ABSTRACT

Title of Document:	THE MISSING VALUE PROBLEM: A REVIEW AND CASE STUDY
	Jing Zhou, M. A., 2006
Directed By:	Professor Paul J. Smith, Statistics Program, Department of Mathematics.

The purpose of this thesis is to review methods of imputation and apply them to data collected by Equal Employment Opportunity Commission (EEOC).

First, I discuss several imputation methods and review theory of multiple imputation (MI). Next, I review aspects of missing data and outline an artificial data simulation. I describe simulation based on EEOC dataset listing numbers of employees by ethnicity in large establishments. Mean imputation and MI are applied to simulated datasets. In the first scenario, we impute data for nonresponding establishments. The more we impute, the higher our resulting population means. In the second scenario, we simulate item nonresponse. I find mean imputation and MI generate similar means. The means are not affected by percentage of missingness regardless of imputation methods. The results suggest MI produces larger standard error than mean imputation. Last the percentage of missingness has no effect on standard error in case of MI.

A DIFFERENT WAY TO SOLVE THE MISSING VALUE PROBLEM: THE CASE OF EQUAL EMPLOYMENT OPPORTUNITY DATA.

By

Jing Zhou

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Arts
2006

Advisory Committee:
Professor Paul J. Smith, Chair
Professor Eric V. Slud
Professor Benjamin Kedem

ACKNOWLEDGEMENTS

I would like to extend a special thank you to my advisor, Dr. Paul Smith and my professors, Dr. Eric Slud and Dr. Benjamin Kedem, for their dedication and guidance, and to my mother, father and brother for their love and faith.

Table of Contents

List of Tables

Chapter 1: Introduction

Missing data is omnipresent in survey research. Although when we collect information for statistical analysis, complete data for all subjects are desired, the possibility that some data will be unavailable should not be ignored. It may be lost, be costly to obtain or be unusable. When missing data means there is no response obtained for a whole unit in the survey, it is called unit nonresponse. When missing data means responses are obtained for some of the items for a unit but not for other items, it is called item nonresponse. Missing data has to be dealt with before we can do anything meaningful with the dataset. Many statistical problems have been viewed as missing data problems in the sense that one has to work with incomplete data. Advances in computer technology have not only made previously long and complicated numerical calculations a simple matter but also advanced the statistical analysis of missing data.

Missing data usually means lack of responses in the data. It is often indicated by "Don't know", "Refused", "Unavailable" and so on. Missing data are problematic because most statistical procedures require a value for each variable. When a data set is incomplete, an analyst has to decide how to deal with it. This requires a missing-value procedure (Graham and Schafer, 2002).

Data may be missing in any type of study due to any reason. For example, subjects in longitudinal studies often drop out before the study is complete. Sometimes this happens because they are not interested anymore, are not able to find time to participate, died or moved out of the area. Whatever the reason is, the study will suffer from missing-data problem.

Missing data causes a variety of problems in data analysis. First, lost data decrease statistical power. Statistical power refers to the ability of an analytic technique to detect a significant effect in a data set. Also, it is well known that a high level of power often requires a large sample. Thus, it appears that missing data may meaningfully diminish sample size and power.

Second, missing data produce biases in parameter estimates and can make the analysis harder to conduct and the results harder to present. The bias may be either upward or downward, which means the true score may be either overestimated or underestimated. In an example of Roth and Switzer (1995), a memory study has a true score validity of 0.7. Research may show that the estimated validity is 0.5, an underestimate introduced when the median of observed values is substituted for missing values. This may happen because substituting the median reduces observed variance in a variable.

The methods examined in this thesis to deal with missingness are mean imputation and multiple imputation. Imputation consists of replacing the missing data with values derived from the respondents or from a relationship between the nonrespondents and respondents. Mean imputation can be used when the missingness is either unit nonresponse or item nonresponse. Multiple imputation is more useful for item than unit nonresponse, but it is possible to use it for the latter as well (Little, 2006).

According to Little and Rubin (2002), the mechanisms leading to missing data can be classified into three subgroups:

- Missing completely at random (MCAR)

- Missing at random (MAR) and
- Not missing at random (NMAR).

Denote the complete data by $Y = \{y_{ij}\}$, $i = 1,...,n, j = 1,...,k$ and the missing-data indicator matrix by $M = \{M_{ij}\}, i = 1,...,n, j = 1,...,k$. Denote the conditional distribution of M given Y by $f(M \mid Y, Q)$, where Q is a vector of unknown parameters.

MCAR means that the missing data mechanism is unrelated to the variables under study, whether missing or observed: a missing response happens purely by chance. That is $f(M \mid Y, Q) = f(M \mid Q)$ for all Y, Q.

Let Y_{obs} denote the observed components of Y and let Y_{mis} denote the missing components.

In the case of MAR, the missingness does not depend on the missing values, but may be related to other observed data. That is, $f(M \mid Y, Q) = f(M \mid Y_{obs}, Q)$ for all Y_{mis}, Q.

For example, consider a study with income as the key variable of interest. If ethnicity is always observed and minority group members tend not to report their income, the missing value mechanism may be MAR because whether a person responds depends on her/his ethnicity.

When data are not missing at random, the missing data are said to be NMAR. In contrast to MAR where the probabilities of missingness are determined entirely by observed data and unknown parameters, NMAR arises due to the data missingness pattern being only explainable by the data which are missing. In other words, the distribution of M depends on the missing values in the data matrix Y. For this reason, NMAR is also called informative missingness.

I demonstrate and compare various imputation methods using a data set from the EEOC which contains reports on numbers of employees by gender, ethnicity and occupational category on all businesses meeting size and other criteria. In the first scenario, I identify companies which responded in 2002 but not in 2003 and treat these companies as nonrespondents. I impute their values—the numbers of employees-g by mean imputation. In the second scenario, I artificially make some of the data values in the 2003 dataset to be missing (the number of minority employees). Then I impute their values by mean imputation and multiple imputation.

This thesis is organized as follows.

The second chapter presents several single imputation methods and deletion methods. Mean imputation, regression imputation, hot deck imputation, listwise deletion and pairwise deletion are outlined. Their advantages and disadvantages are detailed.

The third chapter deals with multiple imputation. First, I list the advantages and the assumptions. Next, the main concepts are pointed out as well as its key features. I also give a general idea how multiple imputation works.

The fourth chapter focuses on an artificial data simulation meant to resemble the EEOC data. At each Monte Carlo replication, I create variables y, the number of minority employees; N, the total number of employees and M, the missingness indicator. The resulting dataset contains 100 observations of each variable. I apply various imputation methods to this dataset.

The fifth chapter explains the structure of the EEOC data and provides some baseline information about the dataset. It includes a discussion of missing data mechanisms for the EEOC data. The method EEOC uses for dealing nonrespondents now is briefly mentioned and in contrast, the listwise deletion method is discussed.

The sixth chapter is devoted to simulations based on the EEOC data. The number of minority employees in each classified job categories is made missing to be imputed by mean imputation and multiple imputation. I use a subset of the data based on SIC codes because the limitation of computer power and time. The results are analyzed and discussed.

The seventh chapter reviews the essence of missing data problem, methods dealing with it and the findings of our analyses. Suggestions to improve EEOC practices are proposed

Chapter 2: Classical Methods Dealing with Missing Values

Generally there are two ways dealing with missing values. One is deletion which includes listwise deletion and pairwise deletion. It discards records with missing values. The other is imputation which includes single imputation and multiple imputation. This method fills in one or more values for each missing value.

Listwise deletion is also called complete case analysis. It restrict analysis to those subjects with no missing data. In other words, it deletes all cases with at least one missing item. It assumes incomplete cases are like complete cases.

Pairwise deletion is also known as available case analysis. The idea is to compute each of the summary statistics using all the cases that are available to compute that one statistic

2.1 Historical Development of Treatment

Rubin (1980) contributed substantially to the study of missing data and developed a framework of inference from incomplete data. These techniques are still in use today. Case deletion and single imputation were well documented by the 1980's. About the same time, Rubin introduced the idea of multiple imputation. In the late 1980's, new methods for Bayesian simulation were developed. Multiple imputation and maximum likelihood imputation are now becoming standard because of modern computer technologies (Graham and Schafer, 2002).

In addition, Rubin not only published the first book on multiple imputation in 1987, he also published a highly influential book entitled *Statistical Analysis with Missing Data* with Little (Little and Rubin, 1987). In this book, they laid the groundwork for development of the EM algorithm for numerous missing data applications. (Graham and Schafer, 2002)

Since 1987, addressing the issue of missing data has been aided by numerous software products, such as SAS, SPSS, ViSta, MicrOsiris and R, that have become available. Best of all, many of them are free. And the money is well spent for those are not free because their value outweighs their cost. Although much improvement and development is still needed in the area of dealing with missing data, statisticians have made tremendous progress. Because missing data is a universal problem in the sense that it may occur in any discipline, missing data analysis should be made accessible to researchers all over the world. This writer believes a good start has been made and that useful and accessible missing data procedures will become an integral part of mainstream statistical packages soon.

2.2 Overview of Single Imputation

When a large database will be analyzed by many users, there is a desire to "clean up" the data, which includes dealing with missing values. The reason is that standard procedures cannot be used when there are missing values and corresponding procedures that adjust for missing values may not be easy to derive. Imputation is one of the most common procedures for handling missing values (Gyimah, 2001). Single imputation is just as the name suggests, filling in a single value for each missing value. Single imputation is attractive for several reasons.

First, it saves a great deal of data that listwise deletion drops since it keeps all the rest of the data for an individual for use in analysis. It also saves more data than pairwise deletion since it preserves the data paired with a previously missing value (Roth & Switzer, 1995). Overall, it retains data in incomplete cases that would have been discarded if the analyses were restricted to complete cases such as when using listwise deletion and pairwise deletion.

Second, when descriptive statistics and other statistical measures are of interest, standard complete-data methods of analysis can be used on the filled-in data set. However, some of the statistical measures are biased using the filled-in data set. For example, mean imputation underestimates the variance. Complete data software seems to keep closer pace with the statistical methodological developments than incomplete data software. An example of complete data software that can be used to process the data is SPSS. Mean imputation, which is one single imputation option, appears in several SPSS procedures.

Third, sometimes data is missing because it is confidential information. Thus, the public cannot view these data. In those events, the data producers, who have access to all the data, are able to incorporate all the knowledge by imputation into the data set for public use. The public sees the imputed dataset, not the original full dataset, because some of the data are confidential.

Last, the nonresponse problem is solved in the same way for all users so the analyses will be consistent across all users. Thanks to the data producers, the researchers will have the imputed data set which is completed and can concentrate on addressing the questions of their interest.

2.3 Listwise Deletion

Because listwise deletion may sacrifice a large amount of data when the information contained in the incomplete cases is discarded, we risk decreasing power due to a large loss of data and risk running our analyses on a partial set of data that are not representative of the population of respondents. In particular, if one has a large data set containing hundreds of variables, there may be relatively few cases with all variables observed. If we simply discard the incomplete cases, we will probably end up with many fewer cases than we hoped to have. And also by deleting the incomplete cases, we are ignoring the information that is contained in the incomplete cases. Resources are invested into gathering these data. If we don't use all of the available information, we are wasting money and time. In addition, even if we don't observe all the variables, the information in the incomplete cases may provide some insight about the outcome of interest. As a result of discarding data, the standard errors will generally be larger in the listwise deleted data set than the original one because less information is utilized. Also, they will tend to be larger than standard errors obtained from mean imputation.

Another problem with listwise deletion is that, if we want to calculate sample statistics, the data must be MCAR (Allison, 2002). If the data are not MCAR, whether MAR or NMAR, the listwise deleted data set will not be a random subsample of the original data. The complete cases are probably not representative of the entire data set. Because of that property, for any parameter of interest, estimates based on the listwise deleted data set may be biased even if they are unbiased when based on the full data set. Thus, the standard errors and other test statistics obtained from the listwise deleted data set will not be as appropriate as they are in the complete data set.

On the other hand, if the MCAR assumption is valid, listwise deletion does have its advantages. The obvious one is that it can be used for any kind of statistical analysis, from structural equation modeling to loglinear analysis. Thus it is the default in most statistical software. Also, it doesn't require special computational methods. And last but not least, when data are MCAR, then the reduced sample will be a random subsample of the original data. Any estimate produced from the reduced sample will be unbiased as long as the estimate is unbiased for the entire data set (Allison, 2002).

2.4 Pairwise Deletion

While single imputation replaces a value for the missing value, pairwise deletion drops pairs of missing observation on the variables under examination. The idea of pairwise deletion is to use all the cases that contain data for a given calculation. For example, to compute the covariance between two variables X and Z, all cases that have data present for both X and Z are used. To compute the covariance between X and Y, all the cases that have data present for both X and Y are used, even the cases with missing values for Z or any other variable (Allison, 2002). Although pairwise deletion still loses data compared to single imputation, it preserves a great deal of information that would be lost when using listwise deletion. As a result, Monte Carlo studies have found that it results in less dispersion around the true score (Roth and Switzer, 1995).

One concern is that under pairwise deletion, the samples used to compute each statistic are slightly different. For example, one might find 60 cases are available to compute the covariance between X and Y while only 40 cases are available to compute the covariance between X and Z (Roth and Switzer, 1995). Because the sample base changes from variable to variable according to the pattern of missing data, interpreting the covariance matrices or correlation matrices may be difficult. This can occasionally lead to a correlation matrix or a covariance matrix that is not positive definite (Roth, 1994) and can also lead to a sample correlation greater than one. When the matrix is not positive definite, there may be less information in the matrix than would be expected based on the number of variables involved.

2.5 Some Single Imputation Methods for Dealing with Missing Data

In this section, I will describe mean imputation, regression imputation and hot deck imputation.

2.5.1 Mean Imputation

Mean imputation is one of the most frequently used imputation methods. Basically, it uses the mean of observed values of a variable in place of missing data values for that same variable. Mean imputation enjoys many of the advantages of single imputation mentioned above. For example, it saves a great deal of data that listwise deletion and pairwise deletion eliminate, although mean imputation may not always give reasonable values.

But while mean imputation preserves data, Maxim (1998) argues that it leads to a biased estimation of the true population parameter in general. Following

Maxim’s argument, I argue, in the EEO-1 report, if the employers who have higher numbers of minority employees are less likely to report their number of minority employees, then substituting the mean number of minority employees of the respondents will certainly underestimate the true mean.

In addition to biased estimation of the true population parameter, there may be underestimation of the variance. This underestimation is caused by not accounting for the variation that would likely be present if the missing values were observed instead of being replaced by the same mean value. It also results from the increased sample size by including cases of the missing observations that being imputed (Roth, 1994).

