Contents

List of Tables

1

Basic Concepts

1.1 Introduction

Data collection and analysis forms the backbone of all empirical research and almost every data analysis involves variables with some missing values (which will be defined later). The missing values may arise due to unit nonresponse where a sampled subject refuses to provide any values for the variables of interest, or due to item nonresponse, where a sampled subject provides information only for some variables.

The complete-case or available-subjects is the common approach that restricts the analysis to subjects without missing values in the relevant variables. This approach, though convenient, can result in biased estimates of the parameters or population quantities because the included and excluded subjects from the analysis may differ systematically. Even if the included subjects are a random subset of the sampled subjects, the sampling error increases due to the reduced sample size.

Many *ad hoc* and naive methods are also used in practice. For example, in a multiple regression analysis with missing values in one categorical covariate, subjects with missing values are treated as a separate category when creating the dummy variables. This analysis uses all of the subjects but may seriously bias the regression coefficients for the other covariates. Another naive method involves substituting a fixed value, such as the mean, median or the mode based on respondents for all subjects, with missing values. This strategy creates artificial "peaks" in the imputed (or the completed) data set, resulting in bias in any analysis involving statistics measuring spread or dispersion, such as regression analysis.

Despite several studies demonstrating bias through theoretical and simulation investigations in using complete-case and *ad hoc* methods, they continue to be used. There may be special circumstances or assumptions under which

some of these approaches may be valid, but those are hard to verify. Why bother with these approaches when better approaches are available to analyze incomplete data?

There is no assumption free method for analyzing incomplete data. Since the analysis involves something that is unknown (missing values) and something that is known (observed values), inference is always going to be conditional on the relationship between what is known and what is unknown. This relationship is often expressed through a missing data mechanism, a probability model or process that leads to missing values and its relationship to the study or analysis variables. It is very important to understand the missing data mechanisms to judge the appropriateness of an analysis procedure using the observed data.

1.2 Definition of Missing Values

A clear definition of a missing value should be established before proceeding further. Operationally (or heuristically), a value is defined to be missing for a variable if a meaningful value for the specific analysis to be performed is hidden. Examples of such situations include variables such as income, blood pressure, education, age, etc.

The definition also depends on the scientific question being answered. Consider an example where the definition of a missing value is not clear cut. Suppose that a survey is being conducted prior to an election where candidates from the major parties (A and B) are contesting for a seat. A question (X) is asked, "Are you going to vote for the party A or B?" and the response options are (1) A, (2) B and (3) Don't know. One may plan an analysis with X as having three response categories. For the projection of a winner, however, "Don't know" responses may be treated as missing values and a mechanism may be needed to classify "Don't knows" into vote for A or B.

The problem can be more complex when the following question is asked, (Y), "Are you planning to vote in the upcoming election?" with the response options (1) Yes, (2) No and (3) Don't know. Again, for some analysis the three-response category variables may be legitimate analytical variables. For the projection purposes, both variables (X and Y) have to be used and the "Don't know" response (3) in both variables have to be treated as missing values. The resolution of missing values in Y, determines the relevant population for handling the missing values in X.

Consider a longitudinal study measuring blood pressure repeatedly and some subjects dropped out of the study. If the subject is known to be alive, then the missing blood pressure measurement is a meaningful value for the analysis. If the subject is known to have died, then the blood pressure measurement is, generally, not a meaningful value for the analysis (dealing with selection due to death in the analysis is a separate issue). Even here, it is possible that some analysis may involve consideration of values such as "Had this person been alive at the time of measurement what would have been his/her blood pressure"? Such questions may arise in the competing risk analysis of two or more diseases.

An intuitive way to define the missing values for a variable in a specific analysis is to consider whether or not one should impute those values for subjects. In general, it is a good idea to flag all the imputed values to provide a flexibility in the analysis and for diagnostic purposes.

1.3 Missing Data Pattern

A pattern of missing data describes the location of the missing values in a potential complete data matrix (that is, data matrix with 100% response rate). For simplicity, consider a rectangular data matrix with rows representing subjects and columns representing variables. The rows and columns in the data matrix can be sorted or rearranged to get special patterns of missing data. The following figure illustrates various patterns of missing data.

Pattern (a) shows a monotone pattern of missing data where the variable $j = 2, 3, \ldots, p$ is observed on a subset of subjects with variable $j-1$ observed. This pattern of missing data typically arises in a longitudinal or a panel study when the drop outs from each wave are not followed in the future (generally, a bad idea).

Another common pattern is shown in Pattern (b) where the data is missing only on one variable in the analysis. Pattern (c) occurs when two files are appended where File 1 provides data on Y_1 and Y_2, and File 2 provides data

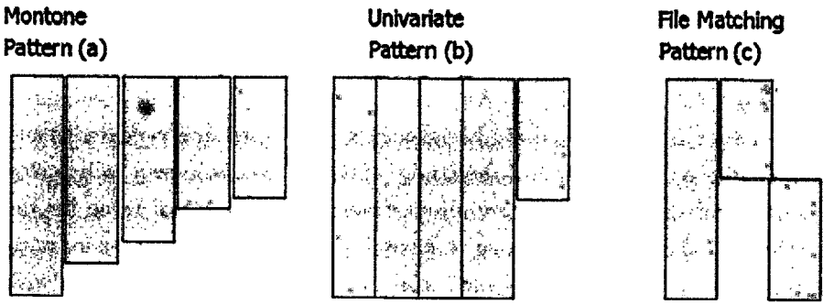


Figure 1.1: Patterns of missing data

on Y_1 and Y_3. This type of pattern also occurs in causal inference where Y_2 and Y_3 are the potential outcomes under two treatments $Y_1 = 1$ or $Y_1 = 0$, where Y_2 is not observed on those receiving treatment $Y_1 = 0$, and Y_3 is not observed on subjects receiving treatment $Y_1 = 1$.

The pattern of missing data could be exploited in the model specification or by breaking the estimation problem into simpler modular tasks. The second use of pattern is to understand the limitation of the data or identify parameters that cannot be estimated. For example, in Pattern (c), there is no information to estimate the partial correlation between Y_2 and Y_3 conditional on Y_1.

In most, if not all, practical situations, the pattern of missing data will be arbitrary or a general pattern of missing data. The methods described in this book are geared towards the general pattern of missing data but could be applied to other patterns of missing data. Whenever possible, alternative methods for a specific pattern of missing data will be suggested.

1.4 Missing Data Mechanism

To understand the concept of missing data mechanism, consider a case with single variable U with some missing values and a set of covariates V with no missing values. Let R be a response indicator taking the value 1 if U is observed and 0 if U is missing. The missing data mechanism is an assumed probabilistic or regression relationship between R and (U, V). One may view R as a "treatment assignment" in an experimental design context. This analogy

will be helpful in understanding the terminology used in the missing data mechanism.

Note that the substantive data (U, V) is not fully observed because U is not fully observed. The substantive data can be decomposed into the observed data, (U_{obs}, V), consisting of the observed set of values U_{obs} on subjects who provided them, and missing data U_{mis}, consisting of unknown values of U on subjects who did not provide information about U. Since R is a binary variable, the relationship between R and $D = (U, V)$ can be expressed through the probability specification, $Pr(R = 1|U, V)$ (the probability of observing U or the response propensity).

The values, U_{mis} are considered to be missing completely at random (MCAR), if the response propensity is a constant number across all the subjects. That is, $Pr(R = 1|U, V) = \text{constant}$. In the experimental design context, this mechanism is similar to the treatment assignment in a completely randomized design (CRD) where every individual in the sample has the same chance of being assigned the treatment $R = 1$.

The MCAR assumption implies that the distribution of the outcome variable is the same for both groups ($R = 0$ and $R = 1$). Thus, under MCAR the subsample corresponding to the complete-case data (U_{obs}, V_{obs}) is a simple random sample or a random subset of the full sample. There is no loss of "representativeness" of the sample for the population by restricting our analysis (U_{obs}, V_{obs}).

Generally, the complete-case analysis is valid under the MCAR mechanism. There are some exceptions. For example, the complete-case analysis yields unbiased estimates in a regression analysis with U as the dependent variable and some or all of V as independent variable even when the data are not MCAR (see Little and Rubin (2002) for more details). It is difficult to establish general conditions for the validity of complete-case analysis. Besides, the complete-case analysis usually will have larger sampling errors due to smaller sample size even if the point estimates are unbiased. Better methods (and software to implement them) are available.

The MCAR is a strong assumption. Often several factors or characteristics of subjects may influence the decision to answer survey questions. A weaker assumption is called missing at random (MAR). In the experimental design context, this mechanism may be viewed as randomized block design. Referring to the same setup with U and V, suppose that there are m individuals with

the same value of $V = v$, forming a block based on V. Let m_1 and m_o be the number of respondents and nonrespondents, respectively, in this block. Under the MAR assumption, the distribution of U is the same for the m_1 respondents and the m_o nonrespondents. That is, within each block, the assignment of the label of respondent/nonrespondent is completely at random. More generally, the response propensity under MAR is $Pr(R = 1|U, V) = Pr(R = 1|U_{obs}, V)$.

Whether MAR is a reasonable assumption depends upon the correlation between U and the blocking variable V. Strong correlation between V and U implies weak MAR assumption. For example, if U is income and V consists of only age and gender, then the MAR assumption conditional on age and gender is stronger in comparison to when V consists of age, gender, race, education, occupation, employment, etc. Hence, it is critical to obtain information on correlates of variables with missing values on the full sample to match the respondents and nonrespondents as much as possible to make residual differences between U_{obs} and U_{mis} within any block minimal and random.

Ignorable missing data mechanism is defined as MCAR or MAR and also "distinctness" of the parameters in the distributions of the substantive variables (U, V) and the response indicator R given (U_{obs}, V). Suppose that $Pr(U|V, \theta)$ or $Pr(U, V|\theta)$ is the statistical model (conditional or joint distribution) that the analyst will be using if he/she had the complete data, and θ is the unknown parameter to be inferred. Suppose that $Pr(R = 1|U_{obs}, V, \phi)$ is the response propensity with unknown parameter ϕ. If θ and ϕ are not functionally related to each other (or that the knowledge of one does not provide any information about the other) then the parameters are called "distinct," and the missing data mechanism is ignorable.

For a technical description, let Θ and Φ be the parameter space for θ and ϕ, respectively. A MAR or MCAR missing data mechanism is ignorable if the joint parameter space for (θ, ϕ) is $\Theta \times \Phi$ or in the Bayesian framework, the parameters θ and ϕ are *a priori* independent. Under this assumption, the model for the response propensity can remain unspecified in order to construct valid likelihood or Bayesian inferences under the posited substantive statistical model.

Finally, missing not at random (MNAR) is reserved for all the mechanisms that are not MCAR or MAR. That is, even after blocking on the variable V, the distributions of U_{obs} and U_{mis} are different. That is, the response propensity depends upon the unknown values U_{mis} (and possibly, U_{obs} and V).

Note that both MAR and MNAR are unverifiable assumptions based on the observed data. Since U_{mis} is not known, one cannot verify that U_{obs} and U_{mis} have similar or different distributions. If one suspects the distributions to be different, then this difference has to be posited, perhaps, based on substantive knowledge or external information to conduct appropriate analysis. There are two possible approaches to model the differences, the selection model and the pattern mixture model. Analysis under these models are discussed in Chapter 7.

1.5 Problems with Complete-Case Analysis

Practitioners often ask, "Is there a level of missing data for which complete-case or available case method is reasonable?". There is no simple answer to this question as it depends upon the fraction of incomplete cases, the parameter being estimated and the amount of information available from the subjects to be discarded from the analysis. Consider a population where a variable has a "bell-shaped" distribution and the goal is to estimate the population median. In a simple random sample from this population, a fairly large amount of missing data would be needed to move the median substantially. For estimating the 90^{th} percentile even a few missing observations can have a large impact. In a multiple regression analysis, each variable may have a small number of missing values but, collectively, a substantial number of observations may be discarded.

One of the best ways to understand the impact of missing data is through a carefully conducted simulation study. Let D be a binary dependent variable, E, a binary exposure variable and X, a single continuous covariate. Assume the following model to create these three variables:

1. $X \sim \text{Normal}(0, 1)$.
2. $E \sim \text{Bernoulli}(1, \pi(X))$ where $\text{logit}(\pi(X)) = 0.25 + 0.75 \times X$.
3. $D \sim \text{Bernoulli}(1, \theta(X, E))$ where $\text{logit}(\theta(X, E)) = -0.5 + 0.5 \times E + 0.5X$.

Suppose that the sample size is 1000. The goal is to infer about the regression coefficient in the logistic regression model (in (3) above) for D with E and X as predictors (the true value here is 0.5).

Delete some values of X with probability governed by the logistic regression model,

$$\text{logit}[Pr(X \text{ is missing})] = -1 - 0.5 \times D - 0.5 \times E + 3 \times D \times E$$

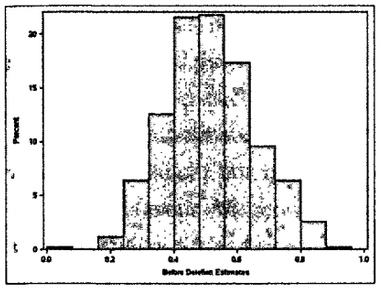
For each subject, calculate the probability based on the model, generate a uniform random number between 0 and 1 and set X to missing if this number is less than or equal to the probability. This particular model deletes values, on the average, for about 25% of the subjects but differs by D and E: 29% for $(D = 0, E = 0)$, 33% for $(D = 1, E = 0)$, 37% for $(D = 0, E = 1)$ and 6% for $(D = 1, E = 1)$. By design the data are missing at random as they depend only on the observed covariates D and E. Data are not missing completely at random as the four cell percentages of missingness are not the same. Table 1.1 provides the sample size and mean (SD) of the covariates X for the four cells formed by D and E in the before and after deletion data sets.

Clearly, the sample sizes in the four cells are different but the distribution of the covariates is similar for the before and after deletion data sets in each cell. This is expected given that the data are missing at random conditional on D and E.

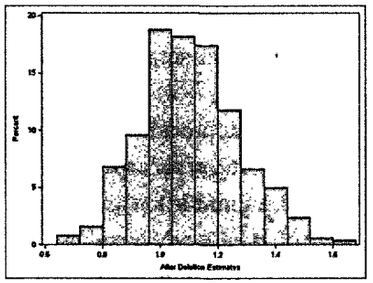
A logistic regression analysis of the before-deletion data set with D as the dependent variable and E and X as predictors results in the following estimates (standard error): Intercept, $\widehat{\beta}_o$ =-0.5372 (0.1059), the coefficient for E, $\widehat{\beta}_1 = 0.4478$ (0.1420), and the coefficient for X, $\widehat{\beta}_2 = 0.5541$ (0.0782). The analysis conducted on the after-deletion data set yielded -0.5904 (0.1281), 0.8850 (0.1683) and 0.5412 (0.0922), as the respective numbers. Estimates of the intercept and the coefficient for X are similar in the before and after

Table 1.1: Descriptive statistics based on simulated before and after deletion data sets

Cell	Before Deletion		After Deletion	
	Sample Size	Mean (SD)	Sample Size	Mean (SD)
$D = 0, E = 0$	299	-0.55 (0.90)	212	-0.57 (0.90)
$D = 0, E = 1$	269	0.06 (0.85)	170	0.11 (0.82)
$D = 1, E = 0$	145	-0.13 (0.84)	97	-0.13 (0.88)
$D = 1, E = 1$	287	0.50 (0.92)	269	0.51 (0.92)



(a) Histogram of the before deletion estimates across the 500 replicates



(b) Histogram of the after deletion estimates across the 500 replicates

Figure 1.2: Results from a simulation study of logistic regression analysis with missing covariates

deletion data sets, but the estimates of the coefficient for E are remarkably different.

The simulation experiment was repeated 500 times, each time generating a complete data, fitting the logistic regression model, setting some values to missing and then fitting the same regression model for the complete cases. Figure 1.2 provides the histograms of the estimates across the 500 replications. As expected, the before-deletion estimates are centered around the true value of 0.5 whereas the complete-case estimates of severely biased and the true value is not even within the realm of 500 complete-case estimates.

