UCONN LIBRARY

University of Connecticut
OpenCommons@UConn

Doctoral Dissertations | University of Connecticut Graduate School

8-22-2018

Advances in the Analysis of Incomplete Data Using Multiple Imputations

Ruochen Zha
University of Connecticut - Storrs, ruochen.zha@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
Zha, Ruochen, "Advances in the Analysis of Incomplete Data Using Multiple Imputations" (2018). *Doctoral Dissertations*. 1944.
https://opencommons.uconn.edu/dissertations/1944

Advances in the Analysis of Incomplete Data Using Multiple Imputations

Ruochen Zha, Ph.D.

University of Connecticut, 2018

While Multiple Imputation (MI) has become one of the most broadly used methods for handling incomplete data, many questions remain unanswered regarding statistical inference when MI is used with incomplete data. One such question is how to calculate statistical power. Although it is widely acknowledged that MI improves estimation efficiency, reduces estimation bias, and partially restores power loss, there is a gap in the literature as far as quantifying the power gained from using MI over complete case analysis (CCA). Furthermore, the rates of missing information are well developed for traditional MI, but not for newly-adjusted MI. This thesis presents methodologies and simulation studies to calculate statistical power when MI is used, to compare the performance of MI with that of CCA, and to examine under which conditions MI can better restore statistical power. We also provide formulas to compute the rates of missing information for an adjusted two-stage MI, and apply them to evaluate the impact of an extra information source.

Advances in the Analysis of Incomplete Data Using Multiple Imputations

Ruochen Zha

B.S., Statistics, Huazhong University of Science and Technology, China, 2012

M.S., Statistics, University of Connecticut, US, 2016

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2018

Copyright by

Ruochen Zha

2018

APPROVAL PAGE

Doctor of Philosophy Dissertation

Advances in the Analysis of Incomplete Data Using Multiple Imputations

Presented by

Ruochen Zha, B.S. Statistics, M.S. Statistics

Major Advisor _____
Ofer Harel

Associate Advisor _____
Haim Bar

Associate Advisor _____
Elizabeth Schifano

University of Connecticut

2018

To my mother Ling Ma and my father Lixin Zha

ACKNOWLEDGEMENTS

My sincere gratitude and appreciation go to my Ph.D Advisor, Dr. Ofer Harel, without whom I couldn't have accomplished this achievement. I cannot thank him enough for all the work he has done helping me write this dissertation and teaching me how to do research. I was just out of undergraduate school when I first came to Uconn, barely knew anything about graduate study, and he was the one who opened the door of research to me. I am truly grateful to have his continuous guidance, support, patience, and understanding through my graduate study and other non-academic aspects of my life.

I must also thank Dr. Elizabeth Schifano and Dr. Haim Bar, who are excellent professors. I'm thankful to have them as my associate advisors in my dissertation committee. I appreciate all the valuable advice they gave to improve the dissertation. I also want to thank Dr. James Grady, for providing me the opportunity to work in UCHC, where I gained valuable experiences working with real-life data and consulting. I would also like to thank Dr. Golda Ginsburg, for kindly allowing me to use their data; Dr.Juned Siddique, for helping me better understand their research on data harmonization.

Many thanks to the Department of Statistics for providing me five years of assistantship, and being a great place to study in. Many thanks to all the fantastic faculties and staff here for helping me with my work, my study, and my life. Many thanks to my friends, whose encouragements are so important to me

Finally, I would like to thank my parents, for their love and support, for being my best role models, and for being my source of courage. It has been a long journey, and I'm glad my family is always there.

The data used in this manuscript is from a grant (R01 MH077312) awarded to Dr. Golda

Ginsburg by the National Institute of Mental Health (ClinicalTrials.gov: NCT00847561).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Overview

Incomplete data are an inevitable hurdle for researchers in all fields. Essential records are regularly and for many different reasons incomplete, missing, or otherwise inaccessible. Although not always considered an important issue, incomplete data can in fact lead to severe complications for researchers who do not deal with them appropriately.

Multiple Imputation (MI) is a well-known technique for dealing with missing data. It creates multiple completed datasets by replacing missing entries with different plausible values, therefore allowing researchers to apply methods initially developed for complete datasets. As compared to other missing data techniques, MI entails several important advantages. Firstly, it makes use of all available data, thereby minimizing unnecessary waste of potentially valuable information. Secondly, since the missing values are replaced by multiple plausible values instead of just one, MI captures the uncertainty introduced by the imputation process. Thirdly, MI is quite flexible in that it allows for the imputation model and the analysis model to differ. This enables the addition of auxiliary variables,

which have been shown to be very beneficial, to the imputation model. Fourthly, since the imputation model is a relatively general model, different types of analysis can be run on the same imputed datasets after the imputation is completed. Lastly, other desirable properties of MI include the utilization of the data collector's knowledge and the retention of database consistency across users [Rubin, 1988].

Although numerous studies have been conducted on MI performance, many questions nevertheless remain unresolved. One such question is related to statistical power. Simulation studies have shown that, as compared with complete case analysis (list-wise deletion), MI provides higher power and better estimation [Rubin, 1987, Graham et al., 2007, Desai et al., 2011, White and Carlin, 2010]. While offering substantial evidence for the attractiveness of MI in practice, these studies lack precise quantification for power gained by using MI instead of complete case analysis (CCA).

Power calculation plays an important role in data analysis and experiment design. Statistical power is defined as the probability of correctly rejecting a null hypothesis, therefore evaluating the likelihood of detecting a statistically significant effect when it is present [Balkin and Sheperis, 2011]. Knowing the statistical power of a hypothesis test can help increase research efficiency, guide research design, and estimate required sample size [Steidl et al., 1997]. To date, most funding agencies, including the National Institutes of Health (NIH) and the National Aeronautics and Space Administration (NASA), require a power calculation segment in their grant application.

In general, there are two different methods to calculate statistical power: closed-form expression and simulation. Researchers can usually compute their statistical power using tables or closed-form formulas provided in textbooks such as those by Murphy et al. [1998] and Cohen [1988]. This method requires researchers to provide significance level α, effect size, and sample size and has been implemented in most standard statistical software including SAS [SAS Institute Inc., 2011], R [R Core Team, 2015], and SPSS [IBM Corp., 2013], among others. As an alternate, Monte Carlo simulation has also been recommended, especially when obtaining the closed-form formulas is impractical [der Sluis S et al., 2008].

However, closed-form methods do have their limitations. Despite the fact that traditional closed-form formulas and tables are widely used in today's research, they were actually developed for use with complete datasets. This means that, as with other complete data methods, traditional power calculations also suffer from incomplete data problems. A traditional power calculation method may lead to biased results or inefficient decisions when naively applied to an incomplete data set, thereby reducing research quality. Although articles discussing how to calculate statistical power for incomplete data do exist [Davey and Savla, 2010, 2009, Tang, 2017, Muthén and Muthén, 2002, Tang and Tang, 2002], none have considered MI. Considering that MI is a commonly used missing data technique, researchers face an important challenge in selecting optimal sample size without a way to calculate statistical power when MI is used for incomplete data. In that case, experiment resources cannot be used most efficiently. In this paper, we provide a general method for calculating statistical power for the MI t-test and the two-sample student's t-test. We are thereby able to quantify the impact on statistical power of using MI and

compare the performance of MI to the performance of CCA and to the performance that would have been obtained had there been no missing data. We also present a way to calculate the optimal sample size derived from the power-calculation method.

Another problem of interest is the rate of missing information. Also known as the fraction of missing information (FMI) [Rubin, 1987, Schafer, 1997, van Burren, 2012], the rate of missing information is defined as the ratio of information lost due to missing data to total information of the complete dataset. It is a useful tool to evaluate the effect of missing data and provides a guide for determining the number of imputations needed for MI. While FMI estimates and their behavior are well studied for traditional MI [Rubin, 1987, Schafer, 1997, Harel, 2007b], they are yet to be developed for a newly adjusted two-stage MI [Reiter, 2008]. We therefore present a method for calculating the rates of missing information for the adjusted two-stage MI, and discuss how it can be used to assess the impact of missing data as well as how it can be a source for extra information.

We now present a literature review of existing research on missing data, multiple imputation, statistical power calculation, and rate of missing information.

1.2 Missing Data

Missing data has long been a challenge for researchers in a range of different fields [Graham, 2009]. In clinical trials, if a patient unexpectedly drops out before the trial is completed, researchers will fail to collect data on this patient after he or she leaves; if a patient stays in the experiment but does not comply with the pre-assigned experiment

plan, the data collected will be unreliable and should not be used. In survey studies, participates commonly choose not to answer certain questions because they are "not applicable" or too sensitive. Incomplete data caused by missing or erroneous entries can occur in many different fields and for many different reasons: incorrect experiment design, defective measurements, equipment malfunction, experiment unit failure to comply with the experiment plan, participants skipping certain questions in a survey, records lost over time, incorrect records caused by human error, or observations that are simply not available due to confidentiality reasons. While researchers can manage to prevent some missingness from happening with better experiment design, other missing values are simply inevitable. In short, datasets are rarely perfect and incomplete datasets can lead to many expected and unexpected results.

Why is missing data of such great concern to us? From a statistician's perspective, incomplete data can severely impact statistical analysis, and may even destroy the validity of a population inference. One major reason for this issue is that most traditional analysis methods are not designed for incomplete datasets [Schafer and Graham, 2002]. Naively applying a complete data method to an incomplete dataset may therefore lead to biased results. For example, let's consider a national income survey. If all low-income individuals choose to not participate and yet the researcher still decides to use the (observed) sample mean as the estimate of population average, his or her conclusion will be positively biased. Furthermore, when the dataset is incomplete, modern statistical software and packages can provide erroneous or unexpected analysis results. A simple example of this is that the R function "mean" is used to calculate the mean of a vector in which non applicable (NA) entries exist, the software will refuse to conduct the calculation unless it is specifically

told to remove the missing entries. Other software and packages will automatically and by default eliminate any incomplete observations.

1.2.1 The Missing Data Mechanism

When dealing with incomplete data, researchers not only study the observed data itself, but also the underlying process that generates the missing data. This process is usually unknown and not testable, but it plays an important role in modern missing-data inference. Before choosing an appropriate missing-data technique, researchers need to make assumptions about the underlying process.

Consider a complete dataset Y_{com}, which can be split into observed and unobserved parts (Y_{obs}, Y_{mis}). The distribution of Y can be represented by $P(Y, \theta)$, with θ denoting the distribution parameter. R is the response indicator, where R=1 means the corresponding value is observed, while R=0 means the corresponding value is missing. Rubin [1976] denoted that R should be viewed as a probabilistic phenomenon, making it feasible to discuss the distribution of R, $P(R, \phi)$, where ϕ is the parameter of the distribution. This distribution describes the underlying process that creates missing values, and is known as the missing data mechanism (MDM) [Schafer and Graham, 2002].

There are three major types of MDM depending on the distribution of R. When the distribution of R relies on neither observed nor unobserved data, the MDM is considered missing completely at random (MCAR). The term fully implies that the propensity to be a non-response is completely irrelevant to the data itself. Mathematically speaking, it

means that $P(R|Y_{com}, \phi) = P(R|\phi)$. When MDM is MCAR, Y_{obs} is eventually a random sample drawn from Y_{com} [Enders, 2010], and Y_{mis} and Y_{obs} therefore have the same distribution, i.e., no systematical difference [Sterne et al., 2009].

If the distribution of R is related to the observed data but not to the unobserved data, we call the MDM missing at random (MAR). A more formal way to describe MAR is that $P(R|Y_{com}, \phi) = P(R|Y_{obs}, \phi)$. The last type of MDM is missing not at random (MNAR). In cases of MNAR, the distribution of R may depend on unobserved data, i.e., whether a missing value is determined by the value itself or by some other unobserved quantities. For an example explaining the three different types of MDM, consider again the national income survey. Researchers record personal annual income with set variables such as age and gender. If there is a computer system failure which causes some individuals to be excluded from the survey, this would be considered MCAR; if, for some reason, response rate depends only on gender, it would be MAR; if all low-income individuals decide not to participate, this would be considered MNAR.

Unfortunately, while MAR is a frequently used assumption for many missing data approaches, the assumption itself is not testable using only the observed data [Schafer and Graham, 2002, Molenberghs et al., 2008, Rhoads, 2012]. Only with additional data, such as follow-up data coming from the experiment units with initially missing data [Graham and Donaldson, 1993, Schafer and Graham, 2002], may the assumption be tested. If it is not available, Schafer and Graham [2002] suggest making explicit assumptions and reporting the sensitivity of the results.

1.2.2 Ignorability

Modeling the missing data requires a specification of the conditional distribution $P(\mathbf{Y_{mis}}|\mathbf{R}, \mathbf{Y_{obs}}, \mathbf{X_{com}})$, where X_{com} is the covariate of the whole dataset. The model can however be simplified using the ignorability assumption. Some believe that ignorability is equivalent to MAR. In fact, it requires more than just MAR. Formally defined [Rubin, 1987], missingness is considered ignorable if 1. θ and ϕ are "distinct", meaning that $P(\theta, \phi) = P(\theta)P(\phi)$, and 2. the MDM is MAR [Rubin, 1976, Harel and Zhou, 2007]. With the above two conditions present we have

$$
\begin{aligned}
P(Y_{mis}|R, Y_{obs}, X_{com}) &= \frac{P(Y_{com}, R)}{P(Y_{obs}, X_{com}, R)} \\
&= \frac{1}{P(Y_{obs}, X_{com}, R)} \int\int P(Y_{com}, X_{com}, R, \phi, \theta) d\phi d\theta \\
&= \frac{1}{P(Y_{mis}, R)} \int\int P(R|Y_{com}, X_{com}, \phi) P(Y_{com}, X_{com}|\theta) P(\phi, \theta) d\phi d\theta \\
&= \frac{1}{P(Y_{obs}, X_{com}, R)} \int\int P(R|Y_{obs}, \phi) P(Y_{com}, X_{com}|\theta) P(\phi) P(\theta) d\phi d\theta \\
&= \frac{P(R|Y_{obs}, X_{com}) P(Y_{mis}, Y_{obs}, X_{com})}{P(Y_{obs}, X_{com}, R)} \\
&= \frac{P(R|Y_{obs}, X_{com}) P(Y_{mis}|Y_{obs}, X_{com}) P(Y_{obs}, X_{com})}{P(Y_{obs}, X_{com}, R)} \\
&= P(Y_{mis}|Y_{obs}, X_{com})
\end{aligned}
\tag{1.1}
$$

.

Equation (1.1) implies that, when the missingness is ignorable, the distribution of $Y_{mis}|Y_{obs}$ does not depend on the distribution of R. In other words, we can safely "ignore" the specific distribution of R. Modern missing data techniques usually require researchers to model the joint distribution of $P(Y_{mis}, R|Y_{obs})$. With the ignorability assumption, the usually complicated joint distribution can be simplified.

1.2.3 Monotone Pattern

A missingness pattern specifies where the missing values are located in the incomplete dataset. We say that missing data with k variables follows a monotone pattern if it can be arranged as $(Y_1, ..., Y_k)$, such that:

1. If a unit i is observed on Y_j, it is also observed in any $Y_{j'}$, where $j' < j$;

2. If a unit i is missing on Y_j, it is also missing in any $Y_{j'}$, where $j' > j$ [Carpenter and Kenward, 2013].

It is easier to conduct a maximum likelihood estimation on an incomplete dataset with a monotone pattern and ignorable MDM. Little and Rubin [2002] denoted that when the MDM is ignorable, ignorable ML estimates can be calculated by maximizing $P(\theta|Y_{obs})$. When the missingness pattern is monotone, the ignorable likelihood function can be partitioned in such a way that each part can be separately maximized to gain the maximum likelihood estimates of the parameters [Little and Rubin, 2002]. In a bi-variate case where Y_1 is partially observed and Y_2 is fully observed, an inference about the conditional distribution $Y_2|Y_1$ can be obtained using the completed cases, and an inference about the marginal distribution of Y_2 can be obtained by analyzing Y_2 in all cases.

1.3 Missing Data Techniques

To obtain valid inference from incomplete data, researchers need to be cautious when choosing a missing data technique. Many different methods have been developed to deal with incomplete data. However, some commonly used methods can themselves result in

misleading conclusions if specific assumptions are not met.

One such example is Complete Case Analysis (CCA), also called list-wise deletion (LD). This method uses only complete cases when performing analysis, discarding all incomplete cases. Due its simplicity, CCA is the most popular missing data approach and is widely used in statistical analysis. It is known that if the data are assumed to be missing completely at random (MCAR), CCA will provide an unbiased estimate for the parameters of interest. There are other situations where it is appropriate to use CCA with MAR or MNAR. Bartlett et al. [2015] found that if a correctly specified logistic regression is provided and the parameter of interest is an exposure odds ratio, then the parameter of interest can be unbiasedly estimated using CCA as long as "the missingness doesn't jointly depend on the exposure and outcome." However, CCA also has several well-known disadvantages. First, when the MDM is MAR or MNAR, CCA has the potential to generate biased results since it ignores the systematic difference between the observed and unobserved data. Second, simply discarding the incomplete cases may result in a waste of information from partially observed subjects, which can then lead to less efficient parameter estimates. In addition, there are situations for which it is impossible to use CCA. Reiter [2008] described a dataset where one variable is completely missing. In such a case, there are zero complete cases and CCA becomes impractical.

Another method for dealing with incomplete data is single imputation. Researchers sometimes use the average of the observed parts of the sample to replace the missing values; while at other times they build a model to impute the missing values. Since, single

imputation completes the dataset, researchers can then use regular software and packages for their analysis of the completed dataset. Mean imputation however has similar disadvantages as CCA in terms of ignoring potential differences between observed and unobserved data. Even if the values are imputed from a plausible model, single imputation will underestimate the standard error of the parameter estimates. This is because single imputation mistakenly treats imputed values in the same way as truly observed data points. In this case, variability introduced by the imputation process itself is ignored.

Other than noticeably flawed incomplete data approaches, modern missing data techniques have been developed. There are Bayesian iterative simulation methods, including data augmentation and the Gibbs' sampler [Little and Rubin, 2002]. Schafer and Graham [2002] recommended two modern missing data approaches: the first is the maximum likelihood estimation (MLE) based on all available data, and the second is multiple imputation (MI), proposed by Donald [1978] and detailed in the following section.

1.4 Multiple Imputation

Multiple imputation (MI), introduced by Rubin [1987], is a conventional and flexible method to deal with incomplete data [Schafer, 1999]. As a simulation-based method, it replaces missing entries with different plausible values, thereby creating multiple complete datasets. MI's practical usefulness and simplicity drew much attention soon after its development and several explorations of particular aspects of MI have since been conducted. Meng [1994] discussed the method as it pertains to cases where the model adopted by analysts is uncongenial to the model used for Multiple Imputation. Schafer and Olsen [1998]

and Schafer [1999] reviewed important features of MI, answered frequently asked questions, and provided guidance for practical usage of MI. Graham [2009] provided a review of the technique through a practical summary, discussing many issues such as inclusion of auxiliary variables, number of imputations, and missing data analysis challenges like analyzing clustered, longitudinal, and categorical data. Schafer [1997] developed several joint modeling techniques to generate imputations for multivariate missing data. White et al. [2011b] discussed another imputation approach: multiple imputation by chained equation, also known as fully conditional specification. Advanced imputation methods (such as nested imputation) and new combining rules are also discussed by Shen [2000], Reiter [2008], Harel [2007a], Marshall et al. [2009], and McGinniss and Harel [2016].

1.4.1 Multiple Imputation: the General Idea

In general, MI consists of three major steps: imputation, post-imputation analysis, and pooling of results. In the imputation stage, each missing value is replaced by m plausible values. There are different ways to obtain the plausible values, including a random draw from a Bayesian posterior predictive distribution. The imputation stage results in m completed datasets. Notice that the completed dataset is different from the "complete dataset" – while the former represents the data with imputed values, the latter represents the dataset we would have seen if there had been no missing values. Consider an example where the population parameter of interest is Q (including population mean, correlation coefficient, etc.). In the analysis stage, a complete data method is applied to each of the m datasets. This stage leads to m estimated parameters $\hat{Q}_i$ and their corresponding variance U_i(i = 1, 2, ... m). The results-pooling stage uses the m analysis results to form the

final inference using Rubin's combining rule. The final estimate is the average of them estimates:

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i. \tag{1.2}$$

The final variance estimated consists of the within-imputation variance $\bar{U}_m$ and between imputation variance B_m. The within-imputation U_m is the average of the individual variance estimates from each imputed dataset:

$$\bar{U}_m = \frac{1}{m}\sum_{i=1}^{m}U_i. \tag{1.3}$$

The between imputation variance B_M is defined as follows:

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q}_m)^2. \tag{1.4}$$

The final estimated variance of the MI estimate $\bar{Q}_m$ is a summation of adjusted between-MI imputation and within-MI imputation. According to Rubin's rule, this is $T_m = (1 + \frac{1}{m})B_m + \bar{U}_m$. We can define $r_m = (1 + \frac{1}{m})B_m/\bar{U}_m$ as the relative increase in variance due to nonresponse [Rubin, 1987], and $\gamma_m = \frac{r_m}{r_m+1}$ as the fraction of missing information due to nonresponse [Rubin, 1987].

1.4.2 Proper Imputation, Congeniality, and Self-Efficiency

Certain rules guarantee that MI generates valid inference. From a frequentist perspective, valid inference should have randomization validity, meaning that actual interval coverage equals nominal interval coverage and that the actual rejection rate equals the nominal rejection rate [Rubin, 1996, 1987]. To achieve randomization validity for MI inference, the complete data inference must be randomized and valid and the MI model must be proper.

Rubin [1987] provided a technical description for proper imputation modeling. Roughly speaking, if the imputations are drawn from a Bayesian posterior predictive model under a posited response mechanism and an appropriate model for the data, the large sample MI is proper [Rubin, 1996, 1987]. Rubin [1996] recommended the use of all variables that are either related to the response mechanism or to $(\hat{Q}, U)$ in the imputation model to avoid improper imputations.

Based on Meng [1994]'s definition of congeniality, an imputation model is congenial with the analysis procedure if there is a Bayesian model which can provide posterior mean and variance asymptotically equal to $\hat{Q}$ and U given both observed and complete data, and if the posterior predictive model of Y_{mis} under that Bayesian model is the same as the imputation model. MI provides valid inference if the imputation model and analysis model are congenial. If they are not, the inference is still valid when the imputer makes less assumptions than the analyst. When however the imputation model is less saturated than the analysis model, the confidence validity is unfortunately not guaranteed [Meng, 1994, Xie and Meng, 2014].

Meng [1994] also mentioned that to have valid MI, the complete data procedure cannot be arbitrary; it must be self-efficient. A self-efficient procedure is one whose efficiency (i.e., the variance of the estimate) cannot be enhanced by deleting part of the data. Thorough discussions about proper MI and self-efficiency can be found in Nielsen [2003a,b], Rubin [2003], and Meng and Romero [2003].

1.4.3 Different MI Algorithms

There are several different ways to multiple-impute incomplete data. Carpenter and Kenward [2013] mentioned several methods for imputing normally-distributed data: sequential regression for data with a monotone missing pattern, joint modeling, and full conditional specification (FCS). These methods are described below.

When missingness has a monotone pattern and MAR MDM can be assumed, the joint likelihood distribution can be partitioned into several conditional distributions, the parameters estimates of which can be validly obtained from the observed data. Under these circumstances, researchers can use a sequence of linear regressions to finish the imputation task without requiring iterations.