For example, in the EEO-1 2003 report database, there are 221,289 employers. Among the 221,289 employers, 219,322 are observed. The 2003 exit companies which are the companies present in 2002 but not in 2003, number 1967. Then mean imputation suggests to impute 1967 additional ones to get 221,289 total records. This would increase the N in the calculation of a sample variance, but would not increase the sum of squared deviations around the mean added by the 1967 additional cases. Although 1967 doesn’t seem to be large enough to make a big difference, the variance is indeed underestimated after all because in the formula for variance the sum of squared deviations around the mean in the numerator won’t increase but the N in the denominator will increase.

Downwardly biased variance estimates bias correlation. And if the same cases are missing for two variables and their overall means are substituted, the magnitude of these estimated correlation can be inflated. Some researchers have reasoned that mean imputation in the case of one variable can lead to bias in estimates of other variables in the regression analysis (Roth, 1994).

2.5.2 Regression Imputation

Regression imputation replaces missing variables by predicted values from a regression of the missing variables on variables observed for that unit. This regression is usually calculated from units with present. It involves the use of one or more independent variables. The regression formula is built on the cases with complete data and a well-fitted model is established by using complete cases. An error term, which is computed from the complete cases, can be included in the prediction to maintain the underlying variability in the data, thus reflecting the uncertainty of the predicted value (Watson & Wooden, 2003).

While mean imputation uses the mean of complete data for the missing values, regression imputation uses the data in a regression model which relates a dependent variable, y_i to the independent variable, $x_{i\,.}$ First, records with complete data from both independent and dependent variables are used to estimate the parameters of the model. Then a predicted value is generated by regressing the missing value (dependent variable) on all other values for independent variables which have no missing data. So the predicted value is produced by using complete data. According to Smith (2005), to apply regression imputation, the underlying missing data mechanism should usually be MAR. The proper regression model depends on the form of the dependent variable. A probit or logit model is used for binary variables, Poisson or other count models for integer-valued variables, and ordinary least squares or related models for continuous variables (Smith, 2005).

An interesting feature of regression imputation is that the estimator that uses regression imputation to represent the unobserved values is model unbiased. In other words, the estimator is unbiased under the regression model used for imputation but may be biased under other models.

Some critics have commented that regression imputation theoretically provides “the best” estimate for a missing value. Little and Rubin (1989) argued that regression imputation works best when most of the variation in y is explained by x. They argued that to provide reasonable estimates of means, we can calculate the sample mean and covariance matrix of the complete data . Then we use these estimates to produce the estimated values of the missing data by substituting the observed values of the case on which data is missing into a regression function. As it turns out, the imputed data for a variable are more consistent with the data for the rest of the variables in the complete dataset than for the “true” data set. Therefore, although this method gives good estimates of means, it underestimates the variance. But the underestimation is less than mean imputation. (Little and Rubin, 2002).

Regression imputation enjoys all the advantages of single imputation generally and has the special feature of being able to make use of many categorical and numeric variables. It works well for numerical data, especially when the correlation between independent variables and dependent variable is high. As a result, under a regression model, as the correlation decreases, the standard error of the estimate will become higher, and the accuracy of replacing missing items with a value predicted from a regression equation will become lower (Roth, 1994).

2.5.3 Hot Deck Imputation

Hot deck imputation is a technique that imputes missing values using sampled values of respondents. It replaces a missing value with an observed value-- the donor. The donor can be the response of a randomly selected case for the variable of interest. An alternative and better way is to choose the donor randomly not from the full set of all respondents' data, but from a similar respondent who closely matches the nonrespondent.

To implement the hot deck, first, we stratify the data into imputation classes or clusters based on key explanatory variables. Then we match the nonrespondents to their imputation classes. After that, a donor is selected at random, by nearest match or by some other way within the imputation classes. Once, a matching donor is found, the values reported by the donor are imputed for the nonrespondent. (Little and Rubin, 2002). Creating a larger number of classes yields improved accuracy of imputation, but it can also lead to very small imputation classes. Small imputation classes may cause difficulty finding a donor. When that happens, we should combine the imputation classes.

For example, in the EEO-1 data, the companies are classified by NAICS code. If for some company, the number of minority employees, say Asian, is missing and since each instance with missing data value is associated with one NAICS code, we might match the case with missing number of Asian employees to an imputation class and randomly chose a donor from that class based on NAICS code. The number of Asian employees for the donor is then inserted in place of the missing number of Asian employees for the nonrespondent.

Hot deck imputation has a long history of use and is common in practice. It is very heavily used with census data (the United States Census Bureau has used it for decades in forms other than I have described) because it has the advantage of using the occurring values for imputation. It can be carried out as the data are being collected because it uses everything in the data set so far. It is conceptually simple and easily programmed in a programming language.

Hot deck imputation can maintain the proper measurement level of variables. That means by using hot deck imputation, categorical variables will remain categorical and continuous variables will remain continuous. Also, hot deck imputation is usually nonparametric and avoids distributional assumptions. When applying hot deck using a random donor, the imputed values will have the same distributional shape as the observed data. So unlike other methods, it reflects both the mean and variance of the underlying data.

There are several disadvantages of hot-deck procedure. First, it is difficult to decide how similar a case should be in order to be a donor case. The users are required to create software to perform the selection of donor cases and the subsequent imputation of missing values in the database. In that sense, hot-deck procedure is not a handy approach to solve incomplete-data problems.

Second, with the hot-deck procedure, categorization of variables is encouraged because each has some effects on the imputed values. But, categorizing variables sacrifices information when it forces continuous variables into categories

Chapter 3: Multiple Imputation

3.1 Historical Development

Rubin (1980) first proposed to use multiple imputation to cope with missing data and extended the method thereafter. As a matter of fact, the first book on multiple imputation was published by Rubin in 1987. In 1987, the multiple imputation scheme was studied in large sample surveys. In those surveys, a large number of investigators would use the data collected in a single study for a number of different analyses. To do multiple imputation, a lot of computing power was needed. And without advanced computer technology and adequate computational facilities, multiple imputation remained little known. However, more recent development of faster and more sophisticated computers made multiple imputation become quite popular in survey and nonsurvey contexts. Multiple imputation has performed well in a number of studies that comparing approaches for handling missing data in the structural equation modeling context in the 1990's (Gold and Bentler, 2000). There is another reason for the recent popularity of multiple imputation. Statisticians began treating missing values as a source of variation to be averaged over after the start of an EM algorithm in the late 1970's. And multiple imputation can do this averaging in a simple way (Sinharay, Stern and Russell, 2001).

3.2 Model Assumptions

According to Rubin and Schenker (1986), " The theoretical justification for multiple imputation is most easily understood from the Bayesian perspective." A particular imputation model is needed for performing multiple imputation, and the correctness of the assumed imputation model determines the success and failure of multiple imputation. Like any statistical method, certain assumptions are required for multiple imputation. Basic understanding of these assumptions is necessary for use of multiple imputation. The three assumptions which are essential in multiple imputation are (a) a model for the data values, (b) a prior distribution for the parameters of the data model, and (c) the nonresponse mechanism (Sinharay, et al., 2001).

3.2.1 Data Model

The first and the most crucial step in performing multiple imputation is to relate the combination of the observed values Y_{obs} and the missing value Y_{mis} –the complete data Y- to a set of parameters. In order to achieve that goal, one has to assume a probability model. If we let $p(Y_{mis} \mid Y_{obs})$ denote the predictive distribution for the missing values conditional on the observed values, one can find this predictive distribution using the probability model and the prior distribution on the parameters. (Sinharay et al., 2001)

The predictive distribution can be written as

$$p(Y_{mis} \mid Y_{obs}) = \int p(Y_{mis}, \theta \mid Y_{obs}) d\theta \qquad (1)$$

$$= \int p(Y_{mis} \mid Y_{obs}, \theta) p(\theta \mid Y_{obs}) d\theta \qquad (2)$$

where

$$p(\theta \mid Y_{obs}) = \int p(\theta, Y_{mis} \mid Y_{obs}) dY_{mis} \qquad (3)$$

$$= \int p(\theta \mid Y_{mis}, Y_{obs}) p(Y_{mis} \mid Y_{obs}) dY_{mis} . \qquad (4)$$

$p(\theta \mid Y_{obs})$ is the observed data posterior and $p(\theta \mid Y_{mis}, Y_{obs})$ is the complete-data posterior:

$$p(\theta \mid Y_{obs}, Y_{mis}) \propto p(\theta) L(\theta \mid Y_{obs}, Y_{mis})$$

where L is the likelihood function and $p(\theta)$ is the prior distribution which will be discussed in the next section.

While the probability model assumed is based on the complete data, in most cases, the model is chosen from a class of multivariate models. For continuous variables, the multivariate normal is the most convenient because it is manageable computationally. For describing the associations among variables in cross-classified data, the loglinear model has been traditionally used. The other model the data analysts have used include a general location model which combines a loglinear model for the categorical variables with a multivariate normal regression for the continuous ones (Sinharay et al., 2001).

Because the real data rarely conform to assumed models, in most applications of multiple imputation, the model used to generate the imputations is at best only approximately true. The most convenient model for continuous variables is the multivariate normal one. However, use of multivariate model is risky. One should check whether the multivariate model fits the data approximately well. But when the variables are binary or categorical, the multivariate normal model also seems to work well in the sense that it gives quite acceptable results. Other models that data analysts have used include a log-linear model for categorical variables, and a mixture of log-linear model and a multivariate normal model for mixed continuous and categorical data sets (Sinharay et al., 2001).

3.2.2 Prior Distribution

Bayes's Theorem is the usual statistical result used to implement the model-based multiple imputation method. Under the Bayesian guidelines, a prior distribution for the parameters denoted as $p(\theta)$ is needed to carry out the analysis for the imputation model. Once we have the prior distribution, by combining it with the complete data model, we can produce the predictive distribution $p(Y_{mis} \mid Y_{obs})$ for the missing values conditional on the observed values from which, in turn, one can generate the imputation. In the Bayesian paradigm, this prior distribution quantifies one's belief or knowledge about model parameters before any data are seen (Sinharay et al., 2001).

Because of the possibilities of different results from imputation models based on different prior distributions, Bayesian methods have at times been criticized as subjective and unscientific. Nevertheless, in practice, the choice of the data model weighs more heavily than the choice of the prior distribution because results of a Bayesian procedure tend to be far more sensitive to the choice of the data model (Sinharay et al., 2001).

Usually, for convenience, a "noninformative" prior distribution is acceptable for multiple imputation. As a matter of fact, it is the default option for much statistical software (Sinharay et al., 2001). Experts have regarded it as corresponding to a state of prior ignorance about model parameters. It works well for vast majority of data analyses (Fay, Meng and Rubin, 1997).

For many data analyses, when the sample size is moderately large, the choices of the prior distributions hardly make a difference because any reasonable prior distribution will produce essentially the same results. In the situations of small samples, the choice of the prior distribution model does affect the results substantially (Sinharay et al., 2001).

.

3.2.3 The Nonresponse Mechanism

For any missing-data method, some statistical assumptions about the manner in which the missing values were lost must be made. The missingness mechanism is assumed to be MAR by most of the techniques presently available for creating multiple imputation because MAR is a convenient starting point for data analysis. On the other hand, addressing the possibility of NMAR missing data will make computations very complicated and will almost certainly be problematic (Sinharay, et al., 2001).

The MAR assumption means the missingness depends on the data values that are observed but not on the ones that are missing. Thus, one can obtain the values to impute for the missing observations based on the observed data. The MAR assumption is mathematically convenient because it allows one to avoid an explicit probability model for nonresponse. Furthermore, one has many variables that are always observed and some other variables that sometimes have missing values. Assuming MAR means that one infers how the missing variable depends on the observed variables by examining the complete cases. In other words, one defines a response indicator M that is equal to one if any of those other variables is observed and zero if it is missing. MAR implies that there exists a relationship between other variables and M only through their mutual association with the observed variables. If MAR is satisfied, then one can regress unobserved variables on the observed variables using only data for the respondents and then use the results to estimate the missing values of other variables.

Several points regarding MAR have to be mentioned. First, MAR is defined relative to the variables present in a dataset. If the observed variable on which the missingness depends is removed from the dataset, then MAR may no longer be satisfied. Second, if information about several good predictor variables that control the missingness mechanism is available, the MAR assumption tends to be more plausible because the MAR assumption depends on the available data. So one should include any characteristics that even remotely affect missingness in the model. Third, there is no way to test the MAR assumption using the data at hand because that requires the knowledge of the missing data themselves.

MAR is a popular convenient starting point for data analysis because the computation involved is relatively simple compared to NMAR. The results of multiple imputation will be invalid if the true missing data mechanism is NMAR and the missing value mechanism is not modeled.

3.3 The General Idea of Multiple Imputation

Multiple imputation is an attractive approach to analyzing incomplete data. Basically, multiple imputation is an extension of the single imputation idea.

Application of this technique requires three steps: imputation, analysis and pooling. It apparently solves the missing-data problem at the beginning of the analysis.

First, several imputed data sets that are possible representations of the data must be created. We fill in the missing entries in the incomplete data sets m $(m > 1)$ times. The imputed values are generated from the predictive distribution which is produced by combining the prior distribution and the complete data model. The predictive distribution can be different for each missing entry. We get m complete data sets at the end of this step.

Second, we analyze each of the m imputed data sets in the same fashion by a complete data method and get m results for each statistic of interest. This step is somewhat simpler than the same analysis without imputation, since there is no need to bother with missing data.

Third, the results from performing identical analyses on each of the imputed data sets are combined, using simple rules provided by Rubin and others, to produce overall estimates and standard errors that reflect missing-data uncertainty. We average the m results from each data set and get the final output. Usually, one reports the mean over m repeated analyses, the estimated error variance, a confidence interval or a p value.

3.4 Advantages of Multiple Imputation (MI)

Multiple imputation improves upon single imputation techniques which use only a single value. While it retains the advantages of single imputation, it remedies some shortcomings of single imputation.

First, multiple imputation provides a solution for one of the limitations incurred by single imputation, which replaces the missing value with some kind of a point estimate. Analyses of the completed data by single imputation don't reflect the fact that the filled-in data were not true values but estimates and contain uncertainty about the imputed values. Even if the assumptions of the models are met, the uncertainty associated with imputing data is still not properly accounted for by single imputation. (Meng, 1994).

One of the reasons that the uncertainty is not properly accounted by single imputation is that the imputed values are treated as truly observed rather than estimated values. Although the missing values may be replaced in such a way that the distributions of the variables and the relationships among the variables were not changed, the data set so obtained will still be unable to account for the uncertainty in the missing data. That uncertainty will cause the variability in the data set to be underestimated. As a result, the standard errors of the parameters would be underestimated and the type I error rate for any hypothesis test would be higher than the intended rate. In other words, the test would be positively biased. While point estimates may be unbiased, confidence intervals will be too narrow, and p-values will be too low. (Sinharay, Stern & Russell, 2001).