If the missing data mechanism logistic model does not involve the interaction term, $D \times E$, then it has been shown that the complete-case analysis is valid (see, for example, Vach (1994)). There may be other special cases for the complete-case analysis to be valid, and it is difficult to ascertain whether those conditions are met for a particular problem. Besides, the complete-case analysis is usually less efficient. Better strategies are available for using the observed data more effectively and incorporate additional variables in the data set that may be correlated with the variables with missing values.

1.6 Analysis Approaches

The major focus of this book is on options for the analysis of incomplete data under the MAR assumption. There are two general purpose approaches for

practice: weighting and imputation. Weighting approach assigns weights to subjects with no missing values to compensate for removing the subjects with missing values.

Imputation approach fills in a plausible set of values for the missing set. Multiple imputation incorporates the uncertainty in the plausible values by repeating the imputation process several times (say, M times). Each imputed set of values, when combined with the observed set of values, yields a completed data set. Each completed data set is analyzed separately. The completed data inferences are combined to form a single inference. The combining rules are relatively simple and easy to implement. Weighting or imputations involve assumptions about the nature of differences between subjects with and without missing values.

For many, imputations conjure "making-up of data." This would be true, if one were to create a single completed data set and then analyze it as though it were a complete data. The complete data cannot be created from the observed data, and the imputations are not the actual values for the nonrespondents. Collectively, however, the imputations under certain assumptions can create a plausible data set sampled from the population, thus resulting in a plausible inference about the population. This plausible inference needs to incorporate uncertainty due to imputations. The multiple imputation through multiple completed data sets incorporates this uncertainty.

The multiple imputation approach solves the missing data problem once because the same set of completed data sets can be used for a variety of analysis. This is especially appealing in a public-use setting where the data producer can use his/her knowledge, background variables and auxiliary variables to create multiple imputed (or completed data sets) and make it available to the users. The user repeats his/her analysis using his/her own complete data software on each completed data set and then combines the point estimates, standard errors, test statistics, etc. In fact, many software packages such SAS, STATA, SUDAAN and R have built-in functions to combine inferences.

Other methods can be used for incorporating the imputation uncertainty. For example, one could repeat the entire imputation process in a repeated replication setting (jackknife, bootstrap, balanced half samples, etc.) and then combine the estimates to arrive at a single inference. Under this approach, one could release all the replicated-imputed data sets to the users. The

replication approach is more computer intensive but may be useful, especially if the complete data inference is going to be based on replication technique.

The third approach works directly with the statistical model and the observed data to construct inferences for the parameters. The estimation can be motivated from the likelihood principles or from a Bayesian perspective. Though this approach may be preferred from a purely theoretical perspective, its implementation may require considerable programming and analytical skills.

The goal of this book is to provide a practical guide for using weighting and imputation approaches. The emphasis is on the multiple imputation (MI) approach because it is versatile in handling many applications. Many software packages are available to implement this method. Given its wider applications, MI may not be the most efficient approach for all the problems but has numerous advantages. For example, the MI approach can be made more efficient than some of the traditional methods such as observed data maximum likelihood with effective use of auxiliary variables.

The crux of the matter, of course, is how to create imputations. The most straightforward and principled approach is to use a Bayesian framework. Start with a model for the observable, conditional on some unknown parameters, and a prior distribution for the parameters. Then, conditional on the observed data, construct the joint posterior predictive distribution of the unobserved (parameters and the set of missing values), and draw values of the unobserved from this joint distribution. The drawn values of the missing set of values are treated as imputations.

Suppose that Y is the potential data matrix with no nonresponse. Let R be the response indicator matrix of the same dimension as Y; with 1 if the corresponding element in Y is observed and 0 if it is missing. A statistical model for the observable is a joint density/mass function of (Y, R) which may involve some unknown parameters θ. Let $\pi(\theta)$ be the prior density of θ. Let Y_{obs} denote all the observed values in Y (correspond to 1s in R) and let Y_{mis} be the unknown values in Y (corresponds to 0s in R). The relevant predictive distribution for generating imputations is

$$Pr(Y_{mis}|Y_{obs}, R) = \frac{\int f(Y, R|\theta)\pi(\theta)d\theta}{\int f(Y, R|\theta)\pi(\theta)d\theta dY_{mis}} \propto \int f(Y, R|\theta)\pi(\theta)d\theta,$$

as the denominator is a function of the known values (Y_{obs}, R).

The idealistic description of the imputation process is difficult, if not impossible, to implement in practice. This book adopts the basic principle but constructs imputations through several approximations of the predictive distribution. The most prominent or emphasized approach is the sequential regression multivariate imputation (SRMI), where each variable is predicted by all others in the data set through a sequence of regression models. Many other sensible approaches are also illustrated as alternatives. To emphasize the dangers of careless imputation, nonsensible but prevalent approaches are used to illustrate the pitfalls.

The book uses actual and simulated data sets to illustrate important aspects of weighting and multiple imputation approaches. The actual data sets arise from randomized trials, observational studies and sample surveys using both longitudinal and cross-sectional study designs. This book is primarily aimed for practitioners with emphasis on discussions of the underlying assumptions, methodology and implementation. As far as the analysis of incomplete data is concerned, there are always assumptions (there is no free lunch!). It is essential to understand the assumptions for each method to judge its appropriateness or make the necessary modifications. Even though the emphasis is on applications, several sections provide theoretical basis and discussion.

It is assumed that the reader is familiar with basic statistical techniques, such as rules of probability, Bayes theorem, the basis for statistical inference, general and generalized linear models, the likelihood based inference, etc. Typically, a two course sequence in statistics or biostatistics should be sufficient to read major portions of this book. Deeper knowledge of statistical inference topics are needed to understand certain theoretical material.

Each chapter contains a section on selected readings on relevant topics. These are not exhaustive as research on missing data is vast and is nearly impossible to include the full scope of published literature. Furthermore, the original articles where the methods or concepts are developed may be highly technical and, hence, alternative references are provided. Thus, the selection of bibliography is subjective but informed by teaching students at various levels.

1.7 Basic Statistical Concepts

There are two approaches for drawing inference about a population or phenomenon: Frequentist repeated sampling and Bayesian. The philosophical difference between these approaches arise because of differing concepts of probability: its definition, construction and use.

For a practitioner, ideas from both frameworks can be useful to solve a particular scientific problem, choosing the most scientifically appropriate interpretation. This book uses ideas from both frequentist and Bayesian perspectives, but, has a definite Bayesian flavor. This section gives a very brief overview of some of the basic statistical concepts useful to understand the material in the book. A course in Bayesian analysis will be helpful.

A desire to answer a substantive research question leads to an experiment which when implemented results in data. For example, the desire to know the prevalence of diabetes in a population (the research question) leads to an experiment that involves conducting a survey of n randomly chosen subjects from the population (experiment) resulting in x people reporting having diabetes and $n-x$ people reporting not having diabetes in the sample (data). The possible values of x are $\{0, 1, 2, \ldots, n\}$ and is called the sample space (a statement of all possible results from the experiment).

A statistical model posits probability of observing a result from any arbitrary portion of the sample space. In the example, if θ were the probability that a random individual in the population has diabetes, then the probability of observing $x = 0$ is $(1-\theta)^n$. In general, observing x diabetic subjects in the sample of size n is given by the binomial distribution,

$$f(x|\theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}$$

where the notation $n!$ stands for the product of the first n integers, $n \times (n-1) \times (n-2) \times \ldots \times 1$.

The goal of inference is to answer the following question: "Given that an experiment described above was implemented which resulted in x subjects with diabetes and $n-x$ without, what can be inferred about θ"? The two approaches, frequentist and Bayesian, differ in terms how this question is answered. Obviously θ can be any number between 0 and 1. The interval (0,1) is called the parameter space (to avoid technical difficulties, assume that

none being diabetic and all being diabetic are impossible scenarios for the population).

Under the frequentist framework, it is assumed that the true value of θ for this population is θ_o. A procedure is invented or developed to estimate θ_o which needs to satisfy some "good" properties across the entire spectrum of data that could result from the experiment. That procedure should then be applied to the data in hand to infer about θ_o.

The two prominent "good" properties are unbiasedness and minimum variance. Both are conceptualized based on a thought experiment. Suppose that a procedure yields an estimate $\widehat{\theta}(x)$ which depends upon x. Assume that experiment is conducted millions of times (or infinite number of times), each time resulting in x and $\widehat{\theta}(x)$. The estimate is called unbiased if the average of the estimates across the repeated experiments is equal to θ_o. This can be stated as

$$E(\widehat{\theta}(x)|\theta_o) = \sum_{x=0}^{n} \widehat{\theta}(x) f(x|\theta_o) = \theta_o.$$

The second property of minimum variance can be described as follows. Suppose that $\tilde{\theta}(x)$ is any other unbiased procedure for estimating θ_o. The variance of $\tilde{\theta}(x)$ across the repeated experiments is larger than the variance of $\widehat{\theta}(x)$. This can be stated as

$$\text{Var}(\widehat{\theta}(x)|\theta_o) = \sum_{x=0}^{n} (\widehat{\theta}(x) - \theta_o)^2 f(x|\theta_o)$$

$$< Var(\tilde{\theta}(x)|\theta_o) = \sum_{x=o}^{n} (\tilde{\theta}(x) - \theta_o)^2 f(x|\theta_o).$$

The sampling variance estimate (or simply the sampling variance) of $\widehat{\theta}(x)$ is defined as an estimate of the $Var(\widehat{\theta}(x)|\theta_o)$. The standard error is the square root of the sampling variance estimate. In the binomial example, the only "good" estimate (unbiased and minimum variance) is the sample proportion $\widehat{\theta}(x) = x/n$ and its sampling variance is $\widehat{\theta}(x)(1 - \widehat{\theta}(x))/n$.

Instead of a single estimate, it may be desirable to construct an interval that covers the true value, θ_o . A procedure results in an interval $(\widehat{\theta}_L(x), \widehat{\theta}_U(x))$ such that $Pr(\widehat{\theta}_L(x) \leq \theta_o \leq \widehat{\theta}_U(x)) \geq 1-\alpha$ for some prespecified α. If $\alpha = 0.05$ then the interval is called a 95% confidence interval. It is preferable to be as close to equality (to $1-\alpha$) as possible and a "good" property of the procedure is to result in a shortest possible interval for every x.

Thus, for every experiment, the idea is to construct a point and interval estimation procedure with good properties for every target quantity of interest and then apply that procedure for the particular data set in hand to construct inferences. The suffix "o" for θ may be omitted because any value in the parameter space can be the true value. Thus, the procedures should have "good" properties for any value in the parameter space.

Another type of inference is testing a preconceived hypothesis about the likely value(s) of the target quantity of interest which will not be discussed in this section but can be found in many basic textbooks (see Section 1.9). It suffices to say that, this question may also be answered by checking whether the preconceived values are included in the interval estimate.

A Bayesian accepts that there is a true value θ_o but, more importantly, the Bayesian view is that since θ_o is not known then there is *a priori* uncertainty about its value, and that should be expressed in a form of a prior distribution for θ, with the density $\pi(\theta)$ defined on the parameter space. This prior distribution could be constructed from pilot data, data from similar populations or the subject matter knowledge.

The statistical model accepted by both, frequentists and Bayesians, specifies $f(x|\theta)$ and, therefore, the product $f(x|\theta)\pi(\theta)$ is the joint distribution of (x, θ). The marginal density is $f(x) = \int_0^1 \pi(\theta, x)d\theta$, and the conditional density is,

$$\pi(\theta|x) = \frac{\pi(\theta, x)}{f(x)} = \frac{L(\theta|x)\pi(\theta)}{f(x)}, \tag{1.1}$$

where $L(\theta|x)$ is the binomial density, $f(x|\theta)$, evaluated at the observed value of x (so, it is the function of the parameter, θ) and is called the likelihood function. The conditional density given in equation (1.1) is called the posterior density of θ given the data x (and n, of course). This is the direct result of Bayes theorem in probability. This binomial problem was considered in the paper by Thomas Bayes published posthumously in 1763, through the efforts of his friend Richard Price.

The posterior density forms the basis for all Bayesian inferences about θ. This density is used to directly answer the questions of the type "What is the probability that θ is in the interval (a, b)"? The answer to this question is the integral $\int_a^b \pi(\theta|x)d\theta$.

The posterior mean defined as

$$E(\theta|x) = \int_0^1 \theta\pi(\theta|x)d\theta$$

is the average value of θ and the posterior variance

$$Var(\theta|x) = \int_0^1 (\theta - E(\theta|x))^2 \pi(\theta|x) d\theta$$

can be used to express the uncertainty.

To construct interval estimates, suppose that (a, b) is such that

$$Pr(a \leq \theta \leq b|x) = \int_a^b \pi(\theta|x) d\theta = 1 - \alpha$$

and for any value of θ in the interval (a, b) and θ^* outside the interval (a, b), $\pi(\theta|x) > \pi(\theta^*|x)$. That is, the values of θ inside the interval have higher values of the posterior density than for those outside the interval. The interval (a, b) is called the $100(1 - \alpha)\%$ highest posterior density credible interval. The interpretation is that the probability of the prevalence rate being between a and b is $1 - \alpha$.

For the binomial example, assume that all values of θ are equally likely, leading to a uniform prior distribution,

$$\pi(\theta) = 1,$$

for interval (0,1) and is set to 0 outside the interval (as they are impossible). The posterior density can be shown to be a beta distribution with parameters $x + 1$ and $n - x + 1$,

$$\pi(\theta|x) = \theta^x (1 - \theta)^{n-x} / B(x + 1, n - x + 1)$$

with the posterior mean $\tilde{\theta}(x) = (x + 1)/(n + 2)$ and the posterior variance

$$Var(\theta|x) = \frac{\tilde{\theta}(x)(1 - \tilde{\theta}(x))}{n + 3}$$

which is not that different from the estimates from the frequentist perspective. In fact, $\pi(\theta) = \theta^{-1}(1-\theta)^{-1}$ will give results identical to the frequentist answer. However, this is not a proper density function as it is not integrable with respect to θ.

The uniform prior is equivalent to adding 1 to both the numerator and the denominator (the number of diabetic and nondiabetic subjects). A diffuse Jeffreys prior, $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$, would add 1/2 instead of 1 which is shown to have better variance properties, though slightly biased.

In practice, the prior distribution is usually diffuse. Heuristically, in the binomial example, if the likelihood function and the prior density were plotted against θ, the prior density will be relatively flat when compared to the likelihood function. For a variety of problems, a diffuse prior distribution is almost equivalent to working directly with the likelihood function. It is an accepted principle that the likelihood function is the best "summary" of information in the data about the parameters, the Bayesian point of view provides the best mechanism for studying the likelihood function and constructing inferences. Thus, the Bayesian approach generally yield methods that have good frequentist properties in a broader definition of "good" procedures. Obviously, this is a point of view of a practitioner who wishes to develop easily interpretable inferences that can be justified from both philosophical perspectives.

All prior distributions in this book are diffuse in the sense that "likelihood function dominates the prior density." The diffuse priors are not always easy to define. Jeffreys prior (improper) is one example where all location parameters are treated as uniform on the real line, and the prior distribution of all scale parameters are uniform on the logarithmic scale. The prior distribution for the binomial parameter, θ, is either uniform on (0,1) or a beta prior $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for small values of α and β. The main emphasis is on using the information in the data set summarized in the likelihood. Of course, methods described can be modified to incorporate information for constructing a proper prior distribution.

Modern computation methods have made simulating values from the posterior density of the parameters as a convenient and efficient way of drawing inferences. Throughout the book, this strategy is adopted as imputations are also simulation of the missing values and fits well within the Bayesian framework.

For the binomial example, the uniform prior, $\pi(\theta) = 1$, results in the posterior density of θ as a beta distribution with parameters $x + 1$ and $n - x + 1$. By simulating a large number of values from this beta distribution, many features of the posterior distribution can be studied. Furthermore, any transformation (for example, logit, $\phi = \log[\theta(1-\theta)^{-1}]$) of the simulated values are the draws from the posterior distribution of the transformed parameter. In complex problems, this feature will be useful where it may be easy to draw from one distribution but not from the other.