Special assumptions for the missingness patterns are not required for joint modeling, which usually uses the Gibbs Sampler method. This allows researchers to impute the missing value as well as estimate the unknown parameters [Schafer, 1997]. Roughly speaking, this method randomly picks from the predictive posterior distribution of the data's condition parameters, then randomly draws values from the distribution of the missing data condition on the parameters and the observed data. This procedure is repeated multiple times until convergence achieved.

FCS, also known as Multiple Imputation using Chained Equation (MICE), is similar to the joint modeling method in that it makes no assumptions about the missingness

pattern and allows researchers to impute missing values and estimating parameters simultaneously. A difference between joint modeling and FCS is that, instead of randomly picking from $(Y_{mis}|Y_{obs}, \theta)$ when imputing missing values, FCS draws each variable from its distribution condition on all the other variables [Schafer, 1997, White et al., 2011a, van Buuren and Groothuis-Oudshoorn, 2011, Azur et al., 2011, van Buuren, 2007]. As such, FCS can be easily extended to non-normal data. It is flexible in handling each variable on a case-by-case basis, and provides an easier way to handle situations where specifying a joint distribution is difficult [Raghunathan et al., 2001]. However, there is limited theoretical basis for FCS. While it is similar to an MCMC procedure, its properties are not generally justified [White et al., 2011a]. There is therefore no guarantee that FCS is a proper MI method.

Predictive mean matching (PMM) [Rubin and Schenker, 1986, Little, 1988] is another MI algorithm that can be used to deal with non-normal data. This method has recently been discussed by Vink et al. [2014], Morris et al. [2014]. Instead of directly using random draws from the plausible model of unobserved values given observed values, PMM uses an observed response whose predicted mean is closest to the random draw. PMM's major disadvantage is similar to that of FCS: while PMM solves the issue that MI may generate impractical values, it unclear whether it is proper or not.

Since its development, different reserachers have proposed other extensions of MI. Nested multiple imputation [Shen, 2000], also known as two-stage multiple imputation [Harel, 2009], is one such extension of traditional MI. Whereas, in traditional MI, all missing data are treated similarly and are imputed together, in nested MI, however, the

unobserved data, Y_{mis}, are split into two parts : $(Y_{mis}^{A}, Y_{mis}^{B})$. This is done either to gain computational efficiency or because the unobserved data are of different types.

1.5 Statistical Power

Statistical power analysis is particularly important in the social, behavioral, and biomedical sciences [Faul et al., 2007b]. Defined as one minus the probability of falsely accepting the alternative hypothesis (type II error rate), it describes the ability of a hypothesis test to detect a significant effect. Researchers conducting a randomized control trial would like to find out whether a treatment has significant effect on the responses; observers in an observational study may want to test whether a factor impacts the observations in which they are interested. In these two situations and in others, the reliability of the hypothesis test conclusions are assessed in terms of statistical power, along with the type I error rate (significance level).

Statistical power provides a way to compare different statistical procedures. For example, researchers are often interested in the most powerful test procedure (MP test). With a fixed significance level, the test procedure that provides a higher power is considered the "better" test. This means that, under the same conditions, the procedure producing higher power will have a better ability to correctly reject the null hypothesis, therefore reducing the probability of generating erroneous test conclusions.

Another (and perhaps most common) role statistical power calculation plays is in helping determine the required sample size at the experiment design stage. High-quality study

pre-planning by researchers is crucial to achieve satisfactory statistical power [Peterman, 1990, Moher et al., 1994]. Studies suffering from insufficient power will have a low chance of detecting the effect in which they are interested, while studies with excessive power may be too costly and entail other ethical issues [Baguley, 2004]. A well-designed study should strive to have a low type I error rate and high statistical power. While the former can be fixed in the hypothesis test procedure, type II error rates can only be reduced by increasing the sample size. Failing to carefully evaluate statistical power before the data is collected may lead to studies with data that are either too small or larger than necessary. Today, most funding agencies require a power calculation segment in their grant application.

1.5.1 Power Calculation for Complete Data

The traditional method to calculate statistical power is to use closed form formulas. By examining the test-statistic distribution under the alternative hypothesis, the probability that the null hypothesis has been rejected can be obtained. Consider a single linear regression $y = \beta_0 + \beta_1 x + \epsilon$. An F test is used to test if $\beta_1 = 0$. The test statistic follows an F distribution $F_{p,q}$ under the null hypothesis and a non-central F distribution $F_{C,p,q}$ under the alternative hypothesis. C is the non-central parameter and p, q are the degrees of freedom. With a significance level equal to α, the rejection region of this test is $(F_{\alpha,p,q}, \infty)$ ($F_{\alpha,p,q}$ is the critical value for an F distribution with significance level α and degree of freedom p and q) . The statistical power corresponding to the F test is then $P(F_{p,q,C} > F_{\alpha,p,q})$, the probability that the test statistic (in this case, a random variable following a non-central F distribution with parameters p, q, and C) falls in the rejection region. More details about calculating statistical power for other types of hypothesis tests

can be found in Cohen [1988] and Murphy et al. [1998].

A more popular structure for the closed form method is presented by Cohen [1988], who indicated that statistical power is determined by effect size, significance level, and sample size. As Cohen [1988] also stated, effect size is a value that describes how large "the degree to which the phenomenon under a study is manifested". In other words, effect size provides a standard way to measure the difference between null and alternative hypotheses. In studies focusing on population average, the effect size is the difference of the mean divided by the population standard deviation. A smaller effect size will be harder to detect, therefore yielding a lower power (with all other conditions fixed). This method is widely used accross almost all fields, and is implemented by most statistical software.

The Monte Carlo simulation is an alternate method that has also been recommended, especially when obtaining the closed-form formula is impractical [der Sluis S et al., 2008]. This is detailed in Muthén and Muthén [2002] and Beaujean [2014]. In general, the procedure includes generating a large population from which a large number of samples are drawn. Each of the samples are then estimated and power is calculated as the proportion of replicates where the null hypothesis is rejected at a certain significance level.

In most circumstances, statistical power can be calculated using statistical software. Power calculation functions/procedures to calculate statistical power for complete datasets have been implemented in most statistical-analytical tools in use today, including Stata [StataCorp, 2013], R package PWR [Champely et al., 2015], SAS procedure GLMPOWER,

POWER, and the POWTABLE macro [SAS, 2008]. Other softwares have also been specifically developed for sample size and power calculation, such as G*power [Faul et al., 2007a], nQuery [Elashoff, 2007], and PASS [NCSS, 2017].

1.5.2 Power calculation for Incomplete Data

As with other methods, traditional power calculation methods were initially designed for complete data. Missing entries lead to power reduction, and incomplete data impact the reliability of power calculation results in general. If some data are missing in the data-collection stage of an experiment, then the final statistical power will be lower than the expected nominal power.

One way to solve this issue is to use complete case analysis (CCA). After deleting all incomplete cases, CCA treats the remaining complete cases as a complete dataset with a smaller sample size. If power is calculated before data are collected, researchers can make assumptions about how much data will be missing. Although quite simple to apply in practice, this method does have certain disadvantages. As mentioned in Section 1.3, incomplete cases can still contain useful information. Deleting them without careful prior evaluation may result in the loss of potentially useful records, thus requiring an unnecessarily large sample size which may likely be a waste of experiment resources especially as it is expensive to conduct experiments on only one experiment unit.

Luckily, there are other ways to calculate statistical power by incorporating more reasonable missing-data techniques. Davey and Savla [2010] introduced a method to calculate statistical power with missing data using a structural equation model. This approach consists of seven steps, including specifying population models for null and alternative hypotheses, generating complete data, specifying an incomplete data model, applying the incomplete data model to the known data structure, and, lastly, estimating and calculating statistical power and sample size. This method takes the MDM into consideration while selecting the incomplete data model and treating each different pattern of complete/incomplete data as a group of its own. It would therefore be more appropriate to use this method over CCA when the MDM is not completely random.

Another method for calculating statistical power with incomplete data is to perform the maximum likelihood method on the incomplete data, then evaluate the estimators and variances and calculate the statistical power directly. For example, Tang [2017] derived the closed-form restricted maximum-likelihood estimator and the Kenward-Roger's variances estimator for fixed effects in mixed-effects models for repeated measures. Then, using the Kenward-Roger's variances estimator, a power calculation formula is derived for a Wald t-test from the interest estimates. An advantage of this method is that by using the closed-form restricted maximum-likelihood (REML) estimator method, all available information are taken into consideration, thereby overcoming the waste-of-resource disadvantage of CCA. However, it requires relatively heavy mathematical calculations and new formulas need to be derived for new data structures and hypothesis tests.

The Monte Carlo simulation provides another solution for calculating statistical power for a partially observed dataset. Muthén and Muthén [2002] presented how statistical power can be calculated using this simulation method. A large number of samples (the article recommends 10000) are randomly drawn from a population with hypothesized parameters and a special MDM. Each of the samples is then statistically analyzed and power is calculated using the number of models correctly rejecting the null hypothesis (i.e., with a p value smaller than the desired significance level α). Other applications of this method can be found in Wolf et al. [2013]. Overall, the Monte Carlo simulation is quite flexible and easy to understand. It can be applied under many conditions, and can help clarify how missing data affects statistical power. However, it does require heavy computation and can be impractical if researchers are interested in a large number of different conditions [Davey and Savla, 2009].

1.6 Rates of Missing information

The rate of missing information [Schafer, 1997], also known as the fraction of missing information [Rubin, 1987], is a quantity measurement for incomplete data analysis that measures relative information loss due to missing data. The rate of missing information is defined as the ratio between the information loss due to incomplete data and the information contained in the complete dataset, and is referenced for both maximum likelihood (ML) algorithms and multiple imputation (MI).

The rate of missing information is of interest to researchers in dealing with incomplete data. It can be used to monitor the quality of survey data and evaluate how unobserved

data affect the inference of the parameter of interest, Q [Wagner, 2010]. Schafer [1997] pointed out that the convergence rate of an EM algorithm is determined by the rate of missing information λ in the scalar case, thereby also representing the convergence rate of data augmentation. Harel [2007a] demonstrated that the rate of missing information is also important to determine the number of imputations required for the multiple imputation method. The use of rates of missing information to determine the number of imputations is further discussed by Bodner [2008] and von Hippel [2018, in press]. Andridge and Thompson [2015] recently proposed a variable selection method using λ to select the best candidate imputation model.

There are different ways to calculate the rate of missing information. One such way is through the maximum likelihood method, as shown by Fraley [1999], Little and Rubin [2002], and Savalei and Rhemtulla [2012]. Another way is to utilize the results obtained from multiple imputation (MI). As proposed by Rubin [1987], the rate of missing information can be calculated using the within-imputation and between-imputation variances.

1.7 Outline

The rest of this dissertation is structured as follows. In Chapter 2, we introduce a methodology which can in most cases calculate statistical power when MI is used for incomplete data. We provide a detailed definition of proper imputation, which is an important condition for the vailidity of Rubin's MI inferences. We derive a general power calculation method from Rubin [1987]'s MI inference. We show that, in addition to differences between the null and alternative hypotheses and the significance level, when the data

is incomplete we require the number of imputations, the expected within-imputation variance, and the between-imputation variance (i.e., the expected MI variances) to calculate statistical power. We then provide closed-form formulas for the expected MI variances and specify the general power calculation method for a one-sample t-test. We then provide simulation studies to compare the Monte Carlo simulated power with the power calculated by our method. These simulation studies are conducted on several different setups and prove that our method generally provides a precise estimation of statistical power. Through the simulations we also examine when MI can better recover lost information as compared to complete case analysis. In addition, we estimate the MI variance expectations using the MI variances if a closed-form expectation cannot be obtained. We use the Child Anxiety Prevention Study to illustrate our methodology.

In Chapter 3, we extend our results from Chapter 2. We propose a method that uses maximum likelihood to obtain the expected MI variances. Since MLE-related inference is widely explored and examined, we believe the MLE-based MI variance expectation estimation method can be of great help. We apply this method to the one-sample t-test discussed in Chapter 2 to show that this method generates similar results as in Chapter 2. We then obtain the expected MI variances for a two-sample t-test and perform a closed-form power calculation equation for the two-sample t-test. A simulation study similar to the one in Chapter 2 is performed to test the validity of our proposed method.

We furthermore propose a way to estimate optimal sample size for the scenarios discussed in Chapter 3. We show that if MDM is MCAR, MI variances are determined by the population variance of the response, the sample size, the number of complete cases, and

the multiple correlation coefficient between the response and the covariates used for imputation. Therefore, statistical power is determined by sample size, effect size, significance level, percent of missing responses, number of imputations, and the multiple correlation coefficient between the response and the covariates. We provide a power/sample-size table to help researchers to determine their optimal sample size at the experiment design stage. Simulation studies conducted on several different setups are provided to examine the sample-size table's performance.

In Chapter 4, we examine the rates of missing information for an adjusted two-stage MI [Reiter, 2008], which is derived for scenarios where an extra information source is introduced only for imputation but not for analysis. This chapter is motivated by Siddique et al. [2015] in which the effect of fluoxetine on depression was evaluated. Their study included multiple clinical trials using different measurements to measure depression levels. Two different depression measurements were cited as two correlated responses and any clinical trial using only one of the measurements was considered incomplete. Since there were no trials in the study that successfully used both measurements, two additional clinical trials were used as the calibration dataset to capture the relationship between the two measurements. The additional two trials were not used for the analysis, as Reiter [2008]'s two-stage MI was used to harmonize the clinical trials. The researchers wanted to evaluate the effect of the extra information source (called the validity dataset) but could not. We proposed a method to calculate the overall rate of missing information, the rate of missing information due to the unknown model, and the rate of missing information if the model is given. These values can later be used to evaluate the impact of both the missing data and the calibration dataset. We examine the behavior of these rates of

missing information in a simple scenario with only the two responses and no covariates. We alter the correlation between the two responses and assess how this change affects the rates. We then conduct a different simulation in which the data structure mimics that of Siddique et al. [2015]'s trials. We estimate the rates of missing information and again assess their behavior when the calibration data have different sizes.

In Chapter 5, we review the results obtained and presented in the dissertation. We discuss the limitations of our work, and provide directions for potential future research.

1.8 Notations

Below are the notations used in the dissertation, following the order of appereance.

General missing data related notations:

$\boldsymbol{X}$: population covariates

$\boldsymbol{Y}$: population response

ρ: population correlation coefficient

Q: parameter of interest

Q_0, Q_1: the values of Q under null and alternative hypothesis

$\boldsymbol{R}$: response indicator

$\boldsymbol{I}$: inclusion indicator

$\boldsymbol{X_{com}}$, $\boldsymbol{Y_{com}}$: the covariates and response of the complete dataset, in which there is no missing data

$\boldsymbol{X_{obs}}$, $\boldsymbol{Y_{obs}}$: the covariates and response of the fully observed part of a dataset. The fully

observed part of a dataset is part of the dataset where both the response and covariates are observed

$\boldsymbol{X_{mis}}$, $\boldsymbol{Y_{mis}}$: the covariates and response of part of the dataset, where the response is not observed

y_i, $\boldsymbol{x_i}$: response and covariate(s) of the i^{th} unit in a sample

$\boldsymbol{\mu_x}$, $\boldsymbol{\Sigma_x}$: mean vector and variance-covariance matrix of $\boldsymbol{x}_i$,

q: a random variable which follows a Chi-square distribution

S^2: variance of response given a population of size N

σ^2: variance of the random error.

$\bar{y}_{obs}, \bar{x}_{obs}$: average of response and covariate of sample fully observed part

$\bar{x}_{com}$: average of covariate of complete sample.

β_i: regression coefficients. i= 0, 1, 2....

$\boldsymbol{\beta}$: $(\beta_1, \beta_2, ..)$. A vector of regression coefficients without the intercept

$\boldsymbol{\beta}^*$: $(\beta_0, \beta_1, \beta_2, ..)$. A vector of regression coefficients including the intercept

b_i: sample esitmate of $beta_i$, i=0, 1, 2,...

$\boldsymbol{X}^*_{com}$: complete design matrix including the intercept.

$\boldsymbol{X}^*_{obs}$: design matrix including the intercept of the fully observed part

$\boldsymbol{R_d}$: diagonal response matrix. $R_d = diag(r_{d1}, ..., r_{dn})$. $r_{di} = 1$ represents the corresponding y_i is observed

p_{mis}: the probablity that the i^{th} reponse is missing

p: dimension of $\boldsymbol{x}$

t_i: a binary indicator showing if a unit is assigned to treatment or control group

$\boldsymbol{t}$: a vector of $t + i$

$\boldsymbol{P}$: projection matrix of $\boldsymbol{X}$

$\boldsymbol{Z}$: $\boldsymbol{Z} = (\boldsymbol{t}, \boldsymbol{X})$

$\boldsymbol{Z}^*$: $\boldsymbol{Z}^* = (\mathbf{1}, \boldsymbol{Z})$

$D_{org} = (X_{org}, Y_{org})$: original dataset with one variable completely missing

$D_{val} = (X_{val}, Y_{val})$: Validation dataset with all variables observed

Multiple imputation related notations:

$\bar{Q}_m$: MI estimate of Q

$\bar{U}_m$: within-imputation variance

r_m: relative increase in the variance due to missing data

T_m: MI variance

B_m: between imputation variance

B, U: expectation of B_m and $\bar{U}_m$ given the sample is fixed

B_E, U_E, T_E: expectation of B_m, $\bar{U}_m$ and T_m given the population

r_E: expectation of r_m given the population

q_M: estimate of Q using nested MI

b_M, $\bar{w}_M$, $\bar{u}_M$: between-nest, within-nest, within-imputation variance of nested MI

T_M: variance estimate of nested MI

λ: overall rate of missing information

λ^{θ}: rate of missing information due to unknown model

$\lambda^{Z_{org}|\theta}$: rate of missing information due to unknown data if the model is given

Chapter 2

Power Calculation in Multiply Imputed Data

2.1 Introduction

In this chapter, we introduce formulas to calculate statistical power when MI is used with incomplete datasets. We focus on situations where the main focus is the hypothesis test of a scalar population parameter Q (such as population mean and regression coefficients).

As mentioned in Chapter 1, there is a lack of literature to date calculating statistical power when MI is used with incomplete data. In response to this gap, we derive the general power calculation equation using the population distribution of MI estimates and variances, given by Rubin [1987]. For this general equation to hold, the imputation must be proper [Rubin, 1987]. We then develop specific power calculation formulas for population mean when a linear relationship exists between response and covariates. We use simulations and data examples to demonstrate our findings.

The rest of the chapter is structured as follows: in Section 2.2, we provide a detailed introduction of the definition of proper MI and display the distribution of MI estimates and variances. We derive a general formula to calculate statistical power when MI is used and specify this formula under several different conditions. Section 2.3 includes a simulation study to demonstrate our method. This method is then applied to the Children Anxiety Prevention Study dataset [Ginsburg et al., 2015] in Section 2.4, and, finally, Section 2.5 contains a discussion.

2.2 Methodology

2.2.1 Conditions for Valid MI Inference

Let's use $\boldsymbol{I}$ to denote the sample inclusion indicator. It is a matrix of the same size as the population. The i^{th} row $\boldsymbol{I}_i = 1$ represents that unit i of the population is included in the sample. Rubin [1987] determined that the following distribution of $\bar{Q}_m$ holds:

$$\frac{\bar{Q}_m - Q}{\sqrt{T_m}} \sim t_\nu \tag{2.1}$$

with $\nu = (m-1)(\frac{T_m}{(1+\frac{1}{m})B_m})^2$ if, first, the posterior distribution of Q is normal and, second, approximately

$$\begin{aligned} \hat{Q}_i|\mathbf{X_{com}}, \mathbf{Y_{obs}}, \mathbf{R}, \mathbf{I} &\sim N(\bar{Q}_\infty, B_\infty) \\ U_i|\mathbf{X_{com}}, \mathbf{Y_{obs}}, \mathbf{R}, \mathbf{I} &\sim (U_\infty \ll B_\infty). \end{aligned} \tag{2.2}$$

Here $T_m = \frac{1}{m}B_m + \bar{U}_m$ is the MI variance; $(U_\infty \ll B_\infty)$ means the distribution is centered around U_∞ with variance substantially less than B_∞. U_∞, B_∞ and $\bar{Q}_\infty$ are the values of U_m, B_m and $\bar{Q}_m$ when $m \to \infty$.

2.2.1.1 MI Inference from a Randomization Perspective

The above MI inference is a Bayesian inference. However, traditional statistical survey methodologies and randomized trials refer more to the frequentist inference since they want to control both statistical power and significance level simultaneously. To make inference from a randomization perspective, it is necessary for us to understand the population distributions of $\bar{Q}_m$, U_m, and B_m. In other words, we want to know how these statistics behave if the population is treated as fixed and if the sampling and missing data generation are treated as random.

Rubin [1987] indicated that under mild conditions, MI statistics have the following distributions:

$$\begin{aligned} \bar{Q}_m|\mathbf{X},\mathbf{Y} &\sim N(Q, U_E + (1+\frac{1}{m})B_E), \\ \bar{U}_m|\mathbf{X},\mathbf{Y} &\sim (U_E, \ll U_E + (1+m^{-1})B_E), \\ ((m-1)\frac{B_m}{B_E}|\mathbf{X},\mathbf{Y}) = q &\sim \chi^2_{m-1}. \end{aligned} \tag{2.3}$$

Here, U_E, B_E and Q_E correspond to the expectations of $\bar{U}_m$, B_m, and $\bar{Q}_m$.
Rubin [1987] ref erred to this as randomization validity, which means that an MI inference is valid over repeated sampling and realization of the missing mechanism.

Rubin [1987] also stated two conditions for valid MI inference. First, the complete data inference needs to be randomization valid. This means that, with the absence of missing data, the complete data statistics $\hat{Q}$ and U should satisfy

$$\begin{aligned} \hat{Q}|\mathbf{X},\mathbf{Y} &\sim N(Q, U_E), \\ U|\mathbf{X},\mathbf{Y} &\sim (U_E, \ll U_E). \end{aligned} \tag{2.4}$$

Second, the imputation itself needs to be proper. This entails:

a) Treating the complete sample as fixed and, under the posited response mechanism, we must have the following distributions with an infinite number of imputations:

$$\begin{aligned} \bar{Q}_\infty|\mathbf{X},\mathbf{Y},\mathbf{I} &\sim N(\hat{Q},B), \\ \bar{U}_\infty|\mathbf{X},\mathbf{Y},\mathbf{I} &\sim (U,\ll B), \\ B_\infty|\mathbf{X},\mathbf{Y},\mathbf{I} &\sim (B,\ll B), \end{aligned} \tag{2.5}$$

with B defined as $var(Q_\infty|\mathbf{X},\mathbf{Y},\mathbf{R})$;

b) Treating the population as fixed,

$$B|\mathbf{X},\mathbf{Y} \sim (B_E,\ll U_E). \tag{2.6}$$

Rubin [1987] stated that if MI is constructed from a Bayesian framework, then the imputation process is more or less proper.

2.2.2 Power Calculation with MI: a General Formula

We are interested in calculating the statistical power of the following hypothesis test: $H_0 : Q = Q_0$ versus $H_a : Q \neq Q_0$. We assume that, under H_a, the value of Q is Q_1. For simplicity, we will only discuss two-sided cases. One-sided cases can be easily obtained by slightly adjusting the two-sided formula.

