_18_sumchronic+_19_sumc
hronic+_20_sumchronic)/20

*Correlation Table and Descriptives*
*without imputation*
pwcorr A1PAGE_M2 A1PRSEX income ExerciseFull DrinkingFull A1PDEPDX
raceclean educlean avgPWB sumchronic, sig
summarize A1PAGE_M2 A1PRSEX income NewSmokeM1 ExerciseFull
DrinkingFull A1PDEPDX raceclean educlean avgPWB sumchronic
*with imputation*
```

```
pwcorr meanage meansex meanincome meanexercise meandrink meandepress
meanrace meanedu meanpsych meanchron, sig
summarize meanage meansex meanincome meanexercise meandrink
meandepress meanrace meanedu meanpsych meanchron


***TRACE PLOTS, CHECKING RANDOMNESS BETWEEN IMPUTATIONS***

mi impute chained (regress) A1PAGE_M2 (logit) raceclean (ologit)
educlean (ologit) A1PRSEX (ologit) income (mlogit) NewSmokeM1
(regress) Exercise (logit) Drinking (poisson) sumchronic (regress)
avgPWB (logit) A1PDEPDX, add(10) rseed(367485) dots savetrace(extrace,
replace) burnin(100)
use extrace, replace
reshape wide *mean *sd, i(iter) j(m)
tsset iter

*Creating Mean of Means for Graphs*
gen
meanincome=(income_mean1+income_mean2+income_mean3+income_mean4+income
_mean5+income_mean6+income_mean7+income_mean8+income_mean9+income_mean
10+income_mean11+income_mean12+income_mean13+income_mean14+income_mean
15+income_mean16+income_mean17+income_mean18+income_mean19+income_mean
20)/20
gen
meanchron=(sumchronic_mean1+sumchronic_mean2+sumchronic_mean3+sumchron
ic_mean4+sumchronic_mean5+sumchronic_mean6+sumchronic_mean7+sumchronic
_mean8+sumchronic_mean9+sumchronic_mean10+sumchronic_mean11+sumchronic
_mean12+sumchronic_mean13+sumchronic_mean14+sumchronic_mean15+sumchron
ic_mean16+sumchronic_mean17+sumchronic_mean18+sumchronic_mean19+sumchr
onic_mean20)/20
***Graphs***
tsline A1SHHTOT_mean* meanincome
tsline sumchroninc_mean* meanchron
```

References

1. Azur, M., Stuart, E., Frangakis, C., & Leaf, P. (2012). Multiple Imputation by Chained Equations: What is it and how does it work?. *Int J Methods Psychiatric Research,* 20(1), 40-49. doi: 10.1002/mpr.329
2. Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in TIMSS: a case study. *Large-scale Assessments in Education*, 1:4. doi:10.1186/2196-0739-1-4.
3. Brim, O. G., Ryff, C. D., & Kessler, R. C. (2004). *How healthy are we?: A national study of well-being at midlife*. University of Chicago Press.
4. Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.
5. Cole, S.R., Chu, H. & Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35, 1074-1081. Doi:10.1093/ije/dyl097
6. Collins, L.M., Schafer, J.L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6(4),* 330-351. Doi:10.1037//1082989x.6.4.330-351
7. Fox, J.A., & Swatt, M.L. (2008). Multiple imputation of Supplementary Homicide Reports, 1976 2005. *J Quant Criminol*, 25, 51-77. doi: 10.1007/s10940-008-90582
8. Galera, C., Melchior, M., Chastang, J.F., Bouvard, M.P. & Fombonne, E. (2009). Childhood and adolescent hyperactivity-inattention symptoms and academic achievement 8 years later: the GAZEL Youth Study. *Psychological Medicine*, 39, 1895-1906. Doi:10.1017/S0033291709005510
9. Lavori, P.W., Dawson, R. & Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, 14, 1913-1925.
10. Little, T.D., T.D. Jorgensen, K.M. Lang, and E.W. Morre. (2013). "On the Joys of Missing Data."Journal of Pediatric Psychology 39(2), 151-162. doi:10.1093/jpepsy/jst048.
11. Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics (italics), 39,906-913. doi:10.1038/ng2088
12. *The multiple imputation FAQ page*. Retrieved from http://sites.stat.psu.edu/~jls/mifaq.html.
13. Patrician, P. A. (2001). Multiple Imputation for Missing Data. *Research in Nursing & Health*, 25, 75-84. doi: 10.1002/nur.10015
14. Phillips Montalto, C., & Sung, J. (1996). Multiple Imputation in the 1992 Survey of Consumer Finances. Journal of Financial Counseling and Planning (italics), 7, 133-146.
15. Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New Jersey: John Wiley &Sons, Inc.
16. Rubin, D. B. (1996). Multiple Imputation After 18 Years. *Journal of the American Statistical Association, 91*(434), 473. doi:10.2307/2291635
17. Sassler, S., & Mcnally, J. (2003). Cohabiting couples' economic circumstances and union

transitions: A re-examination using multiple imputation techniques. *Social Science Research*,553-578.

18. StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.
19. StataCorp. 2013. *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.
20. Statistical Computing Seminars Missing Data Techniques in Stata Part 1. Retrieved from http://www.ats.ucla.edu/stat/stata/seminars/missing_data/Multiple_imputation/mi_in_stata_pt1_new.htm