Extending the simple experiment, suppose that only m of n individuals answered the question about diabetes and w out of m are diabetic. In the absence of any other information, assume the data are MCAR (or MAR). Imputation involves drawing from the predictive distribution of

$$Pr(y_{m+1}, y_{m+2}, \ldots, y_n | w, m, n)$$

where $y_{m+1}, y_{m+2}, \ldots, y_n$ are the binary yes/no variables for the $n-m$ respondents. Note that,

$$Pr(y_{m+1}, y_{m+2}, \ldots, y_n | w, m, n) =$$

$$\int_0^1 Pr(y_{m+1}, y_{m+2}, \ldots, y_n | w, m, n, \theta) \pi(\theta | w, m) d\theta.$$

Now, conditional on θ, $y_j, j = m+1, m+2, \ldots n$ are independent Bernoulli random variables with parameter θ and θ has a beta posterior distribution with parameter $w+1$ and $m-w+1$. Thus, the imputation involves 2 steps:

1. Draw θ, say θ^*, from the beta distribution with parameter $w+1$ and $m-w+1$.
2. Draw $n-m$ independent uniform random numbers, $u_{m+1}, \ldots, u_n$ and set $y_j = 1$ if $u_j \leq \theta*$.

This two stage imputation process of drawing a value of the set of parameters and then conditional on drawn parameters, drawing the missing values is the standard strategy and is routinely used.

Now, suppose that gender (male/female) was observed on all n subjects and, under MAR, the response propensity may depend on gender. In this situation, the imputations are carried out in the same manner, except separately for males and females. When one has a large number of covariates, a regression model may be useful for creating imputations.

The frequentist concepts based on repeated sampling are useful in model checking. The thought experiment underlying the frequentist framework can be made a reality by simulating thousands of copies of data from the model. If the observed data is not within the realm of plausibility among the generated data sets (several examples will be given later in the missing data context), then the model is questionable and should be refined.

1.8 A Chuckle or Two

Many statistical properties of an inference procedure are established under the "true" model (or correctly specified model). Generally, the concept of true model is quite elusive. The thought of a proposed model being true is perhaps wrong almost all the time (a famous quote attributed to George Box is "All models are wrong but some are useful"). The concept of true response propensity is even more elusive than the prediction model. The following story, told in many classes might illustrate the point and give an opportunity to have a chuckle or two.

John, a dashing young man, just broke up with his girlfriend and was feeling depressed. He was watching TV but his mind was wandering, not concentrating on what was on the TV, as his fingers mindlessly push the buttons on the remote control. The doorbell rang and John wondered "who could that be"?

He got up and opened the front door. There was an older looking, affable gentleman who introduced himself as Peter from the University of Michigan, conducting a survey funded by a government agency. Peter explained the topic and all the usual preambles that an excellent interviewer provides to any potential respondent. John thought that this would be a great distraction and invited Peter to his living room. Peter opened his computer and administered the informed consent and thus began the interview.

John was enjoying the interview and was intrigued by the questions. Both John and Peter were developing a good rapport and the survey interview was progressing nicely. Suddenly the phone rang, startling both John and Peter. John, saw that it was from his girlfriend (ex?), Jody. He said, "Excuse me, let me take this phone call, I will be right back." He left the living room to be in private and while walking he heard Jody sobbing. She was mumbling "I am so sorry, so sorry...." John couldn't make out what she was saying and said, "Honey, why are you crying? What is the matter"? His heart was thumping loudly. Jody composed herself, and said "I am sorry; I shouldn't have broken up with you. I want to give another chance to our relationship. Can you come here right now, please"? John melted. Now John felt Peter's company as a nuisance!! John said, "Sure honey, I will be there in few minutes."

3

Imputation

An imputation based approach involves replacing the set of missing values with a plausible set of values. This plausible set is created or estimated based on available information. This task should not be construed as creating "actual" or "real" values for nonrespondents. For a layman, this idea conjures an image of a statistician making up the data. This is true only if one were to analyze the data as if the imputed data are real values. Thus, any single imputation analysis that does not account for the imputation process in the ensuing statistical inferences may inflate the precision and declare effects significant when they are not, and thus invalid.

Consider a simple example to assess the severity of this problem. Suppose that, in a sample of size 100, 70 respondents provided the data yielding a respondent sample mean of 50 and a sample standard deviation of 10. Under the missing completely at random assumption, the estimate of the population mean is 50 and the standard error is $10/\sqrt{70}$. Suppose that the observations on 30 subjects are filled in by drawing values at random with replacement from the 70 subjects. The filled in data, when plugged into any software package, to estimate the mean and its standard error will yield the mean of 100 observations (70 original and 30 filled-in) to be 50, on average and the standard error as the standard deviation of the filled in data (which on average will be 10) divided by $\sqrt{100}$. Of course, this is incorrect because the 30 observations filled are not adding any information that is not already there in the 70 observations. In fact, in the absence of any other information, the correct standard error from the filled-in data set has to be somewhat larger than $10/\sqrt{70}$ as we are adding noise to the observed data. Thus, the analysis of singly imputed data, without reflecting the uncertainty about the imputed values is incorrect and should be avoided in practice.

This chapter discusses approaches for creating plausible values for the missing set of values and Chapter 4 discusses construction of correct inferences from the data set with imputed values. The major emphasis is on multiple

imputation. An alternative approach on replicating the imputation process using bootstrap or jackknife techniques is discussed in Chapter 9.

The goal of imputation is to fill in the missing set of values with **plausible values** to obtain a **plausible sample data** from the population. This can then be used to construct **plausible inferences** about the population quantities of interest. The notion of plausibility can be best explained heuristically. Consider a scenario with two variables Y_1 and Y_2 in a data set with Y_1 fully observed and Y_2 with some missing values. Assume that the data are missing at random. That is, the missing data mechanism depends only upon Y_1. Suppose that the missing values in Y_2 have been imputed using some procedure. How does one check heuristically whether the imputed values are plausible from the population?

A simple approach is to construct a scatter plot based on the **completed data set** (or the filled-in data set) with Y_1 on the horizontal axis, Y_2 on the vertical axis and with observed and imputed values differentiated using colors or symbols. For a plausible data set from the population (under the stated assumptions), the observed and imputed values should be interchangeable. The idea of plausible data set rules out imputing the mean of the observed values (unconditional mean) or the predicted values from a regression equation. These methods may be useful for specific purposes, but they are not general purpose imputation methods.

Suppose that in a sample of size n measuring a variable Y, only m values are observed. The mean of the observed values, $\bar{y}_o$, is imputed for every $n-m$ missing values. Assume that the data are MCAR (or MAR in the absence of any covariates). In this case, the complete-case analysis is correct. The mean of the completed data is $\bar{y}_o$ which is the correct mean. The complete data variance will be inflated as the numerator is still $\sum_i^m (y_i - \bar{y}_o)^2$, but the denominator is $n-1$. That is, the completed-data variance is smaller than s_o^2, the variance based on m observed values; the completed data variance is $s_C^2 = s_o^2(m-1)/(n-1)$. Depending upon the lack of symmetry among the observed values, the median, mode and other percentiles also will be incorrect (that is, different from the complete-case analysis).

Now, consider a problem with two variables Y_1 and Y_2 with Y_1 fully observed, and Y_2 is observed on m out of n individuals. Assume that the data are missing at random. A linear regression model may be developed and the predicted values from this model may be considered as imputation. The

observed values would have error around the regression line, whereas the imputed values do not. Again, the variance computed from the completed data will be smaller affecting other statistics. Statistics that are not affected are the estimated regression coefficients (intercept and slope).

Sometimes the mean imputation is used (rather inappropriately) in the scale construction. Suppose that a scale is defined as summing a set (say, k) of yes/no (or Likert scale items) items. Not all scale items are observed. An arbitrary rule is made: For a given subject, if more than k_o items are missing then his/her scale value is set to missing; otherwise, substitute the mean of the observed item responses for the missing items and then proceed with the scale construction. At the face value, there are two things wrong with this approach. First, it ignores some observed values on a set (may even be a large number) of subjects. Second, it makes the assumption that there will be no variation in the responses of the items that are missing. This artificially decreases the variation in the scale and affects all analyses that involves the spread of the scale (i.e., correlation and regression analysis).

Substitution of mean or predicted values was developed in the era of limited computational power and for situations where the tabulation of means and proportions is of interest such as reports from the government agencies and other official statistics documents. These methods should be avoided in practice now given the methodological developments, vastly improved computational power and software environment. In the later chapters, pitfalls of using such methods will be illustrated through examples and homework problems.

3.1 Generation of Plausible Values

The most straightforward approach to generate a plausible completed-data set is through draws from the predictive distribution of the missing set of values, conditional on the observed set of values (and external information, if any). For the simple bivariate problem, a set of plausible values may be generated as follows:

1. Fit a regression model $Y_2 \doteq f(Y_1) + \epsilon$ and define $\widehat{Y_2} = \widehat{f}(Y_1)$ for both respondents and nonrespondents.

2. Construct residuals for the respondents, $\widehat{\epsilon} = Y_2 - \widehat{Y_2}$.
3. For each nonrespondent, draw a random residual from the respondents and add it to his/her predicted value.

The above approach, though works in practice, does not account for all the uncertainties involved in the construction of the predictive distribution. A principled way to generate plausible values is to use a Bayesian framework. The following modification of the above procedure incorporates the uncertainties:

1. Suppose that $I = \{1, 2, \ldots, m\}$ are the indices of the respondents (that is, both Y_1 and Y_2 observed). Draw a sample of size m from the index set I with replacement and denote it as I^*.
2. Sample m values from the index set I^* with replacement and extract the corresponding values of Y_1 and Y_2.
3. Carry out the steps (1), (2) and (3) in the simple approach discussed previously.

Steps (1) and (2) in the modified approach is the approximate Bayesian bootstrap (ABB) and incorporates the uncertainty in estimating the prediction equation and in the distribution of the residuals. This procedure uses a mix of parametric and nonparametric assumptions to construct the predictive distribution. It uses a parametric model for the functional relationship but arbitrary or unspecified distribution for the residuals.

Now consider the following procedure. Create C strata based on Y_1. Suppose that n_c and m_c are the sample size and the number of respondents, respectively, in cell $c = 1, 2, \ldots, C$. Randomly sample $n_c - m_c$ observations from the m_c observations as imputation set for that cell. This has nonparametric flavor where empirical distribution in each cell is used as the predictive distribution. This is the basic setup of the hot-deck method.

A fully parametric model may be used, if appropriate. Suppose that the following regression model, $Y_2 = \beta_o + \beta_1 Y_1 + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ fits well to the respondent data. Assume that the prior information is diffuse or noninformative as indicated before, $Pr(\beta_o, \beta_1, \log\sigma) \propto c$. Let $\widehat{\beta} = (\widehat{\beta_o}, \widehat{\beta_1})^t$ be the 2×1 vector of regression coefficients, $V = X^t X$ is the 2×2 matrix where X is $m \times 2$ matrix with the first column of 1s and the second column Y_1 and

$\widehat{\sigma}^2$ is the estimated residual variance. The following steps result in plausible values for the missing set:

1. Draw a chi-square random variable, u, with $m-2$ degrees of freedom and define $\sigma_*^2 = (m-2)\widehat{\sigma}^2/u$. This step draws a value from the posterior distribution of σ^2.
2. Let T be the Cholesky decomposition of V such that $TT^t = V$. Let $Z = (z_1, z_2)^t$ be a 2×1 vector of independent standard normal deviates. Define $\beta^* = \widehat{\beta} + \sigma_* TZ$. This step draws from the posterior distribution of β conditional on the observed data and σ_*.
3. For a nonrespondent, define the imputed value as $Y_2^* = \beta_o^* + \beta_1^* Y_1 + \sigma_* z_3$, where z_3 is another independent standard normal deviate. Repeat the process for all nonrespondents (each time drawing a new z_3). This step draws from the posterior predictive distribution of Y_2 given the observed data, β^* and σ^*.

This approach assumes only the observed data [(Y_2, Y_1) on the respondents and Y_1 on the nonrespondents] as known and all the missing values in Y_2 and model parameters β_0, β_1 and σ^2 as unknown, drawing values from their joint posterior distribution, conditional on the observed data. This approach is implemented in many missing data software packages.

Thus, there are many ways of constructing the predictive distribution. Some are described as algorithms, and others are based on explicit model assumptions. Underlying all the algorithms, there is an implicit model. This model needs to be gleaned from the algorithm and checked against the data. That is, the model needs to be "validated" through proper model diagnostics for obtaining valid inferences from the imputed data. Careless imputations can yield biased estimates (even more biased than the complete-case estimates).

3.2 Hot Deck Imputation

The predictive distribution can be created based on explicit or implicit models as explained in the previous section. Hot deck is a popular approach for creating imputations using an implicit model. To be concrete, consider the data in Table 1.1 with four cells based on D and E with missing values in

X. Just like in the weighting procedure, consider the four cells as imputation cells.

A hot deck procedure creates imputations by sampling with replacement the needed number of observations from the respondent pool within each cell. For example, in the cell with $D = 0, E = 0$, sample 87 observations with replacement from 212 respondents. An implicit model is that, for each cell, the predictive distribution for the missing observations is the "empirical distribution" estimated based on the respondents. The empirical distribution function can be expressed as

$$\widehat{F}_{de}(x) = \sum_{i}^{m_{de}} I_{[x_i \leq x]}/m_{de},$$

where m_{de} is the number of respondents in the cell $D = d, E = e; d, e = 0, 1$, and $I_{[x_i \leq x]} = 1$ if $x_i \leq x$ and 0 otherwise is the indicator function.

If the number of respondents in each cell is small, then some smooth estimate of $\widehat{F}$ can be constructed or some cells can be collapsed. This procedure closely aligns with the adjustment cell weighting method. Here instead of assigning weights to the respondents, the values are imputed for the nonrespondents by drawing values from the respondents.

There are many other ways of constructing hot deck imputation cells. For example, stratification based on the propensity score can form imputation cells. In the AHEAD example, the quartiles based on the propensity scores may be considered as imputation cells, the values for the nonrespondents in any given cell can be sampled from the respondents from the same cell.

Another popular approach for creating imputation cells is based on the predicted value of the variable with missing values (predictive mean matching). Suppose that U is the variable with some missing values, and V is the set of predictors with no missing values. Regress U on V using an appropriate, good fitting model based on the respondents. To construct the prediction equation one should use proper exploratory and diagnostic techniques to arrive at a reasonable model. Suppose $\widehat{U}_i$ is the predicted value for subject $i = 1, 2, \ldots, n$, where n is the full sample size. Note that U is observed for the respondents, but $\widehat{U}$ can be computed for the full sample. Create cells based on $\widehat{U}$, and then sample an appropriate number of observations with replacement from the respondent pool in each cell. Unlike the methods based on the cross-classification of V or the propensity score stratification, this approach uses the observed data on U as well.

There are no fixed rules in terms of how to construct the cells. The goal is to make the respondents (called donors of their responses) and nonrespondents (called recipients of the values from the donors) as similar as possible on the variables measured on both. One could use the response propensity, predicted values and any other combination of variables to create imputation cells. This flexibility in using the observed data makes the approach quite useful in practice.

3.2.1 Connection with Weighting

Suppose that C imputation cells or strata have been formed with n_c and m_c as the sample size and respondent size, respectively, in stratum $c = 1, 2, \ldots, C$. Let $y_{ic}, i = 1, 2, \ldots, m_c; c = 1, 2, \ldots, C$ be the observed responses. Let $\bar{y}_c$ be the sample mean in cell c. The weights for observations in cell c is $w_c = n_c/m_c$ and the weighted estimate is $\bar{y}_w = \sum_c w_c \bar{y}_c/n$. The weighted mean can be rewritten as

$$\bar{y}_w = \sum_c \sum_i^{m_c} w_{ic} y_{ic}/n$$

where $w_{ic} = n_c/m_c$.