Based on the results in (2.3), along with the fact that rejecting the null hypothesis is the same as finding $Q_0 \notin C$ when C is the confidence interval, $(\bar{Q}_m - \sqrt{T_m}t_{\nu,\alpha/2}, \bar{Q}_m + \sqrt{T_m}t_{\nu,\alpha/2})$, a general formula to calculate statistical power in this case is:

$$P\left(\bar{Q}_m \notin C|\mathbf{X},\mathbf{Y},Q=Q_1\right) = P(z < \frac{Q_0-Q_1}{T_E^{1/2}} - \frac{(\frac{B_E q}{m-1}(1+\frac{1}{m})+U_E)^{1/2}}{T_E^{1/2}} t_{\nu,\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}) + P(z > \frac{Q_0-Q_1}{T_E^{1/2}} + \frac{(\frac{B_E q}{m-1}(1+\frac{1}{m})+U_E)^{1/2}}{T_E^{1/2}} t_{\nu,\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}), \tag{2.7}$$

with $T_E = (1+\frac{1}{m})B_E + U_E$.

If we define r_E, the ratio of (expected) adjusted between-imputation variance and within-imputation variance as follows:

$$r_E = \frac{(\frac{1}{m}+1)B_E}{U_E} \tag{2.8}$$

so that (2.7)'s new form would be:

$$P(\bar{Q}_m \notin C|\mathbf{X},\mathbf{Y},Q=Q_1) = P(z < \frac{Q_0-Q_1}{(U_E(1+r_E))^{1/2}} - \frac{(\frac{q}{m-1}r_E+1)^{1/2}}{(1+r_E)^{1/2}} t_{\nu,\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}) + P(z > \frac{Q_0-Q_1}{(U_E(1+r_E))^{1/2}} + \frac{(\frac{q}{m-1}r_E+1)^{1/2}}{(1+r_E)^{1/2}} t_{\nu,\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}), \tag{2.9}$$

with $\nu = (m-1)(1+\frac{1}{r_E})^2$.

When the number of imputations goes to infinity ($m \to \infty$), power asymptotically goes to:

$$P(\bar{Q}_m \notin C|\mathbf{X},\mathbf{Y},Q=Q_1) = P(z < \frac{Q_0-Q_1}{(U_E(1+r_E))^{1/2}} - z_{\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}) + P(z > \frac{Q_0-Q_1}{(U_E(1+r_E))^{1/2}} + z_{\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}). \tag{2.10}$$

In comparing Equation (2.10) with the complete data power calculation formula $P(z < \frac{Q_0-Q_1}{U_E^{1/2}} - z_{\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}) + P(z > \frac{Q_0-Q_1}{U_E^{1/2}} + z_{\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y})$, we can see that the power will converge to the complete data power (the ideal power, as we know power will decrease with non-responses) as r becomes smaller. In other words, MI can provide a satisfying statistical power if the fraction of missing information due to non-response is small. The formula also

implies that, when r is small, just a few imputations are enough to make improvements. If r is large, however, a larger m might be necessary to increase the difference in power between MI and CCA.

When calculating power in complete data, researchers have to specify several parameters: type I error rate (α), mean difference ($Q_0 - Q_1$), sample size, and variance estimate for mean difference. With incomplete data, a few additional parameters must be specified to describe the missing data process, in addition to the parameters that need to be specified in complete data. In particular, we need more information about m (the number of imputations) and r (the relative increase in variance due to missing data). See Equation (2.9).

In some situations, the value of r can be directly computed by using other population parameters, as will be shown in Section 2.2.3. When the missing mechanism is not monotone or when the model used is rather complicated, we may not be able to obtain an analytic result for r. In that case, we will need to estimate r using B_m and U_m based on their calculation from the MI procedure. Furthermore, a large number of imputations will be necessary, since the values of B_m and U_m are less stable when m is small [Schafer, 1997, Harel, 2007a].

2.2.3 Specific Power Calculation Formulas for Different Cases

In this section, we obtain several specific power calculation formulas for different cases, based on Equation (2.7) above. The key here is to find an explicit expression for B_E and

U_E, which are the population expectations of the between- and within-imputation variances. We focus on cases where the parameter of interest is the population mean, $Q = \bar{Y}$. We consider a sample of size n with n_1 complete cases. The remaining n_0 cases have fully observed $\boldsymbol{X}$, but are missing response $\boldsymbol{Y}$.

2.2.3.1 MCAR: One-Dimensional Covariates

When Q is the population mean and the hypothesis to test is H_0: $Q = Q_0$, the complete data estimate is $\hat{Q} = \bar{y}$, the sample mean. Therefore, $U_E = E(var(\bar{y} - \bar{Y})) = S_Y^2/n$ when S_Y^2 is the population variance of Y and n represents the sample size. Notice that $B_E = E(var(\bar{Q}_\infty|\mathbf{X}, \mathbf{Y}, \mathbf{I})|\mathbf{X}, \mathbf{Y})$. Consider the general normal linear regression model,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{2.11}$$

where $\epsilon_i \sim N(0, \sigma^2)$. Under a normal imputation method we find that:

$$\bar{Q}_\infty = \bar{y}_{obs} + b_1(\bar{x}_{com} - \bar{x}_{obs}). \tag{2.12}$$

Here, b_1 is the regression slope estimated from observed data and $\bar{x}_{obs}$ represents the mean of x corresponding to the observed responses.. When the population distribution of $(Y_i, X_i) \sim N(\mu_1, \mu_2, \mathbf{\Sigma})$, Equation (2.12) is simply the MLE estimate of μ_1 [Little and Rubin, 2002].

When, according to Hansen et al. [1953], the missing mechanism is MCAR, the variance of $\bar{Q}_\infty$ is approximately equal to the variance of $\bar{y}_{obs} + \boldsymbol{\beta_1}(\bar{x}_{com} - \bar{x}_{obs})$ when the number of complete cases is large enough. This gives us:

$$B_E \approx E(var(\bar{Q}_\infty|\mathbf{X}, \mathbf{Y}, \mathbf{I})|\mathbf{X}, \mathbf{Y}) = (\frac{1}{n_1} - \frac{1}{n})S_Y^2(1 - \rho^2), \tag{2.13}$$

with $\rho = \frac{S^2_{XY}}{S_X S_Y}$. Here S^2_{XY} is the population covariance between X and Y, and S^2_X is the population variance of X. Now, if the samples are drawn from an infinite population, we can easily see that:

$$B_E = (\frac{1}{n_1} - \frac{1}{n})\sigma^2 \tag{2.14}$$

Therefore, (2.9) can now be formulated as:

$$\begin{aligned} P(\bar{Q}_m \notin C|\mathbf{X},\mathbf{Y},Q=Q_1) = P(z < \frac{Q_0 - Q_1}{(S^2_Y(1/n - 1/N)(1+r_E))^{1/2}} - \frac{(\frac{q}{m-1}r_E + 1)^{1/2}}{(1+r_E)^{1/2}} t_{\nu,\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}) \\ + P(z > \frac{Q_0 - Q_1}{(S^2_Y(1/n - 1/N)(1+r_E))^{1/2}} + \frac{(\frac{q}{m-1}r_E + 1)^{1/2}}{(1+r_E)^{1/2}} t_{\nu,\frac{\alpha}{2}}|\mathbf{X},\mathbf{Y}) \end{aligned} \tag{2.15}$$

with $r_E = \frac{(1+1/m)(1/n_1 - 1/n)(1-\rho^2)}{1/n - 1/N}$.

2.2.3.2 MCAR: Multi-Dimensional Covariates

Consider the population model as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i \tag{2.16}$$

Take the coefficients with intercept as $\boldsymbol{\beta^*} = (\beta_0, ..., \beta_p)$, the coefficients without intercept as $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$, $\mathbf{Y} = (Y_1, ..., Y_N)'$, $\mathbf{X_l} = (X_{1l}, X_{2l}, ..., X_{Nl}), l = 1, 2, ..., p$, $\mathbf{X}^* = (\mathbf{1}, \mathbf{X_1}, .., \mathbf{X_p})$, and $\mathbf{X} = (\mathbf{X_1}, .., \mathbf{X_p})t$. Using the reasoning from the one-dimensional case above, we now find that $\bar{Q}_\infty = \bar{y}_{obs} + b_1(\bar{x}_{1,com} - \bar{x}_{1,obs}) + ... + b_p(\bar{x}_{p,com} - \bar{x}_{p,obs})$.

With the same hypothesis test as in Section 2.2.3.1, U_E remains the same as it was in 2.2.3.1. To obtain B_E, we follow the steps outlined in Hansen et al. [1953]. We show that Hansen's approximation still works if $\boldsymbol{X}$ is multi-dimensional instead of one-dimensional.

Therefore, when the number of complete cases is large enough, we have:

$$\begin{aligned} B_E \approx E(var(\bar{Q}_\infty|\mathbf{X},\mathbf{Y},\mathbf{I})|\mathbf{X},\mathbf{Y}) &= (\frac{1}{n_1}-\frac{1}{n})(var(\mathbf{Y})-2\beta^t Cov(\mathbf{X},\mathbf{Y}) \\ &\quad + \boldsymbol{\beta}^t cov(\mathbf{X})\boldsymbol{\beta}) \\ &= (\frac{1}{n_1}-\frac{1}{n})var(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{2.17}$$

When the population is infinite, we can still conclude that $B_E = (\frac{1}{n_1}-\frac{1}{n})\sigma^2$.

2.2.3.3 The MAR Case

Transitioning from an MCAR to an MAR structure entails a loss in the equivalence to random sampling, meaning that Hansen's approximation no longer works to estimate B_E since the missing mechanism is no longer equivalent to 'random sampling'. Instead of calculating B_E directly, we use the difference between T_E and U_E.

Rubin [1987] stated that $var(\bar{Q}_\infty|\mathbf{X},\mathbf{Y}) = T_E$. As we know, $\bar{Q}_\infty = \bar{y}_{obs} + b_1(\bar{x}_{1,com} - \bar{x}_{1,obs}) + ... + b_p(\bar{x}_{p,com} - \bar{x}_{p,obs})$. This can be represented in matrix form as $\frac{1}{n}\mathbf{1}'\mathbf{X}^*_{\mathbf{com}}\mathbf{b}^*$ which is equal to $\frac{1}{n}\mathbf{1}'\mathbf{X}^*_{\mathbf{com}}\mathbf{b}^*(\mathbf{X}^{*\prime}_{\mathbf{obs}}\mathbf{X}^*_{\mathbf{obs}})^{-1}\mathbf{X}^{*\prime}_{\mathbf{obs}}\mathbf{Y}_{\mathbf{obs}}$. Here, $\mathbf{X}^*_{\mathbf{com}}$ is the design matrix for the complete data; $\mathbf{X}^*_{\mathbf{obs}}$ and $\mathbf{Y}_{\mathbf{obs}}$ are the design matrix and response vector, respectively, of the fully observed data (data with both X and Y observed).

$\mathbf{X}^{*\prime}_{\mathbf{obs}}\mathbf{X}^*_{\mathbf{obs}}$ can also be rewritten as $\mathbf{X}^{*\prime}_{\mathbf{com}}\mathbf{R}_{\mathbf{d}}\mathbf{X}^*_{\mathbf{com}}$, with $\mathbf{R}_{\mathbf{d}}$ as a diagonal matrix $diag(r_{d1},...,r_{dn})$ with r_{di} zero or one, where zero represents that the corresponding Y_i is missing. Since the missing mechanism is MAR, $r_{di} \sim Bernoulli(P_i)$ with P_i only relating to $\boldsymbol{X}$.

Noticing that when the infinite population $(\mathbf{X}, \mathbf{Y})$ is given, ϵ_{com} and $\mathbf{X^*_{com}}$ can be viewed as being independently drawn from the population. Therefore,

$$\begin{aligned} var(\bar{Q}_\infty|\mathbf{X}^*, \mathbf{Y}) &= E(var(\bar{Q}_\infty(\mathbf{X^*_{com}}, \boldsymbol{\epsilon_{com}}))|\mathbf{X^*_{com}}, \mathbf{R}) \\ &\quad + var(E(\bar{Q}_\infty(\mathbf{X^*_{com}}, \boldsymbol{\epsilon_{com}}))|\mathbf{X^*_{com}}, \mathbf{R}) \\ &= \sigma^2 E(\frac{1}{n^2}\mathbf{1'X^*_{com}}(\mathbf{X^{*\prime}_{com}RX^*_{com}})^{-1}\mathbf{X^{*\prime}_{com}1}) \\ &\quad + \boldsymbol{\beta^{*\prime}} var(\frac{1}{n}\mathbf{1'X^*_{com}})\boldsymbol{\beta^*}. \end{aligned} \tag{2.18}$$

In Equation (2.18), with a single linear regression (SLR) model, the first and second terms are simply $\sigma^2 E(\frac{1}{n_1} + \frac{(\bar{x}_{obs}-\bar{x})^2}{n_1 s^2_{x_{obs}}})$ and $\frac{1}{n^2} var(\mathbf{X})\beta_1^2$. Here, $s^2_{x_{obs}}$ is the sample variance of the observed covariate. The sum of the two terms is T_E. B_E is the difference between U_E and T_E. U_E is easy to find based on previous work with MCAR cases.

Notice that, in the SLR situation, if we use the observed $(\mathbf{X^*_{com}}(\mathbf{X^{*\prime}_{com}R_d X^*_{com}})^{-1}\mathbf{X^{*\prime}_{com}})$ instead of the expectation and replace other parameters with their MLE estimator, the result will be the same as the large sample variance of $\mu_1 - \hat{\mu}_{1,MLE}$ given by Little and Rubin [2002]. Such replacement implies that we are now working within the naive sampling framework [Verbeke and Molenberghs, 2000]. In other words, the design matrix and missing mechanism are now considered fixed rather than random. Since statistical power is a frequentist property, it is rather important for us to consider the random sampling conditions.

There are situations where analytic results will be difficult to obtain. An alternative is to use B_m and U_m computed from MI as the point estimators of the true B_E and U_E. If the imputation method is proper and the complete-data inference is randomization-valid,

then B_m and U_m are unbiased estimates of B_E and U_E, and the variance can be neglected as m goes to ∞. Therefore, to perform this alternative procedure, we may need to perform additional imputations to guarantee that m is large enough in order to obtain more stable estimates.

2.3 Simulation Study

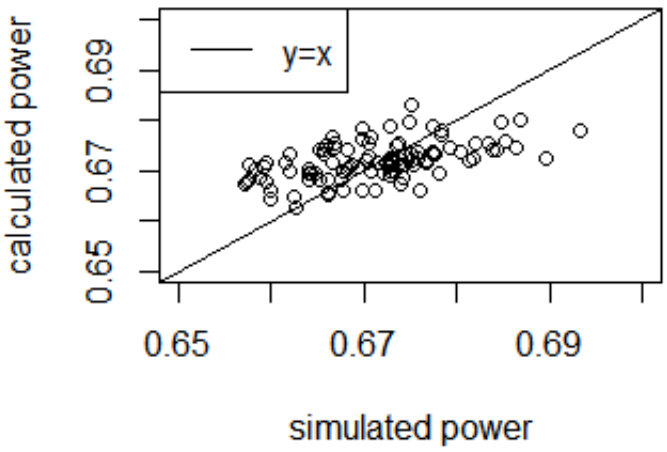
In the following section, we use the Monte Carlo simulation to evaluate our findings. The general power calculation in Equation (2.7) was used in all three simulations. Tpecific equations for MI variances in Section 2.2.3 are used depending on the simulation setups. The first simulation evaluated the general formula's bias (caused by the approximation). For simplicity, the simulation setup is MCAR with one-dimensional X. The second simulation compared the power of MI and CCA. Here, we wanted to find when MI effectively improves statistical power and when it fails. In Simulation 2, we have both MCAR (multi-dimensional) and MAR (one-dimensional) setups. Simulation 3 compared the two methods of estimating B_E and U_E.

2.3.1 Simulation 1

Starting with the simplest case, we considered MCAR with a one-dimensional $\mathbf{X}$. The null hypothesis was that the response's population mean Q is equal to 0 and the alternative hypothesis is $Q = 1$. We generated 100 populations of size 1,000,000 under this alternative hypothesis. For simplicity, we set $\mathbf{X} \sim N(0, 4^2)$, $\boldsymbol{\beta^*} = (1, 1)$, $\boldsymbol{\epsilon} \sim N(0, 1)$, $m = 10$, and the sample size as 100. The missing data proportion was set to 20%. Notice that the true parameters were used in the power calculation. A theoretical power was

calculated using Equation (2.15) for each population. Then, 5000 samples of size 100 were randomly drawn from each population, and 20% of responses from each sample were randomly set to be missing. We applied MI on each incomplete sample and tested whether the population's mean $\bar{Y}$ was equal to 0. The ratio between the number of cases rejecting the null hypothesis and the total number of cases (5000) was our Monte-Carlo simulated power. The results are shown in Figure 2.1.

Figure 2.1: Comparison between theoretical and simulated power under MCAR, one dimensional $\mathbf{X}$



In the simulation, $\mathbf{X} \sim N(0, 4^2)$, $\boldsymbol{\beta}^* = (1, 1)$, $\boldsymbol{\epsilon} \sim N(0, 1)$, and $m = 10$. Each point in the figure represents the theoretical power and the simulated power for one population. Ideally, the points should be clustered along the forty-five-degree line.

The average difference and average squared difference between the simulated power and calculated power are approximately -6.936×10^{-4} and 4.402×10^{-10}, respectively. In Figure 2.1, all points were approximately located around the y = x line, implying that for

each population, the theoretical power and the simulated power were very close to each other.

Next, we used a multi-dimensional covariate $\mathbf{X}$. We tested four different setups: a three-dimensional $\mathbf{X}$ with a 'large' variance-covariance matrix; a three-dimensional $\mathbf{X}$ with a 'small' variance-covariance matrix; a three-dimensional $\mathbf{X}$ with an increased sample size; and a two-dimensional $\mathbf{X}$ with a 'large' variance-covariance matrix. For each, the proportion of missing responses was set to 20%. The variance-covariance matrix was set to $\sigma^2(0.6\boldsymbol{I_n} + 0.4\boldsymbol{J_n})$. Here I_n is the $n \times n$ identity matrix, and J_n is the $n \times n$ matrix of ones.Table 2.1 summarizes the σ^2 and sample size we used for the different setups.

In Table 2.1, setups 1, 3, and 4 used $\sigma^2 = 4$, and setup 2 used $\sigma^2 = 1$. The sample size

Table 2.1: Average difference (diff) and average squared difference (MSE) between the simulated and theoretical powers

setup	σ^2	sample size	dimension of $(\boldsymbol{X^*})$	diff	MSE
1	4	100	4	-0.005	2.348e-05
2	1	100	4	-0.000	4.930e-07
3	4	200	4	0.004	2.024e-05
4	4	100	3	0.006	3.809e-05

was 100 in setups 1, 2, and 4, in setup 3 it was 200. We had the same null and alternative hypothesis as in the one-dimensional $\mathbf{X}$ case. Table 2.1 shows the average difference and the average squared difference between the theoretical and simulated powers.

Table 2.1 indicates that, on average, the difference between the theoretical power and the simulated power was smaller than 0.001.

We also plotted the simulated power versus the theoretical power for each setup (Figure 2.2). In Figure 2.2, we can see that the simulated power was close to the theoretical power. Although some points are not exactly on the 45° line in the figure, the average difference and MSE nevertheless indicate good fit. We believe this deviation from the 45° line is because our results are asymptotic, and they may perform better with larger sample size.

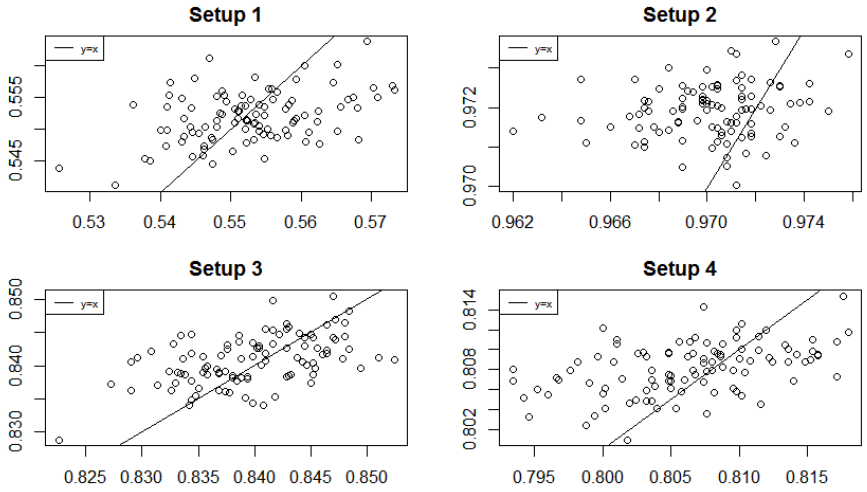


Figure 2.2: Missing Data mechanism MCAR, multi-dimensional $\mathbf{X}$

In all four setups, the x-axis is the simulated power and y-axis is the theoretical power. For setups 1, 3, 4, $\sigma^2 = 4$; for setup 2 it was 1. The sample size for setups 1, 2, 4 was 100 while it was 200 for setup 3. $\boldsymbol{\beta}^* = (1, 1, 1, 1)$ for setups 1, 2, and 3, while for setup 4 it was (1, 1, 1). $\boldsymbol{\epsilon} \sim N(0, 1)$, $m = 10$ for all setups. Each point represents the theoretical power and the simulated power for one population. Ideally, the points should be along the forty-five-degree line.

2.3.2 Simulation 2

Our first simulation confirmed that our formula was working well. In simulation 2 we applied both MAR and MCAR missing data mechanisms with different missing data proportions. We compared the statistical power obtained using MI and CCA with the ideal power (the power that could be obtained if there were no missing data). For simplicity,

we only reported the one-dimensional $\mathbf{X}$ case.

2.3.2.1 Simulation 2-1: MCAR

For the MCAR simulation, we again created a population using model (2.11) with $\boldsymbol{\beta}^* = (1, 1)$. The null hypothesis was that the population mean of the response is 0. To examine how r affects the computed statistical power, we chose two different population settings: large and small B_E. In both settings, the value of U_E was kept roughly unchanged ($var(\mathbf{Y}) = 20$ versus $var(\mathbf{Y}) = 19$). The variance of random error (which determines the value of B_E) was set to be 16 and 1 respectively.

We considered a 200-individual sample. We first calculated the ideal power with no missing data. Using our theoretical formula, we then calculated the statistical power when the dataset was incomplete and MI was used. The missing data proportion p was set to 0.1, 0.15, 0.2,....,0.8 and the number of imputations m was set to 3, 10, and ∞.

We randomly drew 5000 samples from the population and deleted $100p\%$ of the responses. We used both CCA and MI to analyze the incomplete dataset and considered the ratio between the number of successfully rejecting H_0 and the total number of Monte Carlo samples as the Monte Carlo simulated power. In the MI simulation, the m was set to 10, as suggested by Schafer [1999].

After completing the simulation, we had six different estimates: ideal power; theoretical power using MI with m = 3; theoretical power using MI with m = 10; theoretical power

using MI with m equal to ∞; simulated power using MI with $m = 10$; and simulated power using CCA. Of these, we did not evaluate the simulated powers with MI, with $m = 3$, or with m approaching ∞. We did however want to highlight how much power could be gained by increasing the number of imputations.

Figure 2.3 indicates that, when B_E is small, MI had a much better performance compared to CCA, and that increasing m did not make much of a difference. When B_E is large however and when p was greater than 0.3, even MI could not prevent the power and the ideal power from deviating. It is better in this case to have a larger m in this case since a small m caused a noticeable loss in power (when m was 3).

