Comparison of Methods for Handling Missing Data in a Collegiate Survey of Tobacco Use

Liza M. Nirelli[1], Michael D. Larsen[1], Ivana T. Croghan[2], Darrell R. Schroeder[2], Kenneth P. Offord[2], and Richard D. Hurt[2],
Iowa State University[1]
Mayo Clinic[2]
Department of Statistics, Snedecor Hall, Ames, Iowa 50011-1210, larsen@iastate.edu

Keywords: EM algorithm, hot deck imputation, multiple imputation, Rosenberg self-esteem scale, single imputation, smoking cessation.

1. Introduction

The Mayo Clinic in Rochester, Minnesota conducted a convenience sample of 2,005 men and women enrolled in one of three Midwestern undergraduate schools (Minnesota State University in Mankato, MN; Minnesota State University in Winona, MN; and Rochester Community and Technical College, Rochester, MN) in the spring semester of 2003. The "Survey of Young Adults" survey instrument was a booklet consisting of 21 pages of Scales and demographic questions. Included were the Body-Area Satisfaction Scale (BASS; Brown, Cash, and Mikulka 1990), the Perceived Stress Scale (PSS; Cohen, Kmarck, and Mermelstein 1983), the Rosenberg Self-Esteem Scale (SE; Rosenberg 1965) and the Positive and Negative Affect Scale (PANAS; Watson, Clark, and Tellegen 1988). Other questions assessed tobacco use, lifestyle choices, concerns about another person's smoking, and interest in future programs assessing other issues. Generally, response levels were quite good to most questions.

Series of questions were used to form the scales to measure the complex concepts such as self-esteem. Some students skipped individual questions, whereas others skipped entire groups of questions. Some students did not answer the questions concerning tobacco status and demographic variables. A common approach for survey scale score calculation uses only complete cases. Since the level of response was high, using available cases for the analyses probably is a reasonable approach. In many other surveys and studies, however, *missingness* of data is a much more serious problem. Of particular concern are situations in which data are not missing at random, but rather are missing more commonly for subgroups with certain characteristics or responses.

To account for missing data, the complete case results will be compared with results using alternate methods for handling missing values. These alternative methods include single imputation methods, methods that avoid explicit imputation, and multiple imputation methods. Single imputation methods include hot deck, mean, mode, and prediction imputation. Methods not using explicit imputations include weighting and log linear modeling. Multiple imputations can be based on hot deck methods and log linear models. The methods make various implicit or explicit assumptions about why the data are missing. The implementations of the methods will be described in the context of the Survey of Young Adults. Methods that involve an explicit model for a nonignorable missing data mechanism are not considered in this paper. Future work could consider such models and the use of more variables in forming imputations.

The scales included in this survey were used as predictors of smoking status (Croghan et al. 2005). Here we suggest three possible analyses using the scales. Initially we think about the analyses without considering the issue of missing scale measurements or other missing values. The first analysis is a comparison of the average scale scores for tobacco users and non-users. The second analysis is a linear regression to predict scale values based on several factors including tobacco use status. The third analysis is a logistic regression to predict tobacco use status based on several factors including the scale value. The alternatives for dealing with the missing scale values might affect the three analyses differently. No claim is being made that tobacco use and the scale values exhibit a causal relationship. These analyses are chosen because they represent typical analysis that one might conduct in tobacco use studies. Future work could consider analyses involving relationships between or among several scales with missing values.

The original data set, fortunately for the researchers, did not suffer from a high degree of missingness. With little missing data, the various approaches to missingness will likely have small impacts on results. To study the impact of the imputation methods, parts of the original data set will be deleted randomly

under various assumptions. The results will be compared to the results based on available cases.

The rest of this article is organized as follows. Section 2 describes the collegiate smoking survey and background. Sections 3 and 4 discuss the problem of missing data and methods of handling missing data in studies, respectively. Sections 5 and 6 present the simulations and some results, respectively. Section 7 states the implications of the results for research practice.

2. Collegiate Survey on Smoking

The prevalence rates of tobacco use in the U.S. are the highest (27%) among young adults 18 to 24 years-of-age (CDC 2000). College students comprise the largest portion of this age group and report substantial tobacco use (Croghan et al. 2005, Nirelli 2005).