Instead of weighting, suppose that the missing values are imputed by drawing $n_c - m_c$ values from the m_c respondents in each cell. Let r_{ic} be the number of times y_{ic} was selected in the imputation process. The imputation or completed-data estimate is

$$\bar{y}_I = \sum_c \sum_i (1 + r_{ic}) y_{ic}/n.$$

In cell c, $n_c - m_c$ draws are made and each of the m_c observations has a chance of $1/m_c$ of being drawn at each draw. Since the draws are independent, $E(r_{ic}|D) = (n_c - m_c)/m_c$ or $E[(1+r_{ic})|D] = w_{ic}$ where D is the observed data, $\{n_c, m_c, y_{ic}, i = 1, 2, \ldots, m_c; c = 1, 2, \ldots, C\}$. The expected value of $\bar{y}_I$ over repeated imputations of the missing values, conditional on the observed data is $\bar{y}_w$. Thus, across repeated imputations and conditional on the observed data, the hot deck estimate will be distributed around the weighted mean (that is, $E(\bar{y}_I|D) = \bar{y}_w$).

Conditional on the observed data, the variance of $\bar{y}_I$ is

$$Var(\bar{y}_I|D) = \frac{1}{n^2} \sum_c \sum_i y_{ic}^2 Var[(1 + r_{ic})|D].$$

Noting that $Var[(1+r_{ic})|D] = (n_c - m_c)m_c^{-1}(1-m_c^{-1})$. Some algebraic simplification obtains

$$v_D = Var(\bar{y}_I|D) = \sum_c \left[\frac{n_c - m_c}{n^2}\{(1-m_c^{-1})^2 s_{yc}^2 + (1-m_c^{-1})\bar{y}_c^2\}\right],$$

where $\bar{y}_c$ and s_{yc} are mean and the standard deviation of the observations in cell c.

Thus, the unconditional variance of $\bar{y}_I$ is

$$Var(\bar{y}_I) = Var[E(\bar{y}_I)D)] + E[Var(\bar{y}_I|D)]$$

$$= Var(\bar{y}_w) + E(v_D).$$

It is analytically difficult to compute $E(v_D)$ since it is a nonlinear function of the random variables $n_c, m_c, \bar{y}_c$ and s_{yc}^2. It is of the order n^{-1} as can be seen using the following approximation. Assume that $1 - m_c^{-1} \approx 1$, hence,

$$v_D = \frac{1}{n}\left(1 - \frac{m}{n}\right)\sum_c p_c(s_{yc}^2 + \bar{y}_c^2)$$

where $p_c = (n_c - m_c)/(n-m)$, $n = \sum_c n_c$ is the sample size and $m = \sum_c m_c$ is the number of respondents. In this situation, there is no benefit in using the hot deck imputation estimates over the standard weighted estimate as it is less efficient and randomly varies from the weighted estimate.

3.2.2 Bayesian Modification

From a Bayesian point of view, all uncertainties need to be incorporated to obtain valid inferences. In the hot deck procedure, the imputations are drawn from the estimated (or empirical) distribution without incorporating uncertainty in that estimate. A simple modification uses the Bayesian bootstrap in each imputation cell as follows:

1. Draw $m_c - 1$ uniform random numbers between 0 and 1 and order them to obtain $u_o = 0 \leq u_1 \leq u_2 \ldots \leq u_{m_c-1} \leq 1 = u_{m_c}$.
2. Draw a uniform random number, v, and choose y_{ic} as the imputation if $u_{i-1} < v \leq u_i$.
3. Repeat Step 2 to fill all $n_c - m_c$ missing values in the cell.
4. Repeat the procedure for all the cells.

The underlying theory is that the distinct values in the set $\{y_{ic}, i = 1, 2, \ldots, m_c\}$ are modeled as having a multinomial distribution with unknown cell probabilities. These probabilities are assumed to have a noninformative Dirichlet distribution. The above procedure is equivalent to obtaining draws from the posterior predictive distribution of the missing values conditional on the observed values under this model assumption.

An alternative approach is to draw values using the approximate Bayesian bootstrap procedure as follows:

1. Sample m_c values with replacement from the m_c values of $y_{ic}, i = 1, 2, \ldots, m_c$.
2. Sample $n_c - m_c$ with replacement from the sample obtained in step 1.
3. Repeat the process for all the cells.

The process for creating imputation cells is similar to developing adjustment cells for weighting. All model building and diagnostic procedures need to be applied to create sufficiently large and homogenous groups so that missing at random (conditional on the imputation cells) becomes plausible. The variables used to form imputation cells needs to be correlated with the variables being imputed.

The hot deck procedure described so far focusses on imputing a single variable, conditional on fully observed covariates on respondents and nonrespondents. In practice, however, several variables may have missing values and the number of fully observed covariates may be limited or even nonexistent. Thus, a sequential or a multivariate approach may have to be used to impute all the missing values as described later.

3.3 Model Based Imputation

Let y_i be the complete data vector on subject $i = 1, 2, \ldots, n$ which is assumed to be sampled from the population that follows a distribution with the density function $f(y_i|\theta, z_i)$, where z_i is a vector of auxiliary variables with no missing values. The statistical model is explicitly stated (i.e., parametric such as multivariate normal, log-linear model, etc. or semiparametric such as generalized

additive models etc. or nonparameteic such as classification and regression trees, etc.). Let r_i denote a vector (of the same dimension as y) of response indicator variables with 1 if the corresponding element in y_i is observed and 0, if missing. Let $y_{i,obs}$ denote the components of y_i that are observed (corresponding elements in r_i are all equal to 1) and $y_{i,mis}$ denote components of y_i that are missing (corresponding elements in r_i are all equal to 0). The notation can become a bit confusing because r_i is used as a random variable defining the missing data mechanism and also as an indexing set to identify which variables are observed or missing.

The missing data mechanism is the probability specification $g(r_i|y_i, z_i, \phi)$, governing which components in y_i are observed or missing. Thus, the full statistical model is the joint distribution of (y_i, r_i), $f(y_i, r_i|z_i, \theta, \phi)$, which can be decomposed into $f(y_i|z_i, \theta)g(r_i|y_i, z_i, \phi)$. This is called a selection model decomposition. On the other hand, the joint distribution can be decomposed into $h(y_i|r_i, z_i, \alpha)k(r_i|z_i, \beta)$, which is called a pattern-mixture model decomposition. In practice, sometimes the selection model decomposition is easier to formulate and may be useful from a conceptual point of view, and other times the pattern-mixture model may be more appealing. Obviously, when both are defining the same joint distribution, there is a mapping,

$$f(y_i|z_i, \theta)g(r_i|y_i, z_i, \phi) = h(y_i|r_i, z_i, \alpha)k(r_i|z_i, \beta.).$$

Sometimes the mapping may be useful to understand the proposed missing data mechanism, under a pattern mixture model, or to understand the pattern mixture model under the proposed selection model decomposition. The difference between the two formulations will play a crucial role under the NMAR mechanism, discussed in Chapter 7.

Under MAR mechanism, $g(r_i|y_i, z_i, \phi) = g(r_i|y_{i,obs}, z_i, \phi)$, and ignorability condition (that is, ϕ and θ are not functionally (from a frequentist perspective) or probabilistically (from a Bayesian perspective) related to each other), the predictive distribution is

$$f(y_{i,mis}|y_{i,obs}, z_i, r_i, \theta, \phi) \propto f(y_{i,mis}|y_{i,obs}, z_i, \theta).$$

Thus, only the substantive model is needed for imputation purposes and the missing data mechanism can be unspecified (as long as ignorability condition is assumed to be true).

Obviously, θ is not known. If $\widehat{\theta}$ is an estimate then the imputations can be drawn from $f(y_{i,mis}|y_{i,obs}, z_i, \widehat{\theta})$. This approach ignores the uncertainty in the estimate $\widehat{\theta}$. A proper procedure is defined below:

1. Let $\pi(\theta)$ be the prior density for θ.
2. Let $y_{obs} = \{y_{i,obs}, i = 1, 2, \ldots, n\}$, $z = \{z_i, i = 1, 2, \ldots, n\}$. Construct the posterior density

$$\pi(\theta|y_{obs}, z) \propto \pi(\theta) \times \prod_i^n f(y_{i,obs}|z_i, \theta)$$

   where $f(y_{i,obs}|z_i, \theta) = \int f(y_i|z_i, \theta)dy_{i,mis}$ is the marginal distribution of the observed portion of the complete data.
3. Draw a value θ^* from $\pi(\theta|y_{obs}, z)$ and then draw imputations from $f(y_{i,mis}|y_{i,obs}, z_i, \theta^*)$ and thus incorporating uncertainty about θ.

The idealized description may not be achievable in practice. There are many ways to approximate this strategy (some were discussed for the simple bivariate example). Some general methods are listed below:

1. Let $\widehat{\theta}$ be the mode of the posterior distribution and $\widehat{V}$ be the second derivative of the log posterior density. This is easy to implement using the Newton-Raphson method or other numerical optimization techniques, available in SAS, STATA and R packages for various statistical models. Approximate the posterior density of θ by a multivariate normal density with mean $\widehat{\theta}$ and $(-\widehat{V})^{-1}$ as the variance. Draw θ^* from this normal distribution and then draw from $f(y_{i,mis}|y_{i,obs}, z_i, \theta^*)$ to create imputations.
2. When the sample size is small or in the case of skewed situations (such as logistic regression with rare events), the multivariate normal distribution may not be a good approximation. Importance ratios may be used to modify the above algorithm. Suppose that $\pi(\theta|y_{obs}, z)$ can be calculated up to a constant of proportionality. Generate several values $\theta_l, l = 1, 2, \ldots, L$ from the approximate multivariate normal distribution in (1). Calculate the importance ratios, $w_l = \pi(\theta_l|y_{obs}, z)/\phi(\theta_l|\widehat{\theta}, -\widehat{V}^{-1})$ where $\phi(x|\mu, \Sigma)$ is the multivariate normal density function with mean μ and the covariance matrix Σ, evaluated at x. Resample θ^* from $\theta_l, l = 1, 2, \ldots, L$ with

probability proportional to w_l. The following steps describe the implementation:

(a) Normalize the importance ratios so that they add up to 1. That is, define $w_l^* = w_l / \sum_l^L w_l$.

(b) Form cumulative sums, $c_i = \sum_l^i w_l^*, i = 1, 2, \ldots, L-1$, and define $c_o = 0$ and $c_L = 1$.

(c) Generate a uniform random number, u, and define $\theta^* = \theta_i$ if $c_{i-1} < u \leq c_i$.

The above procedure is called the sampling-importance-resampling (SIR) algorithm for generating values from $\pi(\theta|y_{obs}, z)$. Instead of generating from a normal distribution, sometimes it is useful to generate from a longer tail distribution such as multivariate t with 5 to 10 degrees of freedom with location parameter $\widehat{\theta}$ and scale matrix $-\widehat{V}^{-1}$ (note that the denominator in the importance ratio is the density from which the values of θ have been drawn). Approaches discussed in (1) and (2) could be modified by drawing values in a transformed scale $\psi = h(\theta)$ (where the normal approximation may be better) and then retransform to the original scale $\theta = h^{-1}(\psi)$. For example, when θ is a proportion the transformations, $\text{logit}(\theta) = \log[\theta/(1-\theta)]$ or $\sin^{-1}\sqrt{\theta}$ are better suited for normal approximation.

3. Suppose $\widehat{\theta}$ is an estimate of θ (such as posterior mode) that can be easily obtained, but the second derivative of the log posterior is not easy to obtain and, again, suppose that $\pi(\theta|y_{obs}, z)$ is easy to calculate up to a constant of proportionality. The following is an approximation of the SIR algorithm given in (2):

   (a) Generate several values of complete data $y_i^{(l)}$ from $f(y_i|z_i, \widehat{\theta})$ and retain $D_l = \{y_{i,obs}^{(l)}, z_i, i = 1, 2, \ldots, n; l = 1, 2, \ldots, M\}$ and discard all the generated values of missing observations.

   (b) Calculate the estimate θ_l, $l = 1, 2, \ldots, M$ using D_l.

   (c) Calculate $w_l = \pi(\theta_l|y_{obs}, z)$.

   (d) Resample a value, θ^*, from $\{\theta_l, l = 1, 2, \ldots, M\}$ with probability proportional to w_l.

   (e) Generate imputations from $f(y_{i,mis}|y_{i,obs}, z_i, \theta^*)$.

4. **Gibbs sampling**. Often it is easier to draw values from the complete data posterior density of θ, $\pi(\theta|y_{obs}, y_{mis}, z)$ where $y_{mis} = \{y_{i,mis}, i = 1, 2, \ldots, n\}$. The Gibbs sampling is an iterative strategy: At iteration t, draw a value of θ, $\theta^{(t)}$, from $\pi(\theta|y_{obs}, y_{mis}^{(t-1)}, z)$ and then draw $y_{i,mis}^{(t)}$ from $f(y_{i,mis}|y_{i,obs}, z_i, \theta^{(t)})$. This iteration is continued until the effect of starting value is eliminated.

The field of Bayesian computations continues to evolve, as new approaches for drawing values from the posterior distribution are being developed and implemented. Except for some simple models (i.e., multivariate normal, log-linear model, etc.) software is not available in the missing data context. Implementation of these methods is still a challenge and requires considerable programming efforts.

3.4 Example

A case-control study was conducted to assess the relationship between dietary intake of omega-3 fatty acids and primary cardiac arrest (PCA) (defined as a sudden pulseless condition in the absence of any prior history of heart disease). In particular, the two fatty acids docosahexaenoic acid (DHA) and eicosapentaenoic (EPA) are of interest as they are not synthesized by the body and mostly derived through dietary intake of fish. The two omega-3 fatty acids were measured in the red-cell membrane and summed (REDOMEGA3). This measure is relative to all other measurable fatty acids in the membrane. PCA is a binary outcome variable, 1 for cases (disease) and 0 for controls (no disease). The primary model of interest is the logistic regression model,

$$\text{logit}Pr(\text{SCA} = 1|\text{REDOMEGA3}) = \beta_o + \beta_1 \text{REDOMEGA3}.$$

Unfortunately, REDOMEGA3 is missing for several cases and controls. As is typically the case in many practical applications, several covariates are available that are predictive of the variable with missing values. The study, for a different purpose, collected dietary intake of various types of fish using a food frequency questionnaire. The food frequency questionnaire also elicited information about serving size, which allowed computing an "estimate" of

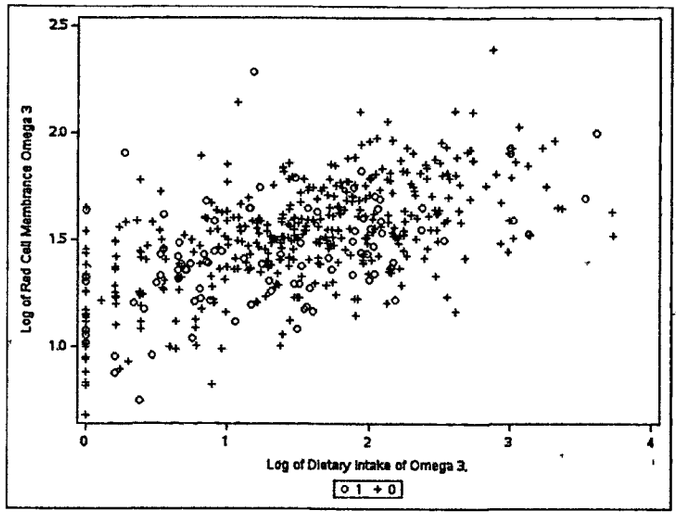


Figure 3.1: Scatter plot of red-cell and dietary values of omega-3 fatty acids on log-scale

omega-3 fatty acid intake. A glitch was that many cases did not survive and the dietary information had to be obtained from the spouse.

The variable DIETOMEGA3 is an estimate from all spouses of cases and controls reporting about their spouses dietary intake. This measure is in absolute intake of DHA+EPA and is not directly comparable to REDOMEGA3. Nevertheless, DIETOMEGA3 may provide information about the missing REDOMEGA3. A few cases and controls with missing values in DIETOMEGA3 are ignored in this analysis.