On the other hand, multiple imputation allows the researchers to make valid assessments of uncertainty. Under multiple imputation, several filled-in datasets are generated and analyzed by researchers separately. Thus, the multiple imputation procedure utilizes a series of complete data analyses and then combines the results. By allowing more than one value of a missing variable to be estimated, multiple imputation corrects for sampling variability. In particular, point estimates such as means, correlation coefficients are calculated by taking averages of these point estimates created by each imputed dataset. Then the variance of these estimates is also calculated by taking the average of the variances from each imputed dataset. The variation of the parameter estimates across imputations is also taken into account by between-imputation variance. Since the parameter estimates are created independently across the imputations, no covariance terms are included. The between-imputation variance or standard error reflects the uncertainty due to missing data and imputation. Also the uncertainty is reflected by wider confidence intervals and larger p-values than under single imputation.

Wayman (2003) has mentioned another advantage of multiple imputation. It is very user friendly and familiar to many researchers. It works in conjunction with standard complete-data methods and software. Basically, multiple imputation represents repeated conditional draws under a model for nonresponse. Valid inferences are obtained simply by combining complete data in a straightforward manner. That means that once the values have been filled in, standard complete data methods of analysis can be used. First, the researchers use multiple imputation to produce full, complete datasets. Then analyses are performed on these datasets. And these analyses can be carried out using procedures in SAS, SPSS or virtually any method or software package the analyst chooses. Although the statistical principles behind multiple imputation may appear nontrivial to some researchers, user-friendly software helps researchers concentrate on learning and implementing the process of multiple imputation rather than the underlying statistical theory.

Multiple imputation can incorporate any of several imputation strategies, such as regression-based methods. But the actual method applied depends on the pattern of the missing data. For example, if the pattern of missing data is complex, one can impute one missing variable by regression using all available information. Then one can use that imputed variable to impute other missing variables.

The assumption in multiple imputation that missing data are MAR rather than MCAR comes in handy because the missing values rarely occur completely at random. In the EEO-1 reports, it is quite reasonable to assume that nonrespondents differ from respondents in some way.

As mentioned before, in many applications, only a relatively few imputations are required—just three to five imputations are sufficient to obtain excellent results. In the process of randomly drawing to impute in an attempt to represent the distribution of the data, multiple imputation increases the efficiency of estimation. Also, even when data collectors don't use explicit models, it is still more computationally efficient for statistically sophisticated data user to simulate the correct inference by using multiple imputation (Rubin, 1987).

3.5 Concepts of Multiple Imputation

The theoretical motivation for multiple imputation is Bayesian. To illustrate the procedure, first introduce some notations. Let Q denote the quantity to be estimated such as a mean, correlation, regression coefficient or odds ratio. Let Y_{mis} denote the missing data and Y_{obs} denote the observed data. Thus, the complete data set is (Y_{obs}, Y_{mis}) . Let $\hat{Q} = \hat{Q}(Y_{obs}, \hat{Y}_{mis})$ denote the imputed data estimate of Q and $\text{var}(\hat{Q}) = \text{var}(\hat{Q}(Y_{obs}, \hat{Y}_{mis}))$ denote the estimated variance of $\hat{Q}$.

There are two steps in imputing Y_{mis}. First, we simulate a parameter value from the observed data posterior $p(\theta \mid Y_{obs})$. It requires carrying out a traditional Bayesian analysis with missing data. Second, we simulate a missing data vector from the conditional posterior distribution $p(Y_{mis} \mid Y_{obs}, \theta)$ using the value θ generated in the first step. So m values are imputed for each data set with missing values and $m > 1$ independent simulated imputed data sets $(Y_{obs}, \hat{Y}_{mis}^{(1)}), (Y_{obs}, \hat{Y}_{mis}^{(2)})...(Y_{obs}, \hat{Y}_{mis}^{(m)})$ are produced.

Once the imputed data sets have been constructed, the analysis is carried out separately for each data set. This analysis can be done by any standard complete-data method. As a matter of fact, this analysis can proceed just as if there are no missing data although a separate analysis is performed on each imputed data set. The analysis enables us to calculate each of the imputed-data estimates $\hat{Q}^{(t)} = \hat{Q}(Y_{obs}, \hat{Y}_{mis}^{(t)})$ along with their estimated variances $\hat{\text{var}}(\hat{Q}^{(t)}) = \hat{\text{var}}\left[\hat{Q}(Y_{obs}, \hat{Y}_{mis}^{(t)})\right]$.

Next, once the analyses have been completed for each imputed data set, what is left is to combine these analyses to construct one overall set of imputed-data estimates. These estimates are produced as in a non-imputation analysis. Following the rules established by Rubin (1987), the point estimate for Q is simply the average (Schafer, 1999):

$$\hat{Q}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\hat{Q}^{(t)}.$$

The total variance of $\hat{Q}_{MI}$ consists of two parts: the “between-imputation” variance and the “within-imputation” variance.

The “between-imputation” variance is

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{Q}^{(t)} - \hat{Q}_{MI})^2.$$

The “within-imputation” variance is

$$W = \frac{1}{m}\sum_{t=1}^{m}\hat{\text{var}}(\hat{Q}^{(t)}),$$

where $\hat{\text{var}}(\hat{Q}^{(t)})$ is an estimated variance based on imputed data set t, $t = 1,...,m$

The estimated total variance is

$$T = (1+\frac{1}{m})B + W .$$

The part of T involving B measures uncertainty introduced by imputing missing data. The statistic B measures the variation of the point estimates from data set to data set. If the estimates vary significantly from data set to data set, then the uncertainty due to imputation is high and B is huge. But if the estimates are very similar to each other, the uncertainty is less and B is small. The statistic W measures the natural variability within the data. It is produced by averaging the variance estimates from each imputed data set. The total T will equal to W if and only if Y_{mis} carries no information about Q, and thus the imputed data estimates $\hat{Q}^{(t)}$ will be equal.

The quantity $(Q-\hat{Q}_{MI})T^{-1/2}$ is approximately distributed as Student's t with v_m degrees of freedom (Rubin 1987), where

$$v_m = (m-1)\left[1+\frac{W}{(1+m^{-1})B}\right]^2 .$$

Theoretical justification for the above analyses can be found in Rubin and Schenker (1986) and the references given there.

When the complete-data degrees of freedom $v_0 = N-1$ is small and there only a modest proportion of missing data, the computed degrees of freedom, v_m, can be much larger than v_0, which is inappropriate. The use of an adjusted degrees of freedom is recommended by Barnard and Rubin (1999):

$$v_m^* = \left[\frac{1}{v_m}+\frac{1}{\hat{v}_{obs}}\right]^{-1}$$

where $\hat{v}_{obs} = (1-\gamma)v_0(v_0+1)/(v_0+3)$ and $\gamma = (1+m^{-1})B/T$.

When v_m^* is large, the distribution of $(Q-\hat{Q}_{MI})T^{-\frac{1}{2}}$ will be approximately normal.

It is said that for multiple imputation, only 3-5 imputations are needed. However, Rubin and Schenker (1986) have pointed out that the required number of imputation is related to the amount of missing information in the dataset. If the amount of missing information is modest (e.g., less than 30%) as shown in Table 1, then 3 to 5 imputations are enough. But as the percentage of missing data increases, more imputations will be needed.

Rubin (1987, p. 114) showed that the efficiency of an estimate based on m imputations is approximately

$$(1+\frac{r}{m})^{-1}$$

where r is the fraction of missing information for the quantity being estimated and $r=\frac{(1+m^{-1})B}{W}$ and efficiency is one measure of desirability of an estimator.

The fraction r measures how much more precise the estimate would have been if no data had been missing. Table 1 shows the efficiency attained by different m and r values. From this table, it is obvious that the gain in efficiency is reduced rapidly after the first few imputations. Consider the column for 30% missing information ($r=0.3$), a usual rate for many applications. With $m=4$ imputations, the efficiency has already reached 0.93. Increasing the number to $m=8$ will only raise the efficiency to 0.96. While the computational effort doubles, the gain is slight. In most situations, there is simply no point to produce and analyze more than 5 imputed datasets.

Table 1: Efficiency of multiple imputation by number of imputations m and fraction of missing information r. (http://support.sas.com/md/app/paper/miv802.pdf.)

		r			
m	0.1	0.3	0.5	0.7	0.9
3	0.96	0.91	0.86	0.81	0.77
4	0.97	0.93	0.89	0.85	0.82
5	0.98	0.94	0.91	0.88	0.85
8	0.98	0.96	0.94	0.91	0.90
10	0.99	0.97	0.95	0.93	0.92
15	0.99	0.98	0.96	0.96	0.94
20	1.00	0.99	0.98	0.97	0.96

3.6 Key Features of Multiple Imputation

The first key feature of multiple imputation is that it incorporates into users' analyses the uncertainty disregarded by single imputation in order to obtain valid statistical inferences. With single imputation, one imputes one and only one logically possible value for each missing observation. One cannot be certain about the one imputed value because if one could, it wouldn't be missing. Multiple imputation deals with that problem by inserting more than one value for each missing observation in a way that reflects the uncertainty (Meng, 1994).

The second key feature of multiple imputation is the separation between the imputation phase and the analysis phase. One might think that the data collector must be the best person to do imputations because he/she has much better knowledge about the data. In reality, in many situations, the imputation is done by one person and the final analysis may be done by some other users who employ the imputed dataset. Although probably only one person collects the data, the imputed dataset may be shared by many other users (Sinharay et al., 2001).

The third key feature of multiple imputation is also true for single imputation: it is a user-friendly procedure. It removes as many burdens as possible from users facing incomplete data so what the users have is the imputed dataset obtained by the data collector. Once the users have the dataset, they can use any complete-data technique to do the final analysis. That procedure is a trivial task for computers—it is available in any standard statistical software, because it only requires repeating the same standard complete-data analyses several times. Although sometimes some additional computation is needed for combining the complete-data analysis, it usually only requires simple arithmetic and looking at standard statistical tables. Most importantly, the users don’t need to worry about the missingness mechanism because they have the access to the completed datasets.

Rubin (2002) has argued that multiple imputation is the ideal choice when the imputer and the ultimate users are totally different persons. The imputer is responsible for obtaining correct imputed values by using his/her knowledge so that the users will more likely to be able to come up with valid results when they employ the complete data.

In addition, some experts say that the replications of artificially completed datasets provide reasonable ground for familiar analysis. If the imputation is done correctly, it will give reliable indications of the directions and sizes of needed adjustments for missingness.

Last, Fay, Meng and Rubin (1992) has investigated the behavior of repeated-imputation inference when the imputer’s and analyst’s models differ. They mention that when the imputer’s model is more general in the sense of making fewer assumptions than the analyst’s, multiple imputation leads to valid inferences with some loss of power caused by additional generality of adding extra variation among the imputes. On the other hand, if the imputer makes more assumptions than the analyst and his assumptions are correct, the final estimate becomes more precise than any estimate derived from the observed data and analyst’s model alone.

3.7 The Methods in the SAS Proc MI procedure

Multiple imputation is available in SAS v.9 in the procedures PROC MI and

PROC MIANALYZE. PROC MI is an experimental procedure in SAS 8.1. The Markov Chain Monte Carlo (MCMC) method, regression method and propensity score method are the three methods available in the PROC MI procedure in SAS 9.1 and 8.1.

In MCMC, there are two steps. In the imputation I-step, I denote the variables with missing values for observation i by $Y_{i(mis)}$ and the variables with observed values by $Y_{i(obs)}$. Then the I-step draws values for $Y_{i(mis)}$ from a conditional distribution $Y_{i(mis)}$ given $Y_{i(obs)}$ (Yuan, 2000). In other words, SAS generates a sample of $Y_{1(mis),}Y_{2(mis).}$ ….. from the approximation to the predictive density $p(Y_{mis} \mid Y_{obs})$.

The posterior P-step simulates the posterior estimates. It updates the current approximation to $p(\theta \mid y_{obs})$ to be the mixture of conditional densities of Q given the augmented data pattern generated in the I-step. That is, (Tanner & Wong, 1987)

$$g_{i+1}(\theta) = m^{-1} \sum_{j=1}^{m} p(\theta \mid y_{mis}^{(j)}, y_{obs})$$

These new estimates that were generated in the P-step are then used in the I-step. Without prior information about the parameters, a noninformative prior is used. Other informative priors also can be used (Yuan, 2000).

The two steps are iterated long enough for the results to be reliable for a multiply imputed data set. With a current parameter estimate $\theta^{(t)}$ at the t^{th} iteration, the I-step draws $Y_{mis}^{(t+1)}$ from $p(Y_{mis} \mid Y_{obs}, \theta^{(t)})$ and the P-step draws $\theta^{(t+1)}$ from

$p(\theta \mid Y_{obs}, Y_{mis}^{(t+1)})$. This creates a Markov Chain $(Y_{mis}^{(1)}, \theta^{(1)})$, $(Y_{mis}^{(2)}, \theta^{(2)})$, …,which converges in distribution to $p(Y_{mis}, \theta \mid Y_{obs})$ (Yuan, 2000).

In the regression method, a regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the resulting model, a new regression model is then simulated and is used to impute the missing values for each variable (Rubin, 1987).

To use the regression method, the data set must have a monotone missing data pattern. The process is repeated sequentially for variables with missing values. That is, for a variable Y_j with missing values, a model

$$Y_j = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_{(j-1)} Y_{(j-1)}$$

is fitted with the nonmissing observations (Yang, 2000).

The fitted model has the regression parameter estimates $(\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_{(j-1)})$ and the associated covariance matrix $\sigma_j^2 V_j$, where V_j is the usual X'X matrix from the intercept and variables $Y_1, Y_2, ..., Y_{(j-1)}$ (Yang, 2000).

For each imputation, new parameters $(\beta_{*0}, \beta_{*1,} ..., \beta_{*(j-1)})$ are simulated from $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{(j-1)}), \sigma_j^2$ and V_j. The missing values are then replaced by

$$\beta_{*0} + \beta_{*1} y_1 + \beta_{*2} y_2 + ... + \beta_{*(j-1)} y_{(j-1)} + z_i \sigma_{*j}$$

where $y_1, y_2, ..., y_{(j-1)}$ are the covariate values of the first $(j-1)$ variables and z_i is a simulated normal deviate.

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariate. In this method, a propensity score is generated for each variable with missing values to indicate the probability of the observation being missing. The observations are then grouped based on these propensity score, and an approximate Bayesian bootstrap imputation is applied to each group (Rubin, 1987).

The following steps are used to impute values for each variable Y_j with missing values: (Yuan, 2000).