Despite the high prevalence, there are few tailored intervention strategies targeting tobacco use in young adults. Moreover, only a handful of studies have examined the association between psychosocial variables and tobacco use. In a Mayo Clinic study of 656 college students, Vickers et al. (2003) assessed the relationship between tobacco use and psychosocial characteristics of depression, coping style, exercise level, and weight concerns. In that study, tobacco users reported lower levels of physical activity, higher scores of scales measuring depressive symptoms, and an increase in maladaptive coping styles in response to emotional distress compared to those not using tobacco. Research also suggests that several psychological factors such as self esteem, body image, general mood, and stress are related to tobacco use (Croghan et al. 2005, Nirelli 2005).

The study population consisted of undergraduate students from Minnesota State University (MSU) at Mankato, Winona State University (WSU), and Rochester Community and Technical College (RCTC). According to institutional websites the respective student enrollments of the institutions in 2004 were 16,079, 7,583, and 7,489. The percentages of female students at MSU, WSU, and RCTC were 54.3, 63.8, and 62.8%. The percentages of Caucasian students were 95.1, 95.8, and 88.0%.

All materials were reviewed by a professional editor and approved by the Mayo Clinic IRB and respective educational institutions. A letter to describe the purpose of the study was sent to classroom instructors in advance. Instructors who indicated interest in participating were asked to provide course information to the investigators concerning the class title, student enrollment, and gender composition of the class. The surveys were distributed in a classroom setting by the class instructor. Most surveys were administered in required courses such as first year courses, but some were administered in upper level courses.

A cover letter explaining the purpose of the study, a survey booklet, and a return envelope were given to all members of participating classes. Neither the class instructor nor the researchers monitored the participants. Participants indicated their refusal by placing the blank survey materials inside the provided envelope. Students were asked to complete the survey only once and students who had completed the survey in another class were asked to refrain from completing it again. The questionnaire took approximately 30 minutes to complete. The participants were assured that identifying information would not be collected and they would not be contacted regarding the study. Their involvement was completely voluntary and they could have withdrawn at any time. Following completion, the survey packets were returned to the Mayo Clinic and were tallied in a single session. These precautions ensured that no information could be traced to survey respondents (Croghan et al. 2005).

Of the 2,057 students invited to complete the study survey, 52 refused participation and one was excluded due to an invalid response pattern. In this study of missing data methods, we focus on analyses involving the Rosenberg self-esteem scale. In data analyses reported here, students in each year of school were restricted to a 6-year age range. For example, freshman could range from 17 to 22 years of age; this eliminates students much older than the typical student. 478 participants could not clearly be defined as current tobacco users or non-users, so none of their responses were considered in analyses. Another 302 respondents neglected to provide demographic information or complete the Rosenberg self-esteem questions. Of the 1,224 complete cases used in analyses here, 403 (32.9%) were categorized as current tobacco users and 821 (67.1%) as never users of any tobacco.

Females represented 62% of the 1,224 respondents. Most (92%) were Caucasian. Demographic characteristics were similar between tobacco users and non-users with the exception of age and year in school. Tobacco users were slightly but statistically significantly older than non-users (20.2.±1.6 versus 19.7±1.4 years of age; P-value <0.001) and were more likely to be college juniors or seniors.

Participants were surveyed about their use of cigarettes, chewing tobacco/snuff, cigars, and pipe tobacco. Researchers used standard questions to ask participants if they used particular tobacco products at least 100 times in their lifetime. Those who reported not having used a tobacco product more than 100 times in their lifetime were categorized as non-users. Those who reported having used any tobacco product 100 times were asked about their use in the past 30 days. Participants who had used tobacco products more than 100 times in their lifetime and within the past 30 days were categorized as current tobacco users. Those who had used tobacco more than 100 times in their life, but not within the past 30 days, were categorized as former tobacco users.

The first scale participants encountered was the Rosenberg Self-Esteem Questionnaire (SE) (Rosenberg, 1965). The ten questions on this scale are rated using a four-point Likert scale ranging from strongly agree to strongly disagree. Higher scores are associated with higher self-esteem; as a result the scores for some questions were reversed before adding to the total when computing the overall scale score. Studies using the SE scale report test-retest reliabilities that range from 0.72 to 0.90.