Figure 3.1 provides a scatter plot of log(REDOMEGA3) and log(DIETOMEGA3 + 1) (1 was added to avoid taking the logarithm of 0's as there were subjects who reported not taking fish at all or insignificant amount to estimate the intake). Given that this is a case-control study, separate indicators for cases ("+") and controls ("o") are used in the scatter plot. The scatter plot shows a linear relationship for both cases and controls and the correlation coefficient between them is about 0.6.

Table 3.1 gives the mean and standard deviation for four subgroups, case/control and whether or not missing red-cell membrane values. The mean and standard deviation of dietary omega 3 for those with and without missing red cell values are similar for both cases and controls. In a complete-case analysis

Table 3.1: Mean (SD) of red-cell membrane and dietary intake of omega-3 fatty acids

Status	Missing Red Cell Value?	Sample Size	Red Cell Omega-3	Dietary Omega-3
Case	Yes	252	—	4.23 (5.87)
	No	95	4.31 (1.16)	4.51 (6.16)
Control	Yes	148	—	4.98 (5.46)
	No	403	4.73 (1.16)	5.41 (5.60)

only 95 cases and 403 controls will be used and 252 cases (73%) and 148 (27%) controls will be discarded. The overall missing data percentage is 44.5%.

The complete-case logistic regression analysis with PCA as the dependent variable and REDOMEGA3 as the predictor results in the estimated regression coefficient of -0.3444 with the standard error of 0.1098. The question is how much information can be recovered through imputation using a reasonable proxy measure, the dietary intake of omega-3 fatty acids.

Imputation of missing values is carried out separately for cases and controls. This will maintain the differences in the observed data distributions between the two groups in the imputation process. Not using separate models will be equivalent to imputing under the null (no difference) which is not appropriate. For controls, a regression analysis of REDOMEGA3 on DIETOMEGA3 results in the estimated regression coefficient as $\widehat{\beta}_o = (1.24568, 0.18039)^t$, residual variance as 0.03999 on 401 degrees of freedom and, for the sake of completeness, the Cholesky decomposition of the inverse of the cross-product matrix as

$$T_o = \begin{bmatrix} 0.10668471 & 0 \\ -0.05419328 & 0.02861489 \end{bmatrix}$$

The following are the steps for generating the imputation of log(*REDOMEGA3*) for the controls:

1. Generate a chi-square random variable, u, with 401 degrees of freedom. Since the sample size is large, one can approximate $u = 401 + z_1\sqrt{2 \times 401}$ where z_1 is a random normal deviate. To be more precise, one can generate 401 random normal deviates and construct u as the sum of squares of these deviates. Define $\sigma_o^{*2} = (401 \times 0.03999)/u$.

2. Generate two random normal deviates, z_2 and z_3, and define

$$\beta_o^* = \begin{bmatrix} \beta_{oo}^* \\ \beta_{o1}^* \end{bmatrix} = \begin{bmatrix} 1.24568 \\ 0.18039 \end{bmatrix} + \sigma_o^* \begin{bmatrix} 0.10668471 z_2 \\ -0.05419328 z_2 + 0.02861489 z_3 \end{bmatrix}.$$

3. For the control subject j with missing value, generate a random normal deviate, z_j to construct

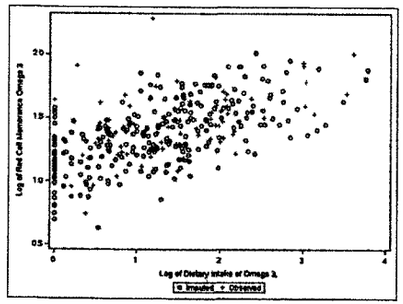
$$\log(REDOMEGA3_j*) = \beta_{oo}^* + \beta_{o1}^* \log(DIETOMEGA3_j+1) + \sigma_o^* z_j.$$

Similarly for the cases, the estimated regression coefficient is $\widehat{\beta}_1 = (1.22511, 0.15199)^t$, the residual variance 0.04856 with 93 degrees of freedom and the Cholesky decomposition of the inverse of the cross-product matrix,

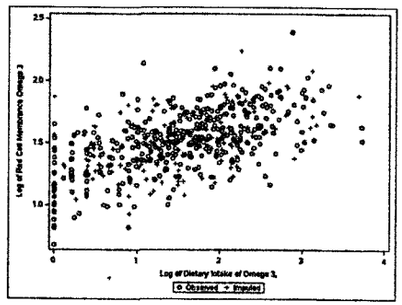
$$\begin{bmatrix} 0.1967918 & 0 \\ -0.1074688 & 0.06565849 \end{bmatrix}.$$

The following two scatter plots in Figure 3.2 provide the observed and imputed values $log(REDOMEGA3)$ plotted against the $\log(DIETOMEGA3 + 1)$ both for cases and controls. Imputed values are indicated by "o" and observed values with "+." For both cases and controls, the imputed and observed values exhibit similar properties, and thus generate a plausible completed-data set from the original population of cases and controls.

Table 3.2 gives the sample size, mean and standard deviation of the observed and imputed values. After imputation, the imputed variables are retransformed to the original scale and the same logistic regression model was



(a) Scatter plot of observed and imputed log(REDOMEGA3) versus observed log(DIETOMEGA3+1) for the cases



(b) Scatter plot of observed and imputed log(REDOMEGA3) versus observed log(DIETOMEGA3+1) for the controls

Figure 3.2: Comparison of observed and imputed values

Table 3.2: Summary statistics of the observed and imputed logarithm of the red-cell membrane omega-3 fatty acids by case-control status

Sample	Status	Sample size	Mean	SD
Case	Observed	95	1.428	0.252
	Imputed	252	1.372	0.259
	Combined	347	1.387	0.258
Control	Observed	403	1.523	0.248
	Imputed	148	1.500	0.248
	Combined	551	1.517	0.248

fit, resulting in the estimated regression coefficient as -0.4579, which is much stronger than the complete-case estimate, -0.3444. The imputed data standard error is meaningless as it assumes that the imputed values are real and, hence, not provided. Also, with the amount of missing data, the difference between the complete-case and imputed estimates may be subject to considerable variation.

This example uses information from an auxiliary variable to impute missing the values in the key variable of interest. In this case, the auxiliary variable is potentially related to the outcome (PCA) as well. In practice there could be many such variables for each variable with missing values. The advantage of the imputation approach is the ability to leverage all such variables in order to recover information lost due to missing values. It is possible that these auxiliary variables may have missing values as well. Methods to handle missing values in many variables and of many types are needed, and the next section describes a generalization of the basic regression method discussed in this section.

3.5 Sequential Regression Imputation

In practice, a data set may have several variables with missing values and differing types such as continuous, categorical, count, etc. Usually, there are structural dependencies between variables in a data set. For example, suppose that the question is asked of a current smoker "How long have you been smoking?" and the person refused to answer, it is known that the years smoked has to be less than age, at the least. Now suppose that the question was asked, "Did you smoke as a teenager?" and the response is no. This gives

a narrower range for years smoked. The data sets to be analyzed in practice may contain many such variables.

Sometimes the question is asked only of a subset and therefore any "not applicable" variables are not to be treated as missing data. For example, if the person was asked "Have you ever smoked?" and the answer is no, then all subsequent questions related to smoking are not relevant for this subject.

The variables may also have to be imputed in a certain order. For example, the subject was asked "Q1: Do you plan to vote in the coming election?" with the response option (1) yes, (2) no and (3) don't know. The follow-up question is "Q2: Do you support the ballot initiative A?" with the response option (1) yes, (2) no, (3) undecided. If the goal is to predict the success of the ballot initiative A, then first Q1 has to be imputed and then Q2 for the subset with observed or imputed Q1 to be "yes".

Developing explicit models (joint distributions) or performing hot deck imputations are difficult, if not impossible. A pragmatic approach is to consider a variable by variable imputation but using all the relevant information as predictors. A sequential regression or chained equations approach is one such pragmatic approach. Various other names have been given to this approach such as fully conditional specification or flexible conditional models, etc.

Suppose that U is a collection of variables with no missing values and $Y_1, Y_2, \ldots, Y_p$ are p variables with missing values. Though it is not necessary, suppose that the variables are ordered by number of missing values from lowest for Y_1 and largest for Y_p. An alternative approach is to order on the basis of dependence on other variables from "least dependent" to "most dependent." As discussed later, the ordering will have no effects as the imputed values on any variable will eventually depend on all other variables. Furthermore, the pattern of missing data is assumed to be arbitrary.

The sequential regression approach is an iterative procedure. In the first iteration, Y_1 is regressed on U and the missing values are imputed. An explicit regression model, a hot deck or predictive mean matching may be used to create imputed values. Let $Y_1^{(1)}$ denote the filled-in version of Y_1. Now Y_2 is imputed using $(U, Y_1^{(1)})$ as covariates. Let $Y_2^{(1)}$ denote the filled-in version of Y_2. This process continues until Y_p is imputed using $(U, Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{p-1}^{(1)})$.

We cannot stop at iteration 1 because imputation of Y_1, for example, fails to exploit the observed information from $Y_2, Y_3, \ldots, Y_p$. Iterations $t = 2, 3, \ldots$, proceed in the same manner except that all other variables (with

some filled at the current and the rest in the previous iterations) are used in imputing each variable. Specifically, at iteration 2, Y_1 is reimputed using $U, Y_2^{(1)}, Y_3^{(1)}, \ldots, Y_p^{(1)}$ as predictors; Y_2 is reimputed using $U, Y_1^{(2)}, Y_3^{(1)}, \ldots, Y_p^{(1)}$ as predictors, etc. In general, at iteration t, Y_j is reimputed using

$$U, Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{j-1}^{(t)}, Y_{j+1}^{(t-1)}, \ldots, Y_p^{(t-1)},$$

as predictors. The iteration is continued for a few times in order to fully use the predictive power of the rest of the variables when imputing each variable. Empirical analysis show that 5 to 10 iterations are sufficient to condition the imputed values on any variable on all other variables.

3.5.1 Details

This section provides a detailed description using a parametric framework. Let Y_o be the observed vector for the variable to be imputed and X be the full sample predictor matrix based on all other variables which is also partitioned into X_o and X_1 where X_o corresponds to subjects in Y_o and X_1 corresponds to subjects with missing values in Y. For the clarity of presentation, suppress the notation for iteration.

1. For a continuous variable, fit a linear regression model of Y_o on X_o and obtain estimated regression coefficient $\widehat{\beta}$, the residual variance $\widehat{\sigma}^2$ and the inverse of the cross-product matrix $(X_o^t X_o)^{-1}$. Let T be the Cholesky decomposition of $(X_o^t X_o)^{-1}$ such that $TT^t = (X_o^t X_o)^{-1}$. Let m be the sample size in this analysis and p be the number of regression coefficients. The imputations are carried out as follows:
   (a) Generate a chi-square random variable, u, with $m - p$ degrees of freedom and define $\sigma_*^2 = (m-p)\widehat{\sigma}^2/u$.
   (b) Generate a vector, z, of p standard normal deviates and define, $\beta_* = \widehat{\beta} + \sigma_* T z$.
   (c) Generate a vector, v, of $n - m$ standard normal deviates and define the imputed values as $Y_{*1} = X_1\beta_* + \sigma_* v$ where n is the sample size, where X_1 is the predictor matrix corresponding to nonrespondents.

X_o (and X_1) is determined using the standard regression diagnostics, such as scatter plots, residual plots and other graphical techniques to develop a good fitting model. The dependent variable may be transformed (like in the example given in the previous section) to achieve normality of the residuals. The variables may be either transformed back at the end of all the iterations or before proceeding to the next variable. Instead of parametric regression model, hot deck imputation based on the propensity score and the predicted values of Y can be used. The goal is to make sure that the model provides a good basis for predicting the missing values.

2. If Y is binary then a logistic regression model may be used. The following are the steps to generate the imputed values:

   (a) Fit a logistic regression model with Y_o as the dependent variable and X_o as independent variables. Let $\widehat{\beta}$ be the maximum likelihood estimate of the regression coefficient and let $\widehat{V}$ be the estimated variance-covariance matrix.

   (b) As before let T be the Cholesky decomposition of $\widehat{V}$ such that $TT^t = \widehat{V}$. Generate a vector, z, of p standard normal deviates and define $\beta_* = \widehat{\beta} + Tz$. If the normal approximation is not reasonable (which can be checked by plotting the likelihood function against the parameters) then SIR algorithm could be used to generate β_*.

   (c) Using β_* compute the predicted probability p_{*1} for the nonrespondents. That is, define the linear predictor $L_{*1} = X_1\beta_*$ and then define $p_{*1} = 1/(1 + \exp(-L_{*1})$.

   (d) Generate a vector, u, of $n - m$ uniform random numbers between 0 and 1. Set Y_{*1} to 1 or 0 depending upon whether the generated value of u is less than or greater than equal to the corresponding predicted probability in the vector p_{*1}.

   Hoshmer-Lemeshaw goodness of fit tests can be used to develop a good prediction model. This is similar to the techniques used while developing a response propensity model. Instead of using parametric model, hot deck with imputation cells based on the response propensity and predicted probability can be used.

3. For a count variable, a Poisson regression model may be used to develop imputation. The Poisson regression model specifies $Y_o \sim Poisson(n_o\lambda_o)$ where $\log(n_o) + \log\lambda_o = X_o\beta$ or $\log\lambda_o = -\log n_o + X_o\beta$. That is $-\log n_o$ is the offset such as person-years of follow-up or any other suitable denominator.

4. For a nominal variable, multinomial logit model may be used. Suppose that Y can take k levels. The model is

$$\log Pr(Y = j|X) = X\beta_j$$

   with $\sum_j Pr(Y = j|X) = 1$. As in the logistic model, obtain the maximum likelihood estimate of the regression coefficients, its covariance matrix and perturbation β_{j*}. Next, compute the predicted values for each nonresponding subject, $p_{j*}, j = 1, 2, \ldots, k$. Generate a uniform random number, u and impute level j if $\sum_l^{j-1} p_{l*} < u \leq \sum_l^j p_{l*}$.

5. Mixed or semi-continuous variables occur frequently in many practical applications. For example, real estate income. Most people may have 0 as the income value (no real estate) and a continuous value for the rest. This type of variable can be handled using a two part model. First, impute 0 or non-zero using the logistic regression model and then, conditional on being non-zero use the linear regression model (that is, subset the data for subjects with those observed or imputed as having real estate income) to impute continuous values.

6. For an ordinal variable with k levels, the following proportional odds model may be used:

$$\pi_j = Pr(Y \leq j) = \frac{\exp(\alpha_j + X\beta)}{1 + \exp(\alpha_j + X\beta)}$$

   with $Pr(Y = j) = \pi_j - \pi_{j-1}$ used to impute the level.

3.5.2 Handling Restrictions

Consider now some restrictions on the imputed values. Suppose that for a continuous variable, the imputed value for the subject i should be between a_i and b_i. The imputed draws are then made from a truncated normal distribution. Similarly, for a categorical variable with k levels, $1, 2, \ldots, k$, if for the

subject i, the imputed values can be either 1 or 2 then adopt the following procedure. Let p_{ij*} be the predicted probability for subject i to belong to level j. Compute $\pi_{i1*} = p_{i1*}/(p_{i1*} + p_{i2*})$ and $\pi_{i2*} = 1 - \pi_{i2*}$. Generate a uniform random number and define the imputed value to be 1, if the uniform number is less than or equal to π_{i1*} and 2 otherwise.

This strategy will also be useful in handling missing data in survival or time-to-event analysis. Suppose that X is a set of covariates with some missing values, Y is time-to-event and C is the censoring indicator. Define a variable T which is equal to Y for actual time-to-event, and set T to missing if the observation is censored. The imputation of T has to be greater than or equal to the corresponding Y for the censored observations. Some transformation of survival times will be needed to achieve normality of the residuals.