1. Create an indicator variable R_j with the value 0 for observations with missing Y_j and 1 otherwise.
2. Fit a logistic regression model

   $$\log it(p_j) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_{(j-1)} Y_{(j-1)}$$

   Where $p_j = \Pr(R_j = 0 \mid Y_1, Y_2, ..., Y_{(j-1)})$ and $\log it(p) = \log(p/(1-p))$.
3. Create a propensity score for each observation to indicate the probability of its being missing.
4. Divide the observations into a fixed number of groups based on these propensity scores.
5. Apply an approximate Bayesian bootstrap imputation to each group.

Chapter 4: Artificial Data Simulation

In this simulation study, I create the missingness artificially. Thus, the "missing" values are known. I impute these using various imputation methods. Since the true values are known, we can compare the performance of imputation methods to see which is closer to the truth. The simulated data setup roughly resembles the EEO-1 data of Chapter 6.

4.1 Simulation Description

First I assume there are $L=10$ strata, indexed by $h=1,...,10$. Then I generate a dataset containing y_{hi}, N_{hi}, and M_{hi}, $i=1,...,n_h=100$. The variable y is the number of minority employees. Thus, $y_h = \sum_i y_{hi}$ is the number of minority employees in each stratum. I assume y_{hi} has a binomial distribution given N_{hi} with stratum probabilities $Q_h = 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14$, respectively. The variable N is the total number of employees. Then, $N_h = \sum_i N_{hi}$ is the total number of employees in each stratum. I assume it follows an approximate normal distribution with mean 100 and standard deviation 20. The variable $M_{hi} = 1$ if y_{hi} is missing and 0 otherwise. Since the model I assume has a MAR mechanism, the probability of missing depends on the stratum; that is $P_h = P[y_{hi}$ is missing |stratum h]= $P[M_{hi} = 1 \mid \text{stratum } h]$ is 0.25, 0.35, 0.4, 0.45, 0.53, 0.51, 0.2, 0.29, 0.1, 0.29, for h=1,…,10, respectively. I generate data y_{hi}, N_{hi} and M_{hi}, i=1,…, n_h with n_h=100. The resulting dataset contains 100 values of y, N, M in each stratum.

Based on the dataset generated, let Q_h be the proportion of minority employees in stratum h and let $Q = \sum W_h * Q_h$, where $W_h = 0.10$, the weight for each stratum. Now let $\hat{Q}_1$ be the estimator of Q from S_1, the dataset after deleting the incomplete observations; let $\hat{Q}_2$ be the estimator of Q from S_2, the complete dataset before causing some y values to be missing; let $\hat{Q}_3$ be the estimator of Q from S_3, the dataset with values imputed by mean imputation, and let $\hat{Q}_4$ be the estimator of Q from S_4, the dataset with values imputed by multiple imputation. The goal is to calculate the bias and variance of $\hat{Q}_1, \hat{Q}_2, \hat{Q}_3$ and $\hat{Q}_4$.

The simulation proceeds as follows.

For each of $R = 100$ generated samples, $r = 1,...,R$:

- Compute $\hat{Q}_{11}, \hat{Q}_{12}.........\hat{Q}_{1r}; \hat{Q}_{21,}\hat{Q}_{22}.........\hat{Q}_{2r}; \hat{Q}_{31,}\hat{Q}_{32}.........\hat{Q}_{3r.}$ and $\hat{Q}_{41,}\hat{Q}_{42}....\hat{Q}_{4r.}$
- For dataset S_j, let $\hat{Q}_{jr} = \sum W_h \hat{Q}_{hjr}$ and

  $\hat{Q}_{hjr} = \sum_{S_j} y_{hi} / \sum_{S_j} N_{hi}$,i=1,2…… n_h=100, h=1,2……10, j=1,2,3.

  $W_h = 0.10$ is the weight for each stratum.

- To estimate bias, let $\frac{1}{R}\sum \hat{Q}_{1r} = \tilde{Q}_1$, $\frac{1}{R}\sum \hat{Q}_{2r} = \tilde{Q}_2$, $\frac{1}{R}\sum \hat{Q}_{3r} = \tilde{Q}_3$.

  So for each of the three datasets, the Monte Carlo bias is $\tilde{Q}_j - Q$, j=1,2,3.

- The Monte Carlo variance is $\text{var}(\hat{Q}_j) = \frac{1}{R-1}\sum(\hat{Q}_{jr} - \tilde{Q}_j)^2$, j=1,2,3.

- We look at the average estimated variance, $\hat{\bar{V}}(\hat{Q}_j) = \frac{1}{R}\sum \hat{V}_r(\hat{Q}_j)$, j=1,2,3,

  and compare them to the Monte Carlo variances above, namely

  $\text{var}(\hat{Q}_1), \text{var}(\hat{Q}_2)$ and $\text{var}(\hat{Q}_3)$. Here, $\hat{V}_r(\hat{Q}_j) = \sum_h W_h^2 \hat{V}_h(\hat{Q}_{hjr})$, j=1,2,3.

  h=1,2…10.

  where $\hat{V}_h(\hat{Q}_{hjr}) = \sum_i (y_{hi} - \hat{Q}_{hjr} N_{hi})^2 / (n_h \bar{N}_h^2)$, i=1,2… $n_h = 100$, h=1,2…10.

  j=1,2,3.

- Given r, for multiple imputation, we generate m=3 imputed data sets

  (k=1,…,m) and compute

  $\hat{Q} = \frac{1}{m}\sum \hat{Q}_k$ and $\hat{Q}_k = \sum_h W_h \hat{Q}_{kh}$, where $\hat{Q}_{kh} = \sum_i y_{hi} / \sum_i N_{hi}$,

  i =1,2… $n_h = 100$;

  $\hat{v}_k = \sum_i (y_{hi} - \hat{Q}_{kh} N_{hi})^2 / (n_h \bar{N}_h^2)$, i =1,2… $n_h = 100$;

  $\hat{V}_w = \sum_h W_h^2 \sum_k \hat{v}_k / m$ and $\hat{V}_b = \sum_h W_h^2 \sum_k (\hat{Q}_{kh} - \hat{\bar{Q}}_h)^2 / (m-1)$.

4.2 Outputs

Table 2 summarizes the output of the simulation exercise.

Table 2 : Output of Simulation Exercise

	Mean	Bias	var (estimate)	$\hat{\hat{V}}_r$	Bias/se*	MSE**
Deleted	0.09502	0.00002	0.005240	0.0007357	0.000874	0.000524
No Missing	0.09496	-0.00004	0.000828	0.0008438	-0.00139	0.000828
Mean	0.09500	0	0.000815	0.0006869	0	0.000815
Multiple	0.09699	0.00199	0.000816	0.000883	0.069664	0.000819

*se = sqrt (var(estimate)).

**MSE=var(estimate) +(Bias)^2

While the true mean is 0.095, the means for the deleted, complete, mean-imputed and multiple-imputed data methods are 0.09502, 0.09496, 0.095 and 0.09699. They are very close to the true mean. Thus, the bias is very small in each case, but much higher for multiple imputation than for other methods.

The estimated variance $\hat{\hat{V}}(\hat{Q}_2)$ is 0.0008438 and the Monte Carlo variance var($\hat{Q}_2$) is 0.000828 for the complete data method. As expected, they tend to agree.

For the complete data method $\hat{\hat{V}}(\hat{Q}_2)$ is 0.0008438 and for the mean-imputation method $\hat{\hat{V}}(\hat{Q}_3)$ is 0.0006869. Thus, one underestimates the variance using mean imputation. In contrast, $\hat{\hat{V}}$ is reasonably accurate if one uses multiple imputation.

Bias/se gives an idea of how important the bias is compared to the sample variability. The bias/se for multiple imputation method is 0.069664, which is the largest. But since all of them are less than 0.2, the calculations of the confidence intervals based on normal distribution can be reasonably assumed to be reliable.

The MSE's correspond to the bias/se. They are all relatively small. And for complete data , mean imputation and multiple imputation methods, they are 0.000828, 0.000815, and 0.00819. They are very similar.

Chapter 5: Equal Employment Opportunity Data

The Equal Employment Opportunity Commission (EEOC) collects periodic (annual and/or biennial) reports from public and private employers, unions and labor organizations in the United States as required by Title VII of the Civil Rights Act of 1964. With its headquarters in Washington, D.C., and through the operations of 51 field offices nationwide, EEOC coordinates all federal equal employment opportunity regulations, practices and policies. The Commission interprets employment discrimination laws, monitors and conducts hearings in the federal sector employment discrimination program, sponsors outreach and technical assistance programs, and provides funding and support to state and local Fair Employment Practices Agencies charged with enforcing anti-discrimination laws on state and local levels.

5.1 Structure of EEO-1 Data

The Equal Employment Opportunity Commission (EEOC) collects periodic reports from firms and organizations all over the United States. The EEO-1 report captures race, ethnicity, and gender information on employees. The data collected is entered into a computerized database. After the data are entered, EEOC makes sure the numerical values are consistent. For example, the total number of employees must be equal to the sum of the numbers of female and male employees and must be equal to the sum of the numbers of employees of each ethnicity. If there is any discrepancy, the report is sent back to the company to be corrected. Because of this “clean-up,” the final result doesn’t have any missing data. All the missing values are zeros.

The EEO-1 report is filed by all private employers with 100 or more employees and by firms with fewer than 100 employees which meet other conditions. The data collected in the EEO-1 report are the following: company ID number, status code which indicates type of reports, unit number, unit name, unit address, city name, state, zip code. It also collects numbers of employees in each combination of ethnic group, gender and job category. There are five ethnic groups: Asian, Indian, Hispanic, Black and White; two genders: male and female; and nine job categories: officials and managers, professionals, technicians, sale workers, office and clerical, craft workers (skilled), operatives (semi-skilled), laborers (unskilled), and service workers.

5.2 Discussion of Missing Data Mechanism

The EEO-1 data have numeric as well as character variables. I am interested in the numeric variables: the numbers of employees in the various racial and occupational categories. I intend to examine the numbers of employees from different ethnicities. But first, I will discuss possible missing data mechanisms.

The MAR assumption assumes that the relationship between missing data mechanism and missing values can be sufficiently explained by data that are observed. For example, in EEO-1, among the observed data is the location of the working place: south, west, middle west, etc and NAICS and SIC codes. It is known that more Asian immigrants are in the west and east coasts, such as California and New York, and more Hispanics are in major metropolitan areas. More Asian immigrants are in the restaurant business and more Hispanics are in construction. If the working places in these locations and industries tend not to report Chinese or Hispanic employees, the number of minority employees may be MAR because it depends on the location of the working place which is an observed value. But one can never be sure. Under the MAR assumption, if the measured cause of missingness is included properly in the analysis, all biases associated with the missing data are adjusted. This discussion holds for a given NAICS or SIC code and region, but not across the board.

If the missing data values are unrelated to the value itself or to any other variables in the data set, the missing data mechanism is MCAR. In other words, the missing data values are a simple random sample of all data values. The number of minority employees may be MCAR if the following two conditions are satisfied. One, the employers who do not report their number of minority employees have, on the average, the same number of minority employees as the employers who report. Two, each of the other variables in the data set such as the number of majority employees would have to be the same, on average, for the employers who do not report their number of minority employees and the employers who report their number of minority employees. Again, there is no certainty. The big advantage of data being MCAR and MAR is that the cause of missingness does not have to be part of the analysis to control for missing data biases.

The missing data for a variable are NMAR or "nonignorable" if the probability of missing data on the variable is related to the value of that variable even if other variables in the analysis are controlled. This situation would arise when the value of the missing variable is itself a cause of missingness. So if the employers who are more likely to have minority employees tend not to report the number of their minority employees, then the number of minority employees may be NMAR. However, we cannot be sure.

The numbers of employees in various ethnicities are collected in the EEO-1 report. But that information might be less likely to be obtained for minority employees. The EEO-1 data may be assumed to be MAR. In the cases where the numbers of minority employees are missing, we can predict the pattern of the missingness from other variables such as locations of the companies or the industries rather than from the numbers of minority employees.

However, we never can be sure because with large number of Hispanic workers, the companies hiring illegal Hispanics may not want to admit it and may tend not to report their numbers of Hispanic workers. In that case, the missingness may be NMAR.

5.3 EEOC's Method for Handling Missing Values

The EEO-1 report is filed by the public and private employers, unions and labor organizations in the United States each year. After the reports are received by the EEOC office, the information on the reports is entered into a database. EEOC's work sometimes concerns the exit companies in the EEO-1 report from one year to the next. For example, the companies whose identification numbers were present in 2002 but not in 2003 are classified as 2003 exits. The number of employees working at the exit companies in 2003 is unknown. According to a report written by Cartwright (2005), EEOC's method for handling this missing value is to substitute employee numbers from the companies with the same identification number in 2002 for 2003. The problem with this method is that the workforce estimates from 2002 may overestimate the number of employees in 2003 due to possible loss of employees through layoffs, mergers and buy-outs.

Sometimes the absence of a firm is analogous to "death:" a firm is liquidated. Sometimes parts of a firm are reorganized as the results of mergers, or buy-outs. Sometimes a firm decreases total employment over time, not fulfilling the EEO-1 reporting requirements anymore. Sometimes a firm changes its name. Some are nonrespondents—those simply not reporting. There is no way for EEOC to know the cause for each exit. However, EEOC does do follow-ups, first by telephone and then by letters. While some firms will reply, others won't. For those that won't, EEOC simply labels them as exit firms.

The EEO-1 data set of year 2003 contains 221,289 companies and 252 variables. It is unlikely that each of 252 variables for all companies is observed. If a value is missing, for example, if the column in the EEO-1 report for Hispanic male employees is missing, when it is converted into computer database, the value for this column will be recorded as zero. It has been suggested to eliminate cases with missing value which is zero. This method is called listwise deletion or complete case analysis. Although it is often the default option for analysis in many statistical software packages, there are several drawbacks of this method.

Also, pairwise deletion appears to be making use of all available data, but it is not a desirable procedure. Overall, the disadvantages of listwise deletion and pairwise deletion outweigh their advantages as discussed in Chapter 3.

Chapter 6: Simulations Based on EEO-1 Data

Since the EEO-1 data published by EEOC is already cleaned up, there is no missing data in the data set. I artificially introduce missingness by simulating unit nonresponse and item nonresponse. Then I apply various imputation schemes to study their performance.

6.1 SIC Code

In the 2003 EEO-1 data, there are SIC codes which range from two-digits to four digits. SIC code stands for Standard Industrial Classification code. It indicates the company's type of business and is used as a basis for assigning review responsibility for the company's filing. I use the two-digit SIC code as the basis of my imputation scheme. Table 3 shows which industry each SIC code represents.