Overall, the Survey of Young Adults was relatively complete. For example, of the 2004 surveys, every question involved in the SE scale was observed in 1,802 of them. In the 202 remaining, there were two main types of missing data: unit nonresponse and item nonresponse. *Unit nonresponse* on a set of questions occurs when a member of the sample does not participate in the survey. A skip pattern error in the design of the survey instrument caused 177 students to skip all the SE questions. *Dropout* is used to describe students who got tired of filling out the survey, and at a certain point in the survey, stopped filling it out. After completing some initial sections, these students left the remainder of their survey blank. *Item nonresponse* describes a situation in which a student skipped a particular question, but continued filling out the rest of the survey. Table 1 indicates the level of nonresponse for SE scale in terms of tobacco use and gender groups. A survey instrument skip pattern that was fixed partway through the survey caused some missing data on the SE questions among the smokers.

3. Missing Data Mechanisms

Missing data is a problem that occurs frequently in the practice of statistical analysis. Given any large dataset, it is likely that there will be missing values scattered throughout. It is also possible that there will be a large block of missing values. For example, in the Survey of Young Adults, the individual question, "I take a positive attitude towards myself," was blank 8.9% of the time and, as mentioned above, a block of 10 questions from the SE scale, was missing 8.8% (177) of the time due to a skip pattern error. Probability mechanisms that lead to data being missing are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Rubin 1976, Little and Rubin 2002).

The easiest scenario to address is data that are MCAR. The 177 students who left the SE questions blank due to a skip pattern error in the survey would be considered an example of MCAR. MCAR means that the probability that an observation (Y) is missing is unrelated to the value of Y or to the value of any other variables (Little and Rubin 2002). Thus data on age would *not* be considered MCAR if older people were more likely to omit reporting their age than younger people. In a dataset that contains responses to several survey scales, such as in the case of the Young Adult Survey, someone who did not complete the SE 10 questions, would be missing an SE score, but that would not affect whether or not the data can be classified as MCAR. Researchers must know or be willing to believe that responses to the 10 SE questions are simply randomly deleted in order to consider them MCAR.

A more realistic mechanism in many situations is MAR. In MAR, data are missing due to another variable, e.g., age, sex, race, geography or other observable values. If females fail to report self-esteem and their decision to report or not report is unrelated to their self-esteem score, then the data are MAR. This is not a problem as long as analyses are conducted for females and males separately. Technically, data are MAR if the chance of missingness does not depend on the value of Y after controlling for other variables (Little and Rubin 2002).

If data are MCAR or MAR and parameters in data analysis models and missingness mechanisms are distinct, then the missing data are ignorably missing (Rubin 1976). As the term 'ignorable' suggests, when missingness is ignorable, the missing data can be ignored when performing an inference. If, however, data are NMAR (missing as a function of the value of the variable that is missing), then they are nonignorable. It is then necessary to model both the missingness mechanism and the observed values. Data not missing at random are the hardest to deal with, which is unfortunate since it is the most

realistic assumption in many situations. For example, if females are less likely to answer the self-esteem questions and the probability that their self-esteem score is recorded varies according to scores within each gender group, then the data are nonignorable.

Missingness mechanisms affect the performance of statistical inference methods. For illustration purposes, let's define *X*=gender and *Y*= Rosenberg SE score and say that an investigator is interested in three distributions: the marginal distribution of *X* (*f(X)*), the marginal distribution of *Y* (*f(Y)*), and the conditional distribution of *Y* given *X* (*f(Y|X)*). Suppose that *X* is fully observed on the sample, but some values of *Y* are missing. If *X* is fully observed, then inference for the marginal distribution of *X* is not affected by missing data and standard methods should yield valid inferences. Inferences for the distribution of *Y* will potentially be affected by the reason some *Y*-values are missing. If *Y* is MCAR, then missing data are ignorable. As a result, inferences for *f(Y)* and *f(Y|X)* using complete cases with both *X* and *Y* observed should not be biased due to the missing values. They will be, however, less precise due to a reduced sample size.

In another case, suppose *Y* instead of being MCAR is MAR with the probability that *Y* is missing depending on *X*. Using only complete cases can lead to some bias in the inference for *f(Y)*. Imagine a scenario in which *X* and *Y* are positively correlated and units with large values of *X* are more likely to be missing on *X*. The average of the observed *Y*-values will necessarily be biased low for the mean of *Y*. Inference for the conditional distribution *f(Y|X)*, however, should not be biased. Inference for the conditional distribution will still be negatively impacted by the reduced sample size. If *X* is known for more cases than is *Y*, then variable X can be useful for the purpose of increasing the efficiency of the estimated mean of *Y* and reducing the effects of selection bias (Little and Rubin 2002). If missing data are not MCAR or MAR, they are NMAR. In this case, analysis of *f(Y)* or *f(Y|X)* without modeling of the missing data mechanism will lead to biased inferences; adjusting analyses of *Y* for *X* is not sufficient to remove all bias.