Sometimes a particular variable is applicable only for a subset of cases. Suppose that a question asked is "Q1: Have you ever smoked cigarettes?" with yes/no response options. A follow-up question is "Q2: Do you smoke cigarettes now?" with yes/no response options. Q2 is applicable to subjects who responded yes to Q1. Imputation of Q2 has to be restricted to those who responded yes to Q1. Thus, the missing values in Q1 has to be imputed first, and then impute missing values in Q2. Obviously if Q1 is missing then Q2 is also missing.

There are two possible ways to handle this situation. Create a new variable by combining Q1 and Q2, into a three category variable with coding 1: never smoker (Q1=no), 2: former smoker (Q1=yes, Q2=no) and 3: current smoker (Q1=yes , Q2=yes). Set all subjects with missing Q1 or Q2 to missing for the newly created variable. The restriction is imposed as discussed above by imputing levels 2 or 3 if Q1=yes.

The second option is to impute the missing value in Q1 first by using a logistic regression model, then subset the data with subjects observed or imputed Q1=yes, and then use another logistic regression model to impute the missing values in Q2. This strategy maintains the question structure and gives more flexibility to the analyst.

Now that Q1 and Q2 have been imputed, the question arises, "How to use Q1 and Q2 as predictors in the imputation of the next variable, say, blood pressure"? One strategy is to recode Q1 and Q2 to create a new three category variable with never, former and current smoker as categories and use two dummy variables as predictors. This requires recoding of variables before

each regression which may be inconvenient. The second option is to treat Q2 with three categories yes, no and not applicable. Use one dummy variable for Q1 and two dummy variables for Q2 (treating not applicable as the reference category). Though we are using more dummy variables than necessary in the prediction equation, the advantage of this approach is that it avoids recoding and facilitates automating the imputation procedure. Also, the goal in the imputation is to get a good prediction equation that uses all the variables. The model need not be parsimonious and not purely viewed through a substantive research perspective (for example, the individual regression coefficients may not be interpretable, but a combination of them may be interpretable).

One of the variables of interest may be a scale that is constructed by summing, say, k items (which may be binary 0/1 or 3 to 5 point Likert type item). Consider two situations. In the first situation, subjects responded either to all the items or none of the items and in the second, responses from subjects are mixed. Some provided all, some provided none and the rest provided some and not the others. In the first situation, the scale may be directly imputed (perhaps bounded by the minimum and maximum value possible).

In the second situation, it is better to impute the individual items and then construct the scale by summing after the imputation. This strategy uses partial information and reflects both intra-subject and inter-subject variation in the imputation process. This strategy may not be feasible when the number items and the scales are large. An intermediate solution is to impute the scale, but use the sum of the observed responses as the lower bound and the maximum possible value as the upper bound.

3.5.3 Model Fitting Issues

The advantage of the sequential regression approach is that it reduces the problem of imputation modeling to finding a sequence of good fitting regression models using all available information. All exploratory data analysis techniques can be used to find these regression models. Developing a good fitting model is an iterative process. First, develop a working model based on exploratory data analysis and substantive understanding, check the residuals and perform other model diagnostics, refine the model, if necessary and again perform all the checks, refine the model, etc. Some of these steps can be automated (like generating plots and outputs from model diagnostics).

Sometimes the number of variables may be large making such model building task difficult and time consuming. One potential way to save on model building tasks is to reduce the covariate space for each regression model using a principal component analysis (PCA) of the covariates. Suppose that Y_j is the variable being imputed and $X = (U, Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ are the predictors. Create principle components $P_1, P_2, \ldots, P_k$ where k is the number of columns in X. Since the principle components are orthogonal to each other, the problem reduces to finding the best fitting k univariate regression models. This may be automated to some extent. One can drop a few principal components corresponding to small percentage of variance explained. This strategy works for all regression models and greatly simplifies the imputation task.

When the number of covariates is large, the covariance matrix of the parameter estimates may be less stable, and the perturbations β_* may be far from the center of the distribution. This can be a significant problem when the predictors in the models are highly correlated. Here also PCA can be helpful. Another option is to use a ridge regression which amounts to shrinking each regression coefficient towards 0. Many software packages are available to fit these models and provide the needed output to carry out the imputation task: Estimated regression coefficients and their covariance matrix for the specified regression model.

Transformation of the variables, or rescaling, is a powerful tool in model building strategy. The Box-Cox transformation of the dependent variable is a powerful method to achieve normality of the residuals. Suppose that Y_j is a continuous variable to be imputed, and X is the covariate matrix. A Box-Cox power transformation fits the regression model,

$$\frac{Y_j^{\lambda} - 1}{\lambda} = X\beta + \epsilon$$

where the transformation is $\log(Y_j)$ when $\lambda = 0$. When there are some negative values in Y_j, add a constant (can be empirically determined while choosing λ) to make all of them positive to avoid problems with taking logarithms or odd number power transformation. The parameter λ can be estimated from data or determined by trial and error. There are software packages that will provide an estimate of λ using maximum likelihood. Typically, the variables, once transformed, are used in that scale throughout the imputation process and then retransformed to the original scale after imputations of all the variables and all the iterations are completed.

One has to be careful when transforming and retransforming the variables. For example, when using the logarithmic transformation, a large nonsensical value may be obtained by exponentiating what seemingly appears to be a reasonable value on the logarithmic scale. This is especially problematic when log transformation is used for variables such as income, assets or wealth. A cube-root transformation might be better for such variables. However, a hot deck approach using the response propensity and predicted value might be a more reasonable approach for variables that are difficult to model.

When a regression diagnostics shows heteroscedasticity, it can be incorporated in the modeling process as follows. As before, let Y be the continuous variable being imputed and suppose that the residual plot against a particular predictor x_1 shows evidence of increasing or decreasing variance. The model may be modified as $Y \sim N(X\beta, \sigma^2 x_1^{\alpha})$. Suppose $\widehat{\beta}$ is the ordinary least square estimate, and e is the residual $Y - \widehat{Y}$. A simple approach to estimate α is to regress $\log(e^2)$ on $\log(x_1)$, and the slope provides an estimate of α, $\widehat{\alpha}$. Now impute the missing values using the refined regression model, $Y^* = Y/x_1^{\widehat{\alpha}/2} \sim N(X\beta/x_1^{\widehat{\alpha}/2}, \sigma^2)$. After the imputation has been carried out for Y^*, multiply by $x_1^{\widehat{\alpha}/2}$ to get the imputation on the original scale, Y.

3.6 Bibliographic Note

Imputation, just like weighting, may have originated in 1930s or 1940s in many government agencies. The hot deck method was perhaps developed in the U. S. Census Bureau or some other official statistical agency. Ford (1983) provides an overview of the hot deck procedure. A more recent review is Andridge and Little (2010). Single imputation is still prevalent in many statistical agencies despite the problems associated with it. Mean imputation is used less. Regression imputation (imputing the predicted values from a regression model) may be more common and can be appropriate in some cases. Many of these imputation techniques were developed when computational power was limited. For more technical details about hot deck method, regression imputation method and some other variance estimation techniques see Kim and Shao (2014).

The sequential regression approach was first proposed by Kennickell (1991) for continuous variables in the survey of consumer finances. Brand (1999) in a doctoral dissertation develops this methodology further (names it variable-by-variable imputation). Van Buuren and Oudshoorn (1999) in a technical report introduced a version of the method and called it multivariate imputation by chained equations and introduced a software MICE. Raghunathan et al (2001) developed the methodology for several types of variables, incorporated bounds and restrictions and termed it, sequential regression multivariate imputation (SRMI). A SAS based and stand-alone software for implementing this procedure is available via web site (www.iveware.org). Another excellent source for software for performing imputation is (www.multiple-imputation.com). The sequential regression approach has been implemented in STATA, another popular statistical software. See Royston (2004). For using R package implementation see van Buuren (2012).

3.7 Exercises

1. **Project.** For this project use the NHANES data downloaded for the exercise in Chapter 2. Perform a hot deck imputation of the missing values using the methods discussed in this chapter. Compare the hot deck and weighted estimators.

2. Compare the sampling variance of the single imputation hot deck estimator (treating the imputed values as real) and compare it with the sampling variance of the weighted estimator. Discuss the extent of under estimation of the sampling variance due to ignoring the imputation uncertainty.

3. Perform a regression analysis of Hemoglobin A1c on the covariates and develop an imputation model. Using the available software or by writing your own code, generate imputation of the missing values. Compute the sample mean from the imputed data and its sampling variance (ignoring the imputation uncertainty). Compare the hot deck and model-based imputation estimates and their standard errors to the weighted estimate and its standard error.

4

Multiple Imputation

4.1 Introduction

It has been repeatedly emphasized that the imputed values should not be treated as the real or actual values for the nonrespondents. However, if a singly imputed data set is analyzed using any standard software, then they are treated as real values, therefore, the stated standard errors will understate the uncertainty. The fact that imputations are guesses (or plausible values) created from the observed data should be reflected in the standard error calculations. Multiple imputation is a mechanism for adding this additional uncertainty.

Under this approach, the imputation process is repeated several times. Each set of imputed values, when combined with the observed set of values, results in a completed data set. Each completed data set is analyzed to obtain estimates of the parameters, their completed-data standard errors and test statistics. The variation across the completed data quantities is used to capture the additional uncertainty due to imputation. The completed data estimates, standard errors, test statistics, etc. are combined to form a single inference.

4.2 Basic Combining Rule

Suppose that M imputed data sets have been created. Let $e_1, e_2, \ldots, e_M$ be the estimates of a parameter, θ, and $U_1, U_2, \ldots, U_M$ be the corresponding estimated sampling variances (square of the standard errors). The multiple

imputation estimate is defined as the average of the M estimates,

$$\bar{e}_{MI} = \sum_{l}^{M} e_l/M. \tag{4.1}$$

The variance across the M estimates, $B_M = \sum_l^M (e_l - \bar{e}_{MI})^2/(M-1)$, actually measures the uncertainty due to imputations. The multiple imputation sampling variance is defined as

$$T_M = \bar{U}_M + (1 + 1/M)B_M \tag{4.2}$$

where $\bar{U}_M = \sum_i^M U_i/M$ is the average of completed-data sampling variances. One can interpret $\bar{U}_M$ as an estimate of the sampling variance that would have been obtained if there had been no missing data and $(1 + 1/M)B_M$ is the additional variance due to missing data. Thus, the ratio

$$r_M = (1 + 1/M)B_M/T_M \tag{4.3}$$

can be viewed as the proportionate increase in the variance due to missing data. This quantity is called the fraction of missing information (FMI). The ratio $a_M = T_M/\bar{U}_M$ is the efficiency of the multiple imputation estimate relative to complete data estimate. Both these quantities are useful to determine the number of imputations needed, the impact of missing data on inferences and to judge the recovery of information from subjects with partial information through imputation. For example, if FMI is smaller than the proportion of subjects that would be discarded in the complete-case analysis then information has been recovered by including the partially observed subjects. This recovery of information depends on the parameter being estimated and the information in the partially observed subjects.

To construct confidence intervals for θ, a t-distribution with degrees of freedom

$$\nu_M = (M-1)/r_M^2 \tag{4.4}$$

is used. Let $t_{\alpha/2,\nu_M}$ be the appropriate percentile of the t distribution with ν_M degrees of freedom, then $100(1-\alpha)$ % confidence interval for the parameter, θ, is computed as $\bar{e}_{MI} \pm t_{\alpha/2,\nu_M}\sqrt{T_M}$. To test the null hypothesis, $H_o : \theta = \theta_o$, the test statistic is $(\bar{e}_{MI} - \theta_o)/\sqrt{T_M}$ and is referred to a t distribution with ν_M degrees of freedom.

The degrees of freedom, ν_M, is derived assuming that the complete data inference (in the absence any missing values) are based on large samples (infinite degrees of freedom for the complete data analysis). Sometimes, the complete data analysis (that is, without any missing values) has a finite degrees of freedom. For example, the complete data inferences in a regression analysis based on say, $n = 30$ observations with, say, $p = 5$ predictors has $n - p = 25$ degrees of freedom. In such situations, a correction is needed. Suppose ν_{com} is the complete data degrees of freedom. Define,

$$c_M = \frac{\nu_{com} + 1}{\nu_{com} + 3}\nu_{com}(1 - r_M).$$

The revised degrees of freedom is

$$\nu_M^* = \frac{\nu_M c_M}{\nu_M + c_M}. \tag{4.5}$$

Given the t sampling distribution, the sampling variance is actually a multiple of T_M. The variance of a t distribution, with scale σ and degrees of freedom ν, is $\nu\sigma^2/(\nu-2)$. Thus, the standard error of the multiple imputation estimate may also be defined as $\sqrt{\nu_M^* T_M/(\nu_M^* - 2)}$. When ν_M^* is large, this modification may not make a difference.

4.3 Multivariate Hypothesis Testing

The combining rules in the previous section were for a scalar parameter, θ. In many analyses inferential quantity of interest may be a K-dimenstion vector of parameters, θ. Let $e_1, e_2, \ldots, e_M$ be the estimated vectors from the completed data sets with $U_1, U_2, \ldots, U_M$ as the corresponding variance-covariance matrices. Let $\bar{e}_{MI} = \sum_l e_l/M$ be the mean vector of the completed data estimates and $\bar{U}_M = \sum_l U_l/M$ be the average covariance matrix. Let $B_M = \sum_l (e_l - \bar{e}_{MI})(e_l - \bar{e}_{MI})^t/(M-1)$ be the between imputation variance-covariance matrix. The multiple imputation variance estimate is $T_M = \bar{U}_M + (1 + M^{-1})B_M$. Wald's chi-square statistics for testing the null hypothesis $H_o : \theta = \theta_o$ is $D_M = (\bar{e}_{MI} - \theta_o)^t T_M^{-1}(\bar{e}_{MI} - \theta_o)$.

When M is small relative to k, the matrix B_M may not be of full rank and the inverse of the matrix, T_M, may be unstable. If the effect of missing data on all the parameters are roughly the same, an approximation is $\tilde{T}_M =$

$(1+g_M)\bar{U}_M$ and the revised test statistic is

$$\tilde{D}_M = (\bar{e}_{MI} - \theta_o)^t \bar{U}_M^{-1} (\bar{e}_{MI} - \theta_o)/(1+g_M) \tag{4.6}$$

where $g_M = (1+M^{-1})Tr(B_M \bar{U}_M^{-1})/k$ and tr is the trace of the matrix or the sum of the diagonal elements of the matrix.

An approximate sampling distribution of $\tilde{D}_M/k$ is an F distribution with k and w_M degrees of freedom where

$$w_M = 4 + (t-4)[1 + (1-2t^{-1})g_M^{-1}]^2 \tag{4.7}$$

provided $t = k(M-1) > 4$. If $t \leq 4$ then define $w_M = t(1+k)(1+g_M^{-1})^2/2$.

The derivation of w_m assumes that the complete data analysis is based on large samples (infinite degrees of freedom). When the degrees of freedom for the complete data analysis is ν_{com} then the modification of w_M (similar to equation (4.5)) is

$$w_M^* = \frac{w_M c_M}{w_M + c_M} \tag{4.8}$$

where

$$c_M = \frac{(\nu_{com}+1)\nu_{com}}{(\nu_{com}+3)(1+g_M)}.$$

4.4 Combining Test Statistics

The combining rules, discussed so far, requires estimates and standard errors for the single parameter inference and the estimates and their covariance matrices for the multiparameter inference. There are many situations where the completed data inferences are based on test statistics or the p-values. For example, suppose that a goodness of fit or association test has been performed on each completed data set yielding M chi-square statistics $d_1, d_2, \ldots, d_M$. A combining rule for these test statistics is as follows. Let $\bar{d}_{MI} = \sum_l^M d_l/M$ be the average of the test statistics. Let $a = \sum_l \sqrt{d_l}/M$ and $b = \sum_l (\sqrt{d_l} - a)^2/(M-1)$. The test statistic is defined as

$$D_d = \frac{\bar{d}_{MI}/k - (M+1)r_d/(M-1)}{1+r_d} \tag{4.9}$$

where $r_d = (1+M^{-1})b$ and is referred to an F distribution with $\nu_d = k^{-3/M}(M-1)(1+r_d^{-2})$ degrees of freedom.