Table 3. Classification of Two Digit SIC codes

SIC code	Industry
01 – 09	Agriculture, Forestry, and Fishing
10 –14	Mining
15 – 17	Construction
20 – 39	Manufacturing
40 – 49	Transportation, Communication, Electric, Gas and Sanitary

	Service
50 – 51	Wholesale Trade
52 – 59	Retail Trade
60 – 67	Finance, Insurance and Real Estate
70 – 89	Service
91 – 97	Public Administration
99	Nonclassifiable Establishments

Although EEOC collects reports from government agencies in the United States, the EEO-1 dataset doesn't contain the SIC codes from 91 to 97, which correspond to public administration.

6.2 Simulation Descriptions and Outputs

6.2.1 First Scenario: Simulated Unit Nonresponse

The EEO-1 report identifies each company by establishment identification numbers (UNIT_NBR). Establishment identification numbers present in 2002 but not in 2003 are classified as 2003 exits. The number of 2003 exits is 6557. Recall from the previous chapter that exits are imputed to have the same numbers of employees as in the previous year. Among the 6557, some are out of business—no longer exist; some change names;others are simply not reporting. Those simply not reporting are treated as nonrespondents. Because we don't know the number of exits due to each cause, we first arbitrarily select a random sample of 30% of the 6557 total 2003 exit companies, which is rounded to 1967, to be treated as nonrespondents. Then we divide the companies into imputation classes based on their two-digit SIC code. The size of the companies, namely the number of employees, varies greatly from 100 to more than 20000. The number of large companies is less than the number of small companies, but their numbers of employees are significantly bigger. So to count that impact, we further subdivide the companies within each SIC code into the top 10% and the remaining 90% based on total employment. Then we impute the average number of employees of respondents for the simulated nonrespondents within each imputation class, and we calculate the sum, mean and standard deviation for each variable.

Since 30% nonresponding is an arbitrary value we pick, we do the same simulation based on 50% and 70%, or 3279 and 4590 nonrespondents, respectively, and apply the same procedure as above. The computer I used runs Windows XP 2002 and has 1 Gbyte system memory. The software I used is SAS 8.1. The output of the three different percentages for Asian employees is listed in Table 4., Table 5 and Table 6.

Table 4 Mean Imputation by SIC and Size for Asians with 30% nonrespondents

Label	Sum	Mean	Std Dev
Asian Off and MGRS	351489.36	1.5743781	17.2311871
Asian Professionals	1283724.06	5.7500092	72.366424
Asian Technicians	311691.63	1.3961176	17.6945123
Asian Sales Workers	454461.17	2.0356056	65.9986294
Asian Office and Clericals	517885.23	2.3196923	37.6980647
Asian Craft Workers	148418.08	0.6647888	9.9577902
Asian Operatives	372053.85	1.6664898	16.9330711
Asian Laborers	222757.77	0.9977683	20.6443460
Asian Service Workers	440359.99	1.9724441	36.9315209

Table 5 Mean Imputation by SIC and Size for Asians with 50% nonrespondents

Lable	Sum	Mean	Std Dev
Asian Off and MGRS	357621.85	1.5924880	17.1838797
Asian Professionals	1305340.74	5.8126747	72.1631148
Asian Technicians	316943.07	1.4113456	17.6447769
Asian Sales Workers	462727.52	2.0605230	65.8073188
Asian Office and Clericals	526966.66	2.3465795	37.5910398

Asian Craft Workers	150910.84	0.6720051	9.9294775
Asian Operatives	378271.08	1.6844389	16.8864875
Asian Laborers	226395.30	1.0081370	20.5848253
Asian Service Workers	447799.77	1.9940498	36.8255874

Table 6 Mean Imputation by SIC and Size for Asians with 70% nonrespondents

Label	Sum	Mean	Std Dev
Asian Off and MGRS	364322.08	1.6129082	17.1373542
Asian Professionals	1328728.67	5.8824799	71.9626244
Asian Technicians	322620.09	1.4282872	17.5957299
Asian Sales Workers	471868.61	2.0890327	65.6180051
Asian Office and Clericals	536859.12	2.3767553	37.4853756
Asian Craft Workers	153643.78	0.6802039	9.9015169
Asian Operatives	385096.58	1.7048800	16.8406297
Asian Laborers	230388.60	1.0199647	20.5259572
Asian Service Workers	455820.35	2.0179846	36.7208771

The common trend is that the more missing values—thus, the more we impute, the higher the resulting population means. The explanation is that we did mean imputation separately by SIC code , then size. The imputed companies happen to be mostly from high employment industries. Therefore, our imputed means are greater than our original overall population average. So the more we impute, the higher our resulting population means.

As a check, in the 2003 EEO-1 dataset, for each industry group by two-digit SIC code and by size, we find the means of employments for each group before imputation. And for the 6557 companies exiting in 2003, we find the proportion of companies in each group by SIC, then size. As a result, proportion of exit companies is larger in the groups which have higher means of employments before imputation. It confirms the mean imputation results.

Imputation requires modeling of the individual outcomes, which is more work particularly when there are a lot of survey variables. Here, we have unit nonresponse but not item nonresponse. I only use mean imputation but not multiple imputation because multiple imputation is more useful for item nonresponse than unit nonresponse, although it is possible to use it for the latter as well. The reason is that multiple imputation yields gains in efficiency, and the gains are greatest when there is good predictive covariate information, which is typically the case with item nonresponse rather than unit nonresponse (Little, 2006).

6.2.2 Second Scenario: Simulated Item Nonresponse

The EEO-1 dataset of 2003 has 221,289 observations, namely the companies, and 252 variables. Using SAS, I delete the character variables such as unit_nm, address, etc. and only concentrate on the numerical variables which are the employment of all races and ethnicities. There are 198 of them. I am particularly interested in the numbers of minority employees who are Asian, Indian, Black and Hispanic, categorized by each occupational category, such as officer, professional, technician, etc, for every minority group. Arbitrarily, I randomly selected 10% of 221,289 which is rounded to 22,129 observations on those 36 variables represent the minority employees and replace the actual data as missing values. Now I have a dataset that has item nonresponse because 22,129 of 221, 289 observations on 36 out of 198 variables are replaced by missing value codes.

The dataset has 221,289 observations and 252 variables. The computer I used runs Windows XP 2002 and has 1 Gbyte system memory. The software I used is SAS

8.1 which has PROC MI and PROC MIANALYZE as experimental procedures. Since the companies are divided based on two-digits SIC code, I arbitrarily picked codes in the 40’s which correspond to transportation, communication, electric, gas and sanitary service and also codes in the 60’s which correspond to finance, insurance and real estate. It takes about 45 minutes to complete one run for the SIC codes I pick. It would take excessively long to apply imputation to the entire dataset, especially multiple imputation. I chose to average the results of 22 datasets because 22 is manageable in terms of computer time (about 9 hours), but nevertheless big enough to show difference from 1 or 2 datasets if there is any.

Then, I used multiple imputation on the resulting dataset containing SIC code 40’s and 60’s. The multiple imputation procedure, PROC MI, was new and experimental in SAS Release 8.1 but is part of the current SAS Release 9.1. There are three methods available in the MI procedure. They are regression method, propensity score method and Markov Chain Monte Carlo (MCMC) method. The method of choice depends on the type of missing data pattern. Since I have an arbitrary missing data pattern, I used the MCMC method, which creates multiple imputation by using simulations from a Bayesian prediction distribution for normal data.

PROC MI creates an output data set containing 3 imputed versions of the original dataset. In each version, the missing values are replaced with imputed values. I repeated the same procedure 22 times. Then I combined the results and calculated the average of the 22 means and 22 standard errors for each variable as well as the standard deviation of the average.

Since 10% of the 221,289 observations, which counts to 22,129, is an arbitrary value we pick, we might as well also choose 30% and 50%, which amounts to 66,386 and 110,645 missing values, respectively, and then apply the same procedure as above.

In order to compare the results of mean imputation to multiple imputation, I only select the output of mean imputation for SIC code 40's and 60's. The output of the mean imputation and multiple imputation for 10% missingness for Asian employees is listed in Table 7, Table 8, Table 9 and Table 10.

Table 7 Mean of Asians for Mean Imputation of SIC 40 and 60 with 10% Missing

Label	Mean	Std Dev
Asian Off and MGRS	1.0816300	0.00140750
Asian Professionals	4.0070435	0.0723972
Asian Technicians	0.8966541	0.0110980
Asian Sales Workers	0.3976459	0.0098066
Asian Office and Clericals	0.9323619	0.0075525
Asian Craft Workers	0.9216872	0.0151852
Asian Operatives	2.69979983	0.0202666
Asian Laborers	0.7804796	0.0105382
Asian Service Workers	0.2710785	0.0156026

Table 8 Standard Error of Asians for Mean Imputation of SIC 40 and 60 with 10% Missing

Label	Mean	Std Dev
Asian Off and MGRS	0.0366470	0.0023455
Asian Professionals	0.2086630	0.0196810
Asian Technicians	0.0332877	0.0014245
Asian Sales Workers	0.0229449	0.000826929
Asian Office and Clericals	0.0241572	0.000312928
Asian Craft Workers	0.0353726	0.0019168
Asian Operatives	0.0706273	0.0013654
Asian Laborers	0.0310738	0.000636506
Asian Service Workers	0.0511671	0.0093649

Table 9 Mean of Asians for Multiple Imputation of SIC 40 and 60 with 10%Missing

Label	MonteCarlo Average	Std Dev
Asian Off and MGRS	1.0831580	0.0073777
Asian Professionals	4.0213184	0.0306867
Asian Technicians	0.9014850	0.0082997
Asian Sales Workers	0.3976574	0.0076026
Asian Office and Clericals	0.9318953	0.0084804
Asian Craft Worker	0.9231473	0.0158133
Asian Operatives	2.7054532	0.0291950
Asian Laborers	0.7800523	0.0119454
Asian Service Workers	0.2665542	0.0124755

Table 10 Standard Error of Asians for Multiple Imputation of SIC 40 and 60 with 10% Missing

Label	MonteCarlo Average	Std Dev
Asian Off and MGRS	0.0390964	0.0011000
Asian Professionals	0.2219334	0.0117617
Asian Technicians	0.0365895	0.0018677
Asian Sales Workers	0.0252045	0.0016182
Asian Office and Clericals	0.0262856	0.000862865
Asian Craft Workers	0.0392547	0.0028408
Asian Operatives	0.0787201	0.0042130
Asian Laborers	0.0346617	0.0020712
Asian Service Workers	0.0538644	0.006179

First, by looking at the tables with 10% missingness, the Monte Carlo average of means for the 22 datasets for Asian officials and managers is 1.08163 using mean imputation (Table 11) and is 1.083158 using multiple imputation(Table 9). The Monte Carlo average of means for the 22 datasets for Asian service workers is 0.2710785 using mean imputation and is 0.2665542 using multiple imputation. They are very similar. This pattern extends to the rest of the outcomes. I find that the estimated means for each variable are very similar using mean imputation and multiple imputation. Also, the means are not affected by the percentage of missingness regardless of the imputation methods. For example, in the case of mean imputation, the mean for Asian officials and managers of 10% missingness is very close to the mean for Asian officials and managers of 30% and 50% which are 1.08163, 1.0868443 and 1.0921558 accordingly as shown in Table 11.

Table 11 Means of Asians for Mean Imputation with 10%, 30% and 50% Missing

Variable	Label		Mean	
		10%	30%	50%
maa1	ASIAN OFF AND MGRS	1.08163	1.0868443	1.0921558
maa2	ASIAN PROFFESSIONAL	4.0070435	4.0373046	4.0677730
maa3	ASIAN TECHNICIAN	0.8966541	0.898136	0.9014156
maa4	ASIAN SALES WORKERS	0.3976459	0.3946122	0.3998887
maa5	ASIAN OFFICE AND CLERICALS	0.9323619	0.9325244	0.9311849
maa6	ASIAN CRAFT WORKERS	0.9216872	0.9252825	0.9227964
maa7	ASIAN OPERATIVES	2.6997983	2.6902987	2.6857777
maa8	ASIAN LABORERS	0.7804796	0.7778254	0.7794497
maa9	ASIAN SERIVCE WORKERS	0.2710785	0.2678413	0.2729399

Second, the estimated standard errors are larger when using multiple imputation instead of mean imputation. For example, if the missingness is 30%, the standard error is larger by a factor of approximately 1.3 to 1.4. The estimated standard error seems to go down as the percentage of missing goes up when mean imputation is applied, as shown in Table 12. In addition, the estimated standard error doesn’t change much as the percentage of missing changes in the case of multiple imputation.

Table 12 Standard Errors of Asians for Mean Imputation with 10%, 30% and 50% Missing

Variable	Label		Standard Error	
		10%	30%	50%
sraa1	ASIAN OFF AND MGRS	0.036647	0.0327128	0.0283966
sraa2	ASIAN PROFFESSIONAL	0.208663	0.1845939	0.1586868
sraa3	ASIAN TECHNICIAN	0.0332877	0.0295306	0.0257335
sraa4	ASIAN SALES WORKERS	0.0229449	0.0201951	0.017537
sraa5	ASIAN OFFICE AND CLERICALS	0.0241572	0.021438	0.0183654
sraa6	ASIAN CRAFT WORKERS	0.0353726	0.0317334	0.0263739
sraa7	ASIAN OPERATIVES	0.0706273	0.0621859	0.0529784
sraa8	ASIAN LABORERS	0.0310738	0.0274955	0.0234204
sraa9	ASIAN SERIVCE WORKERS	0.0511671	0.0421378	0.034364

Third, the standard deviation of the Monte Carlo average of the estimated means for the 22 datasets and the Monte Carlo average of the estimated standard error for the 22 datasets is small for either mean imputation or multiple imputation. We can conclude that the variability from sample to sample within an imputation method is not large compared to the estimates of means and standard errors. One can see that most of the standard deviations for all the variables in all missing categories are less than 0.01 and some are even less than 0.001.

Chapter 7: Summary and Conclusion

This thesis describes various methods dealing with missing values and discusses their advantages and disadvantages. It also applies the methods to the artificial data sets and to a real world survey conducted by EEOC.

7.1 Handling Missing Data

Listwise deletion deletes all cases with at least one missing item. It may sacrifice a large amount of data, thus decrease power,waste resources and cause larger standard errors. Second, the data must be MCAR; otherwise the estimates may be biased. On the other hand, listwise deletion can use for any kind of statistical analysis and doesn't require special computational methods.

Pairwise deletion drops cases with missing observations on the variables under examination. It saves more information than using listwise deletion. But because the sample used to compute each statistic is different, it causes difficulty in interpreting the covariance matrices.

Single imputation means filling in a single value for each missing value. It has four attractive features. First, it saves a great deal of data comparing to deletion methods. Second, standard complete-data methods of analysis can be used on the filled-in data set. Third, it is used when the data is confidential. Last, it produces consistent analyses.