4. Missing Data Methods

There are many ways to approach missing data. The five methods in this paper are complete case analysis, single imputation, log linear models and estimation using the EM algorithm, propensity score matching, (Rosenbaum and Rubin 1983, 1985), and multiple imputation (Rubin 1987). See Little and Rubin (2002) for further discussion.

In complete case analysis, one confines attention to cases for which all variables are observed. Complete case analysis can be low in efficiency due to loss of sample size. It can be biased in MAR for some and NMAR for most analyses. Imputation, or filling-in values for missing observations, is a flexible method for handling missing data that has a wide range of implementations. Single imputation can be thought of as explicit or implicit modeling. Common explicit imputation methods include mean imputation, mode imputation, and regression imputation. Mean imputation can be either unconditional (overall) or conditional (within cells defined by other observed variables). The same process can be repeated for the mode. Regression imputation, or estimating a missing value using linear regression, is a form of conditional mean imputation.

A single implicit imputation method is hot deck imputation. Here, missing values are replaced by values from similar responding units in the sample. Similarity is determined by looking at variables observed for both respondents and non-respondents. Hot deck imputations can be generated unconditionally and conditionally within cells. Hot deck imputation methods are sometimes referred to as imputation matching methods or nearest neighbor imputation methods. Sometimes matching is implemented using propensity scores (Rosenbaum and Rubin 1985).

If a single value is imputed for each missing value and analysis are conducted as if all values were real and observed directly, then estimated variances of results can be understated. The understatement of variance due to single imputation can be reduced by applying stochastic methods. A stochastic procedure includes a random element. Hot deck imputation can be stochastic if a random respondent is chosen from among the similar respondents to donate values for imputation. Regression imputation can be stochastic if a random residual is added to an expected conditional mean before the value is imputed.

Multiple imputation is the process of replacing each missing value by a vector of at least two imputed values from at least two draws. These draws typically come from stochastic imputation procedures. MI can account for inflation in the variance due to imputation by repeatedly imputing and generating multiple sets of new data whose coefficients vary from set to set. Thus, the MI procedure is able to capture the variability due to the missingness. This is

beneficial in that once this captured variability is added to the original variance estimate the new variance estimate is no longer biased low (Allison 2002). See Little and Rubin (2002) and Rubin (1987) for further discussion.

In log linear models, cell counts of a contingency table are modeled directly. The usual assumption is that, given expected values for each cell, the cell counts follow independent multivariate Poisson distributions. Conditional on the total sample size, the counts then follow a multinomial distribution. Models that involve fewer parameters than the number of cells in a table are used to express relationships between expected cell counts. For simplicity and ease of interpretation, models typically are specified in a hierarchical manner. In the analysis section, only low order unsaturated models, models with much fewer parameters than cells in the table, will be used. The EM algorithm can be used to estimate log linear parameters in the presence of missing values (Little and Rubin 2002).

A propensity score is the probability that a record has a missing value on a particular variable. The score is estimated using logistic regression. Two units with the same propensity score are expected to have the same distribution of background variables, so matching on the propensity score is a way to create balanced or distributionally comparable groups. See Rosenbaum and Rubin (1985) and Larsen (1999) for examples.

In practice it is common for investigators to treat the imputed values as if they were observed in the first place and then compute the variance estimates using standard formulas for a sampling design. Doing so can lead to inaccurate inferences. In particular, variance estimates from the imputed dataset tend to underestimate the true variance. This underestimation occurs because the additional variability due to the missing values is not taken into account (Rao 1996).

For complete case analysis, if the data is truly MCAR then estimates are not biased. The variance calculated using the reduced sample is correct; as a result, the variance is usually larger than it would be with no missing data. Complete case analysis would not produce correct inferences for MAR or NMAR data (Little and Rubin 2002).