Table 4.1: Maternal smoking and child wheeze status from the six city study

Maternal Smoking	Child's Wheeze Status			
	No Wheeze	Wheeze with Cold	Wheeze without Cold	Missing
None	287	39	38	279
Moderate	18	6	4	27
Heavy	91	22	23	201
Missing	59	18	26	

The procedure given above can be used for testing hypothesis involving a vector of parameters. For example, assessing the significance of regression coefficients for a block of predictors in a regression model. The above formula can also be applied to the likelihood ratio statistics, $d_1, d_2, \ldots, d_M$ (logistic regression model, for example). Chapter 5 provides alternatives that can be applied for the analysis of variance (ANOVA).

The data summarized in Table 4.1 was collected as a part of a six city study and is reported in Lipsitz, Parzen and Molenberghs (1998). Two categorical variables are maternal smoking status (X_1) classified as (none, moderate and heavy) and the child's wheezing status (X_2) (none, wheeze with cold, wheeze without cold). Total sample size is $n = 1138$ with 528 subjects having both variables measured, 507 missing wheeze status (smoking status available) and 103 missing smoking status (wheeze status available).

One of the goals is to assess the strength of association between these two variables. The missing values were imputed using the sequential regression approach: X_1 was imputed using a multinomial logit model with X_2 as a predictor, and X_2 was imputed using a multinomial logit model with X_1 as a predictor. A total of $M = 20$ imputations were created. The sequential regression was run for 50 iterations.

From each completed table, the Pearson chi-square statistic with $k = 4$ degrees of freedom was computed. The average of these 20 chi-square statistics, $\bar{d}_{MI}$, is 34.784. The variance of the square root of the test statistics, b, is 1.681. Thus $r_d = (1+1/20) \times 1.681 = 1.765$. The value of the test statistics D_d is

$$D_d = \frac{34.78/4 - (20+1)/(20-1) \times 1.765}{1+1.765} = 2.439.$$

This statistic is referred to as an F distribution with the numerator degrees of freedom k and the denominator degrees of freedom, $\nu_d = 4^{-3/20} \times (20 -$

Table 4.2: Cell specific percentages (SE) [FMI%] for the data from the six city study

Maternal Smoking	Child's Wheeze Status		
	No Wheeze	Wheeze with Cold	Wheeze apart from Cold
None	47.42 (1.7)[24.2]	7.19 (1.08)[49.8]	7.33 (1.02)[42.4]
Moderate	3.19 (0.63)[30.6]	1.16 (0.43)[46.0]	0.98 (0.48)[62.9]
Heavy	20.50 (2.03)[65.2]	5.63 (1.09)[61.0]	6.60 (1.26)[66.1]

$1) \times (1 + 1/1.765^2) = 20.387$. For the significance level 0.05, the cut-off or the critical value is 2.856. For the significance level 0.10, the cut-off value is 2.243. Based on this analysis, there is evidence for the association at $\alpha = 0.10$ level but not at $\alpha = 0.05$ level.

The cell probability estimates were also obtained by averaging the 20 completed-data proportions with the standard error computed using the formula for T_M. Table 4.2 summarizes the estimates, their standard errors and the fraction of missing information. The estimates are practically the same as the maximum likelihood estimate given in the original article. A reference for FMI is the fraction of incomplete cases, 54%. For some cells, the complete-case and the MI estimates are quite different and including the partial information from the incomplete cases reduces the bias and also decreases the variance for some cells but not for others.

4.5 Basic Theory of Multiple Imputation

As mentioned earlier, imputations are generally drawn from a predictive distribution of the missing values, and there are many ways to construct it. The combining rules and theoretical framework makes a broad set of assumptions, and one needs to make sure that these assumptions are generally met.

1. The analyst is assumed to use the most appropriate and efficient estimation or inference procedure on complete data (that is, with missing data). Suppose that $\widehat{\theta}_C$ is the complete data estimate and U_C is its complete data sampling variance (square of the complete data standard error). It is assumed that $U_C^{-1/2}(\theta - \theta_C)$ has a normal distribution (approximately) with mean 0 and variance 1.

2. Denote the average of $\widehat{\theta}_C$ over the predictive distribution of the missing values results as $\widehat{\theta}_O$ and its sampling variance with respect to the predictive distribution is assumed to be $E(U_C) + B_O$, where B_O is the increase in variance due to missing values and $E(U_C)$ denotes the expectation over the predictive distribution of the missing values.

3. The imputation procedure is such that when M, the number of imputations tends to infinity, the average $\bar{e}_{MI}$ tends to $\bar{e}_\infty = \widehat{\theta}_O$ and B_M tends to $B_\infty = B_O$.

4. The imputation procedure yields $\bar{U}_M$ to be approximately equal to $E(U_C)$.

5. When the number of imputations, M, is small the following two conditions hold:

   (a) $MB_\infty^{-1}(\bar{e}_\infty - \bar{e}_{MI}) \sim N(0, 1)$.

   (b) $(M-1)B_M/B_\infty \sim \chi^2_{M-1}$.

Under the stated assumptions,

$$\theta | B_\infty, \bar{e}_{MI}, \bar{U}_M \sim N[\bar{e}_{MI}, \bar{U}_M + (1 + M^{-1})B_\infty].$$

The multiple imputation estimate, its variance, t reference distribution and various test statistics are based on the approximation that combines the above equation and 5(b) to create the posterior distribution of θ given the complete data.

4.6 Extended Combining Rules

The asymptotic normality of the distribution of the complete and completed data estimates is a critical assumption in deriving the combining rules. All combining rules are based on some approximation of $Pr(\theta|\mathcal{S})$ where $\mathcal{S} = \{\mathcal{S}_l, l = 1, 2, \ldots, M\})$, $\mathcal{S}_l$ is a set of statistics from the l^{th} completed data set and θ is the parameter of interest. The point estimate $\bar{e}_{MI}$ and the variance T_M approximate, the mean $E(\theta|\mathcal{S})$ and variance, $Var(\theta|\mathcal{S})$, respectively.

There may be situations where the asymptotic normality assumption is not reasonable. For example, let θ be the population proportion and $\widehat{\theta}_l, l = 1, 2, \ldots$ be the sample estimates. The completed data variance is $U_l = \widehat{\theta}_l(1 - \widehat{\theta}_l)/n_l$ where n_l is the sample size. Note that the sample size may vary across the completed data sets, especially for the proportions defined for a subpopulation defined by variables with missing values. The sample size has to be fairly large for the normal approximation, $(\theta - \widehat{\theta}_l)/\sqrt{U_l} \sim N(0, 1)$, to be reasonable.

Two possible options are considered in this section. The first strategy considers a transformation of the parameter to achieve approximate normality. Construct inferences (e.g., confidence intervals) on the transformed scale and then retransform to the original scale. The second strategy uses the mean $E(\theta|\mathcal{S})$ and $var(\theta|\mathcal{S})$ and develops nonnormal approximations for $Pr(\theta|\mathcal{S})$.

4.6.1 Transformation

Consider two possible transformations for inference about a proportion parameter, θ. Let $\phi = \log(\theta/(1-\theta))$ and $\widehat{\phi}_l = \log(\widehat{\theta}_l/(1-\widehat{\theta}_l))$. The completed-data variance is, $var(\widehat{\phi}_l) \approx U_l = [n_l\widehat{\theta}_l(1-\widehat{\theta}_l)]^{-1}$. The confidence interval and significance tests can be calculated on the transformed scale using the standard combining rule described in Section 4.2. If (ϕ^*_L, ϕ^*_U) is the confidence interval for ϕ then the confidence interval for θ is $[\theta^*_L = \exp(\phi^*_L)/(1+\exp(\phi^*_L)),$ $\theta^*_U = \exp(\phi^*_U)/(1+\exp(\phi^*_U))]$.

An alternative is to define $\phi = \sin^{-1}(\sqrt{\theta})$ for which $U_l = (4n_l)^{-1}$. If (ϕ^*_L, ϕ^*_U) is the confidence interval for ϕ then $[\sin^2(\phi^*_L), \sin^2(\phi^*_U)]$ is the corresponding confidence interval for θ. There are many other transformations of proportions to achieve normality. The logit and arc-sine-square root transformations are most popular and work well in practice.

Inference about the odds ratio in a 2×2 table is another example where it is better to draw inference on the logarithmic scale (since the normal approximation works better on this scale). Let θ be the odds ratio and define $\phi = \log\theta$. In this case $U_l = (1/a_l + 1/b_l + 1/c_l + 1/d_l)$ where a, b, c and d are the four cell sizes. Once the confidence interval for ϕ is constructed then $[\exp(\phi^*_L), \exp(\phi^*_U)]$ is the confidence interval for θ. Testing the null hypothesis $H_o : \theta = 1$ is equivalent to testing $H_o : \phi = 0$.

As a final example, consider a correlation coefficient. Suppose that r_l is the correlation coefficient between variables X and Y based on the completed data set $l = 1, 2, \ldots, M$. Let $z_l = \log[(1+r_l)/(1-r_l)]/2$ be the

Fisher's z transformation of the correlation coefficient. The completed-data variance is $1/(n_l - 3)$ where n_l is the completed data sample size. As before, compute $\bar{z}_{MI}$, T_M and ν_M using the the transformed values. If (z_{*L}, z_{*U}) is the multiple imputation confidence interval on the z scale, then $r_{*L} = [\exp(2z_{*L}) - 1]/[\exp(2z_{*L}) + 1]$ and $r_{*U} = [\exp(2z_{*U}) - 1]/[\exp(2z_{*U} + 1]$ forms the confidence interval for the correlation coefficient.

The same strategy may be applied for inferring about the partial correlation coefficient between X and Y given Z, which is defined as the correlation coefficient between the residuals $e_{X.Z}$ from the regression of X on Z and the residuals $e_{Y.Z}$ from the regression of Y on Z. There are only $n_l - 2$ independent residuals (in both $e_{X.Z}$ and $e_{Y.Z}$). Accordingly, the completed data variance is $U_l = 1/(n_l - 5)$. In general, if p variables are included in Z, then $U_l = 1/(n_l - p - 4)$.

4.6.2 Nonnormal Approximation

For the binomial proportion, assuming a noninformative prior, $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, the completed data posterior density is

$$\pi(\theta|x_o, n_o, x_l, n_l) \sim Beta(x_o + x_l + 1/2, n_o + n_l - x_o - x_l + 1/2)$$

where (x_o, n_o) and (x_l, n_l) are the observed and imputed values in the completed data posterior distribution of θ. Thus, defining $e_l = (x_o + x_l + 1/2)/(n_o + n_l + 1)$ and $U_l = e_l(1-e_l)/(n_o + n_l + 2)$, and defining T_M as usual, a beta distribution may be fitted by matching the mean and variance. That is, $Pr(\theta|\mathcal{S}) \approx Beta(a, b)$ where $a/(a+b) = \bar{e}_{MI}$ and $ab(a+b)^{-2}(a+b+1)^{-1} = T_M$. Solving the two equations obtains,

$$a = \bar{e}_{MI}\left(\frac{T_M}{\bar{e}_{MI}(1-\bar{e}_{MI})} - 1\right),$$

and

$$b = (1-\bar{e}_{MI})\left(\frac{T_M}{\bar{e}_{MI}(1-\bar{e}_{MI})} - 1\right).$$

The confidence interval for θ may be constructed using this approximate beta distribution. The R-package (BINOM, function BINOM.BAYES) provides a convenient tool for computing the highest posterior density interval.

Both transformation and nonnormal approximation may be useful to construct inference about the variance parameter (the variance components). Let

σ^2 be the variance parameter, and let s_l^2 be the completed-data estimate based on $n_l - 1$ degrees of freedom. From standard statistical theory,

$$\frac{(n_l - 1)s_l^2}{\sigma^2} \sim \chi^2_{n_l - 1}.$$

Defining $\tau^2 = \sigma^{-2}$, the posterior distribution τ^2 is the scaled chi-square distribution $\chi^2_{n_l-1}/[(n_l - 1)s_l^2]$. The completed-data posterior expectation $e_l = 1/s_l^2$ and $U_l = 2/[(n_l - 1)s_l^4]$. Approximate the distribution of τ^2 as scaled chi-square, $a\chi^2_b$, with mean $\bar{e}_{MI}$ and variance T_M. Thus, $ab = \bar{e}_{MI}$ and $2a^2b = T_M$ resulting in

$$a = T_M/(2\bar{e}_{MI}) \; and \; b = 2\bar{e}^2_{MI}/T_M. \tag{4.10}$$

This strategy will be useful while developing analysis of variance based on multiple imputed data sets (See Chapter 5).

4.7 Some Practical Issues

4.7.1 Number of Imputations

How many imputations are needed? The choice of the number of imputations, M, depends on the fraction of missing information. Note that, the degrees of freedom expression in equation 4.4, $\nu_M = (M-1)/r_M^2$ where r_M is the fraction of missing information. For the same degrees of freedom, the parameter with a larger fraction of missing information (relative to a parameter with smaller fraction of information) will need a larger number of imputation. The fraction of missing information varies by the parameter. Generally, ten imputations may suffice if the fraction of missing information is about 20% and five may be sufficient for a smaller fraction of missing information.

The multiple imputation is inherently a simulation technique and thus, the larger the value of M, the more stable the approximation of the posterior distribution of θ. Given the computational power and availability software, generating a large number of imputations is not difficult. Two critical quantities to monitor are $\bar{e}_{MI}$ (defined in equation 4.1) and T_M (defined in equation 4.2). One can be adaptive in choosing M to ensure that these two quantities stabilize.

Many investigators have developed methods for choosing M. A simple rule is to choose $100r$ where r is the largest fraction of missing information. This number may not be available easily. One possibility is to choose $M = 5$, estimate r_M based on this pilot number of imputations, and then increase the number of imputations to $100r_{max}$. Another possibility is to choose the fraction of incomplete observations for a particular analysis as the fraction of missing information. Thus, if the analysis involves discarding 30% of the subjects then choose $M = 30$.

4.7.2 Diagnostics

How to check whether the imputations are reasonable? A general approach is to compare summary statistics (such as mean and standard deviation, skewness, kurtosis, etc.) of the imputed and observed values. These statistics should be similar under missing completely at random and may not be the same if the data are not MCAR. Nevertheless, under modest missing at random assumption, observed and imputed data statistics should be similar.

Now consider a specific approach. Suppose that U is the variable that has some imputed values and $V_1, V_2, \ldots, V_p$ are the predictors used in the imputation model and, for now, assume that they are fully observed. The imputation assumes that the data are missing at random. Suppose that U_{obs} are the observed values, and U^*_{mis} are the imputed values. Under, MAR, if the imputation model is reasonable then

$$f(U_{obs}|V_1, V_2, \ldots, V_p) \equiv f(U^*_{mis}|V_1, V_2, \ldots, V_p).$$

This equivalence can be checked using the response propensity methods as follows. Let R be the response indicator for U, and $p = Pr(R = 1|V_1, V_2, \ldots, V_p)$ be the response propensity which is estimated using, say the best fitting logistic regression model to obtain $\widehat{p}$. Using the property of the balancing through propensity score, under the correct imputation model $f(U_{obs}|\widehat{p}) \approx f(U^*_{mis}|\widehat{p})$.

The balance checking diagnostics that were used in Chapter 2 for the weight construction can be used to asses whether or not, conditional on the propensity score, the observed and imputed values are similar. Specifically, suppose that U^* is a completed data vector with observed and imputed values. Regress U^* on $\widehat{p}$ and store the residuals, $\widehat{e}$. Construct kernel densities of

the residuals separately for the observed and imputed values. If the kernel densities overlap then the imputations are reasonable.

An alternative is to plot U^* versus $\widehat{p}$ and identify the observed and imputed values using different symbols. Under the correct imputation model, the symbols for the observed and imputed values should have be exchangeable across the full spectrum of the propensity score. The scatter plots of the observed and imputed values of log(REDOMEGA3) versus log(DIETOMEGA3 + 1) discussed in Chapter 3 are a simple example of this procedure as there was only one covariate.