Mean imputation, regression imputation and hot deck imputation are three single imputation methods that mentioned here. They enjoy many of the advantages of single imputation.

Mean imputation uses the mean of observed values of a variable in place of missing data values for the same variable. It leads to biased estimates and distort the empirical distribution of the variable whose missing values are imputed.

Regression imputation fills missing values by predict values from a regression of the missing item on items observed for that unit. It can produces model unbiased estimators. The higher correlation between independent variables and depend variables, the better regression imputation works.

Hot deck imputation replaces a missing value with a sampled value of the respondent. It can be carried out as the data are being collected and can maintain the proper measurement level of variables. However, it can be difficult to decide a donor case.

Multiple imputation has three steps: imputation, analysis and pooling. First, create m complete data sets by filling the missing entries in the incomplete data sets m times. The prior distribution and the complete data model combine to generate the predictive distribution $p(Y_{mis} \mid Y_{obs})$ which produces the imputed values. Second, the m imputed data sets are analyzed to give m results for each statistic of interest. Third, the results from each imputed data sets are combined.

Multiple imputation corrects for sampling variability. It works well with standard complete-data methods and software. And in many applications, only relative few imputations are required to obtain excellent results.

7.2 Simulation Results

Artificial data simulation was conducted. I created the missingness and imputed using various imputation methods.

As results, the bias for the mean is very small in each case but much higher for multiple imputation. The estimated variance and the Monte Carlo variance for the complete data method tends to agree. Mean imputation underestimates the variance while the multiple imputation is more accurate.

Using EEO-1 data set, I artificially introduced missingness by simulating unit nonresponse and item nonresponse and applied mean imputation and multiple imputation schemes to study their performance.

For first scenario, unit response was simulated. I selected a random sample of 30%, 50% and 70% of 2003 exit companies to be nonrespondents. Then I imputed the average number of employees of respondents for the simulated nonrespondents within each imputation class. I found that the more I imputed, the higher resulting population means.

For the second scenario, item nonresponse was simulated. EEO-1 dataset of 2003 has 221,289 observations. And I randomly selected 10%, 30% and 50% on 36 variables represent the minority employees and replaced the actual data as missing values. I ran mean imputation and multiple imputation on the dataset containing SIC code 40's and 60's and averaged the results of 22 datasets.

I found that the estimated means for each variable are very similar using mean imputation and multiple imputation. The means are not affected by the percentage of missingness regardless of the imputation methods. The estimated standard errors are larger when using multiple imputation instead of mean imputation. The variability from sample to sample within an imputation method is not large compared to the estimates of means and standard errors.

7.3 Suggestions for Future Work

The missing data is an ongoing problem in EEO-1 dataset. If possible, I recommend re-contacting nonresponding companies and requesting submission of the survey. Obtaining more complete and accurate data is especially important for large companies with unexplained exits in a previous year. If this is not feasible, the following proceedings are suggested.

For the companies present in both last year's and this year's EEO-1 files, I propose investigating a sample of NAICS industrial classification and race/ethnic frequencies to determine whether major discrepancies exist between the entries for 2002 and 2003.

Then, using data from last year's EEO-1 survey, I propose imputing revised estimates for missing values in the current dataset. Then one should compare the statistical characteristics of various labor markets, with and without the imputed values to determine whether the adjustments have an effect on the estimated proportion of women and race/ethnic minorities.

Appendix

Table A-1 Mean Imputation by SIC, Size with 30% Nonrespondents

Label	sum	Mean	Std Dev
ASIAN OFF AND MGRS	351489.36	1.5743781	17.231187
ASIAN PROFESSIONALS	1283724.06	5.7500092	72.366426
ASIAN TECHNICIANS	311691.63	1.3961176	17.694512
ASIAN SALES WORKERS	454461.17	2.0356056	65.998629
ASIAN OFFICE AND CLERICALS	517885.23	2.3196923	37.698065
ASIAN CRAFT WORKERS	148418.08	0.6647888	9.9577902
ASIAN OPERATIVES	372053.85	1.6664898	16.933071
ASIAN LABORERS	222757.77	0.9977683	20.644346
ASIAN SERVICE WORKERS	440359.99	1.9724441	36.931521
INDIAN OFF AND MGRS	40020.37	0.1792578	2.0982431
INDIAN PROFESSIONALS	57783.66	0.2588224	3.4130986
INDIAN TECHNICIANS	31548.92	0.1413127	1.7650953
INDIAN SALES WORKERS	85386.83	0.3824615	17.780325
INDIAN OFFICE AND CLERICALS	72004.6	0.3225203	4.163288
INDIAN CRAFT WORKERS	45245.5	0.202662	3.3647394
INDIAN OPERATIVES	72699.75	0.325634	5.0023068
INDIAN LABORERS	47952.38	0.2147865	7.293481
INDIAN SERVICE WORKERS	76546.13	0.3428626	9.2135113
BLACK OFF AND MGRS	677804.22	3.0359955	37.946753
BLACK PROFESSIONALS	1037721.35	4.648123	47.775861
BLACK TECHNICIANS	621291.64	2.7828665	32.221053
BLACK SALES WORKERS	1682822.84	7.5376377	296.98578
BLACK OFFICE AND CLERICALS	2133089.04	9.5544533	114.67109
BLACK CRAFT WORKERS	610108.41	2.732775	46.378593
BLACK OPERATIVES	1807773.31	8.0973112	98.778021
BLACK LABORERS	1228250.47	5.501534	145.52087
BLACK SERVICE WORKERS	2545010.19	11.3995153	182.00028
HISPANIC OFF AND MGRS	526182.71	2.3568581	27.653662
HISPANIC PROFESSIONALS	605594.32	2.7125556	29.000066
HISPANIC TECHNICIANS	379134.66	1.6982059	21.893906
HISPANIC SALES WORKERS	1302979.95	5.8362595	211.55412
HISPANIC OFFICE AND CLERICALS	1290890.9	5.7821107	76.635625
HISPANIC CRAFT WORKERS	681965.72	3.0546356	36.2815
HISPANIC OPERATIVES	1511465.35	6.7700996	63.037209
HISPANIC LABORERS	1665464.5	7.4598868	134.70357
HISPANIC SERVICE WORKERS	2031124.69	9.0977384	192.15168

Table A-2 Mean Imputation by SIC, Size with 50% Nonrespondents

Lable	Sum	Mean	Std Dev
ASIAN OFF AND MGRS	357621.85	1.592488	17.1838797
ASIAN PROFESSIONALS	1305340.74	5.8126747	72.1631148
ASIAN TECHNICIANS	316943.07	1.4113456	17.6447769
ASIAN SALES WORKERS	462727.52	2.060523	65.8073188
ASIAN OFFICE AND CLERICALS	526966.66	2.3465795	37.5910398
ASIAN CRAFT WORKERS	150910.84	0.6720051	9.9294775
ASIAN OPERATIVES	378271.08	1.6844389	16.8864875
ASIAN LABORERS	226395.3	1.008137	20.5848253
ASIAN SERVICE WORKERS	447799.77	1.9940498	36.8255874
INDIAN OFF AND MGRS	40735.42	0.1813946	2.0924664
INDIAN PROFESSIONALS	58768.18	0.2616944	3.4034944
INDIAN TECHNICIANS	32091.14	0.1429017	1.760155
INDIAN SALES WORKERS	86917.08	0.3870413	17.7285346
INDIAN OFFICE AND CLERICALS	73255.94	0.3262083	4.1516617
INDIAN CRAFT WORKERS	46022.91	0.2049398	3.355141
INDIAN OPERATIVES	73952.71	0.329311	4.9881006
INDIAN LABORERS	48756.71	0.2171133	7.272269
INDIAN SERVICE WORKERS	77900.3	0.3468896	9.1868567
BLACK OFF AND MGRS	689923.7	3.0722262	37.8416386
BLACK PROFESSIONALS	1055123.3	4.6984579	47.6445914
BLACK TECHNICIANS	631718.74	2.8130399	32.1312327
BLACK SALES WORKERS	1713149.57	7.6286451	296.122294
BLACK OFFICE AND CLERICALS	2169100.45	9.6589917	114.352013
BLACK CRAFT WORKERS	620478.77	2.7629884	46.2460247
BLACK OPERATIVES	1838524.14	8.1869373	98.5020509
BLACK LABORERS	1248730.77	5.5605909	145.099212
BLACK SERVICE WORKERS	2588918.61	1.5284395	181.482935
HISPANIC OFF AND MGRS	535702.23	2.3854789	27.5777764
HISPANIC PROFESSIONALS	615874.37	0.742485	28.9201263
HISPANIC TECHNICIANS	385605.66	7170997	21.8324066
HISPANIC SALES WORKERS	1326590.40	0.9072993	210.93977
HISPANIC OFFICE AND CLERICALS	1313288.49	0.848066	76.4210658
HISPANIC CRAFT WORKERS	693587.35	0.0885404	36.1802939
HISPANIC OPERATIVES	1536889.46	0.8437598	62.8662766
HISPANIC LABORERS	1692838.82	0.5382014	134.316987
HISPANIC SERVICE WORKERS	2066894.16	0.2038677	191.599541

Table A-3 Mean Imputation by SIC, Size with 70% Nonrespondents

Lable	Sum	Mean	Std Dev
ASIAN OFF AND MGRS	364322.08	1.6129082	17.1373542
ASIAN PROFESSIONALS	1328728.67	5.8824799	71.9626244
ASIAN TECHNICIANS	322620.09	1.4282872	17.5957299
ASIAN SALES WORKERS	471868.61	2.0890327	65.6180051
ASIAN OFFICE AND CLERICALS	536859.12	2.3767553	37.4853756
ASIAN CRAFT WORKERS	153643.78	0.6802039	9.9015169
ASIAN OPERATIVES	385096.58	1.70488	16.8406297
ASIAN LABORERS	230388.6	1.0199647	20.5259572
ASIAN SERVICE WORKERS	455820.35	2.0179846	36.7208771
INDIAN OFF AND MGRS	41519.91	0.1838148	2.0867833
INDIAN PROFESSIONALS	59835.97	0.2649028	3.3940218
INDIAN TECHNICIANS	32679.28	0.144676	1.7552854
INDIAN SALES WORKERS	88610.55	0.3922921	17.6772587
INDIAN OFFICE AND CLERICALS	74621.72	0.3303615	4.1402049
INDIAN CRAFT WORKERS	46873.87	0.2075176	3.3456568
INDIAN OPERATIVES	75333.86	0.3335142	4.9740732
INDIAN LABORERS	49640.06	0.2197639	7.251271
INDIAN SERVICE WORKERS	79370.58	0.3513854	9.1604869
BLACK OFF AND MGRS	703215.65	3.1132405	37.7381544
BLACK PROFESSIONALS	1073910.95	4.7543638	47.5154662
BLACK TECHNICIANS	642940.98	2.8463956	32.0427239
BLACK SALES WORKERS	1746714.09	7.7329636	295.267545
BLACK OFFICE AND CLERICALS	2208291.17	9.7764341	114.0377
BLACK CRAFT WORKERS	631853.27	2.7973086	46.1150263
BLACK OPERATIVES	1872394.78	8.2893708	98.2299966
BLACK LABORERS	1271246.98	5.6279999	144.681976
BLACK SERVICE WORKERS	2636240.34	11.6710289	180.972054
HISPANIC OFF AND MGRS	546145.46	2.4178673	27.5031502
HISPANIC PROFESSIONALS	627021.91	2.7759195	28.8414703
HISPANIC TECHNICIANS	392619.67	1.7381858	21.7717636
HISPANIC SALES WORKERS	1352723.43	5.9887083	210.331715
HISPANIC OFFICE AND CLERICALS	1337711.43	5.9222479	76.209573
HISPANIC CRAFT WORKERS	706184.11	3.1263823	36.0804424
HISPANIC OPERATIVES	1564765.2	6.9274488	62.6983078
HISPANIC LABORERS	1722579.1	7.6261144	133.934657
HISPANIC SERVICE WORKERS	2105724.07	9.3223543	191.053709

Table A-4 Mean for Mean Imputation of SIC 40 and 60 with 10% Missin

Lable	Mean	Std
ASIAN OFF AND MGRS	1.08163	0.014075
ASIAN PROFESSIONALS	4.0070435	0.0723972
ASIAN TECHNICIANS	0.8966541	0.011098
ASIAN SALES WORKERS	0.3976459	0.0098066
ASIAN OFFICE AND CLERICALS	0.9323619	0.0075525
ASIAN CRAFT WORKERS	0.9216872	0.0151852
ASIAN OPERATIVES	2.6997983	0.0202666
ASIAN LABORERS	0.7804796	0.0105382
ASIAN SERVICE WORKERS	0.2710785	0.0156026
INDIAN OFF AND MGRS	0.1224314	0.0010991
INDIAN PROFESSIONALS	0.1758014	0.0028569
INDIAN TECHNICIANS	0.094307	0.0015977
INDIAN SALES WORKERS	0.0539205	0.0010582
INDIAN OFFICE AND CLERICALS	0.1827186	0.002215
INDIAN CRAFT WORKERS	0.2321297	0.0027917
INDIAN OPERATIVES	0.4280038	0.0053277
INDIAN LABORERS	0.1558319	0.0029134
INDIAN SERVICE WORKERS	0.0331814	0.0011357
BLACK OFF AND MGRS	1.844734	0.0107423
BLACK PROFESSIONALS	2.1005867	0.0167456
BLACK TECHNICIANS	1.1472905	0.0112949
BLACK SALES WORKERS	1.1949405	0.0272728
BLACK OFFICE AND CLERICALS	0.0436568	
BLACK CRAFT WORKERS	3.5899348	0.0343545
BLACK OPERATIVES	11.2090058	0.0509844
BLACK LABORERS	4.6709746	0.0600361
BLACK SERVICE WORKERS	1.2288654	0.0291557
HISPANIC OFF AND MGRS	1.2844085	0.011696
HISPANIC PROFESSIONALS	1.7086481	0.0162297
HISPANIC TECHNICIANS	1.106279	0.016138
HISPANIC SALES WORKERS	0.8095266	0.0239978
HISPANIC OFFICE AND CLERICALS	3.0054366	0.0400011
HISPANIC CRAFT WORKERS	3.0251844	0.0239913
HISPANIC OPERATIVES	7.755669	0.0452373
HISPANIC LABORERS	4.2949734	0.0500952
HISPANIC SERVICE WORKERS	0.7394331	0.020398