2.3.2.2 Simulation 2-2: MAR

For MAR, we repeated the simulation in Section 2.3.2.1, only changing the missing mechanism from MCAR to MAR. We generated missing data in the following way. For each sample, the data was separated into two parts according to the value of $\mathbf{X}$. The upper $100 \times p_1$ percent had $100 \times p_2$ percent chance to miss $\mathbf{Y}$, and the lower $100 \times (1 - p_1)$ had p_3 percent chance to miss $\mathbf{Y}$. We set p_1 as 0.5, p_3 as 0, and p_2 as (0.4, 0.6, 0.8). Overall, the missing proportion p was approximately (0.2, 0.3, 0.4). We did not try a higher p because the simulated power using CCA became too small as p increased. Again, we use the true missing mechanism for the expectation rather than using the observed value.

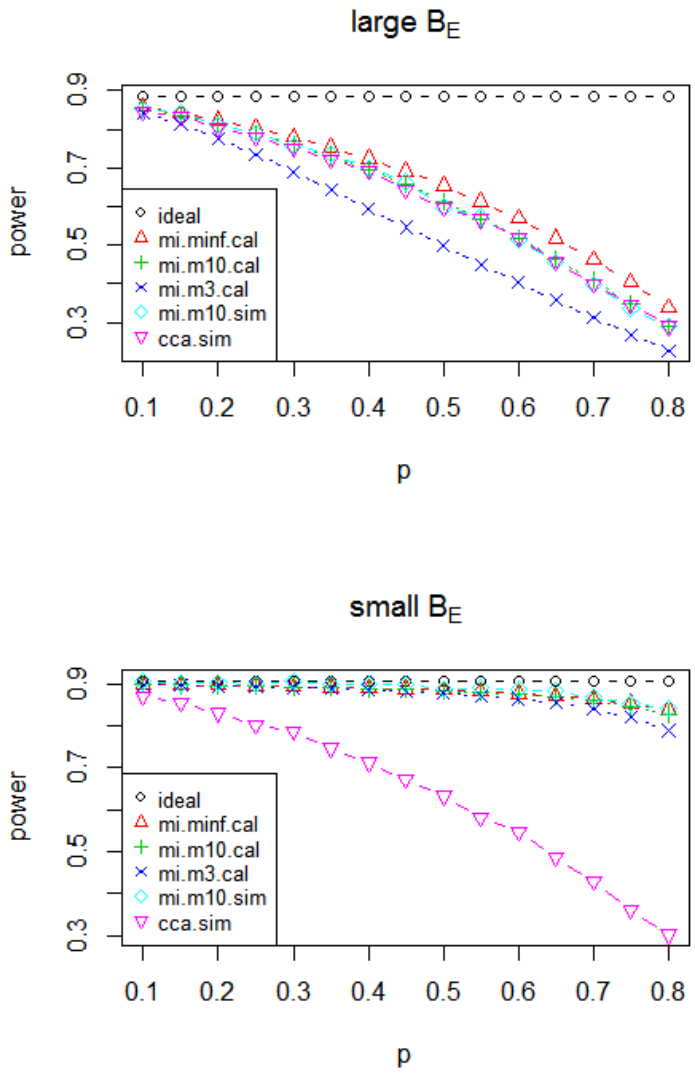


Figure 2.3: MCAR missing data mechanism with different percent of missing response

For the 'large B_E' setup, $\mathbf{X} \sim N(0, 2^2)$, $\boldsymbol{\epsilon} \sim N(0, 4^2)$. For the 'small B_E' setup $\mathbf{X} \sim N(0, 18)$, $\boldsymbol{\epsilon} \sim N(0, 1)$. The ideal power is the power calculated from the complete dataset; the mi.minf.cal is the power calculated using the general formula with $m = \infty$; mi.m10.cal and mi.m3.cal are also powers calculated using the general formula with $m = 10$ and $m = 3$ correspondingly; mi.m10.sim is the power calculated from the Monte Carlo simulation, using MI with $m = 10$; cca.sim is the power calculated from the Monte Carlo simulation using complete case analysis.

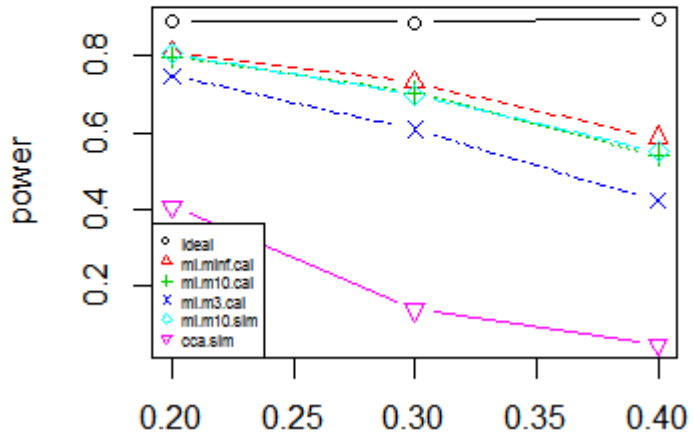


Figure 2.4: MAR missing data mechanism with different percent of missing response $\mathbf{X} \sim N(0, 2^2)$, $\boldsymbol{\epsilon} \sim N(0, 4^2)$. As we did in the simulation for MCAR, we calculated the theoretical MI power when m was 3, 10, and ∞. We only simulated the power for m as 10.

Figure 2.4 indicates that, when the missing mechanism is not MCAR, CCA provides a much smaller statistical power compared with the ideal case and the MI method. Again, we could see that since this was a 'large B_E' setup, we could improve the statistical power by increasing the number of imputations.

2.4 Simulation 3

The Child Anxiety Prevention Study sought to examine the efficacy of family-based intervention in preventing children with parents diagnosed with anxiety from developing anxiety disorders. It is a longitudinal study with 136 subjects enrolled, 70 of which were assigned to the treatment (CAPS program) group, with the rest assigned to the control group. Anxiety levels were measured by the total Anxiety Disorder Interview Schedule

score (total ADIS) at four different time points: pre-randomization, post-intervention, six months after evaluation, and one year after evaluation. Covariates include treatment group, parental anxiety, age, gender, family income, etc.

The dataset, as is the case with many other longitudinal studies, is incomplete. Its missing-data structure is as follows. Except for one subject, all incomplete cases are missing response rather than covariates. Among the 59 incomplete cases, only 23 are part of a monotone missing pattern. We used the R-package VIM [Templ and Filzmoser, 2008] to show the missing pattern of the data (see Figure 2.5). As such, general missing data structure is assumed.

Our main purpose is to examine whether receiving family-based intervention helps with anxiety disorders. In other words, test to see if if the treatment has a statistically significant effect on the response variable (total ADIS score). A random intercept model is used to analyze the outcomes over four time periods. The analysis model is therefore a linear mixed regression model as in Equation (2.19), and the parameter of interest is the regression coefficient of treatment group β_{tx}.

$$y_{ij} = b_{0i} + \beta_{tx} tx + \beta_{time} time_j + \epsilon_{ij}. \tag{2.19}$$

In Equation (2.19), y_{ij} represents the response of the i^{th} subject measured at time point j. j = 1, 2, 3, 4, representing the four different time points: prior to randomization (pr), post-intervention (po), 6 months after target post-evaluation (f6), and 12 months after target post-evaluation.(1yr). b_{0i} is the random intercept which varies among subjects.

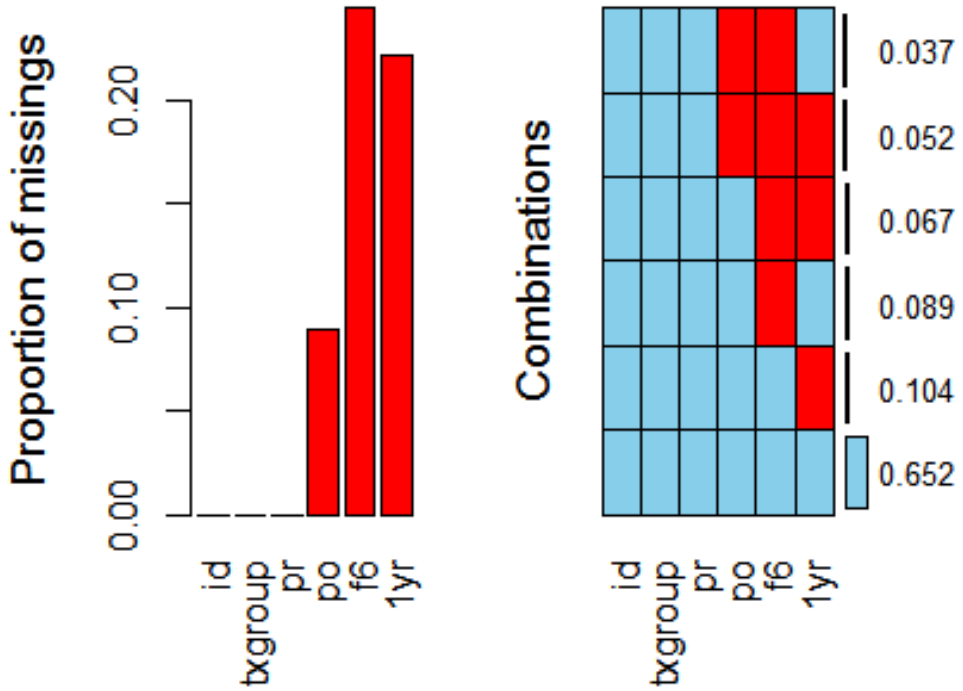


Figure 2.5: Proportion of missing values in total ADIS of CAPS data

Proportion of missing values in covariates and total ADIS measured at four different time points: pre-randomization (pr), post-intervention (po), six months after evaluation (f6), and one year after evaluation (1yr). "Txgroup" is the treatment group indicator. "Tx" represents treatment. Left plot: proportions of missing values in each variable. Proportion of missingness in 1yr, f6, and po are approximately 0.22, 0.24, and 0.09. Right plot: proportion of each combination of observed (light) and unobserved (dark) values.

β_{tx} and β_{time} are the fixed effects of treatment and time.

We use MI to deal with the incomplete data. The imputation is done using the R-package MICE [van Buuren and Groothuis-Oudshoorn, 2011]. Similar to the analysis model, the imputation model also includes time and treatment group as covariates. The purpose of using the chained equation method and including the same covariates is to make sure that our imputation model is at least as general as the analysis model [Meng, 1994, Collins et al., 2001]. We tried different numbers of imputations (m) to see when

stable results could be achieved. Both the predictive mean matching (PMM) method and the Bayesian linear regression method (norm) were used for the multiple imputation procedure. Details of the two methods can be found in Rubin [1987] and van Burren [2012].

After imputation, a linear mixed model was applied to each completed dataset. The final results are pooled using Rubin's rule and are shown in Table 2.2. Herein, we only calculate statistical power when the significance level is 0.05 and the differences between β_{tx}^0 (the value of β_{tx} under the null hypothesis) and β_{tx}^1 (the value of β_{tx}, d, under the alternative hypothesis) are 1, 1.5, and 2.

Table 2.2: Power for the test of β_{tx}

method	m	B_m	$\bar{U}_m$	r		power	
					d=1	d=1.5	d=2
pmm	50	0.019	0.353	0.054	0.373	0.692	0.906
	100	0.018	0.350	0.051	0.379	0.698	0.909
	500	0.017	0.350	0.048	0.375	0.697	0.911
	1000	0.018	0.352	0.051	0.376	0.695	0.908
norm	50	0.015	0.376	0.04	0.362	0.670	0.893
	100	0.016	0.377	0.042	0.354	0.666	0.890
	500	0.018	0.378	0.048	0.356	0.665	0.888
	1000	0.018	0.378	0.047	0.355	0.666	0.889

This table shows the values of B_m,U_m, r, and statistical power calculated for the test of β_{tx}. Four different m (50, 100, 500 and 1000), two different imputation methods (pmm and norm), and three different effect sizes (d) were used for the imputation and to calculate power for each combination.

We can see from Table 2.2 that the power value changes along with the number of imputations. Fluctuation decreases as the number of imputations increases. When m was increased from 500 to 1000, the statistical power for the PMM method only decreased by about 0.001 (d = 1), 0.002 (d = 1.5), and 0.003 (d = 2), and the statistical power of the norm method only changed by about 0.001 for all three d. We therefore decided

that $m = 1000$ is sufficiently large for this dataset. Even though the data is not normally distributed, the results from PMM and from norm are similar to each other. The PMM method yields a slightly higher power. Such difference is predictable. PMM generates imputation using observed values, which should lead to lower variability.

2.5 Discussion and Conclusion

In this chapter, we studied the statistical power of a hypothesis test if MI is used for incomplete datasets. We obtained a general formula and extended it to several particular cases. We found that the amount of power that can be gained by using MI instead of list-wise deletion is determined by the ratio of the between-imputation variance to the within-imputation variance and by the number of imputations. The formula for r is provided when the model is linear and Q is the population average. In a linear regression setting, it is proportional to ρ^2_{XY}. This result coincides with the suggestion made by van Buuren and Groothuis-Oudshoorn [2011] that: "Such predictors help to reduce the uncertainty of the imputations. They are crudely identified by their correlation with the target variable."

Knowing r and m helps us understand how much power we can gain by using MI. It quantifies the power increase and may help show whether we should use MI and, if we do, how many imputations should be enough. With an exact formula for r in linear model cases, we can also build a relationship between power and sample size as we do for complete data analysis. Since MI helps improve statistical power, the required sample size

can be lower than the sample size we estimate using CCA.

Although we found the value of r when the model is linear, this becomes rather complicated with a more advanced model or even for a more complex Q. We might be able to approximate this value since it is a function of γ, which, according to Rubin [1987], can be roughly considered as the fraction of missing observations. However, since covariates are almost always involved in models and the reduction is not clear, how this approximation affects power remains to be examined. The value of r can also be estimated using the B_m and U_m, as computed from the MI procedure. When such estimation is utilized, we may need to pay extra attention to whether variance estimates are stable. We would want the number of imputations, m, to be large enough to ensure that the variability in the estimates will be small enough not to severely affect the calculation of power [Harel, 2007a].

The power calculation procedure described in this chapter is based on two main conditions. First, that the MI procedure is proper. Second, that the large sample distribution of Q is approximately normal. While imputations generated within a Bayesian framework are usually proper, whether those nonparametric methods such as PMM are also proper is yet to be examined. Given the advantage of methods such as PMM, it would be interesting to see how such imputation methods can fit into the picture. On the other hand, when the parameter of interest is clearly nonnormal, for example a variance or an F-test statistic, new inference may need to be constructed.

Chapter 3

Power Calculation with Multiply Imputed Data: Testing the Slope of a Binary Indicator

3.1 Introduction

In the previous chapter, we developed a general method for calculating the statistical power of a scalar hypothesis when MI is used to deal with incomplete data. We also specified the method to a one-sample t-test of population average. In this chapter, we extend these results to focus on a two-sample student's t-test.

Two-sample student's t-tests, which require equal variances across two populations, may be the most commonly used statistical test for two-sample comparison. In psychology, in fact, it is the default method for comparing two groups [Delacre et al., 2017]. A survey by Ruxton [2006] looked at 33 papers in the journal Behavioral Ecology (Volume 16, issue 1-5) that conducted statistical tests to compare the central tendencies of two groups. He found that the two-sample student's t-test (67 occasions; 26 papers) is the most popular two-sample comparison method, followed by the Mann-Whitney U test (43

occasions; 21 papers) and the t-test for unequal variance (Welch's t-test; 9 occasions; 4 papers).

Although the Welch's t-test for two sample comparison was a recommended method, it does not require variance homogeneity and as such we have decided that the student's t-test is of more interest to us. Regarding the former, the British Medical Journal (BMJ) makes the following suggestions: "However, it (Welch's t-test) should not be used indiscriminately because, if the standard deviations are different, how can we interpret a nonsignificant difference in means, for example? Often a better strategy is to try a data transformation, such as taking logarithms [...]. Transformations that render distributions closer to Normality often also make the standard deviations similar." [British Medical Journal] In addition, if the variance heterogeneity issue exists, it will be of less concern if the two samples are of equal size.

Since a two-sample student's t-test yields exactly the same result as a t-test for the slope of a binary indicator within the regression framework, we use the regression framework throughout the rest of the chapter.

In Chapter 2, we showed that when MI is used for incomplete data, statistical power is primarily determined by the difference between null and alternative hypotheses, the number of imputations, the significance level, and the MI variances, B_E and U_E. These last values, the MI variances, are of most interest to us when calculating our statistical power, not only because they contain important information about the sample characteristics and the severity of messiness, but also because obtaining these variances is usually

not so straightforward.

One way to estimate the values of B_E and U_E is to perform MI, attain B_m and U_m, and use them to estimate their expectations. This method is generally suitable for post-hoc power calculation since a dataset is required on which we can perform MI and run the analysis. However, such a dataset is not available at the experiment design stage. If the goal of power calculation is to guide the experiment design (i.e., power must be calculated before a dataset is collected), we need a different way to estimate the MI variances.

As such, a second way to calculate the values of B_E and U_E is by using MI-related theories given by Rubin [1987]. Notice that according to Equations (2.4), (2.5), and (2.6), $U_E = var(\hat{Q})$ is the variance of the complete data estimate, and $B_E = E(V(\bar{Q}_\infty|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{I})|\boldsymbol{X}, \boldsymbol{Y})$ is only related to $\bar{Q}_\infty$. Therefore, the calculation of B_E and U_E can be separated into two steps. First we obtain $\bar{Q}_\infty$ and $\hat{Q}$. Then we can calculate U_E and B_E accordingly.

In Chapter 2, we followed the above procedure to calculate B_E and U_E for a one-sample t-test with auxiliary variables. We showed that if (1) the MI is created from Bayesian linear regression, and (2) the population has a distribution as follows:

$$y_i|\boldsymbol{x_i} \sim N(\boldsymbol{x_i^*}\boldsymbol{\beta}^*, \sigma^2)$$

$$\boldsymbol{x_i} \sim (\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x}) \ \ i = 1, 2, 3, ..., n,$$

then the values of B_E, T_E and U_E corresponding to $\hat{Q} = \bar{y}$ can be calculated using the following equations:

$$\begin{aligned} U_E &= S_Y^2/n \\ &= \frac{\sigma^2}{n} + \frac{\boldsymbol{\beta^{*\prime}\Sigma_x\beta^*}}{n}, \\ T_E &= \sigma^2 E(\frac{1}{n^2}\mathbf{1}'\mathbf{X^*_{com}}(\mathbf{X^{*\prime}_{obs}X^*_{obs}})^{-1}\mathbf{X^{*\prime}_{com}1}) \\ &\quad + \boldsymbol{\beta^{*\prime}} var(\frac{1}{n}\mathbf{1}'\mathbf{X^*_{com}})\boldsymbol{\beta^*}, \\ B_E &= T_E - U_E \\ &= \sigma^2 E(\frac{1}{n^2}\mathbf{1}'\mathbf{X^*_{com}}(\mathbf{X^{*\prime}_{obs}X^*_{obs}})^{-1}\mathbf{X^{*\prime}_{com}1}) - \frac{\sigma^2}{n}. \end{aligned} \tag{3.1}$$

Here, n represents the sample size, $\boldsymbol{X}$ represents the design matrix of random variable X, a variable used only for the imputation procedure, and σ^2 and β represent the regression residual error and coefficient when the response Y on X is regressed.

The method itself is "direct" in that it is directly derived from the MI inference which we use to calculate the statistical power. However, it may sometimes be difficult to obtain $\bar{Q}_\infty$. As such, in this chapter we propose another way to calculate U_E and B_E which is based on the maximum likelihood method. We use this maximum likelihood based method to obtain a closed-form equation for statistical power for a two-sample student's t-test. We also propose a method to calculate sample size using the power-related results.

This chapter is organized as follows. In Section 3.2, we provide a general method to calculate B_E and U_E using the variances of the maximum likelihood (ML) and the complete case analysis (CCA) estimates. We apply the ML-based method to the one sample t-test. We compare the equations derived from the ML-based method with Equations (3.1) to

show that the ML-based method leads to similar results. Since we have previously shown by simulation that Equations (3.1) are effective for power calculation, we are able to prove that the ML-based method is valid. In Section 3.3, we apply the method to a two-sample t-test. We provide different formulas for the MAR and MCAR missing mechanisms, and present a simulation study in Section 3.4 to investigate the performance of our proposed method and to show that it holds up well under different simulation settings. Section 3.5 provides a methodology for sample size calculation and power tables using a required power of 0.8. We selected several specific setups from the table to examine the table's validity.

3.2 Calculating Multiple Imputation Variances

3.2.1 A Method Based on the Maximum Likelihood Estimate

Here, we provide a general method to calculate B_E and U_E based on the maximum likelihood estimator (MLE). The method is mainly derived from Nielsen [2003a]'s MI results of MI variance. As such, most of the notations remain similar to those in the original paper.

Let's use Q to denote the parameter of interest, and $I_c(Q)$ to denote the expected Fisher information of the complete data, $I_o(Q)$ to denote the expected Fisher information of the observed data. Remember that complete data represents the dataset we would have seen if there were no missing values. Nielsen [2003a] demonstrated that the MI variance can be expressed using $I_c(Q)$ and $I_o(Q)$ as in Equation (3.2), if the following three conditions hold: (1) the sample is large; (2) the MI is proper; and (3) the estimates

are MLE.

$$var(\bar{Q}_m) = I_o(Q)^{-1} + \frac{1}{m}(I_o(Q)^{-1} - I_c(Q)^{-1}). \tag{3.2}$$

By rearranging Equation (3.2), we obtain:

$$var(\bar{Q}_m) = I_c(Q)^{-1} + (1 + \frac{1}{m})(I_o(Q)^{-1} - I_c(Q)^{-1}). \tag{3.3}$$

At the same time, according to Equation (2.3), the variance of MI estimate $\bar{Q}_m$ given the whole population can be expressed by the following equation:

$$var(\bar{Q}_m)|\mathbf{X}, \mathbf{Y} = U_E + (1 + \frac{1}{m})B_E \tag{3.4}$$

Notice that U_E the complete data variance estimate. Due to that the estimate method is MLE, with a large sample we should have:

$$U_E = I_c(Q)^{-1}. \tag{3.5}$$

Combining Equations (3.3), (3.4), and (3.6) leads to:

$$B_E = I_o(Q)^{-1} - I_c(Q)^{-1}. \tag{3.6}$$

Finally, the variance of the MLE estimator of Q is asymptotically the information matrix of Q, therefore the following formula can be used to approximately calculate the MI variances:

$$\begin{aligned} U_E &= var(\hat{Q}^{MLE_c}), \\ B_E &= var(\hat{Q}^{MLE_o}) - var(\hat{Q}^{MLE_c}). \end{aligned} \tag{3.7}$$

Here, $var(\hat{Q}^{MLE_c})$ is the variance of the MLE computed from complete data, and $var(\hat{Q}^{MLE_o})$ is the variance of the MLE computed from the observed data.

3.2.2 Method Validation

In this section, we use the ML-based method in Section 3.2 to recalculate B_E and U_E for a one-sample t-test. We show that the ML-based results lead to the same results as in Equations (3.1). Previous simulations show that Equations (3.1) performed well for power calculation, therefore we verify the validity of the ML-based method.