If the data are MCAR, low estimates of the variance are expected when using single imputation, because standard complete data analysis methods do not automatically adjust for the fact that values are imputed. Bias results if conditional imputation is not considered for a variable that is important to MAR assumptions. If a variable is MAR and assumed to be MCAR, then single imputation can distort correlations and relationships. Mean and mode imputation can create a very large overstatement of certainty. Regression imputation can be used to incorporate more variation into conditional means. Stochastic imputation methods, such as hot deck imputation, can create data sets with more realistic variation than deterministic methods.

Missing data analyses using log linear models or propensity scores can produce accurate analyses when variables related to the missing values are used in the models. It still is necessary to try to express suitable variability in analyses when using either of these methods. In general this is challenging to do. One option is the jackknife approach (Rao 1996) which requires conducting analyses multiple times for different subsets of the data.

5. Simulations

Parts of the original data were deleted under various assumptions to study the impact of different imputation methods. Four versions of the original dataset were used to examine different methods of imputation for the SE scale. The first dataset used only complete cases of the original dataset. The other three datasets were simulated to correspond to the three missing data assumptions: MCAR, MAR, and NMAR.

In particular, the original data subset was comprised of 1,224 complete cases. Out of the 2,004 respondents, 1,345 respondents' tobacco use status, age, year in school and gender were known. Of these 1,345, there were 121 respondents each missing some component of the SE-scale. One hundred six participants skipped the entire SE section and 15 participants missed between 1 and 9 questions. The MCAR dataset was created by adding an additional 20% unit nonresponse and another 20% item nonresponse.

The MAR dataset was simulated so that tobacco users were more likely to be missing SE score than non-users. For this scenario, the outcome was the SE scale score and the predicator variable was the tobacco use status. Tobacco users had additional 30% chances of unit and item nonresponse. Non-users had additional 15% chances of both types of nonresponse. Given the proportion of smokers, this corresponds to approximately a 40% nonresponse rate overall.

The NMAR dataset was simulated under the specific assumption that lower values of the SE score are more likely to be missing than higher values of the SE score. The function used to determine the probabilities of missingness was $1/(1+\exp(-SE/10))$. Averaged over the complete cases, there was approximately a 60% probability of an SE score being observed. Thus, in the NMAR dataset, there was approximately an additional 20% chance of both unit and item nonresponse.

6. Results

Results are presented for a linear regression model to predict self-esteem (SE) based on tobacco use status (tobacco user versus non-user), year in school (as a factor variable), gender (f/m), and an interaction between gender and school year. Similar analyses for mean SE and for a logistic regression to predict smoking status are reported in Nirelli (2005). Table 2 presents coefficients and standard errors (s.e.'s) for the coefficient of tobacco status for predicting self-esteem (SE) score.

MCAR, MAR, and NMAR usually lead to different estimates for a given imputation method. However, 4 of 12 imputation methods consistently reduced the regression coefficient for all three missing data mechanisms. These were *overall scale mean, impute mode for each question, conditional hot deck* and *loglinear model.* The other eight imputation methods' regression coefficients had no pattern across the MCAR, MAR, and NMAR columns. The largest differences in standard errors between the original and two missing data mechanisms, MCAR and MAR, were for *impute mean within cells* and *complete case. Multiple imputation hot deck* had the largest difference in standard error between the original and NMAR datasets. The change in significance levels of the P-values across the rows was a result of MCAR, MAR and NMAR producing different estimates. For example, the *conditional hot deck* P-value was significant for the original data (0.027), but not for any of the three missingness mechanisms. The *log linear model* imputation P-value was significant for the original data (0.018), MCAR mechanism (0.045) and NMAR mechanism (0.010), but not for the MAR mechanism (p=0.534).

Imputation methods usually lead to different estimates for a given missing data mechanism. The only missing data mechanism that consistently reduced the regression coefficient for all imputation methods was MAR. The missing data mechanism MCAR reduced the regression coefficient in 8 of the 12 imputation methods, while NMAR decreased the regression coefficient in less than half of the imputation methods. Standard errors for a given missing data mechanism do not exhibit an increasing or decreasing pattern across imputation methods in comparison to the original data. Standard errors increased in 25% of the imputation methods under MCAR, in 42% of the imputation methods under MAR, and in 50% of the imputation methods under NMAR.

The ordering of the standard errors stayed constant down columns and across rows in Table 2. The four smallest standard errors across the missingness assumptions columns were for the *regression prediction*, *regression modified, overall scale mean,* and *loglinear model* imputation methods.