Table A-5 Standard Error for Mean Imputation of SIC 40 and 60 with 10%

Lable	Mean	Std Dev
ASIAN OFF AND MGRS	0.036647	0.0023455
ASIAN PROFESSIONALS	0.208663	0.019681
ASIAN TECHNICIANS	0.0332877	0.0014245
ASIAN SALES WORKERS	0.0229449	0.000826929
ASIAN OFFICE AND CLERICALS	0.0241572	0.000312928
ASIAN CRAFT WORKERS	0.0353726	0.0019168
ASIAN OPERATIVES	0.0706273	0.0013654
ASIAN LABORERS	0.0310738	0.000636506
ASIAN SERVICE WORKERS	0.0511671	0.0093649
INDIAN OFF AND MGRS	0.0032412	0.000066599
INDIAN PROFESSIONALS	0.0088203	0.000945509
INDIAN TECHNICIANS	0.0043691	0.000297589
INDIAN SALES WORKERS	0.0028649	0.000091131
INDIAN OFFICE AND CLERICALS	0.0068207	0.000148617
INDIAN CRAFT WORKERS	0.0099342	0.000752384
INDIAN OPERATIVES	0.016165	0.000577081
INDIAN LABORERS	0.0088243	0.000300353
INDIAN SERVICE WORKERS	0.002932	0.000174864
BLACK OFF AND MGRS	0.0487952	0.000958624
BLACK PROFESSIONALS	0.0593371	0.0015886
BLACK TECHNICIANS	0.0324031	0.0012864
BLACK SALES WORKERS	0.0759876	0.0034063
BLACK OFFICE AND CLERICALS	0.137052	0.0054378
BLACK CRAFT WORKERS	0.1403044	0.0066643
BLACK OPERATIVES	0.2377423	0.0023735
BLACK LABORERS	0.2580996	0.0229582
BLACK SERVICE WORKERS	0.0871233	0.004193
HISPANIC OFF AND MGRS	0.0306984	0.0010454
HISPANIC PROFESSIONALS	0.0538253	0.0011191
HISPANIC TECHNICIANS	0.0412034	0.0031221
HISPANIC SALES WORKERS	0.0591357	0.0041216
HISPANIC OFFICE AND CLERICALS	0.0925819	0.0036181
HISPANIC CRAFT WORKERS	0.0810314	0.002182
HISPANIC OPERATIVES	0.1381343	0.0021327
HISPANIC LABORERS	0.1428239	0.005837
HISPANIC SERVICE WORKERS	0.0653963	0.0048995

Table A-6 Mean for Multiple Imputation of SIC 40 and 60 with 10% Mis

Label	Mean	Std Dev
ASIAN OFF AND MGRS	1.083158	0.0073777
ASIAN PROFESSIONALS	4.0213184	0.0306867
ASIAN TECHNICIANS	0.901485	0.0082997
ASIAN SALES WORKERS	0.3976574	0.0075026
ASIAN OFFICE AND CLERICALS	0.9318953	0.0084804
ASIAN CRAFT WORKERS	0.9231473	0.0158133
ASIAN OPERATIVES	2.7054532	0.029195
ASIAN LABORERS	0.7800523	0.0119454
ASIAN SERVICE WORKERS	0.2665542	0.0124755
INDIAN OFF AND MGRS	0.122439	0.00094889
INDIAN PROFESSIONALS	0.1767646	0.0026821
INDIAN TECHNICIANS	0.0945376	0.0013144
INDIAN SALES WORKERS	0.0538895	0.000916679
INDIAN OFFICE AND CLERICALS	0.1830566	0.002208
INDIAN CRAFT WORKERS	0.2324844	0.0032346
INDIAN OPERATIVES	0.4285815	0.0033952
INDIAN LABORERS	0.1555887	0.0022237
INDIAN SERVICE WORKERS	0.033087	0.0010932
BLACK OFF AND MGRS	1.8387958	0.0088312
BLACK PROFESSIONALS	2.0986537	0.011528
BLACK TECHNICIANS	1.1463838	0.0096528
BLACK SALES WORKERS	1.1878502	0.0230118
BLACK OFFICE AND CLERICALS	5.0542776	0.0244377
BLACK CRAFT WORKERS	3.5873672	0.0273479
BLACK OPERATIVES	11.262916	0.0583778
BLACK LABORERS	4.6495003	0.0566049
BLACK SERVICE WORKERS	1.225502	0.0270548
HISPANIC OFF AND MGRS	1.2821631	0.0064382
HISPANIC PROFESSIONALS	1.7120016	0.0120667
HISPANIC TECHNICIANS	1.1075879	0.0123464
HISPANIC SALES WORKERS	0.8070143	0.0188424
HISPANIC OFFICE AND CLERICALS	3.0018281	0.0250629
HISPANIC CRAFT WORKERS	3.0326351	0.0256947
HISPANIC OPERATIVES	7.7486362	0.0300259
HISPANIC LABORERS	4.2853423	0.0390823
HISPANIC SERVICE WORKERS	0.7357332	0.0175378

Table A-7 Standard Error for Multiple Imputation of SIC 40 and 60 with 10%

Label	Mean	Std Dev
ASIAN OFF AND MGRS	0.0390964	0.0011
ASIAN PROFESSIONALS	0.2219334	0.0117617
ASIAN TECHNICIANS	0.0365895	0.0018677
ASIAN SALES WORKERS	0.0252045	0.0016182
ASIAN OFFICE AND CLERICALS	0.0262856	0.000862865
ASIAN CRAFT WORKERS	0.0392547	0.0028408
ASIAN OPERATIVES	0.0787201	0.004213
ASIAN LABORERS	0.0345517	0.0020712
ASIAN SERVICE WORKERS	0.0538644	0.0067179
INDIAN OFF AND MGRS	0.0036085	0.000199555
INDIAN PROFESSIONALS	0.009661	0.0010562
INDIAN TECHNICIANS	0.0047383	0.000324334
INDIAN SALES WORKERS	0.0031458	0.000134705
INDIAN OFFICE AND CLERICALS	0.0075079	0.000539742
INDIAN CRAFT WORKERS	0.0107306	0.000749116
INDIAN OPERATIVES	0.0173301	0.000613047
INDIAN LABORERS	0.0097098	0.000643966
INDIAN SERVICE WORKERS	0.0032791	0.000302279
BLACK OFF AND MGRS	0.0514204	0.00085096
BLACK PROFESSIONALS	0.0637579	0.0014131
BLACK TECHNICIANS	0.0352544	0.0016716
BLACK SALES WORKERS	0.0818821	0.0036343
BLACK OFFICE AND CLERICALS	0.1459309	0.0034803
BLACK CRAFT WORKERS	0.1508692	0.0065274
BLACK OPERATIVES	0.2552603	0.005727
BLACK LABORERS	0.2712791	0.0192341
BLACK SERVICE WORKERS	0.0940687	0.006649
HISPANIC OFF AND MGRS	0.0326403	0.000736684
HISPANIC PROFESSIONALS	0.0586445	0.0017699
HISPANIC TECHNICIANS	0.0453113	0.003746
HISPANIC SALES WORKERS	0.0645439	0.0041813
HISPANIC OFFICE AND CLERICALS	0.1008022	0.0034247
HISPANIC CRAFT WORKERS	0.0888423	0.0384672
HISPANIC OPERATIVES	0.1491198	0.0030148
HISPANIC LABORERS	0.1559968	0.0085964
HISPANIC SERVICE WORKERS	0.0714415	0.0050557

Table A-8 Mean for Mean Imputation of SIC 40 and 60 with 30% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	1.0868443	0.0304507
ASIAN PROFESSIONALS	4.0373046	0.1864749
ASIAN TECHNICIANS	0.898136	0.0210018
ASIAN SALES WORKERS	0.3946122	0.0150992
ASIAN OFFICE AND CLERICALS	0.9325244	0.0201723
ASIAN CRAFT WORKERS	0.9252825	0.0201948
ASIAN OPERATIVES	2.6902987	0.0437319
ASIAN LABORERS	0.7778254	0.0227945
ASIAN SERVICE WORKERS	0.2678413	0.0386364
INDIAN OFF AND MGRS	0.1231143	0.0026046
INDIAN PROFESSIONALS	0.1749667	0.0058177
INDIAN TECHNICIANS	0.0941522	0.003499
INDIAN SALES WORKERS	0.053955	0.0018904
INDIAN OFFICE AND CLERICALS	0.183742	0.0036438
INDIAN CRAFT WORKERS	0.2325631	0.0054051
INDIAN OPERATIVES	0.4281988	0.011433
INDIAN LABORERS	0.1562963	0.007271
INDIAN SERVICE WORKERS	0.0329293	0.0018979
BLACK OFF AND MGRS	1.843486	0.0271273
BLACK PROFESSIONALS	2.1114668	0.0407903
BLACK TECHNICIANS	1.1494903	0.0260507
BLACK SALES WORKERS	1.1978699	0.0562859
BLACK OFFICE AND CLERICALS	5.0416443	0.08515
BLACK CRAFT WORKERS	3.5846596	0.0831409
BLACK OPERATIVES	11.2220859	0.1360135
BLACK LABORERS	4.6604703	0.1654592
BLACK SERVICE WORKERS	1.2181361	0.0707541
HISPANIC OFF AND MGRS	1.285655	0.0185686
HISPANIC PROFESSIONALS	1.7172006	0.0385179
HISPANIC TECHNICIANS	1.1077452	0.0333255
HISPANIC SALES WORKERS	0.8129535	0.0423457
HISPANIC OFFICE AND CLERICALS	3.0048877	0.0693814
HISPANIC CRAFT WORKERS	3.0170536	0.0463045
HISPANIC OPERATIVES	7.7435544	0.0967281
HISPANIC LABORERS	4.2834397	0.0808966
HISPANIC SERVICE WORKERS	0.7334007	0.0481394

Table A-9 Standard Error for Mean Imputation of SIC 40 and 60 with 30% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	0.0327128	0.0040455
ASIAN PROFESSIONALS	0.1845939	0.0379108
ASIAN TECHNICIANS	0.0295306	0.0028107
ASIAN SALES WORKERS	0.0201951	0.0010362
ASIAN OFFICE AND CLERICALS	0.021438	0.000904271
ASIAN CRAFT WORKERS	0.0317334	0.0019698
ASIAN OPERATIVES	0.0621859	0.0023241
ASIAN ABORERS	0.0274955	0.0012397
ASIAN SERVICE WORKERS	0.0421378	0.0167652
INDIAN OFF AND MGRS	0.0029188	0.00010663
INDIAN PROFESSIONALS	0.0074823	0.0016898
INDIAN TECHNICIANS	0.0038486	0.000536893
INDIAN SALES WORKERS	0.0025568	0.000167691
INDIAN OFFICE AND CLERICALS	0.0060812	0.00043262
INDIAN CRAFT WORKERS	0.0089508	0.0012141
INDIAN OPERATIVES	0.0142817	0.0011155
INDIAN LABORERS	0.0078786	0.000666708
INDIAN SERVICE WORKERS	0.0025514	0.000283707
BLACK OFF AND MGRS	0.0432229	0.0027807
BLACK PROFESSIONALS	0.0536899	0.0026698
BLACK TECHNICIANS	0.0287542	0.0024016
BLACK SALES WORKERS	0.0676821	0.0051266
BLACK OFFICE AND CLERICALS	0.1220539	0.0097115
BLACK CRAFT WORKERS	0.1223029	0.0159672
BLACK OPERATIVES	0.2123261	0.0052213
BLACK LABORERS	0.2247616	0.040058
BLACK SERVICE WORKERS	0.0756141	0.0090673
HISPANIC OFF AND MGRS	0.0274422	0.0016661
HISPANIC PROFESSIONALS	0.0487061	0.0020572
HISPANIC TECHNICIANS	0 .0362768	0.0061719
HISPANIC SALES WORKERS	0.0530059	0.0051331
HISPANIC OFFICE AND CLERICALS	0.0825223	0.0057402
HISPANIC CRAFT WORKERS	0.0714361	0.0036344
HISPANIC OPERATIVES	0.122405	0.0033423
HISPANIC LABORERS	0.1266031	0.0087597
HISPANIC SERVICE WORKERS	0.0560968	0.0102175

Table A-10 Mean for Multiple Imputation of SIC 40 and 60 with 30% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	1.0825622	0.0114775
ASIAN PROFESSIONALS	4.0245711	0.0854513
ASIAN TECHNICIANS	0.8960141	0.0155565
ASIAN SALES WORKERS	0.3946633	0.0115729
ASIAN OFFICE AND CLERICALS	0.9326639	0.0177167
ASIAN CRAFT WORKERS	0.931464	0.0279342
ASIAN OPERATIVES	2.7026934	0.0543924
ASIAN LABORERS	0.7810532	0.0241653
ASIAN SERVICE WORKERS	0.2156297	0.0217999
INDIAN OFF AND MGRS	0.1227261	0.0026073
INDIAN PROFESSIONALS	0.1743427	0.0054653
INDIAN TECHNICIANS	0.094319	0.003231
INDIAN SALES WORKERS	0.0536117	0.0016065
INDIAN OFFICE AND CLERICALS	0.1845562	0.0034819
INDIAN CRAFT WORKERS	0.2333336	0.0054243
INDIAN OPERATIVES	0.4276397	0.0093587
INDIAN LABORERS	0.1566225	0.0065535
INDIAN SERVICE WORKERS	0.0332794	0.0019949
BLACK OFF AND MGRS	1.8370061	0.0155031
BLACK PROFESSIONALS	2.1022897	0.0200055
BLACK TECHNICIANS	1.1501653	0.0209947
BLACK SALES WORKERS	1.1959862	0.0400349
BLACK OFFICE AND CLERICALS	5.0206556	0.044723
BLACK CRAFT WORKERS	3.5952173	0.0775422
BLACK OPERATIVES	11.2450427	0.0858947
BLACK LABORERS	4.6305875	0.1205877
BLACK SERVICE WORKERS	1.2244233	0.058249
HISPANIC OFF AND MGRS	1.2820231	0.0136887
HISPANIC PROFESSIONALS	1.7120948	0.0208302
HISPANIC TECHNICIANS	1.1076003	0.030679
HISPANIC SALES WORKERS	0.810098	0.038197
HISPANIC OFFICE AND CLERICALS	3.0099021	0.0539168
HISPANIC CRAFT WORKERS	3.0263215	0.0498989
HISPANIC OPERATIVES	7.7445917	0.0600859
HISPANIC LABORERS	4.2808683	0.062454
HISPANIC SERVICE WORKERS	0.7395175	0.0350422