Let's assume our population follows this distribution:

$$y_i|\boldsymbol{x_i} \sim N(\boldsymbol{x_i^*}\boldsymbol{\beta}^*, \sigma^2),$$

$$\boldsymbol{x_i} \sim N(\boldsymbol{\mu_x}, \Sigma_x) \;\; i = 1, 2, 3, ..., n.$$

Here, y represents the response variable, $\boldsymbol{\beta}^* = (\beta_0, \beta_1, .., \beta_p)$ is the regression coefficient vector with intercept, covariate $\boldsymbol{x}$ is a p-dimensional random vector, and $\boldsymbol{x_i^{*\prime}} = (1, \boldsymbol{x_i'})$. Let $\boldsymbol{X^*_{com}} = (\mathbf{1}, (\boldsymbol{x_1}, .., \boldsymbol{x_n})')$ be the sample covariate matrix and $\boldsymbol{y}_{com} = (y_1, ..., y_n)'$ be the sample response vector. Then, $(\boldsymbol{X_{com}}, \boldsymbol{y_{com}})$ denotes the complete data (the sample we would have observed if there were no missing entries). Assume that the missingness occurs in the response variable. $\boldsymbol{y_{obs}} = (y_{i1}, .., y_{in_1})$ represents the n_1 observed responses, $\boldsymbol{X_{obs}} = (\boldsymbol{x}_{i1}, .., \boldsymbol{x}_{in_1})$ denotes their corresponding covariates. We use $(\boldsymbol{y_{obs}}, \boldsymbol{X_{obs}})$ to denote the fully observed part (FO) of the data (i.e, all cases with both response and covariates observed) and $(\boldsymbol{y_{obs}}, \boldsymbol{X_{com}})$ to denote the observed part of the data (i.e., all cases with either response or covariates observed).

Consider a one sample t-test with $H_0 : \mu_y = \mu_0$ and $H_1 : \mu_y \neq \mu_0$. Apparently, $\mu_y = \boldsymbol{\mu}_{\boldsymbol{x}}^{*'}\boldsymbol{\beta}^*$, with $\boldsymbol{\mu}_{\boldsymbol{x}}^{*\prime} = (1, \boldsymbol{\mu}_{\boldsymbol{x}}')$. To estimate the corresponding B_E and U_E, we now derive the variance of $\hat{\boldsymbol{\mu}}_{\boldsymbol{y}}^{\boldsymbol{MLE}}$ estimated from the complete data and the observed data. Notice that due to the invariance principle of MLE, we have $\hat{\mu_y}^{MLE} = \hat{\boldsymbol{\beta}}^{*MLE}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE}$. Therefore, to obtain the variance of $\hat{\mu_y}^{MLE}$ is to obtain the variance of $\hat{\boldsymbol{\beta}}^{*MLE}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE}$.

We first derive the variance of the MLE obtained from the complete dataset. With knowledge of multivariate distribution, we have:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{MLE_c} = \bar{\boldsymbol{X}}^*_{com} = \frac{1}{n}\mathbf{1}'\boldsymbol{X}^*_{\boldsymbol{com}}$$

$$\hat{\boldsymbol{\beta}}^{*MLE_c} = (\boldsymbol{X}^{*\prime}_{\boldsymbol{com}}\boldsymbol{X}^*_{\boldsymbol{com}})^{-1}\boldsymbol{X}^{*\prime}_{\boldsymbol{com}}\boldsymbol{y}_{\boldsymbol{com}}.$$

Therefore

$$var(\hat{\boldsymbol{\beta}}^{*MLE_c}|\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c}) = (\boldsymbol{X}^{*\prime}_{\boldsymbol{com}}\boldsymbol{X}^*_{\boldsymbol{com}})^{-1}\sigma^2$$

$$E(\hat{\boldsymbol{\beta}}^{*MLE_c}|\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c}) = \boldsymbol{\beta}^*$$

and

$$\begin{aligned}
var(\hat{\boldsymbol{\beta}}^{*\prime MLE_c}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c}) &= E(var(\hat{\boldsymbol{\beta}}^{*MLE_c}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c})|\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c}) + var(E(\hat{\boldsymbol{\beta}}^{*MLE_c}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c})|\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c}) \\
&= E(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c})(\boldsymbol{X}^{*\prime}_{\boldsymbol{com}}\boldsymbol{X}^*_{\boldsymbol{com}})^{-1}\sigma^2\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c})) + \boldsymbol{\beta}^{*\prime}var(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_c}))\boldsymbol{\beta}^* \\
&= \sigma^2\frac{\mathbf{1}'\boldsymbol{P}_{\boldsymbol{com}}\mathbf{1}}{n^2} + \frac{\boldsymbol{\beta}^{*\prime}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}^*}{n} \\
&= \frac{\sigma^2}{n} + \frac{\boldsymbol{\beta}^{*\prime}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}^*}{n} \\
&= var(\bar{\boldsymbol{Y}}_{\boldsymbol{com}}).
\end{aligned} \tag{3.8}$$

Here, $\boldsymbol{P}_{\boldsymbol{com}}$ is the projection matrix corresponding to $\boldsymbol{X}^*_{\boldsymbol{com}}$.

Next we move to the variance of the MLE of the observed dataset. Notice that the density function of (x, y) is

$$f(\boldsymbol{x}, y) = f(y|\boldsymbol{x}, \boldsymbol{\beta}^*, \sigma^2)f(\boldsymbol{x}|\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}).$$

Since we assume that the missingness only occurs in the response variable y, with MDM ignorable, the ignorable log likelihood function of the observed data is

$$l = \sum_{i=1}^{n} ln(f(\boldsymbol{x}_i|\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x})) + \sum_{i \in FO} ln(f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta^*}, \sigma^2)). \tag{3.9}$$

Using the idea behind the factored likelihood method [Little and Rubin, 2002], we have

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{MLE_o} = \bar{\boldsymbol{X}}^*_{com} = \frac{1}{n}\mathbf{1}'\boldsymbol{X}^*_{com}$$

$$\hat{\boldsymbol{\beta}}^{*MLE_o} = (\boldsymbol{X^{*\prime}_{obs}X^*_{obs}})^{-1}\boldsymbol{X^{*\prime}_{obs}y_{obs}}$$

Following the same logic of Equation (3.8),

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}^{*MLE_o}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_o}) &= E(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_o}(\boldsymbol{X^{*\prime}_{obs}X^*_{obs}})^{-1}\sigma^2\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_o}) + \boldsymbol{\beta}^{*\prime}var(\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^{*MLE_o})\boldsymbol{\beta}^* \\ &= \sigma^2\frac{1}{n^2}E(\mathbf{1'X^*_{com}}(\boldsymbol{X^{*\prime}_{obs}X^*_{obs}})^{-1}\boldsymbol{X^*_{com}}\mathbf{1}) + \frac{\boldsymbol{\beta^{*\prime}\Sigma_x\beta^*}}{n}. \end{aligned} \tag{3.10}$$

Combining Equation (3.8) and Equation (3.10) leads to

$$\begin{aligned} U_E &= var(\hat{Q}^{MLE_c}) \\ &= \frac{\sigma^2}{n} + \frac{\boldsymbol{\beta^{*\prime}\Sigma_x\beta^*}}{n} \\ &= var(\bar{\boldsymbol{Y}}_{\boldsymbol{com}}) \\ B_E &= var(\hat{Q}^{MLE_o}) - var(\hat{Q}^{MLE_c}) \\ &= \sigma^2\frac{1}{n^2}E(\mathbf{1'X^*_{com}}(\boldsymbol{X^{*\prime}_{obs}X^*_{obs}})^{-1}\boldsymbol{X^*_{com}}\mathbf{1}) - \sigma^2\frac{1}{n} \end{aligned}$$

which is exactly the same as in Equation 3.1. We have thus shown that the ML-based method is valid for calculating B_E and U_E.

Another way to calculate the variance of $\mu_y^{MLE_o}$ is to use the Delta method, which leads to the following result:

$$\begin{aligned} var(\mu_y^{MLE_o}) &= (\boldsymbol{\mu_x^*}, \boldsymbol{\beta^*})' cov(\hat{\boldsymbol{\beta}^*}^{MLE_o}, \hat{\boldsymbol{\mu}_{\boldsymbol{x}}^*}^{MLE_o})(\boldsymbol{\mu_x^*}, \boldsymbol{\beta^*}) \\ &= E(\boldsymbol{\mu_x^{*\prime}}(\boldsymbol{X_{obs}^{*\prime}}\boldsymbol{X_{obs}^*})^{-1}\boldsymbol{\mu_x^*})\sigma^2 + var(\boldsymbol{\beta^{*\prime}}\boldsymbol{X_{com}^{*\prime}}\mathbf{1}/n). \end{aligned} \tag{3.11}$$

This should differ slightly from the result obtained in Equation 3.1, since it uses $\boldsymbol{\mu_x}$ instead of $\mathbf{1'}\boldsymbol{X_{com}^*}/n$

3.3 Calculating Statistical Power for a Two-Sample Student's T-Test

3.3.1 Calculating U and B Without Specifying the MDM

In this section, we derive the power calculation function for a two sample test.

Consider a sample of size n with three variables, $(y, t, \boldsymbol{x})$. y is the response of interest; $\boldsymbol{x}$ is an additional information source used only for imputation, but not for analysis; t is the group indicator. In a randomized controlled trial, t can be the indicator of whether an experiment unit is assigned to the control or the treatment group. The i^{th} observation in the sample has the following i.i.d distribution (i = 1, 2, .., n):

$$\begin{aligned} y_i|t_i, \boldsymbol{x_i} &\sim N(\boldsymbol{x_i}\boldsymbol{\alpha_2'} + \alpha_1 t_i + \alpha_0, \sigma^2_{y|\boldsymbol{x},t}) \\ \boldsymbol{x_i}|t_i &\sim N(\boldsymbol{\gamma_0} + \boldsymbol{\gamma_1} t_i, \boldsymbol{\Sigma_{x|t)}} \\ t_i &\sim Bernoulli(p_t) \end{aligned} \tag{3.12}$$

The covariate $\boldsymbol{x_i}$ corresponding to unit i, $\boldsymbol{x_i} = (x_{i1}, ..., x_{ip})'$, is a p-dimensional vector. Let's use $\boldsymbol{t}$ to represent $(t_1, ..., t_n)'$, $\boldsymbol{X}$ to represent $(\boldsymbol{x_1}, ..., \boldsymbol{x_n})'$, $\boldsymbol{Z}$ to represent $(\boldsymbol{t}, \boldsymbol{X})$ (the design matrix without intercept), and $\boldsymbol{Z^*}$ to represent $(\mathbf{1}, \boldsymbol{t}, \boldsymbol{X})$ (the design matrix with

intercept).

Say we are interested in comparing the difference between $E(y_i|t_i = 1)$ and $E(y_i|t_i = 0)$. Since $E(y_i|t_i = 1) = \alpha_0 + \alpha_1 + (\boldsymbol{\gamma_0} + \boldsymbol{\gamma_1})'\boldsymbol{\alpha_2}$, $E(y_i|t_i = 0) = \alpha_0 + \boldsymbol{\gamma_0'}\boldsymbol{\alpha_2}$, the difference we are estimating can be expressed as $Q = \alpha_1 + \gamma_1'\boldsymbol{\alpha_2}$.

We now build on our reasoning from Section 3.2 to develop the corresponding U_E and B_E. Of course, the MLE estimate of Q with complete data is

$$\hat{Q}^{MLE_c} = \hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}_{\mathbf{1}}'^{MLE_c}\hat{\boldsymbol{\alpha}}_2.$$

Let's use $\boldsymbol{\alpha}' = (\alpha_0, \alpha_1, \boldsymbol{\alpha_2'})$ to represent covariate coefficients for y given $\boldsymbol{x}$ and t, and $\boldsymbol{\gamma} = (\boldsymbol{\gamma_0}, \boldsymbol{\gamma_1})$ represent the covariate coefficients of $\boldsymbol{x}$ given t. Recall that $\boldsymbol{t}' = (t_1, t_2, ..., t_n)$, $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n})'$, and $\boldsymbol{Z^*} = (\mathbf{1}, \boldsymbol{t}, \boldsymbol{X})$. We also use $\boldsymbol{T^*}$ to denote $(\mathbf{1}, \boldsymbol{t})$. $\boldsymbol{Z^*_{com}}$, $\boldsymbol{T^*_{com}}$ are the $\boldsymbol{Z^*}$, $\boldsymbol{T^*}$ of the whole dataset. $\boldsymbol{Z^*_{obs}}$ is the $\boldsymbol{Z^*}$ of the fully observed part of the dataset.

For the complete dataset, using multivariate linear regression we obtain:

$$\hat{\boldsymbol{\gamma}}^{MLE_c} = (\boldsymbol{T^*_{com}}'\boldsymbol{T^*_{com}})^{-1}\boldsymbol{T^*_{com}}'\boldsymbol{X_{com}},$$

$$\hat{\boldsymbol{\alpha}}^{MLE_c} = (\boldsymbol{Z^*_{com}}'\boldsymbol{Z^*_{com}})^{-1}\boldsymbol{Z^*_{com}}'\boldsymbol{y_{com}}.$$

For the observed data, if the MDM is MAR, then, using the factored likelihood method [Little and Rubin, 2002], the MLE estimate of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}}^{MLE_o} = (\boldsymbol{Z^*_{obs}}'\boldsymbol{Z^*_{obs}})^{-1}\boldsymbol{Z^*_{obs}}'\boldsymbol{y_{obs}},$$

and the estimate of $\boldsymbol{\gamma}$ remains the same. Therefore,

$$\begin{aligned}
var(\hat{Q}^{MLE_c}) &= E(var(\hat{\alpha_1}^{MLE_c} + \hat{\boldsymbol{\gamma_1}}'^{MLE_c}\hat{\boldsymbol{\alpha_2}}^{MLE_c}|\boldsymbol{x},t)) + var(E(\hat{\alpha_1}^{MLE_c} + \hat{\boldsymbol{\gamma_1}}'^{MLE_c}\hat{\boldsymbol{\alpha_2}}^{MLE_c}|\boldsymbol{x},t)) \\
&= \sigma^2_{y|\boldsymbol{x},t} E((0,1,\hat{\boldsymbol{\gamma_1}}'^{MLE_c})(\boldsymbol{Z^{*\prime}_{com}Z^*_{com}})^{-1}(0,1,\hat{\boldsymbol{\gamma_1}}'^{MLE_c})') \\
&+ (\mathbf{0},\boldsymbol{\alpha}'_{\mathbf{2}})\boldsymbol{\Sigma_{x|t}} \otimes E((\boldsymbol{T^{*\prime}_{com}T^*_{com}})^{-1})(\mathbf{0},\boldsymbol{\alpha}'_{\mathbf{2}})', \\
var(\hat{Q}^{MLE_o}) &= E(var(\hat{\alpha_1}^{MLE_o} + \hat{\boldsymbol{\gamma_1}}'^{MLE_c}\hat{\boldsymbol{\alpha_2}}^{MLE_o}|\boldsymbol{x},t)) + var(E(\hat{\alpha_1}^{MLE_o} + \hat{\boldsymbol{\gamma_1}}'^{MLE_c}\hat{\boldsymbol{\alpha_2}}|\boldsymbol{x},t)) \\
&= \sigma^2_{y|\boldsymbol{x},t} E((0,1,\hat{\boldsymbol{\gamma_1}}'^{MLE_c})(\boldsymbol{Z^{*\prime}_{obs}Z^*_{obs}})^{-1}(0,1,\hat{\boldsymbol{\gamma_1}}'^{MLE_c})') \\
&+ (\mathbf{0},\boldsymbol{\alpha}'_{\mathbf{2}})\boldsymbol{\Sigma_{x|t}} \otimes E((\boldsymbol{T^{*\prime}_{com}T^*_{com}})^{-1})(\mathbf{0},\boldsymbol{\alpha}'_{\mathbf{2}})'.
\end{aligned} \tag{3.13}$$

Furthermore, the MLE of Q, the difference between $E(y_i|t_i=1)$ and $E(y_i|t_i=0)$ can also be formulated as $\bar{y}_{t=1} - \bar{y}_{t=0}$, Therefore, if n_1 of the n observations have $t=1$ and n_2 of them have $t=0$, we have:

$$var(\hat{Q}^{MLE_c}) = (1/n_1 + 1/n_2)(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha}'_{\mathbf{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}). \tag{3.14}$$

From Equation (3.13) and (3.14) we find that:

$$\begin{aligned}
B_E &= var_c\hat{Q}^{MLE} - var_o\hat{Q}^{MLE} \\
&= \sigma^2_{y|\boldsymbol{x},t} E((0,1,\hat{\boldsymbol{\gamma_1}}'^{MLE_c})[(\boldsymbol{Z^{*\prime}_{com}Z^*_{com}})^{-1} - (\boldsymbol{Z^{*\prime}_{obs}Z^*_{obs}})^{-1}](0,1,\hat{\boldsymbol{\gamma_1}}'^{MLE_c})'), \\
U_E &= var(\hat{Q}^{MLE_c}) \\
&= (1/n_1 + 1/n_2)(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha}'_{\mathbf{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}).
\end{aligned} \tag{3.15}$$

Finally, if we instead use the Delta method to derive the variances of the complete and observed data MLEs, we have:

$$\begin{aligned}
B_E &= \sigma^2_{y|\boldsymbol{x},t}(0,1,\boldsymbol{\gamma}'_{\mathbf{1}})E[(\boldsymbol{Z^{*\prime}_{com}Z^*_{com}})^{-1} - (\boldsymbol{Z^{*\prime}_{obs}Z^*_{obs}})^{-1}](0,1,\boldsymbol{\gamma}'_{\mathbf{1}})') \\
U_E &= (1/n_1 + 1/n_2)(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha}'_{\mathbf{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}.
\end{aligned} \tag{3.16}$$

Replacing B_E and U_E in Equation(2.7) with the values from Equations (3.15) or (3.16), we are able to obtain the power calculation formula for a two sample student's

t-test. When the MDM is MAR, it might be difficult to obtain a closed-form result for $E[(\boldsymbol{Z}^{*\prime}_{com}\boldsymbol{Z}^{*}_{com})^{-1} - (\boldsymbol{Z}^{*\prime}_{obs}\boldsymbol{Z}^{*}_{obs})^{-1}]$. In this case, we suggest performing a Monte Carlo simulation to estimate the expectation needed. If we have information about the MDM and the population distribution (that is, parameter values, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$), U_E and B_E can be precisely estimated. This information may be obtained from a pilot study or from past research records. If such information is not available, researchers may need to make different assumptions and report a sensitivity analysis. When the MDM is MCAR, we provide a closed-form result which can estimate the $E[(\boldsymbol{Z}^{*\prime}_{com}\boldsymbol{Z}^{*}_{com})^{-1} - (\boldsymbol{Z}^{*\prime}_{obs}\boldsymbol{Z}^{*}_{obs})^{-1}]$ without using a Monte-Carlo simulation. The details of this are presented in the next section.

3.3.2 Calculating Power when the MDM is MCAR

Equation (3.16) can be further simplified, if the missing data mechanism (MDM) is MCAR.

From Equation (3.16), we have $B_E = \sigma^2_{y|\boldsymbol{x},t}(0, 1, \boldsymbol{\gamma}'_{\mathbf{1}})E([(\boldsymbol{Z}^{*\prime}_{com}\boldsymbol{Z}^{*}_{com})^{-1} - (\boldsymbol{Z}^{*\prime}_{obs}\boldsymbol{Z}^{*}_{obs})^{-1}])(0, 1, \boldsymbol{\gamma}'_{\mathbf{1}})'$.

As mentioned in Section 3.3.1, $\boldsymbol{Z}^{*} = (\mathbf{1}, \boldsymbol{t}, \boldsymbol{X})$. $\boldsymbol{Z} = (\boldsymbol{t}, \boldsymbol{X})$, then $\boldsymbol{Z}^{*} = (\mathbf{1}, \boldsymbol{Z})$.

Let

$$\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^{*} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

$$(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^{*})^{-1} = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{G} & \boldsymbol{H} \end{bmatrix}$$

Here, $\boldsymbol{A} = n$, $\boldsymbol{B} = \mathbf{1}'\boldsymbol{Z}$, $\boldsymbol{C} = \boldsymbol{Z}'\mathbf{1}$, $\boldsymbol{D} = \boldsymbol{Z}'\boldsymbol{Z}$. Using our knowledge of block matrices and inverse block matrices, $\boldsymbol{H} = (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} = (\boldsymbol{Z}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Z})^{-1}$. Therefore, $(0, 1, \boldsymbol{\gamma_1'})'E((\boldsymbol{Z^{*\prime}Z^*})^{-1})(0, 1, \boldsymbol{\gamma_1'}) = (1, \boldsymbol{\gamma_1'})E((\boldsymbol{Z}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Z})^{-1})(1, \boldsymbol{\gamma_1'})'$.

To compute $E((\boldsymbol{Z}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Z})^{-1})$, we present the following lemma.

Lemma 1 *[] Let $n_1 = \sum_{i=1}^{n} I(t_i = 1)$, $n_2 = \sum_{i=1}^{n} I(t_i = 0)$. G_1 denotes the subjects whose $t_i = 1$, and G_2 denotes the subjects whose $t_i = 0$. Let $\boldsymbol{X_1}$ be an $n_1 \times p$ sub matrix of $\boldsymbol{X}$, with each row representing an $x_i \in G_1$, and $\boldsymbol{X_2}$ be an $n_2 \times p$ sub matrix of $\boldsymbol{X}$, with each row representing an $x_i \in G_2$. $\bar{\boldsymbol{x}}_\mathbf{1}$, a p dimensional vector, denotes the average of $\boldsymbol{x_i} \in G_1$; $\bar{\boldsymbol{x}}_\mathbf{2}$ denotes the average of $\boldsymbol{x_i} \in G_2$. $\Delta_{\boldsymbol{x}} = \bar{\boldsymbol{x}}_\mathbf{1} - \bar{\boldsymbol{x}}_\mathbf{2}$ denotes the difference between $\bar{\boldsymbol{x}}_\mathbf{1}$ and $\bar{\boldsymbol{x}}_\mathbf{2}$. Let $\boldsymbol{S}_1 = \sum_{i \in G_1}(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\mathbf{1})(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\mathbf{1})'$ be the sample covariance matrix among subjects belonging to G_1, and $\boldsymbol{S}_2 = \sum_{i \in G_2}(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\mathbf{2})(\boldsymbol{x_i} - \bar{\boldsymbol{x}}_\mathbf{2})'$ be the sample covariance matrix among subject belonging to G_2. $\boldsymbol{S_x} = \boldsymbol{S}_1 + \boldsymbol{S}_2$. Then we have:*

$$(\boldsymbol{Z}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Z})^{-1} = \begin{bmatrix} \frac{1}{n_1} + \frac{1}{n_2} + \Delta_{\boldsymbol{x}}'\boldsymbol{S_x}^{-1}\Delta_{\boldsymbol{x}} & -\Delta_{\boldsymbol{x}}'\boldsymbol{S_x}^{-1} \\ -\boldsymbol{S_x}^{-1}\Delta_{\boldsymbol{x}} & \boldsymbol{S_x}^{-1} \end{bmatrix} \tag{3.17}$$

The proof of Lemma 1 can be found in the appendix.

We now derive the expectation of $(\boldsymbol{Z}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Z})^{-1}$. We first compute $E((\boldsymbol{Z}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Z})^{-1}|\boldsymbol{t})$, then we compute the overall expectation without fixing $\boldsymbol{t}$.