What if some covariates $V_1, V_2, \ldots, V_p$ have missing values? Let V_{obs} denote the observed values. Under the correct imputation model, $f(U_{obs}|V_{obs}) \equiv f(U^*_{mis}|V_{obs})$ or, equivalently, $f(U_{obs}|\widehat{p}_{obs}) \approx f(U^*_{mis}|\widehat{p}_{obs})$ where $p_{obs} = Pr(R = 1|V_{obs})$. One way to estimate $\widehat{p}_{obs}$ is as follows. Suppose that the covariates have been multiply imputed along with U, say M times. Let $\widehat{p}_l$ be the propensity score estimate based on l^{th} completed data on $V_1, V_2, \ldots, V_p$. Assuming that the imputations of $V_1, V_2, \ldots, V_p$ are reasonable,

$$\widehat{p}_{obs} = \sum_{l}^{M} \widehat{p}_l / M.$$

The diagnostics proceed as before with using U^* and $\widehat{p}_{obs}$.

4.7.3 To Impute or Not to Impute

Sometimes a question is raised "Q1: I have a lot of missing values in the variable X. Should I impute or not"? For example, suppose that 75% of the observations are missing. Before answering this question, consider the following question "Q2: What is the alternative"? If the answer to Q2 is complete-case analysis, then it does not make sense because one will be discarding 75% of the observations on other variables, an enormous sacrifice to accommodate not imputing the missing values in X. Furthermore, the result may be severely biased, if the missing data mechanism is not MCAR. If the answer to Q2 is to ignore X from the analysis then imputation is no longer necessary, but this answer has altered the scientific question.

As long as the variable is going to be included in the analysis, it is better to deal with missing values through imputation using as much information as possible. The goal is to maintain the original sample in the analysis and do the best one can with the missing values. In this case, the number of imputations have to be large to get stable results.

4.8 Revisiting Examples

4.8.1 Data in Table 1.1

For the data in Table 1.1, cell-specific imputations may be performed using the normal model with cell-specific mean and cell-specific standard deviation. Consider the cell $D = 0, E = 0$ with 299 sample observations (n_{oo}) but only 212 respondents (r_{oo}). The assumed model is $x \sim N(\mu_{oo}, \sigma^2_{oo})$ with a diffuse prior distribution for the parameters, $\pi(\mu_{oo}\sigma_{oo}) \propto \sigma^{-1}_{oo}$. The sample mean is $\bar{x}_{oo} = -0.57$ and the sample variance $s^2_{oo} = 0.9^2$. Note from the basic theory that

1. $\sqrt{r_{oo}}(\bar{x}_{oo} - \mu_{oo})/\sigma_{oo} \sim N(0, 1)$
2. $(r_{oo} - 1)s^2_{oo}/\sigma^2_{oo} \sim \chi^2_{r_{oo}-1}$.

The eighty seven observations can be multiply imputed as follows:

1. Generate a chi-square random variable, u, with $r_{oo} - 1$ degrees of freedom and define $\sigma^2_{oo}* = 211 \times 0.9^2/u$.
2. Generate a random normal deviate z, and define $\mu^*_{oo} = -0.57 + \sigma^*_{oo}z/\sqrt{r_{oo}}$.
3. Generate 87 random normal deviates, $z_1, z_2, \ldots, z_{87}$, and define $x^*_j = \mu^*_{oo} + \sigma^*_{oo}z_j$ as the imputed values.
4. Repeat steps (1)-(3) M times.

The same procedure can be applied for the other three cells. After imputation, fit the logistic regression model with D as the outcome, E and X as the predictors. Table 4.2 provides five imputed data estimates and their standard errors for the regression coefficient of E in the logistic regression model with D as the dependent variable and E and X as predictors. The mean of the regression coefficients for E is $\bar{e}_{MI} = 0.4369$. The average of the squared standard error is $\bar{U}_M = 0.0205$ and the variance of the five estimates is $B_M = 0.0190$ and, thus, $T_M = 0.0205 + (1 + 1/5)0.0190 = 0.0433$ or the multiple imputation standard error is 0.2081. The fraction of missing information is $r_M = (1 + 1/5)0.0190/0.0433 = 52.7\%$ and the degrees of freedom $\nu_M = 4/.527^2 = 14.4$. This leads to a 95% confidence interval as $(.4369 \pm 2.14 \times 0.2081)$.

Table 4.3: Estimates and their standard errors for logistic regression example using the simulated data in Table 1.1

Imputation	E		X	
	Estimate	Standard Error	Estimate	Standard Error
1	0.4119	0.1434	0.5674	0.0794
2	0.4465	0.1435	0.4861	0.0776
3	0.4558	0.1420	0.5299	0.0780
4	0.4487	0.1423	0.5391	0.0798
5	0.4215	0.1441	0.5129	0.0781

4.8.2 Case-Control Study

Revisit the case-control analysis where the missing values in log(REDOMEGA3) was imputed using a model based procedure using the dietary value log(DIETOMEGA3 + 1). The number of missing values is rather large in this example (73% for cases and 27% for controls). Since imputations are being carried out separately for cases and controls, a large number of imputations may be needed to get stable answers. The problem is simple enough to repeat the imputation a large number of times using a variety of software packages.

A total of $M = 100$ imputations were carried out and after each imputation, the variable was retransformed back to the original scale, and then the substantive logistic regression model with PCA as the dependent variable and imputed (or completed) REDOMEGA3 as a predictor was fit. The second column in Table 4.4 provides the estimated regression coefficient, its standard error and the fraction of missing information.

In this example, an auxiliary variable was used to impute the missing values with the hope of recovering information about the missing red cell membrane values. Perform multiple imputation analysis without using the auxiliary variable as a way to assess its usefulness. The last column in Table 4.4 provides the estimate, standard error and the fraction of missing information based

Table 4.4: Estimated regression coefficient, the standard error (SE) and the fraction of missing information (FMI) for the case-control study example based on $M = 100$ imputations

Quantity	Using Dietary Intake	Ignoring Dietary Intake
Estimate	-0.344	-0.330
SE	0.104	0.111
FMI	61.4%	66.3%

Table 4.5: Estimates, standard errors and the fraction of missing information for estimating the mean

Status	Quantity	Controls	Cases
Include Dietary intake	Estimate	4.710	4.278
	SE	0.056	0.106
	FMI	20.5%	67.0%
Exclude Dietary intake	Estimate	4.727	4.313
	SE	0.059	0.115
	FMI	27.9%	73.1%

on $M = 100$ imputations. Including dietary intake in the imputation model reduces the fraction of missing information by about five percentage points (6.4% reduction in the standard error of the regression coefficient).

The efficiency gain in using dietary intake in the imputation process can differ by the type of parameters. Table 4.5 provides information about estimating the mean of red-cell membrane values for cases and controls. Reduction in the standard error and the fraction of missing information are comparable to those given in Table 4.4.

4.9 Example: St. Louis Risk Research Project

Consider data from St. Louis Risk Research Project described in Little and Rubin (2002). One of the objectives is to evaluate the effects of parental psychological disorders on the development of their children. This data involved 69 families classified into one of three categories, $G = 1$ (normal or control families from local community), $G = 2$ (a moderate risk group where one parent was diagnosed to have a secondary schizo-affective or some other psychiatric illness or one parent having a chronic physical illness), $G = 3$ (a high risk group where one parent suffers from schizophrenia or an affective mental disorder). There are 27 families in the normal group, 24 families in the moderate risk group and 18 families in the high risk group. From each family, two children were evaluated for standardized reading (R) and verbal comprehension (V) (both yielding continuous scores) and a binary variable (S) with low or high number of psychological symptoms. The goal is to compare the six variables (reading scores on two children (R_1,R_2), two verbal comprehension

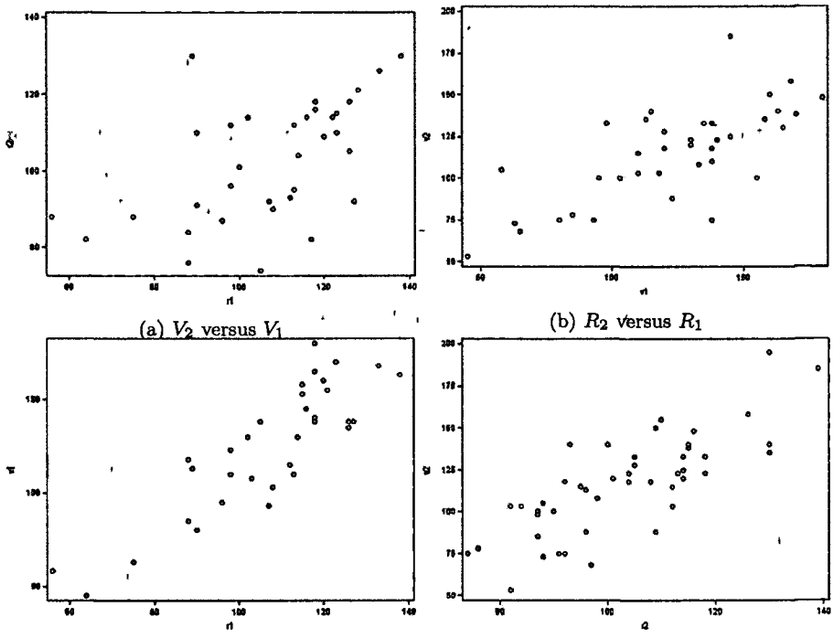
course (V_1, V_2) and two binary symptom classification (S_1, S_2)) across the three family categories. Not all variables are available for each family.

A sequential regression approach will be used to multiply impute the missing values in these six variables and then perform a series of analysis comparing the distribution of these six variables across the three groups. A prediction model is needed for each variable using all others as predictors. Since the goal is to compare the three groups, one may not want to pool models across the three groups to avoid biasing the imputations towards the null, i.e., if V_1 were imputed using two dummy variables for G, and using all other variables as predictors. The imputation model assumes that the relationship among the variables is the same across the three groups except for different intercepts. Separate imputation by group may be less efficient as it fails to capitalize on information from one group to impute in others. Here, however, maintaining the group difference is important and thus separate imputations will be carried out (similar to the case-control study example).

In developing models, the first step is to investigate general relationship between these six variables. Figure 4.1 provides scatter plots (pooling data across all three groups) relating V_1 to V_2, R_1 to R_2, V_1 to R_1 and V_2 to R_2. All scatter plots show that, in general, linear regression models are good candidates. Linear regression models were used to impute the missing values in V_1, V_2, R_1 and R_2 and a logistic regression model for S_1 and S_2.

Given a large number of missing observations, a total of 20 imputations were created for each group. Also, it is important to make sure that R_1, R_2, V_1 and V_2 are imputed positive values (normal linear regression with small sample size can have large variance for the predictive distribution, and hence may result in negative values for positive random variables). The imputations were carried out using the sequential regression approach as implemented in IVEWARE, bounds were set for the variables to be positive and the sequence was run for 50 iterations. The imputations could have been carried out using the package MICE in R or MI in STATA.

After imputations, many analysis could be performed to answer the basic question, "Are there differences in the reading and verbal measures and the symptoms across the three groups"? As a preliminary analysis, construct a family level score through averages $R = (R_1 + R_2)/2$, $V = (V_1 + V_2)/2$ and $S = (S_1 + S_2)/2$. The two contrasts of interest for each variable are $E(Y|G = 2) - E(Y|G = 1)$, $E(Y|G = 3) - E(Y|G = 1)$ where $Y = R, V, S$.



(a) V_2 versus V_1 (b) R_2 versus R_1

(c) V_1 versus R_1 (d) V_2 versus R_2

Figure 4.1: Scatter plots of variables in St. Louis risk research project

One may also be interested in a multivariate hypothesis of jointly comparing all three variables between the two groups.

After the imputation, three variables R, V and S were created and the mean and the covariance matrices were computed for these three variables in each group. Table 4.6 provides the means and their standard errors for each variable and group.

Table 4.6: Multiple imputation mean, standard error and the degrees of freedom for three composite variables R, V and S for three groups

Variable	$G = 1$			$G = 2$			$G = 3$		
	Mean	SE	df	Mean	SE	df	Mean	SE	df
R	112.5	2.6	23.8	102.6	2.6	21.1	102.2	5.1	18.2
V	135.5	4.6	19.4	108.9	6.4	17.9	112.8	9.2	15.5
$S(in\%)$	40.4	8.6	14.1	70.6	8.4	14.4	58.2	11.2	15.1

The differences are not statistically significant. For example, the difference between $G = 1$ versus $G = 3$ for R is $112.5 - 102.2 = 10.3$. That is, children in the high risk group score about 10 points lower in the standardized reading. Since the imputations were carried out independently, the standard error of the difference is $\sqrt{2.6^2 + 5.2^2} = 6.1$. The difference in the mean is about 1.7 times the standard error. Similarly, for V, the difference is $135.5 - 112.8 = 22.7$ and its standard error is $\sqrt{4.6^2 + 9.2^2} = 10.3$. The difference in the verbal comprehension score is slightly more than two standard errors lower for the high risk group than the normal group. For the percentage of symptoms, the difference is $40.4.5 - 58.2 = -17.8$ percentage points with the standard error, $\sqrt{8.7^2 + 11.2^2} = 14.2$. The proportion of children having a high number of symptoms in the high risk group is about 1.3 standard error larger than the normal group. Similar analysis comparing $G = 2$ and $G = 3$ shows that these variables have similar mean between these two groups. More formal tests can be performed by formulating these contrasts as regression coefficients in a general linear model or multivariate analysis of variance as discussed in Chapter 6.

Without imputations, data from many families would have been discarded. For example, of the 27 families in $G = 1$ group, only ten families have both R_1 and R_2 measured. That is, a loss of 17 out of 27 observations or 63%. However using imputations, the fraction of missing information is 13.7%. Table 4.7 provides the fraction missing observations (for the composite measures R, V and S and the fraction of missing information by group).

From the table, it is clear that through multiple imputations, considerable information has been "recovered" compared to the complete case analysis based on families with both children providing the data. The recovery of information could be improved by pooling $G = 2$ and $G = 3$ groups and using

Table 4.7: Fraction of missing observations and missing information: St. Louis risk research study

Variable	$G = 1 (n = 27)$		$G = 2 (n = 24)$		$G = 3 (n = 18)$	
	% Missing	FMI (%)	% Missing	FMI (%)	% Missing	FMI (%)
R	63.0	13.7	41.7	21.5	44.4	30.2
V	44.4	26.7	54.2	31.1	44.4	38.6
S	66.7	43.5	50.0	42.2	55.6	40.1

the dummy variable indicating $G = 1$ as a covariate in addition to the six variables R_1, V_1, S_1, R_2, V_2 and S_2.

4.10 Bibliographic Note

Multiple imputation was proposed by Rubin (1978a) and most theoretical and practical considerations are described in the classic book Rubin (1987). The theoretical justification given in this chapter is derived in Rubin (1987) and Raghunathan (1987). Rubin and Schenker (1986) develops the combining rules for a scalar parameter. Multiparameter case is developed in Raghunathan (1987) and refined in Li et al. (1991a). Combining chi-square statistics is also developed in Raghunathan (1987) and extended in Li et al. (1991b).

Barnard and Rubin (1999) developed the modification in equation (4.5). Reiter (2007) further refined the combining rule. Meng and Rubin (1992) developed a method for performing the likelihood ratio test with multiply imputed data sets.

4.11 Exercises

1. The data from the case-control study analyzed in Chapters 3 and 4 is available on the website www.iveware.org as a part of example data sets with IVEWARE package. It is an SAS data set with file name "Test." Extract the variables DHA_EPA, CASECNT and REDTOT. The goal is to fit the logistic model,

   $$\text{logit}Pr(\text{CASECNT} = 1|\text{REDTOT}) = \beta_o + \beta_1 \log(\text{REDTOT}).$$

   (a) Perform multiple imputation of the missing values using the package of your choice.

   (b) Fit the model given above and construct inferences for β_1.

   (c) Assess the benefit of using DHA_EPA in estimating β_1.