7. Conclusions

The need to impute data in survey analyses is an all too common problem. Fortunately, for the researchers at the Mayo clinic, missing data were not widespread in their survey of college students' tobacco behavior. If there had been a more severe missing data problem, the answer to the question, "Is there an association between low self esteem and tobacco use?" could have been affected. As Rubin and Dempster (1983) warned, researchers must exercise caution when imputing data and reporting conclusions. A researcher should make every effort to understand why data are missing. Making incorrect assumptions about why data are missing can lead to lost efficiency and biased estimates. The fact that assumptions about missing data can affect results is evident in the simulation study because of the fluctuation of the P-values for a given imputation method across different mechanisms.

A limitation of all statistical methods of dealing with missing data in studies is that approaches to handling missing data can be conditioned only on available data unless one is willing to use models that cannot be checked. Often a researcher has only a few basic variables available for all cases. Imputation within cells defined by categorical variables is an effective method of imputation when there are multiple observations within each cell. Regression imputation using covariates is an effective method of imputation when cases with missing outcomes have covariate values similar to those for cases with observed outcomes. Both mean imputation within cells and regression imputation using covariates were effective in this study. Stochastic imputation approaches can produce realistic variability in imputed data and

thereby add appropriate variability to the estimates. Both matching and stochastic imputation methods can increase the accuracy of results even when there are few predictor variables.

Bias in estimates can be hard to discover. How does one know whether or not, for example, the available case mean estimates are reasonable? How does one know why data are missing? Without a larger, more expensive study to gather better data, the only way to assess bias in estimates would be for the researcher to have good *a priori* knowledge of estimates that should be expected and of relationships between outcomes and predictor variables before starting the analysis. As a result, efforts should be planned to reduce levels of missing data, collect variables that are strongly correlated with important outcomes, and learn as much as possible about why there is nonresponse.

Acknowledgments

Work in this paper is the basis for Liza Nirelli's creative component paper for her master's degree in statistics at Iowa State University. Michael Larsen was her major advisor. Ivana Croghan supplied the data from the Survey of Young Adults. Liza Nirelli was employed at the Mayo Clinic in Rochester, Minnesota, prior to a year of residence in Ames, Iowa. She would like thank the investigators with whom she worked closely at Mayo: Ivana Croghan, Darrel Schroeder, Kenneth Offord and Richard Hurt. She would also like thank Irene Faass of Iowa State's Rhetoric, Composition, and Professional Communication graduate program for her expertise in proofreading the master's creative component paper.

References

Allison P.D. (2002) Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research* 28, 301-309.

Brown, T.A., Cash, T.F.and Mikulka, P.J. (1990). Attitudinal body image assessment: Factor analysis of the Body-Self Relations Questionnaire. *Journal of Personality Assessment,* 55, 135-144.

Centers for Disease Control and Prevention. (2000). Cigarette smoking among adults: United States, 1998. *MMWR Morbidity and Mortality Weekly Report, 49*, 881-884.

Cohen, S., Kmarck, T. and Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Social Behavior, 24*, 385-396.

Croghan I.T., Bronars C., Patten C.A., Schroeder D.R., Nirelli, L.M., Thomas J.L., Clarck M.M., Vickers K.S., Foraker R., Lane K., Houlihan D., Offord K.P. and Hurt R.D. (2005) "Is Smoking Related to Body Image, perceived stress, and self-esteem in young adults?" *Preventative Medicine,* submitted.

Dempster A.P. and Rubin D.B. (1983). Overview, *Incomplete Data in Sample Surveys*, Vol II: Theory and *Annotated Bibliography*. New York: Academic Press, 3-10.

Larsen, M.D. (1999). Analysis of a survey on smoking using propensity scores. *Sankhya B, Special Issue on Survey Sampling*, 61, 91-105.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data,* second edition. J. Wiley & Sons, New York.

Nirelli, L.M. (2005). Comparison of Methods for Handling Missing Data in a Collegiate Survey of Tobacco Use, MS creative component paper, Department of Statistics, Iowa State University.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score observational studies for causal effects. *Biometrika,* 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1985). Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statistician*, 39, 33-38.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63, 581-592.

Rubin, D.B. (1987). *Multiple Imputation*. New York: John Wiley & Sons.