Within each group G_g (g = 1, 2), $\boldsymbol{x_i}$ follows an i.i.d multivariate normal distribution. $\bar{\boldsymbol{x}}_{\boldsymbol{g}}|\boldsymbol{t}$ therefore follows a multivariate normal distribution

$$\bar{\boldsymbol{x}}_{\boldsymbol{g}}|\boldsymbol{t} \sim N(\boldsymbol{\gamma_1}, \frac{1}{n_g}\boldsymbol{\Sigma_x})$$

$\boldsymbol{S_g}|\boldsymbol{t}$ follows a p-dimensional Wishart distribution,

$$\bar{\boldsymbol{S}}_{\boldsymbol{g}}|\boldsymbol{t} \sim W(n_g - 1, \boldsymbol{\Sigma_x}).$$

Because of the multivariate normality, $\bar{\boldsymbol{x}}_{\boldsymbol{g}}|\boldsymbol{t}$ and $\boldsymbol{S_g}|\boldsymbol{t}$ are independent. The sampling procedure implies that the two groups are independent. Therefore, given $\boldsymbol{t}$, $\boldsymbol{S_1}$, $\boldsymbol{S_2}$, $\bar{\boldsymbol{x}}_{\boldsymbol{1}}$, and $\bar{\boldsymbol{x}}_{\boldsymbol{2}}$ are independent of each other. The above information leads us to:

$$\Delta_{\boldsymbol{x}} \sim N(\boldsymbol{\gamma_1}, (\frac{1}{n_1} + \frac{1}{n_2})\boldsymbol{\Sigma_x})$$

$$\boldsymbol{S_x}|\boldsymbol{t} \sim W(n_1 + n_2 - 2, \boldsymbol{\Sigma_x})$$

$\Delta_{\boldsymbol{x}}$ and $\boldsymbol{S_x}|\boldsymbol{t}$ are independent; $E(\boldsymbol{S_x^{-1}}|\boldsymbol{t}) = \frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}$, recall that p is the dimension of $\boldsymbol{x}_i$.

Since $\Delta_{\boldsymbol{x}}$ and $\boldsymbol{S_x}$ are independent of each other, we have

$$\begin{aligned}
E(\Delta'_{\boldsymbol{x}}\boldsymbol{S_x^{-1}}) &= E(\Delta'_{\boldsymbol{x}})E(\boldsymbol{S_x^{-1}}) \\
&= \boldsymbol{\gamma'_1}\frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}; \\
E(\boldsymbol{S_x^{-1}}\Delta_{\boldsymbol{x}}) &= E(\boldsymbol{S_x^{-1}})E(\Delta_{\boldsymbol{x}}) \\
&= \frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}\boldsymbol{\gamma_1}; \\
E(\Delta'_{\boldsymbol{x}}\boldsymbol{S_x^{-1}}\Delta_{\boldsymbol{x}}) &= E((\Delta'_{\boldsymbol{x}}\boldsymbol{S_x^{-1}}\Delta_{\boldsymbol{x}})|\Delta_{\boldsymbol{x}})) \\
&= E(\Delta'_{\boldsymbol{x}}\frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}\Delta_{\boldsymbol{x}}).
\end{aligned}$$

Using our knowledge of the expectation of quadratic forms,

$$E(\Delta'_{\boldsymbol{x}}\frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}\Delta_{\boldsymbol{x}}|\boldsymbol{t}) = \boldsymbol{\gamma'_1}\frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}\boldsymbol{\gamma_1} + (\frac{1}{n_1} + \frac{1}{n_2})\frac{p}{n-p-3}.$$

With sample size n sufficiently large, $n_1 \approx np_t$, $n_1 \approx n(1-p_t)$. Recall that p_t is the probability that $t_i = 1$. Finally, we have:

$$E((\boldsymbol{Z}'(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J})\boldsymbol{Z})^{-1}) \approx \begin{bmatrix} \boldsymbol{\gamma_1'}\frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}\boldsymbol{\gamma_1} + (\frac{1}{np_t}+\frac{1}{n(1-p_t)})(1+\frac{p}{n-p-3}) & -\boldsymbol{\gamma_1'}\frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3} \\ -\frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3}\boldsymbol{\gamma_1} & \frac{\boldsymbol{\Sigma_x^{-1}}}{n-p-3} \end{bmatrix} \tag{3.18}$$

From Equation (3.18), we find that:

$$\begin{aligned} (0,1,\boldsymbol{\gamma_1'})E((\boldsymbol{Z_{com}^{*\prime}Z_{com}^{*}})^{-1})(0,1,\boldsymbol{\gamma_1'})' &= (\frac{1}{n_1}+\frac{1}{n_2})(\frac{p}{n-p-3}+1) \\ &\approx \frac{1}{np_t(1-p_t)}(\frac{p}{n-p-3}+1). \end{aligned} \tag{3.19}$$

Since the MDM is MCAR, the fully observed part can be viewed as a smaller sample of size n_o, n_o being the number of fully observed individuals. Following this same logic, we have

$$(0,1,\boldsymbol{\gamma_1'})E((\boldsymbol{Z_{obs}^{*\prime}Z_{obs}^{*}})^{-1})(0,1,\boldsymbol{\gamma_1'})' \approx \frac{1}{n_op_t(1-p_t)}(\frac{p}{n_o-p-3}+1) \tag{3.20}$$

.

With Equations (3.19), (3.20), and (3.16), we have:

$$\begin{aligned} U_E &= \frac{1}{np_t(1-p_t)}(\sigma^2_{y|\boldsymbol{x},t}+\alpha_2'\Sigma_{\boldsymbol{x}|t}\alpha_2), \\ B_E &= \sigma^2_{y|\boldsymbol{x},t}(\frac{1}{n_o}\frac{n_o-3}{n_o-p-3}-\frac{1}{n}\frac{n-3}{n-p-3})\frac{1}{p_t(1-p_t)}. \end{aligned} \tag{3.21}$$

With Equation (3.21), we are able to approximately calculate the values of B_E and U_E using only the population parameters, the sample size, and the number of fully observed responses. We therefore do not need to perform a Monte Carlo simulation to calculate these two values, which results in an a easier power calculation process.

Notice that in the calculation, we use $1/E(n_1)$ to approximate $E(1/n_1)$. A small sample size could lead to an upward bias in the calculated statistical power.

3.4 Simulation Study: MCAR

In Section 3.3.2, we propose a method to calculate B_E and U_E which minimizes computational cost. In this section, we provide Monte Carlo simulation studies to obtain empirical statistical power under different setups. We then calculate the statistical power using the proposed method from Section 3.3.2 and compare the Monte Carlo simulated power with our estimated statistical power to evaluate the performance of the proposed method.

The simulated power is obtained through the following steps: (1) Generate a large number of incomplete samples under a specified simulation setup; (2) Apply MI to each of the sample, test whether the parameter of interest is equal to zero, and record the test result. A P-value smaller than α is considered a correct rejection; and (3) Calculate the simulated power as the ratio of correct rejections to the to the total number of Monte Carlo samples.

The data is generated using the distribution described in Equation (3.12). For our simulation, we set $(\boldsymbol{\gamma_0}, \boldsymbol{\gamma_1}) = (0, 0.5)$, and $(\alpha_0, \alpha_1, \boldsymbol{\alpha_2}) = (1, 1, 1)$. p_t is the probability that $t = 1$ is 0.5. $\sigma^2_{y|\boldsymbol{x},t} = 4$, $\Sigma_{\boldsymbol{x}|t} = 2$. The MDM is set to be MCAR. The response is incomplete and the covariates are all observed. The proportion of missing values in the response variable (y) is set to 20%, 40%, and 60%. This setup of parameters and missing

data proportion is mainly to simplify the situation, and to guarantee an MI power that is neither too high (=1) nor too low. The number of imputations, m, is set to 50, 100, and 200.

The main goal of this simulation is to test if the treatment has a significant effect, i.e., if $E(y|t = 1) = E(y|t = 0)$. In other words, we are interested in whether $\alpha_1 + \gamma_1 \boldsymbol{\alpha_2'} = 0$. The difference between the null and alternative hypotheses is therefore $d = Q_1 - Q_0 = \alpha_1 + \gamma_1 \boldsymbol{\alpha_2'} - 0 = 1.5$. The significance level, α, is set to 0.05. The values of B_E and U_E are calculated from Equation (3.21). Replacing m, d, B_E, U_E, and significance level α in Equation (2.7) with the actual values according to our setup, we are able to calculate the theoretical MI power (according to our method). To evaluate the performance of our method, we compare the theoretical power with the simulated power, as well as with the simulated power using CCA instead of MI. The results are presented in the following table.

Table 3.1: Statistical Power Comparison: $\rho(y, x) = 0.7$

		MI		CCA		complete	
		Calculated	Simulated	Calculated	Simulated	Calculated	Simulated
$r = 0.2$	m = 50	0.9386	0.9347	0.9152	0.9089	0.9618	0.9581
	m = 100	0.9386	0.9361	0.9152	0.9089	0.9618	0.9581
	m = 200	0.9388	0.9356	0.9152	0.9089	0.9618	0.9581
$r = 0.4$	m = 50	0.8932	0.8864	0.8214	0.8120	0.9618	0.9581
	m = 100	0.8944	0.8880	0.8214	0.8120	0.9618	0.9581
	m = 200	0.8949	0.8880	0.8214	0.8120	0.9618	0.9581
$r = 0.6$	m = 50	0.7928	0.7727	0.6489	0.6435	0.9618	0.9581
	m = 100	0.7963	0.7793	0.6489	0.6435	0.9618	0.9581
	m = 200	0.7979	0.7795	0.6489	0.6435	0.9618	0.9581
$r = 0.8$	m = 50	0.5487	0.5143	0.3727	0.3738	0.9618	0.9581
	m = 100	0.5553	0.5218	0.3727	0.3738	0.9618	0.9581
	m = 200	0.5594	0.5223	0.3727	0.3738	0.9618	0.9581

In Table 3.1, we can see that our method slightly overestimates the power, with a difference smaller than 0.01 when the proportion of missing responses is small to moderate (0.2 and 0.4), and with a difference smaller than 0.04 when the proportion of missing responses is large (0.6 and 0.8). This shows that our method estimates statistical power relatively well when MI is used for incomplete data. The overestimation could come from that we ignored the variance coming from n_1 and n_2. However, according to the simulation result, it doesn't affect the result seriously.

In addition, the simulation results show that, with an extra information source for imputation (x), MI clearly recovers more statistical power than does CCA, which in turn lowers experiment costs by requiring less samples. For example, when the proportion of missing responses is 0.4, MI can achieve a statistical power of about 0.9 with a sample size of 200; CCA on the other hand would require about 250 samples to achieve the same goal.

3.5 Sample Size Calculation when MDM is MCAR

Cohen [1988] indicated that statistical power is determined by effect size, significance level, and sample size. Effect size is a unit-free measurement that describes how large the effect to be detected is. An advantage of using effect size is that it combines multiple population characteristics into a single value that is meaningful for power calculation. For example, for a two-sample student t-test, while the population characteristics needed include population means, μ_A, μ_B, and their common standard deviation σ, what we really

need is the effect size $\frac{\mu_A - \mu_B}{\sigma}$ that we really need. In this case, it is easier for researchers who are less familiar with the original power calculation formula to simply present the power/sample size table.

In this section, we put Equation (2.7) into the framework proposed by Cohen [1988]. We show that, for incomplete data, statistical power is still determined by effect size, sample size, and significance level. In addition, two other values describing population characteristics are required: the percent of missing responses and the multiple correlation between the response y and the auxiliary imputation variable $\boldsymbol{x}$. We present sample size calculation tables to better guide researchers using our findings, and display simulation results for several selected setups.

3.5.1 Sample-Size Tables

We first obtain the power calculation formula included in Cohen [1988]'s framework. Chapter 2 provides a second form for calculating statistical power using U_E and $r_E = \frac{(\frac{1}{m}+1)B_E}{U_E}$, which is the ratio of adjusted between-imputation and within-imputation variances.

$$\begin{aligned} P(\bar{Q}_m \notin C | alternative) = P(z < & \frac{d}{(U_E(1+r_E))^{1/2}} - \frac{(\frac{q}{m-1}r_E + 1)^{1/2}}{(1+r_E)^{1/2}} t_{\nu,\frac{\alpha}{2}} | \mathbf{X}, \mathbf{Y}) \\ + P(z > & \frac{d}{(U_E(1+r_E))^{1/2}} + \frac{(\frac{q}{m-1}r_E + 1)^{1/2}}{(1+r_E)^{1/2}} t_{\nu,\frac{\alpha}{2}} | \mathbf{X}, \mathbf{Y}). \end{aligned} \tag{3.22}$$

Here d is the difference between Q_0 and Q_1. Notice that for the two-sample student's t-test described in Chapter 3, we have:

$$U_E = \frac{1}{np_t(1-p_t)}(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}),$$
$$B_E = \sigma^2_{y|\boldsymbol{x},t}(\frac{1}{n_o}\frac{n_o-3}{n_o-p-3} - \frac{1}{n}\frac{n-3}{n-p-3})\frac{1}{p_t(1-p_t)}.$$

Therefore,

$$\begin{aligned} r &= \frac{\sigma^2_{y|\boldsymbol{x},t}(\frac{1}{n_o}\frac{n_o-3}{n_o-p-3} - \frac{1}{n}\frac{n-3}{n-p-3})\frac{1}{p_t(1-p_t)}}{\frac{1}{np_t(1-p_t)}(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2})}(1+\frac{1}{m}) \\ &= \frac{\sigma^2_{y|\boldsymbol{x},t}}{(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2})}(\frac{1}{p_{obs}}\frac{n_o-3}{n_o-p-3} - \frac{n-3}{n-p-3})(1+\frac{1}{m}). \end{aligned} \tag{3.23}$$

Since $y = \alpha_0 + \alpha_1 t + \boldsymbol{x'\alpha_2} + \epsilon$, the multiple covariance matrix between y and $\boldsymbol{x}$ given t is:

$$cov(\boldsymbol{x}, y) = \begin{bmatrix} \sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}) & \boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t} \\ \Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}) & \Sigma_{\boldsymbol{x}|t} \end{bmatrix}.$$

Therefore the squared conditional multiple correlation coefficient is

$$\begin{aligned} \rho^2_{\boldsymbol{x}y|t} &= \frac{\boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t}\Sigma^{-1}_{\boldsymbol{x}|t}\Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}}{\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}} \\ &= 1 - \frac{\sigma^2_{y|\boldsymbol{x},t}}{(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha_2'}\Sigma_{\boldsymbol{x}|t}\boldsymbol{\alpha_2})}. \end{aligned}$$

And we have

$$r = (1-\rho^2_{\boldsymbol{x}y|t})f(n, p_{mis}, p, m). \tag{3.24}$$

Here $f(n, p_{mis}, p)$ is a function of the sample size n, the percentage of missing response $p_{mis} = \frac{n_1}{n}$, the number of imputations m, and p, the dimension of $\boldsymbol{x}$.

In addition, with Equation (3.16), we have

$$\begin{aligned} U_E &= (1/n_1 + 1/n_2)(\sigma^2_{y|\boldsymbol{x},t} + \boldsymbol{\alpha_2'}\boldsymbol{\Sigma}_{\boldsymbol{x}|t}\boldsymbol{\alpha_2}) \\ &= (1/n_1 + 1/n_2)\sigma^2_{y|t} \end{aligned}$$

so that

$$\frac{d}{U_E^{1/2}} = \frac{\delta}{(1/n_1 + 1/n_2)^{1/2}}. \tag{3.25}$$

Here δ is Cohen [1988]'s effect size for a two-sample student t-test.

With Equations (3.22), (3.24) and (3.25), we have

$$\begin{aligned}
power &= p(z < \frac{d}{(\sigma^2_{y|t}(\frac{1}{n_1}+\frac{1}{n_2})(1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis}))^{1/2}} - \frac{\frac{q}{m-1}((1-\rho_{\boldsymbol{x}y})f(n,p_{mis})+1)}{1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis})} \\
&+ p(z > \frac{d}{(\sigma^2_{y|t}(\frac{1}{n_1}+\frac{1}{n_2})(1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis}))^{1/2}} + \frac{\frac{q}{m-1}((1-\rho_{\boldsymbol{x}y})f(n,p_{mis})+1)}{1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis})} \\
&= p(z < \frac{\delta}{((\frac{1}{n_1}+\frac{1}{n_2})(1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis}))^{1/2}} - \frac{\frac{q}{m-1}((1-\rho_{\boldsymbol{x}y})f(n,p_{mis})+1)}{1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis})} \\
&+ p(z > \frac{\delta}{((\frac{1}{n_1}+\frac{1}{n_2})(1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis}))^{1/2}} + \frac{\frac{q}{m-1}((1-\rho_{\boldsymbol{x}y})f(n,p_{mis})+1)}{1+(1-\rho^2_{\boldsymbol{x}y})f(n,p_{mis})} \\
&= f_{power}(n_1, n_2, m, \rho_{\boldsymbol{x}y}, \delta, p, p_{mis}, \alpha).
\end{aligned} \tag{3.26}$$

Recall that n is the total number of individuals in both samples, n_1 and n_2 are the sizes of the two samples, and $n = n_1 + n_2$. n_o is the total number of observed individuals. q is a random variable which follows a χ^2 distribution with a degree of freedom equal to $m-1$.

Equation (3.26) implies that if the data is MCAR and MI is used as the missing data technique, statistical power is determined by Cohen's distance δ, group sizes n_1 and n_2, the percentage of missing responses p_{mis}, the dimension of $\boldsymbol{x}$ p, the number of imputations m, and $\rho_{\boldsymbol{x}y}$, and the multiple correlation between $\boldsymbol{x}$ and y. Notice that the effect of p can be neglected if $p \ll n_o$, and the effect of m becomes stable when m is large enough. Therefore, the statistical power primarily depends on the size of the effect a researcher seeks to detect (the effect size), the amount of missing data (the missing response percentage), the strength of the relationship between the response and the auxiliary imputation variable $\boldsymbol{x}$,

and the sample sizes. This relationship can be reversed to calculate the required sample size(s) for a pre-determined statistical power. For example, with all necessary information, the required sample size for a power of 0.8 is the root of

$$f_{power}(n_1, n_2, m, \rho_{\boldsymbol{x}y}, \delta, p, p_{mis}, \alpha) = 0.8.$$

While there is no closed-form solution, modern statistical software and packages can help us obtain the root of the above equation.

Next we present several sample size tables. For simplicity, we only present tables here for situations where (1) the two group are the same size, i.e., ($n_1 = n_2 = \frac{n}{2}$); and (2) the number of imputations is 50. These tables can however be easily extended to other situations. More results can be found in the appendix.

Roughly speaking, with less missing responses and with $\boldsymbol{x}$ and y being stronger related, the required sample size decreases. In addition, the stronger the relationship between $\boldsymbol{x}$ and y, the more information that can be recovered using MI as compared to CCA. MI's advantage becomes more pronounced as the percent of missing responses increases. When $\rho^2_{\boldsymbol{x}y|t} = 0.9$ and $p_{mis} = 0.6$, MI requires only about half the sample size as that needed for CCA.

3.5.2 Simulation Results for Selected Setups

In this section, we choose several setups from Table 3.2 to examine whether our table works. For each $\delta = 0.2, 0.5, 0.6$, we conducted two different simulations for the two most extreme setups: the "best" setup, where $p_{mis} = 0.2$, $\rho^2_{\boldsymbol{x}y|t} = 0.9$, and the "worst" setup, where $p_{mis} = 0.6$, $\rho^2_{\boldsymbol{x}y|t} = 0.3$.

The simulation was conducted in a way similar to Section 3.4. For simplicity, we set $(\boldsymbol{\gamma_0}, \boldsymbol{\gamma_1}) = (1, 1)$, $(\alpha_0, \alpha_1, \boldsymbol{\alpha_2}) = (1, 1, 1)$. These values are kept fixed throughout the simulations for different setups. We varied δ and $\rho^2_{\boldsymbol{x}y|t}$ by altering the covariance matrix of $\boldsymbol{x}$ and the variance of y given $\boldsymbol{x}$ and t.

The simulation results are shown in Tables 3.3 and 3.4.

There is a slight difference between the setups in Tables 3.3 and 3.4. For Table 3.3, each response has a p_{mis} probability of being missing which has a 0.5 chance of being assigned to each group. Therefore, though the overall percent of missing values is p_{mis}, in each iteration, the number of missing values is not fixed and is determined by a Binomial distribution with parameters n and p_{mis}. Similarly, the number of observations in each group follows a Binomial distribution with parameters n and $p = 0.5$. For Table 3.4, the number of missing values is fixed for each iteration, and each group will always have the same number of observations. For example, if $p_{mis} = 0.2$ and $n = 110$, then each group will always have 55 observations and each group will always have 11 missing observations.

The tables show that, although simulated power is slightly below nominal power when n is small to medium (difference of about 0.01), we are generally able to achieve a power that is "good enough" with the sample size provided by the table. When the sample size is large, this difference become negligible.

3.6 Conclusion and Discussion

In Chapter 2, we presented a situation where MI provides more efficient parameter estimates and thus improves statistical power. In this chapter, we extend the power calculation formula to a two-sample pooled variance test (which is equal to testing the treatment effect using linear regression), and provide a setting where MI improves performance. The formula in Section 3.3.2, though slightly over-estimating the statistical power compared to the simulation results, still provides a reliable method for calculating the MI statistical power. We show that for a two-sample t-test, when the MDM is MCAR, the power is determined by the number of imputation, the percentage of missing responses, and the multiple correlation between the response and the auxiliary imputation variables, in addition to the values required for complete data power calculation (i.e., the effect size, the sample size, and the significance level). We provide sample size tables to display the optimal sample size an experiment needs to obtain a satisfactory power. This chapter only shows tables for powers equal to 0.8 and m-values equal to 50, but other tables can be easily obtained using the methodology presented in this chapter.

Furthermore, our method also provides a solution for evaluating whether MI can provide better results as compared to CCA. In the situation discussed above, the parameter of interest is $\alpha_1 + \boldsymbol{\gamma_1}\boldsymbol{\alpha_2'}$. Since MI uses the entire dataset to estimate $\boldsymbol{\gamma_1}$ and CCA only uses

the fully-observed part of the dataset, MI should provide a smaller variance, and thus a better statistical power, as m increases. However, as shown by Graham et al. [2007], a pre-condition is that the number of imputations not be too small. Despite having a sufficient number of imputations, the amount of improvement can be small and the result can still be much worse compared to estimates calculated from the complete data. This happens when r_E has a large value, meaning that much variability was added during the imputation process. In this situation, r_E can be decreased by increasing $\Sigma_{\boldsymbol{x},t}/\sigma^2_{y|\boldsymbol{x},t}$. In other words, MI can provide a better result if the relationship between y and $\boldsymbol{x}$ becomes stronger.

Worth noting is the fact that if there are no covariates, the power calculated from MI will be no better than that from CCA. This is because if there is no extra information source (in this case, $\boldsymbol{x}$), MI and CCA both use the fully-observed part of the data to obtain information on the effect of the group indicator. This is one reason why we tend to include more variables in the imputation model: to obtain a more efficient estimate and a higher statistical power.