Table A-11 Standard Error for Multiple Imputation of SIC 40 and 60 with 30% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	0.0401749	0.0024457
ASIAN PROFESSIONALS	0.2221412	0.0258599
ASIAN TECHNICIANS	0.0386191	0.004
ASIAN SALES WORKERS	0.0273241	0.0033086
ASIAN OFFICE AND CLERICALS	0.0287883	0.0030223
ASIAN CRAFT WORKERS	0.0438967	0.0047447
ASIAN OPERATIVES	0.0875541	0.0110968
ASIAN LABORERS	0.0383986	0.0048967
ASIAN SERVICE WORKERS	0.0515151	0.012965
INDIAN OFF AND MGRS	0.0037833	0.000327335
INDIAN PROFESSIONALS	0.0099533	0.0028377
INDIAN TECHNICIANS	0.0051334	0.000662
INDIAN SALES WORKERS	0.0034643	0.000488017
INDIAN OFFICE AND CLERICALS	0.0083827	0.000973475
INDIAN CRAFT WORKERS	0.0124219	0.0023251
INDIAN OPERATIVES	0.0183914	0.0020699
INDIAN LABORERS	0.010886	0.0013582
INDIAN SERVICE WORKERS	0.0034939	0.000558081
BLACK OFF AND MGRS	0.0523911	0.0027462
BLACK PROFESSIONALS	0.0685325	0.0043628
BLACK TECHNICIANS	0.0380724	0.0045882
BLACK SALES WORKERS	0.0902412	0.0110448
BLACK OFFICE AND CLERICALS	0.1492558	0.0080518
BLACK CRAFT WORKERS	0.1553668	0.0171637
BLACK OPERATIVES	0.2674966	0.0114858
BLACK LABORERS	0.2666127	0.0362291
BLACK SERVICE WORKERS	0.1033659	0.0152774
HISPANIC OFF AND MGRS	0.0341703	0.0023112
HISPANIC PROFESSIONALS	0.0630032	0.0039918
HISPANIC TECHNICIANS	0.048906	0.009163
HISPANIC SALES WORKERS	0.0701243	0.0093826
HISPANIC OFFICE AND CLERICALS	0.1099521	0.0134563
HISPANIC CRAFT WORKERS	0.0951564	0.0087998
HISPANIC OPERATIVES	0.1592495	0.0161501
HISPANIC LABORERS	0.1629251	0.0131519
HISPANIC SERVICE WORKERS	0.0794418	0.0127597

Table A-12 Mean for Mean Imputation of SIC 40 and 60 with 50% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	1.0921558	0.0339846
ASIAN PROFESSIONALS	4.067773	0.2315698
ASIAN TECHNICIANS	0.9014156	0.0268646
ASIAN SALES WORKERS	0.3998887	0.0219262
ASIAN OFFICE AND CLERICALS	0.9311849	0.0266304
ASIAN CRAFT WORKERS	0.9227964	0.027644
ASIAN OPERATIVES	2.6857777	0.0831955
ASIAN LABORERS	0.7794497	0.0290334
ASIAN SERVICE WORKERS	0.2729399	0.050897
INDIAN OFF AND MGRS	0.123322	0.0037773
INDIAN PROFESSIONALS	0.174391	0.0062617
INDIAN TECHNICIANS	0.0945671	0.005081
INDIAN SALES WORKERS	0.0534308	0.0030303
INDIAN OFFICE AND CLERICALS	0.1833851	0.0070475
INDIAN CRAFT WORKERS	0.2335147	0.0098903
INDIAN OPERATIVES	0.4285913	0.0172243
INDIAN LABORERS	0.1571317	0.0096253
INDIAN SERVICE WORKERS	0.0332621	0.0021586
BLACK OFF AND MGRS	1.8499793	0.0527349
BLACK PROFESSIONALS	2.1094474	0.0613512
BLACK TECHNICIANS	1.1434664	0.0372062
BLACK SALES WORKERS	1.1965428	0.0766121
BLACK OFFICE AND CLERICALS	5.0488074	0.133286
BLACK CRAFT WORKERS	3.5972078	0.1397739
BLACK OPERATIVES	11.2183212	0.2555855
BLACK LABORERS	4.7103412	0.2393305
BLACK SERVICE WORKERS	1.2188024	0.0909178
HISPANIC OFF AND MGRS	1.2844729	0.0227193
HISPANIC PROFESSIONALS	1.7121711	0.0529589
HISPANIC TECHNICIANS	1.1015388	0.0354883
HISPANIC SALES WORKERS	0.8115036	0.0788882
HISPANIC OFFICE AND CLERICALS	2.989925	0.1012007
HISPANIC CRAFT WORKERS	3.0247699	0.0681554
HISPANIC OPERATIVES	7.7448207	0.1380661
HISPANIC LABORERS	4.2773345	0.118563
HISPANIC SERVICE WORKERS	0.7439091	0.0783265

Table A-13 Standard Error for Mean Imputation of SIC 40 and 60 with 50% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	0.0283966	0.004323
ASIAN PROFESSIONALS	0.1586868	0.0421814
ASIAN TECHNICIANS	0.0257335	0.0029475
ASIAN SALES WORKERS	0.017537	0.0013118
ASIAN OFFICE AND CLERICALS	0.0183654	0.0011351
ASIAN CRAFT WORKERS	0.0263739	0.0024598
ASIAN OPERATIVES	0.0529784	0.0034923
ASIAN LABORERS	0.0234204	0.001157
ASIAN SERVICE WORKERS	0.034364	0.0185037
INDIAN OFF AND MGRS	0.0025046	0.000128505
INDIAN PROFESSIONALS	0.0062752	0.0018577
INDIAN TECHNICIANS	0.0032584	0.000589072
INDIAN SALES WORKERS	0.0021547	0.000197023
INDIAN OFFICE AND CLERICALS	0.0052226	0.000716731
INDIAN CRAFT WORKERS	0.0074636	0.0018286
INDIAN OPERATIVES	0.0120796	0.0015107
INDIAN LABORERS	0.0067462	0.000771799
INDIAN SERVICE WORKERS	0.0021943	0.000303804
BLACK OFF AND MGRS	0.0381244	0.0038482
BLACK PROFESSIONALS	0.04632	0.0035615
BLACK TECHNICIANS	0.0241049	0.0032093
BLACK SALES WORKERS	0.0577659	0.0066521
BLACK OFFICE AND CLERICALS	0.1072252	0.0124344
BLACK CRAFT WORKERS	0.1031273	0.0235346
BLACK OPERATIVES	0.1829005	0.0090388
BLACK LABORERS	0.1939694	0.0478224
BLACK SERVICE WORKERS	0.0638691	0.0097155
HISPANIC OFF AND MGRS	0.0236123	0.0018904
HISPANIC PROFESSIONALS	0.041745	0.0026275
HISPANIC TECHNICIANS	0.0300899	0.0066493
HISPANIC SALES WORKERS	0.0445044	0.0091039
HISPANIC OFFICE AND CLERICALS	0.0701412	0.0068552
HISPANIC CRAFT WORKERS	0.0612589	0.0040763
HISPANIC OPERATIVES	0.1042996	0.0046079
HISPANIC LABORERS	0.1064993	0.010723
HISPANIC SERVICE WORKERS	0.0473338	0.0130797

Table A-14 Mean for Multiple Imputation of SIC 40 and 60 with 50% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	1.0849549	0.0137417
ASIAN PROFESSIONALS	4.0352595	0.1147877
ASIAN TECHNICIANS	0.8969574	0.0304354
ASIAN SALES WORKERS	0.4036071	0.0175444
ASIAN OFFICE AND CLERICALS	0.932376	0.0247577
ASIAN CRAFT WORKERS	0.9304558	0.0337406
ASIAN OPERATIVES	2.6830144	0.1048978
ASIAN LABORERS	0.7821873	0.030917
ASIAN SERVICE WORKERS	0.2620415	0.0272273
INDIAN OFF AND MGRS	0.1228141	0.0041722
INDIAN PROFESSIONALS	0.1732795	0.0085321
INDIAN TECHNICIANS	0.0955243	0.004547
INDIAN SALES WORKERS	0.0534809	0.0027169
INDIAN OFFICE AND CLERICALS	0.1828629	0.0069358
INDIAN CRAFT WORKERS	0.2315023	0.0084316
INDIAN OPERATIVES	0.4307157	0.0126303
INDIAN LABORERS	0.1563423	0.0074786
INDIAN SERVICE WORKERS	0.0339215	0.0022181
BLACK OFF AND MGRS	1.837912	0.0298635
BLACK PROFESSIONALS	2.1118412	0.0314929
BLACK TECHNICIANS	1.1437073	0.0301431
BLACK SALES WORKERS	1.2059761	0.0527474
BLACK OFFICE AND CLERICALS	5.0461963	0.0864868
BLACK CRAFT WORKERS	3.5823446	0.1214057
BLACK OPERATIVES	11.244675	0.169185
BLACK LABORERS	4.5945498	0.1358657
BLACK SERVICE WORKERS	1.2345258	0.0912038
HISPANIC OFF AND MGRS	1.2837259	0.0204533
HISPANIC PROFESSIONALS	1.7245618	0.0383396
HISPANIC TECHNICIANS	1.1107027	0.0356002
HISPANIC SALES WORKERS	0.8092986	0.0657937
HISPANIC OFFICE AND CLERICALS	2.9866708	0.0861374
HISPANIC CRAFT WORKERS	3.0418383	0.0600338
HISPANIC OPERATIVES	7.748646	0.1393354
HISPANIC LABORERS	4.2564768	0.1215956
HISPANIC SERVICE WORKERS	0.7512609	0.0582854

Table A-15 Standard Error for Multiple Imputation of SIC 40 and 60 with 50% Missing

Label	Mean	Std Dev
ASIAN OFF AND MGRS	0.0418995	0.0032211
ASIAN PROFESSIONALS	0.2341154	0.0431798
ASIAN TECHNICIANS	0.0488053	0.0097761
ASIAN SALES WORKERS	0.0312259	0.0049241
ASIAN OFFICE AND CLERICALS	0.0328292	0.0071798
ASIAN CRAFT WORKERS	0.0499683	0.0094465
ASIAN OPERATIVES	0.0987496	0.0196884
ASIAN LABORERS	0.0458944	0.0078385
ASIAN SERVICE WORKERS	0.0504145	0.0158453
INDIAN OFF AND MGRS	0.0044279	0.000801308
INDIAN PROFESSIONALS	0.0114807	0.0054119
INDIAN TECHNICIANS	0.0059854	0.0015782
INDIAN SALES WORKERS	0.0037007	0.000642029
INDIAN OFFICE AND CLERICALS	0.0093337	0.0028389
INDIAN CRAFT WORKERS	0.0123457	0.0032808
INDIAN OPERATIVES	0.0201776	0.0031648
INDIAN LABORERS	0.0121465	0.0026049
INDIAN SERVICE WORKERS	0.0041229	0.000561668
BLACK OFF AND MGRS	0.0551877	0.0050477
BLACK PROFESSIONALS	0.0768332	0.0088513
BLACK TECHNICIANS	0.0414101	0.0095097
BLACK SALES WORKERS	0.0997331	0.0136934
BLACK OFFICE AND CLERICALS	0.1689281	0.0229143
BLACK CRAFT WORKERS	0.1633287	0.0391232
BLACK OPERATIVES	0.291786	0.0280246
BLACK LABORERS	0.2771232	0.0425605
BLACK SERVICE WORKERS	0.1173553	0.0208904
HISPANIC OFF AND MGRS	0.03656	0.0037357
HISPANIC PROFESSIONALS	0.0670765	0.0094085
HISPANIC TECHNICIANS	0.0522356	0.0128727
HISPANIC SALES WORKERS	0.0744781	0.0176218
HISPANIC OFFICE AND CLERICALS	0.1157543	0.0190789
HISPANIC CRAFT WORKERS	0.1048462	0.0219965
HISPANIC OPERATIVES	0.1641627	0.0203515
HISPANIC LABORERS	0.1829904	0.0273422
HISPANIC SERVICE WORKERS	0.0823973	0.0156266

Bibliography

Allison, Paul D. Missing Data. CA: Sage., 2002.

Barnard, J. and Rubin, D.B. “Small-Sample Degrees of Freedom with Multiple Imputation.” Biometrika, 86 (1999): 948-955.

Cartwright, Bliss. “Quality Control Issues in the 2003 EEO-1 Statistical File.” Internal Report in EEOC, May 2005.

Fay, R.E., Meng, X.L., & Rubin, D.B. “Multiple Imputation Methodology for Missing Data, Non-Random Response, and Panel Attrition.” March 1, 1997.

Graham, John W & Schafer, Joseph L. “Missing Data: Our View of the State of Art.” Psychological Methods. 7 (2002): 147-177.

Hunter, J.E., Schmidt, F.L. & Urry, V.W. “Statistical Power in Criterion-related Validation Studies.” Journal of Applied Psychology. 61 (1976): 473-485.

Little, R.J.A., & Rubin, D.B. Statistical Analysis with Missing Data.(2nd ed.) New York: John Wiley & Sons, 2002.

Maxim, Paul. Lecture Notes on Quantitative and Empirical Sociology. London: University Of Western Ontario, 1998.

Meng, Xiao-Li. “Multiple Imputation Inferences with Uncongenial Sources of Input.” Statistical Science 9 (1994): 538-558.

Roth, Philip L. “Missing Data: A Conceptual Review for Applied Psychologists.” Personnel Psychology 47 (1994): 537-560.

Roth, Philip L. & Switzer. “A Monte Carlo Analysis of Missing Data Techniques In HRM Settings.” Journal of Management 21 (1995): 1003-1023.

Rubin, D.B. Handling Nonresponse in Sample Surveys by Multiple Imputations. Washington, D.C: U.S Bureau of Census , 1980.

Rubin, D.B. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, 1987.

Rubin, D.B. & Schenker, N. “Multiple Imputation from Random Samples with Ignorable Nonresponse.” Journal of the American Statistical Association, 81, (1986): 366-374.

Schafer, Joseph L. “Multiple Imputation: A Primer.” Statistical Methods in Medical Research, 8 (1999): 3-15.

Sinharay, Sandip , Stern, Hal S. & Russell, Daniel. “The Use of Multiple Imputation for the Analysis of Missing Data.” Psychological Methods 6 (2001): 317-329.

Smith, Mark. “Finding and Using Health Data.” Health Economics Resource Center, 2005.

Tanner, Martin A. & Wong, Wing Hung. “The Calculation of Posterior Distributions by Data Augmentation.” Journal of the American Statistical Association 82 (1987): 528-540.

Watson, Nicole, & Wooden, Mark. “Towards an Imputation Strategy for Wave 1 of the Hilda Survey.” University of Melbourne, 2003.

Wayman, Jeffrey C. “Multiple Imputation for Missing Data: What is It and How Can I Use It?” University of Missour, 2003.

Yuan, Yang C. “Multiple Imputation for Missing Data: Concepts and New Developments.” 2000. SAS Institute Inc., May 2006. http://www.sas.com/rnd/app/papers/multipleimputation.pd