Vickers K.S., Patten C.A., Lane K., Clark M.M., Croghan I.T., Schroeder D.R. and Hurt R.D. (2003). Depressed versus nondepressed young adult tobacco users: differences in coping style, weight concerns and exercise level. *Health Psychology, 22*(5), 498-503.

Watson, D., Clark, L.A. and Tellegen, A. (1988). Development and validation of a brief measure of positive and negative affect: The PANAS Scales. *Journal of Social Behavior and Personality, 13*, 399-420.

Table 1: Levels of nonresponse for Self Esteem (SE) scale by tobacco use and gender group.

Scale	Tobacco user (N=409)		Non-users (N=936)	
	Male (N=180)	Female (N=229)	Male (N=321)	Female (N=615)
SE (10 questions)				
Complete	177 (98%)	226 (99%)	287 (89%)	534 (87%)
Missing Some	3 (2%)	3 (1%)	3 (<1%)	6 (<1%)
Missing All	0 (0)	0 (0%)	31 (10%)	75 (12%)

Table 2. Coefficient (standard error) of tobacco status in multiple linear regression predicting self-esteem score.

	Original Coef. (s.e.)	30% MCAR Coef. (s.e.)	30% MAR Coef. (s.e.)	30% NMAR Coef. (s.e.)
1. Complete case	-0.447 (0.186)	-0.497 (0.244)	-0.248 (0.275)	-0.711 (0.241)
2. Overall scale mean	-0.448 (0.174)	-0.274 (0.133)	-0.121 (0.135)	-0.388 (0.136)
3. Impute mode for each question	-0.509 (0.175)	-0.292 (0.150)	-0.053 (0.153)	-0.385 (0.141)
4. Conditional hot deck	-0.405 (0.182)	-0.121 (0.182)	-0.092 (0.188)	-0.258 (0.186)
5. Hot deck	-0.423 (0.184)	-0.378 (0.178)	-0.128 (0.177)	-0.578 (0.180)
6. Impute mean within cells	-0.447 (0.186)	-0.497 (0.244)	-0.248 (0.275)	-0.711 (0.241)
7.a Regression prediction	-0.448 (0.173)	-0.500 (0.131)	-0.296 (0.134)	-0.706 (0.135)
7.b Regression modified	-0.409 (0.174)	-0.516 (0.132)	-0.369 (0.135)	-0.699 (0.136)
8.a Stochastic regression	-0.447 (0.182)	-0.434 (0.176)	-0.389 (0.178)	-0.718 (0.177)
8.b Stochastic regression modified	-0.460 (0.182)	-0.351 (0.177)	-0.220 (0.185)	-0.647 (0.183)
9. Loglinear model	-0.414 (0.174)	-0.297 (0.148)	-0.093 (0.150)	-0.357 (0.139)
10. Multiple imputation hot deck	-0.436 (0.183)	-0.324 (0.185)	-0.019 (0.203)	-0.449 (0.254)

1. *Complete case*- Restricted to the 1224 observations where SE score, demographic variables and tobacco status were observed. 2. *Overall scale mean*- Overall SE scale mean for all observations with missing SE score was imputed. 3. *Impute model for each question*- SE score mode for each question was imputed and SE score was computed. 4. *Conditional hot deck*- Matched on observed SE questions, then the missing SE questions based on matched observations were imputed. 5. *Hot deck*- Imputed missing SE scores by selecting any other case with observed SE score. 6. *Impute mean within cells*- Sixteen cells were defined by tobacco status (2), gender (2), and year in school (4). Then within a given cell, SE score mean for missing SE scores was imputed. 7.a. *Regression prediction*- Linear regression with main effects (year in school, gender, and tobacco status) model was used to predict SE score. Then the regression coefficients were used to predict SE score for missing SE scores. 7.b. *Regression modified*- Same as *Regression prediction*, but linear regression was used with main effects and gender*year in school interaction. 8.a. *Stochastic regression prediction*- Linear regression with main effects (year in school, gender, and tobacco status) model was used to predict SE score. Then the regression coefficients with some residual error were used to predict SE score for missing SE scores. 8.b. *Stochastic regression modified*- Same as *Stochastic regression prediction*, but linear regression was used with main effects and gender*year in school interaction. 9. *Log linear model*- The EM algorithm was used to fit the log linear main effects model for observed and missing ten SE questions. Then the SE score was computed. *10. Multiple imputation hot deck- Conditional hot deck* imputation repeated three times. Then estimates and standard errors were computed.