Power = 0.7

δ	0.2			0.5			0.8		
p_{mis}	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6
$\rho^2_{\boldsymbol{xy}\|t}$=0.3	732	920	1298	120	152	214	50	62	88
	(776)	(1034)	(1552)	(128)	(170)	(256)	(54)	(70)	(106)
$\rho^2_{\boldsymbol{xy}\|t}$=0.6	684	790	1006	112	130	166	46	54	68
	(776)	(1034)	(1552)	(128)	(170)	(256)	(54)	(70)	(106)
$\rho^2_{\boldsymbol{xy}\|t}$ = 0.9	636	662	716	104	108	118	42	44	48
	(776)	(1034)	(1552)	(128)	(170)	(256)	(54)	(70)	(106)

Power = 0.8

δ	0.2			0.5			0.8		
p_{mis}	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6
$\rho^2_{\boldsymbol{xy}\|t}$=0.3	930	1168	1650	152	190	270	62	78	110
	(986)	(1314)	(1972)	(160)	(214)	(322)	(66)	(88)	(132)
$\rho^2_{\boldsymbol{xy}\|t}$=0.6	868	1004	1278	142	164	208	58	66	84
	(986)	(1314)	(1972)	(160)	(214)	(322)	(66)	(88)	(132)
$\rho^2_{\boldsymbol{xy}\|t}$ = 0.9	808	842	908	132	138	148	54	56	60
	(986)	(1314)	(1972)	(160)	(214)	(322)	(66)	(88)	(132)

Power = 0.9

δ	0.2			0.5			0.8		
p_{mis}	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6
$\rho^2_{\boldsymbol{xy}\|t}$=0.3	1242	1562	2206	202	254	358	80	102	144
	(1318)	(1758)	(2636)	(216)	(288)	(432)	(86)	(114)	(172)
$\rho^2_{\boldsymbol{xy}\|t}$=0.6	1160	1342	1708	188	218	278	76	88	112
	(1318)	(1758)	(2636)	(216)	(288)	(432)	(86)	(114)	(172)
$\rho^2_{\boldsymbol{xy}\|t}$ = 0.9	1080	1126	1216	176	182	198	70	72	80
	(1318)	(1758)	(2636)	(216)	(288)	(432)	(86)	(114)	(172)

Table 3.2: Some sample size tables

These are the sample size tables for required powers equal to 0.7, 0.8, and 0.9. The values in the tables are the required sample sizes (n) when MI is used; the values in the parentheses are the corresponding required sample size under the same conditions if CCA is used instead. For this table, $m = 50$ and the two samples are of the same size $n/2$.

	δ								
	0.8			0.5			0.2		
$(p_{mis}, \rho^2_{\boldsymbol{xy}\|t})$	nominal	simulated	n	nominal	simulated	n	nominal	simulated	n
(0.2, 0.9)	0.8	0.8	54	0.8	0.79	131	0.8	0.8	808
(0.6, 0.3)	0.8	0.78	110	0.8	0.79	269	0.8	0.8	1650

Table 3.3: Simulation results for selected setups

Here are the simulation results for the selected setups. Values in the "nominal" column represent the nominal powers (i.e., predetermined statistical power); the values in the "simulated" column represent the actual powers calculated through the Monte-Carlo simulation.

	δ								
	0.8			0.5			0.2		
$(p_{mis}, \rho^2_{\boldsymbol{x}y\|t})$	nominal	simulated	n	nominal	simulated	n	nominal	simulated	n
(0.2, 0.9)	0.8	0.8	54	0.8	0.8	131	0.8	0.8	808
(0.6, 0.3)	0.8	0.79	110	0.8	0.8	269	0.8	0.8	1650

Table 3.4: Simulation results for selected setups

Here are the simulation results for the selected setups. Values in the "nominal" column represent the nominal powers (i.e., predetermined statistical power); the values in the "simulated" column represent the actual powers calculated through the Monte-Carlo simulation.

Chapter 4

Missing Information Rates for Adjusted Two-Stage Multiple Imputation

4.1 Introduction

Nested multiple imputation [Shen, 2000], also known as two-stage multiple imputation [Harel, 2009], is an extension of traditional MI. Whereas all missing data are treated similarly and are imputed together in traditional MI, the unobserved data, Y_{mis}, are split into two parts (Y^A_{mis}, Y^B_{mis}) in nested MI. This is done either to gain computational efficiency or because the unobserved data are of different types. In the imputation process, Y^A_{mis} is first imputed m times. Then for each of the imputed Y^A_{mis}, Y^B_{mis} is imputed n times. Since there are two different types of unobserved data, the overall rate of missing information can correspondingly be divided into two different parts: the missing information rate due solely to Y^A_{mis}, and the missing information rate due to Y^B_{mis} if Y^A_{mis} were known. These two rates of missing information, their definitions, interpretations, and behaviors are studied by Harel [2007b].

Reiter [2008] discussed an extension for two-stage MI where part of the observed data, also known as the calibration data, is only used for imputation and not for analysis. He showed that Rubin [1987]'s conventional MI (henceforth referred to as conventional MI) creates a positive bias in such a situation, in response to which Rubin introduced a new two-stage MI (adjusted two-stage MI) to solve this complication. In the first stage, researchers construct the posterior distribution of model parameter θ by simply using the observed data. Then m sets of $\theta^{(l)}$ (l = 1, 2,...,m) are independently drawn from this model. In the second stage, n sets of Y_{mis} are imputed for each $\theta^{(l)}$. This method allows researchers to obtain an unbiased estimation of variance and calibrate the inference of Q. An application of this method was implemented by Siddique et al. [2015].

Reiter's adjusted nested MI procedure is different from that used by Shen [2000], and their calculated rates of missing information should subsequently also differ. However, these rates of missing information have not yet been developed. In this chapter we introduce the rates of missing information for this setup, then study and interpret the results. We also use these new results to describe how the missing data and the calibration data affect the final inference of Q. We demonstrate our findings using simulation studies.

The rest of the chapter will be arranged as follows. Methodology is discussed in Section 4.2. A detailed description of Reiter's nested imputation is given in Section 4.2.1. In Section 4.2.2, we derive the general formulas for the rates of missing information and interpret them. In Section 4.3, we discuss the results in more detail and in the context of a specific setup. We explore changes in rates of missing information by changing the relationship between our responses. We present a simulation study based on a data example

from Siddique et al. [2015] and provide a discussion in the final section.

4.2 Methodology

4.2.1 Reiter's Adjusted Nested MI

Reiter [2008] proposed a new nested MI method where some records are only used in the imputation process and are not used or available for analysis. Such records are known as calibration or validation data, as cited Reiter [2008]. Since the setup is different from conventional MI, the use of the conventional MI procedure may result in biased estimates.

Consider a scenario with three variables. Z is the gold standard response, which is measured without error. Y, on the other hand, is an outcome measured with error. X represents the covariates and is also measured without error. In the original dataset D_{org}, only the covariate and the response are measured with error, $D_{org} = (X_{org}, Y_{org})$, while Z_{org} is missing completely. To adjust the inference for measurement error, a validation dataset (if available) can be used to help impute the gold-standard response. The validation data D_{val} includes all necessary variables, $D_{val} = (X_{val}, Y_{val}, Z_{val})$. However, for practical reasons, such validation datasets can only be used for imputation procedures and not released for analysis.

Disseminating the validation dataset may lead to biased inference. This was shown by Reiter [2008] with a simple example in which the parameter of interest, Q, is the population mean of Z. Reiter [2008] theoretically showed that in such case, without using D_{val} for the analysis, conventional MI will result in a positive bias estimating the variance of

Q. Therefore, he suggested using an adjusted two-stage imputation to avoid this bias.

Let θ be the parameters of the model describing the relationship between Z_{val} and (X_{org}, Y_{org}). The imputation process is separated into two different stages. First, m values of θ are drawn from the model $P(\theta|D_{val})$. We denote them as $\theta^{(l)}$, l = 1, 2, ...,m. Then for each $\theta^{(l)}$, n versions of Z are drawn from the the model $P(Z_{org}|X_{org}, Y_{org}, \theta^{(l)})$. They are denoted as $Z^{(l,i)}$, i = 1, 2, 3,....,n. After the imputation, the imputer will release the mn imputed datasets. Let $D^{(l,i)} = (Z^{(l,i)}, D_{org})$, meaning the i^{th} completed dataset within the l^{th} nest.

In the analysis stage, analysts will perform data analysis methods on each of the imputed datasets. Let $q^{(l,i)}$ and $u^{(l,i)}$ be the estimate of Q and its posterior variance based on $D^{(l,i)}$. The final inference is then based on the new combining rules as follows:

$$\begin{aligned}
\bar{q}_M &= \sum_{l=1}^{m}\sum_{i=1}^{n} q^{(l,i)}/(mn) = \sum_{l=1}^{m} \bar{q}_n^{(l)}/m, \\
\bar{w}_M &= \sum_{l=1}^{m}\sum_{i=1}^{n} (q^{(l,i)} - \bar{q}_n^{(l)})^2/[m(n-1)] = \sum_{l=1}^{m} \bar{w}_n^{(l)}/m \\
b_M &= \sum_{l=1}^{m} (\bar{q}_n^{(l)} - \bar{q}_M)^2/(m-1) \\
\bar{u}_M &= \sum_{l=1}^{m}\sum_{i=1}^{n} u^{(l,i)}/(mn).
\end{aligned} \tag{4.1}$$

Using the new combining rules, $\bar{q}_M$ is the two-stage estimate of Q. The three variance estimates $\bar{w}_M$, b_M, and $\bar{u}_M$ are correspondingly within-nest, between-nest, and within-imputation variance. Reiter [2008] showed that the final estimated variance can be obtained as $T_M = \bar{u}_M - (1 + 1/n)\bar{w}_M + (1 + 1/m)b_M$. When the number of records in D_{org}

is large, tests and CIs can be done using the approximation $(Q - \bar{q}_M)/T_M \sim t_{\nu_M}$, with

$$\nu_M^{-1} = (1 + 1/n)^2 \bar{w}_M^2/(T_M^2 m(n-1)) + (1 + 1/m)^2 b_M^2/(T_M^2(m-1). \quad (4.2)$$

4.2.2 Rates of Missing Information

Rubin [1987] introduced the concept of the fraction of missing information and used it to evaluate the relative efficiency of MI. Later, Schafer [1997] argued that the fraction of missing information governs the convergence rate of the EM algorithm [Dempster et al., 1977]. Schafer [1997] recommended calculating the fraction of missing information, since it is a useful diagnostic for assessing how the inferential uncertainty about Q is affected by the missing data. Harel [2007a] developed estimates for the rates of missing information in conventional and two-stage MI. He stated that, as with conventional MI, rates of missing information are still interesting in that they help evaluate how different types of missing data affect overall uncertainty and they provide insight for choosing the number of imputations.

Following the method provided by Harel [2007a], we develop the overall missing information rate and split it into two parts: the rate of missing information due to the unknown model and the rate of missing information due to the missing gold standard, if the model were known. To validate these results, we also use a second method to derive the rates of missing information, in this case doing so directly [Reiter, 2008]. We show that the two methods lead to the same results.

We first derive the overall rate of missing information. According to Harel [2007a], when the goal of MI is to estimate the rate of missing information, m and n need to be large

in order to assure stable results. In Equation (4.2), the degrees of freedom increase as m and n increase. Therefore, to simplify the calculation we assume that nm is large enough to make the student-t distribution close to a normal distribution, so that the information of Q contained in the incomplete data will be T_M^{-1}. The variance of Q would also decrease to $\bar{u}_M$ when there are no missing data. As such, the information of Q contained in the complete data is u_M^{-1}, with $\bar{u}_M^{-1} - T_M^{-1}$ representing the information loss due to missing data, so that the overall rate of missing information should be:

$$\begin{aligned} \hat{\lambda} &= \frac{\bar{u}_M^{-1} - T_m^{-1}}{\bar{u}_M^{-1}} \\ &= \frac{-(1+1/n)\bar{w}_M + b_M}{\bar{u}_M - (1+1/n)\bar{w}_M + b_M}. \end{aligned} \tag{4.3}$$

We next derive the rate of missing information due to missing data if the model is given. If the model is known, each $\theta^{(l)}$ used to impute Z_{org} would be identical to the true value of θ. The between-nest variance would then vanish and the total variance of Q would collapse to $(\bar{u}_M - \bar{w}_M)$. Therefore, the partial rate of missing information due to Z_{org} can be estimated as:

$$\begin{aligned} \hat{\lambda}^{Z_{org}|\theta} &= \frac{\bar{u}_M^{-1} - (\bar{u}_M - \bar{w}_M)^{-1}}{\bar{u}_M^{-1}} \\ &= \frac{\bar{u}_M - \bar{w}_M - \bar{u}_M}{\bar{u}_M - \bar{w}_M} = \frac{-\bar{w}_M}{\bar{u}_M - \bar{w}_M}. \end{aligned} \tag{4.4}$$

We can also derive $\hat{\lambda}^{Z_{org}|\theta}$ directly using the results in Reiter [2008]. Following the notation used by Reiter, let $Q^{(\theta)}$ denote the estimate of Q if the true θ is given to the researchers to impute Z_{org}. Its variance $V^{(\theta)} = var(Q|D_{org}, Q^{(\theta)})$, with $(Q|D_{org}, Q^{(\theta)}, V^{(\theta)}) \sim N(Q^{(\theta)}, V^{(\theta)})$. $Q^{(l)}$ is the estimate of Q if $\theta^{(l)}$ is known and used by the analysts to impute Z_{org}. Let

$$\bar{Q}_M = \sum_{l=1}^{m} Q^{(l)}/m, B_\infty = \lim_{m\to\infty} (m-1)^{-1} \sum_{l=1}^{m} (Q^{(l)} - \bar{Q}_M)^2,$$

$$W_\infty^{(l)} = \lim_{n\to\infty} (n-1)^{-1} \sum_{i=1}^{n} (Q^{(l)} - q^{(l,i)})^2, \bar{W}_\infty = \lim_{m\to\infty} \sum_{l=1}^{m} W_\infty^{(l)}/m.$$

$$Q^* = (Q^{(1)}, ..., Q^{(m)}), W^*_\infty = (W^{(1)}_\infty, ..., W^{(m)}_\infty).$$

The following results (4.5) build on Section 3.2 in Reiter [2008]:

$$\begin{aligned} E(V^{(\theta)}|D^*, B_\infty, W^*_\infty) &= \bar{u}_M - \bar{W}_\infty \\ E(\bar{W}_\infty|D^*) &= \bar{w}_M. \end{aligned} \tag{4.5}$$

Using these results we can directly conclude that the unbiased estimate of $V(\theta)$ is $\bar{u}_M - \bar{w}_M$. Given θ, the information of Q is therefore still estimated by $(\bar{u}_M - \bar{w}_M)^{-1}$ and Result (4.4) also holds.

According to Equation (4.5), since $\bar{u}_M - \bar{W}_\infty$ is the expectation of variance ($V^{(\theta)}$), it should have a positive value. Its estimate, $\bar{u}_M - \bar{w}$, should thus also be positive. This leads to the conclusion that $\hat{\lambda}^{Z_{org}|\theta}$ is negative by definition. Since $\hat{\lambda}^{Z_{org}|\theta}$ represents the amount of information lost due to missing values, a negative value represents information gain. Specifically, it means that with an accurate model, despite some missing data, we still gain additional information.

The last rate of missing information is the difference between $\hat{\lambda}$ and $\hat{\lambda}^{Z_{org}|\theta}$, the amount by which the rate of missing information will drop if θ was known. We define this as:

$$\lambda^\theta = \hat{\lambda} - \hat{\lambda}^{Z_{org}|\theta}. \tag{4.6}$$

4.3 Rates of Information in a Simple Situation

At first sight, a negative rate of missing information seems to be counter-intuitive. However, we use a simplified situation to demonstrate the logic behind this result and its

interpretation. In this section, we derive formulas for the variances, explain the seemingly counter-intuitive negative value, and explore how the rates of missing information change according to data structures.

4.3.1 Data Structures

Consider a simple situation where only two responses (y_1, y_2) are included in a dataset and there are no other covariates. For simplicity, we assume that the responses are following a bi-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (\mu_1, \mu_2)$, and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & {}_{22} \end{bmatrix}$. Let's use $\boldsymbol{\psi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Table 4.1: The Data Structure

	y_1	y_2
part a	missing	observed
part b	observed	missing
part v	observed	observed

Consider that our dataset consists of three different parts: a, b, and v. Part a is missing response y_1, part b is missing response y_2, and the last part, part v, is completely observed. See Table 4.1 for the data structure. The sizes of each part are correspondingly n_a, n_b, and n_v. The third part of the data set, part v, is the calibration, or validation dataset.

4.3.2 The Adjusted Two-Stage MI Process

Let's rewrite the likelihood function of (y_1, y_2) into the product of the distribution of y_2 and the distribution of y_1 given y_2, such that:

$$\begin{aligned} y_1|y_2 &\sim N(\beta_1 y_2 + \beta_0, \sigma^2_{1|2}), \\ y_2 &\sim N(\mu_2, \sigma_{22}). \end{aligned} \tag{4.7}$$

We use $\boldsymbol{\theta} = (\mu_2, \sigma_{22}, \beta_0, \beta_1, \sigma_{1|2})$ to denote the parameter vector. It clearly has a one-one relationship with the bivariate normal distribution parameter $\boldsymbol{\psi} = (\boldsymbol{\mu}, \Sigma)$.

The adjusted nested MI includes two steps in the imputation stage. In the first step, a posterior distribution of $\boldsymbol{\psi}|y_2, y_1$ is calculated from the dataset. Then, m $\boldsymbol{\psi}$'s are drawn from the distribution, creating m nests/models $\boldsymbol{\psi}^l, l = 1, 2, ..., m$. In the second step, n imputations of y_{1a} and y_{2b} are drawn within each nest to create a completed dataset. Here we use y_{2b} to denote the y_2 of part b. Part v is then dropped from each completed dataset for analysis purposes. The imputation process generates mn completed datasets, $D^{l,i}, l = 1, 2, ..., m; i = 1, 2, .., n$.

4.3.3 Derivation of U, B, and W

The parameter of interest, Q, is the mean of the first response y_1.

We know that for each $D^{l,i}$, the i^{th} completed dataset in the l^{th} nest,

$$\begin{aligned} q^{(l,i)} = \bar{y_1}^{(l,i)} &= \frac{n_a \bar{y}_{1b} + n_b \bar{y}^{(l,i)}_{2a}}{n_a + n_b} \\ &= c_1 + c_2 \bar{y_1}^{(l,i)}_a \\ &= c_1 + c_2(\beta_0^{(l)} + \beta_1^{(l)} \bar{y}_{2a} + \sigma^{(l)}_{1|2} \bar{z}^{(l,i)}). \end{aligned} \tag{4.8}$$

$$\bar{q}^{(l)} = c_1 + c_2(\beta_0^{(l)} + \beta_1^{(l)}\bar{y}_{2a} + \sigma_{1|2}^{(l)}\bar{z}^{(l,\cdot)}). \tag{4.9}$$

Here $c_1 = \frac{n_a\bar{y}_b}{n_a+n_b}$, $c_2 = \frac{n_b}{n_a+n_b}$; $\beta_k^{(l)}$ and $\sigma_{1|2}^{(l)}$ are correspondingly the β_k and $\sigma_{1|2}$ for the l^{th} model, k = 0, 1; $\bar{z}^{(l,i)} \sim N(0, 1/n_b)$ is the average of $\boldsymbol{Z^{(l,i)}} \sim N(\mathbf{0}, \boldsymbol{I_{n_b}})$. $\bar{z}^{(l,\cdot)} \sim N(0, \frac{1}{n\times n_b})$ is the average of $\bar{z}^{(l,i)}$, i = 1,...,n.

It is clear that when $\bar{y_1}^{(l,i)} = q^{(l,i)}$,

$$\begin{aligned} q^{(l,i)}|\boldsymbol{\theta}^{(l)} &\sim N(c_1 + c_2(\beta_0^{(l)} + \beta_1^{(l)}\bar{y}_{2a}), c_2^2\sigma_{1|2^{2(l)}/n_b)} \\ \bar{q}^{(l)}|\boldsymbol{\theta}^{(l)} &\sim N(c_1 + c_2(\beta_0^{(l)} + \beta_1^{(l)}\bar{y}_{2a}), c_2^2\sigma_{1|2}^{2(l)}\frac{1}{n\times n_b}). \end{aligned} \tag{4.10}$$

Therefore, we have

$$\begin{aligned} var(\bar{q}^{(l)}) &= E(var(\bar{q}^{(l)}|\boldsymbol{\theta}^{(l)})) + var(E(\bar{q}^{(l)}|\boldsymbol{\theta}^{(l)})) \\ &= \frac{c_2^2}{n\times n_b}E(\sigma_{1|2}^2) + c_2^2(1, \bar{y}_{2a})var(\boldsymbol{\beta^{(l)}})(1, \bar{y}_{2a})', \end{aligned} \tag{4.11}$$

with $\boldsymbol{\beta^{(l)}} = (\beta_0^{(l)}, \beta_1^{(l)})$. Since $\bar{w}_M$ is used to estimate $E(var(q^{(l,i)}|\boldsymbol{\theta^{(l)}}))$, b_M is used to estimate $var(\bar{q}^{(l)})$, $\bar{u}_M$ is used to estimate the complete data variance of Q. We therefore have approximately

$$\begin{aligned} \bar{w}_M &= c_2^2\sigma_{1|2}^2/n_b \\ &= \frac{c_2^2}{n_b}(\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}) \\ b_M &= \frac{c_2^2}{n\times n_b}(\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}) + c_2^2(1, \bar{y}_{2a})var(\boldsymbol{\beta^{(l)}})(1, \bar{y}_{2a})' \\ \bar{u}_W &= \sigma_{11}/(n_a + n_b). \end{aligned} \tag{4.12}$$

4.3.4 Simulation Results

In the simulation study, we follow the data structure discussed in Section 4.3.1. We do so in order to illustrate the results found in Section 4.3.3, and to show how the rates of

missing information behave in this simplified setup. To explore how the rates of missing information change according to the correlation between the two responses, we used low and high correlation coefficients between y_1 and y_2, $\rho = 0.1$ and $\rho = 0.9$. To explore how the results are affected by the number of imputations within each nest, we also vary the values of n, the number of within-nest imputations, from 5 to 100.

Simulation 1: Adjusted Nested MI with Complete Model Knowledge

In this simulation, we focus on a situation where the true model is known. By "given model", we mean that the m models required for the two-stage MI are no longer drawn from posterior distribution. Instead, we simply used the actual parameter chosen to generate the dataset.

To bring our simulation closer to that of Siddique et al. [2015], the motivating example for this work, we set $\mu = (54.0, 22.6)$, $\sigma_{11} = 7.23^2, \sigma_{22} = 3.97^2, \sigma_{12} = 1.23 \times 3.97 \times \rho$, with ρ as the correlation coefficient between the two responses. The number of subjects in parts a and b are both 100. Part v is ignored, as it is not necessary here to calculate the model. The number of models, m, is set to 100.

Since the model is given, $var(\boldsymbol{\beta}^{(l)})$ shrinks to 0. Based on the formulas we described above and the initial settings of the simulation, the true variance, the within-nest variance, and the between-nest variance (i.e., the true parameters MI is meant to estimate) can be found in Table 4.2 (in parentheses). The n in the table is the number of imputations within each nest, and the value ρ is the correlation coefficient between the responses y_1

and y_2.

We can now compare the calculated variances with the simulation results in Table 4.2. We can also evaluate the estimates, the standard error, and the rates of missing information. In Table 4.2, $\bar{q}_M$ is the MI estimate of the parameter of interest (the mean of the first response) and se is its standard error; $\bar{u}_M$, b_M and $\bar{w}_M$ are the within-imputation, between-nest, and within-nest variances; the λ's are the rates of missing information. Notice that here λ and $\lambda^{Z_{org}|\theta}$ are approximately equal to each other. That is because in this setup, the model is given to us, so the overall rate of missing information (λ) is only the rate of missing information if the model is given ($\lambda^{Z_{org}|\theta}$). The values in the parentheses are correspondingly the true parameters $\bar{u}_M$, b_M, and $\bar{w}_M$ that are supposed to be estimated, and which are calculated from Equation (4.12).

To explain the negative values in Table 4.2, we provide Table 4.3, which contains the complete-data analysis results. We obtained these results after we analyzed the original dataset consisting of Trial 1 and Trial 2 without any missing data assigned.

In Table 4.3, $\hat{\mu_1}$ is the complete data estimate of μ_1, the mean of the first response. SE is its standard error. As can be seen from Table 4.2, the simulated imputation variances and the theoretical variances are quite close to each other. As the number of imputations within each nest (n) increases, b decreases towards zero. This is exactly what we

Table 4.2: The simulated results when the model is given

		$\bar{q}_M$	se	$\bar{u}_M$	$\bar{w}_M$	b_M	λ	$\lambda^{Z_{org}\|\theta}$	λ^{θ}
$\rho = 0.1$	n = 5	53.4489	0.3648	0.2519	0.1249	0.0308	-0.8967	-0.983	0.0864
				(0.2614)	(0.1294)	(0.0259)			
	n = 50	53.4348	0.3494	0.2520	0.1297	0.0024	-1.064	-1.0608	-0.0032
				(0.2614)	(0.1294)	(0.0026)			
	n=100	54.166	0.3443	0.247	0.1282	0.001	-1.0838	-1.0784	-0.0054
				(0.2614)	(0.1294)	(0.0013)			
$\rho = 0.9$	n = 5	53.8799	0.4809	0.2541	0.0240	0.0059	-0.0988	-0.1041	0.0053
				(0.2614)	(0.0248)	(0.0050)			
	n = 50	53.8737	0.4787	0.2540	0.0249	5.00E-04	-0.1088	-0.1086	-2.00E-04
				(0.2614)	(0.0248)	(0.0005)			
	n=100	53.873	0.4790	0.2540	0.0246	2.00E-04	-0.1073	-0.1072	-1.00E-04
				(0.2614)	(0.0248)	(0.0002)			

Table 4.3: Complete-data-analysis Results

$\hat{\mu_1}$	SE
53.0129	0.5073

expected: as the number of the imputations increases, the estimate becomes more accurate and therefore moves closer to the true between-nest variance. Since the model here is given to us, the information loss due to the missing model should be 0, which is also demonstrated by the simulation results. In this simulation, the rate of missing information due to the missing model, λ^{θ}, is not that close to zero when n is small (n = 5). It is logical to believe that this deviation from zero occurs because the relative variances are not stable when n is small. We can see that this value moves closer to zero when n is large (50 or 100).

The simulations imply that, even though some data are missing, knowledge of the model will result in information gain. This can be shown by comparing the MI results with the complete data results in Table 4.3. When there is neither missing data nor an extra information source (in this case, the extra information source is the model itself), the standard error of the estimate is approximately 0.51. This value is greater than all

the MI standard errors in Table 4.2.

When comparing the correlation effects, there is a gain of information when the correlation between the two responses is low ($\rho = 0.1$ setup). This is because if the two responses have a perfect linear relationship, the missing response can be directly computed from the observed response, and hence introducing an extra information source will not add any additional information.

Simulation 2: Adjusted Nested MI Without Model Knowledge

In this section, we focus on a more realistic situation where the model is not given to us, and we follow the adjusted two-stage MI procedure. We use the same simulation setup as in Section 4.3.4, except that this time we have additional data of size 100. By combining data parts a, b, and v, we obtain a combined dataset from which we draw the posterior distributions $P(\boldsymbol{\theta}|\boldsymbol{D})$ and $P(\boldsymbol{\psi}|\boldsymbol{D})$. The former is used to calculate the variance-covariance matrix of $\boldsymbol{\beta}$; the later is used to draw the m models required for the two-stage MI, m = 100.

Table 4.4 contains the results calculated from Equation (4.12) (with the values in parentheses).

We can now compare the computed variances with the simulation results in Table 4.4. The final estimate, the standard error of the estimate, and the rates of missing information are also presented in Table 4.4.

0

Table 4.4: The simulated results when model is not given

		$\bar{q}_M$	se	$\bar{u}_M$	$\bar{w}_M$	b_M	λ	$\lambda^{Z_{org}\|\theta}$	λ^{θ}
$\rho = 0.1$	n = 5	53.1933	0.4375	0.2621	0.1465	0.1041	-0.3764	-1.2679	0.8914
				(0.2614)	(0.1294)	(0.1003)			
	n = 50	53.1716	0.4501	0.2620	0.1348	0.0774	-0.2979	-1.0606	0.7627
				(0.2614)	(0.1294)	(0.0770)			
	n = 100	53.1726	0.4439	0.2620	0.1396	0.0753	-0.3352	-1.1410	0.8059
				(0.2614)	(0.1294)	(0.0757)			
$\rho = 0.9$	n = 5	53.5981	0.5191	0.2714	0.0231	0.0256	-0.0082	-0.0932	0.0851
				(0.2614)	(0.0248)	(0.0266)			
	n = 50	53.6003	0.5204	0.2718	0.0223	0.0216	-0.0044	-0.0896	0.0852
				(0.2614)	(0.0248)	(0.0221)			
	n = 100	53.6008	0.5193	0.2717	0.0231	0.0211	-0.0083	-0.0931	0.0847
				(0.2614)	(0.0248)	(0.0219)			

Table 4.4 indicates that when $\rho = 0.1$, we still gain information with the extra information source (in this case, the calibration trial). This can be demonstrated by comparing the MI standard errors with the complete-data-analysis standard error in Table 4.3. However, because of the inaccuracy of the model estimated from the calibration trial, part of the information is still lost due to estimation. This can be seen in the positive value of λ^{θ}, or by comparing the standard errors in Table 4.4 to those in Table 4.2.

4.3.5 Simulation Discussion

From the simulated results, we can see that using Reiter's adjusted two-stage MI to deal with incomplete data entails a final variance T_{mn} that is not necessarily greater than $\bar{u}$. According to Rubin [1987], $\bar{u}$ in MI is approximately equal to $U = var(Q|D_{complete})$, the variance of Q if the data were complete. Therefore, the information of Q carried by the imputed datasets is not necessarily smaller than the information of Q carried by the complete dataset. The change of information is thus sometimes a negative value instead

of a positive one, showing that incomplete data actually carries more information. This seemingly counter-intuitive phenomenon can be verified by comparing T_{mn} with the actual u of the complete original dataset. It is true that if we only had the original incomplete dataset (parts a and b, as in Table (4.1)), this information would always be less than the information we would have had if there were no missing data. However, Reiter's method uses a second information source, i.e., the calibration dataset (part v of Table (4.1)). It is therefore possible to gain more information after performing the adjusted two-stage MI.

4.4 Simulation Study Based on a Data Example

4.4.1 The Data Problem

In this section, a more complex simulation is developed based on the motivating data example shown by Siddique et al. [2015]. In that article, due to the nature of the model chosen and the original dataset, a calibration dataset has to be used for the imputation and later abandoned for the analysis, therefore providing a good situation to use Reiter's adjusted nested MI. We would like to apply our results to this more complicated dataset, to evaluate how the calibration dataset impacts the imputation performance.

Siddique et al. [2015] combined five trials for a meta analysis. The original dataset consists of five different randomized controlled trials (RCT), all designed to study the effectiveness of fluoxetine as a treatment for depression among depressed adolescents. However, these five trials have two different outcomes to measure depression levels, the Children's Depression Rating Scale (CDRS) and the Hamilton Depression Rating Scale (HDRS), and no trial uses both measures. In that case, it would be difficult to build a joint model using

both CDRS and HDRS as the response variables. The researchers treated the original dataset as an incomplete dataset. Each trial was either missing CDRS or HDRS. The structure of the data is shown in Table 4.5

Original Dataset

trials/variables	y_1 (CDRS)	y_2 (HDRS)
1,2,3,4	observed	missing
5	missing	observed

Calibration Dataset

trials/variables	y_1 (CDRS)	y_2 (HDRS)
6,7	observed	observed

Table 4.5: data structure of the actual data set

Siddique et al. [2015] used a mixed-effects model to jointly model the two measurements. Because of the lack of overlap, the relationship between them cannot be properly captured by the original five trials. For this reason, two extra trials (trials 6 and 7) which referenced both CDRS and HDRS were used for the imputation process. The two external trials were then excluded from the post-imputation analysis, as they were trials for a different treatment. Since part of the data were only used for imputation but not analysis, they chose to use Reiter's adjusted imputation to deal with the incomplete data problem. We would like to use the simulation study to apply our methods of rates of missing information for similar study design and evaluate how the calibration dataset affects the analysis results.

4.4.2 The Simulation

Our simulation used Siddique et al. [2015]'s imputation model for both imputation and analysis so that the congeniality between the imputation and analysis models would be retained [Meng, 1994]. This imputation model is shown in Equation (4.13).

$$\begin{aligned} y_{1ij} &= \beta_0 + \beta_1 age_i + \beta_2 gender + \beta_3 time_{ij} + \beta_4 time_{ij} \times T_i + \eta_{y_1 0i} + \eta_{y_1 1i} time_{ij} + \epsilon_{y_1 ij} \\ y_{2ij} &= \alpha_0 + \alpha_1 age_i + \alpha_2 gender + \alpha_3 time_{ij} + \alpha_4 time_{ij} \times T_i + \eta_{y_2 0i} + \eta_{y_2 1i} time_{ij} + \epsilon_{y_2 ij}. \end{aligned} \tag{4.13}$$

In Equation (4.13), the two responses y_1 and y_2 are correspondingly the two depression measurements, CDRS and HDRS. y_{1ij} is the j^{the} CDRS measurement of the i^{th} individual; y_{2ij} is the j^{the} HDRS measurement of the i^{th} individual. $\boldsymbol{\eta} = (\eta_{y_1 0i}, \eta_{y_1 1i}, \eta_{y_2 0i}, \eta_{y_2 1i})$ is the individual-level random-effect vector following a multivariate normal distribution. $\boldsymbol{\epsilon} = (\eta_{y_1 ij}, \epsilon_{y_2 ij})$ is the two dimensional, normally distributed random error vector. T_i represents which treatment group unit i belongs to.

To simplify the process, we generated three trials under the same model. The first two trials were of size 100 and the third was of size 200. After generating the three trials, all y_2 (HDRS) were assigned to be missing for the first trial. Similarly, all y_1 (CDRS) were assigned to be missing for the second trial. The third trial, which was supposed to be the external calibration dataset, had no missing values. We used Reiter's adjusted MI to impute the incomplete data, and calculated the within-nest variance, w, the between-nest variance, b, and the within-imputation variance, u. The parameter of interest, the interaction between time and treatment effect, was β_4, . We also calculated the three rates of missing information. This process was then repeated 100 times to create boxplots of the

variances and the rates of missing information. Finally, we shrunk the size of the third trial to 100, and replicated the whole procedure, to see how the size of the calibration dataset can affect the rates of missing information. The simulation results are in Figure 4.1 and 4.2.

In Figure 4.1, the middle fifty percent of the within-imputation variance, $\bar{u}_M$, and the within-nest variance, $\bar{w}_M$, are approximately on the same level for the two different calibration trial sizes. When the calibration trial is of size two hundred (trial 3 size: 200), u has a smaller variability (a smaller range shown in the boxplot) . This makes sense since we are able to gain more accurate model information when we have a larger calibration trial. Based on this same reasoning, as the model estimated from the data becomes more accurate, the variance between the nests will shrink. This can also be seen in Figure 4.1 where the middle fifty percent of b, the between-nest variance, increases as the size of the calibration trial decreases to 100.

In Figure 4.2, the middle fifty percent of $\lambda^{Z_{org}|\theta}$ are approximately the same for the two different calibration trial sizes. This is because the difference between the two setups is only the size of the calibration trial. Since the λ^{θ} can be considered the overall rate of missing information if we had an infinitely large calibration dataset, it cancels the difference between the two setups. The mean and the middle fifty percent of the overall rate of missing information increases slightly while the calibration trial size decrease to 100. This occurs because the information gained from the extra dataset decreases when we have a smaller calibration dataset. Since the final missing information is a combination of the loss due to missing responses and the gain due to the extra dataset, the overall rate of

information should be larger when the information gain decreases.

4.5 Conclusion

In this chapter, we derived the rates of missing information for Reiter's adjusted nested multiple imputation, a method that was designed primarily for situations where part of the data is only available for imputation and not for analysis. This part of dataset is known as the calibration dataset. We derived the rates of missing information as a tool to evaluate how both the calibration dataset and the missing data impact the performance of MI.

We found that when a calibration dataset is provided, the rates of missing information are not always positive. We argue that such a situation occurs because of an extra information source, i.e., the model, or the calibration dataset. We used a two-dimensional normal example to demonstrate our findings.

Since this chapter was motivated by Siddique et al. [2015]'s application of Reiter's adjusted nested MI, we performed a simulation based on Siddique et al. [2015]'s study to understand how calibration datasets affect imputation performance. We simplified the study by using only three trials instead of seven, and calculated the rates of missing information for the simplified setup. We used two different calibration data sizes to evaluate the impact of the accuracy of the calibration data.

The work presented here has certain limitations. In the first simulation study (Section 4.3.4), we only studied the mean of the first response, not the relationship between the two variables; while in the second simulation study in Section 4.4, we only examined how the size would affect the imputation performance, due to the complexity of the model and

of altering the correlation between the CDRS and HDRS. We performed all simulations under the assumption that the MI model is identical to the population distribution. This was done because our major goal is to construct the rates of missing information and evaluate their performance, then assess if they provide the information needed. It would be interesting to see the results from an inaccurate imputation model.

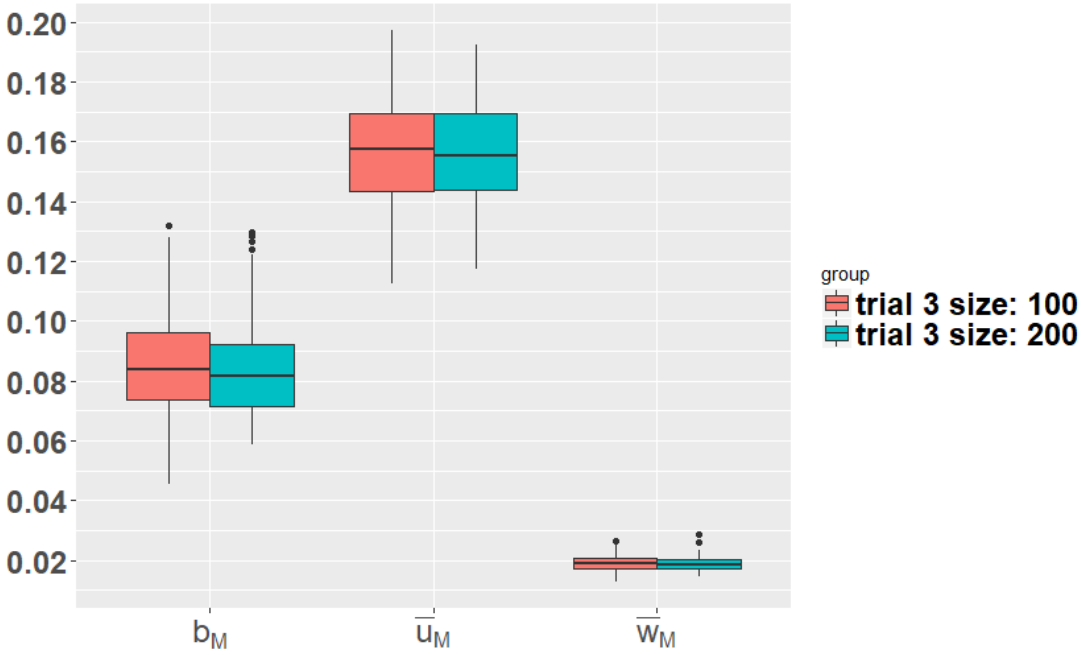


Figure 4.1: Variance box plots

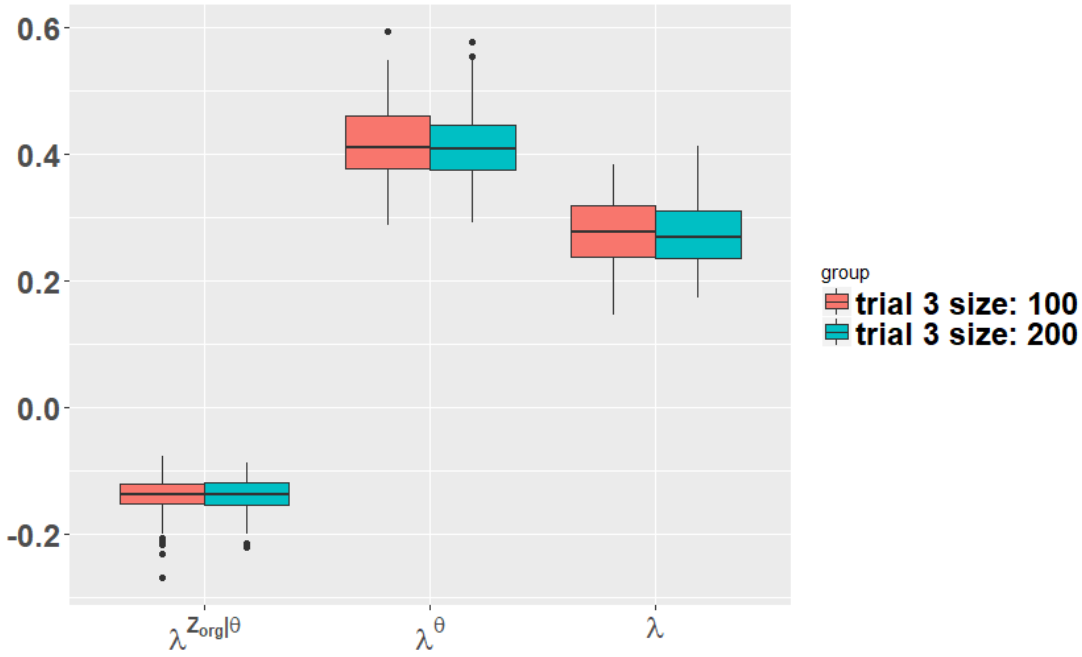


Figure 4.2: Rates of missing information box plots

Chapter 5

Conclusions

This dissertation studies several aspects of MI inference. First, we consider how to calculate statistical power when data are incomplete and MI is used to address the incomplete datasets. Second, we evaluate when MI recovers more power than CCA. Thirdly, we look at how to evaluate the impact of missing data and calibration data (i.e., an extra information source) when calibration data is used for imputation but not for analysis.

The first two questions are examined in Chapters 2 and 3. In Chapter 2, we constructed a power calculation formula using expected MI variances derived from Rubin [1987]. Our result implies that MI can recover more information as the ratio between between-imputation variance to within-imputation variance decreases. When this ratio, which is a function of the missing information rate, approaches zero, the incomplete data power becomes equal to the complete data power. We then specify the closed-form estimates for expected MI variances when a one-sample t-test for population mean is of interest. In this case, power can be calculated at the experiment design stage. We provide simulation results to examine these findings. In general, MI provides better power than

CCA, even when the MDM is MCAR. When MDM is MAR, our results show that the difference between the performance of MI and CCA increases. Notice that, under MAR, the increased power provided by MI occurs not only because MI provides a more efficient estimate (i.e., less standard error of the estimate), but also because the MI estimate is unbiased. Our results further show that when the number of imputations is too small (m = 3), CCA will under some conditions outperform MI. This is an interesting finding. We already know that too few imputations can decrease statistical power, as was addressed in Graham et al. [2007], but we did not yet know that it can in fact go lower than that of CCA. This emphasizes the fact that, if used, MI must be used correctly, and with enough imputations. Otherwise, it may be worse than not using it at all.

Chapter 3 extends the results of Chapter 2. It provides an ML-based method to estimate the expected between-imputation and within-imputation variances. We compare the ML-based MI variance-estimating method with the method used in Chapter 2 and show that the ML-based method can estimate equally well the required variance expectations for statistical power calculation. We then apply this method to a two-sample student's t-test and obtain a closed-form power calculation equation. We use simulation studies to validate our result. We also provide a method to estimate optimal sample size when the MDM is MCAR. We show that, for the two-sample t-test discussed in this chapter, power is determined by the number of imputations, the correlation between response and the imputation covariate, and the percentage of missing responses, in addition to the values required for complete data power analysis (i.e., effect size, significance level, and sample size). We use the R function "uniroot" to estimate an optimal sample size using our power calculation function for the two-sample t-test and provide some sample size tables to guide

researchers with their experiment design. Several setups are selected from the sample size table as the basis for simulation studies to show that the sample size table performs well.

In 4, we develop a method to calculate the rates of missing information for Reiter [2008]'s adjusted two-stage MI. The motivation for this was that, when part of the data used for imputation are not included for analysis, Reiter [2008]'s adjusted two-stage MI is required. In these cases, researchers are interested in measurements that can evaluate the impact of missing entries and of data only used for imputation (this data is called the calibration data). We use MI rates of missing information to serve this purpose. We examine the behavior of the rates of missing information for a two response situation which mimics the study described in Siddique et al. [2015]. We provide closed-form formulas to obtain rates when the corresponding parameter of interest is the average of one response. We use a simplified simulation along with a more complex simulation following Siddique et al. [2015]'s data structure to study how the change in the correlation of response and the size of the calibration data affect the rates of missing information for the adjusted MI.

There are several limitations to our research on statistical power, which are to some degree an indicator of the direction future research may take. The first such limitation is that the results are limited to proper MIs. For improper MIs, there is no guarantee that Rubin [1987]'s MI inferences hold and therefore no guarantee for our power calculation method. Since FCS and PMM are two widely used MI methods for continuous data, it would be of interest to examine how our method works for these MI algorithms.

We focus on population means when we specify the general power calculation function. Although assessing population means is an important task for many researchers, there are also other parameters of interest such as population proportion or correlation coefficient between variables,. Another direction for future research is to extend our results to multiple population mean comparison, i.e., ANOVA.

In this dissertation, we make several assumptions about missing values. We assume that only responses have the potential to be missing and that covariates are fully observed. We chose to make these assumptions because MI was originally developed for survey data and that missing-in-response is an assumption previously used by Rubin [1987]. However, missing covariates can be a common issue, especially when researchers cannot control covariate levels. If the missing data pattern is monotone, it is possible to obtain closed-form results using the ML-based method in Chapter 3. If the pattern is arbitrary, closed-form results may not be available. One way to deal with this issue can be to remove some observations so that the rest of the data can be arranged into a monotone pattern. Similar to CCA, this discards potentially useful data and the estimated power, and means that the required sample size will be conservative. More studies can be done to evaluate the performance of this solution. We also assume that the MDM is MAR. It would be interesting to extend our work to MNAR cases.

We also make assumptions about the data structure. All the results presented in this dissertation are based on large-sample MI inference. We would like to see how it this would work with medium or small samples. Furthermore, the research focuses on multivariate normal data. Multivariate normality is a commonly-used assumption for many statistical

methods, such as the one-sample t-test, general linear regression, and linear discriminant analysis. For these methods, transformation is a potential solution when the normality assumption is violated. The violation of normality may not be a serious issue for t-tests, since population mean still follows an approximately normal distribution when sample size is large. However, we may still want to assess the robustness of our methods when applied to non-normal data, and whether transformation can be a solution if non-normality does lead to severe bias.

We also consistently assume that the imputation model is a "good" one, meaning that it precisely represents the process under which the data is generated. There is always the possibility that the imputer has a "bad" imputation model which may mis-specify the probability distribution of the data or include irrelevant covariates. More studies could be done to explore MI performance regarding statistical power when a flawed imputation model is used.

Further limitations have to do with the study of missing information rates. The calculation of rates relies on the nested-MI variances ($\bar{u}_M$, b_M, and $\bar{w}_M$). While Reiter [2008] showed that a large n may not be required to obtain a type-I error rate close to the nominal α, our research showed that MI variances are unstable with a small number of n (within nest imputation). We would like to study the behavior of MI variances and provide a better guide as to ideal numbers of imputations for researchers who are interested in missing information rates. Another potential issue is that when $\bar{u}_M$ and $\bar{w}_M$ are close to each other, $\lambda^{Z_{org}|\theta}$ can have some unexpected values. This is because the denominator is $\bar{u}_M - \bar{w}_M$ and a small change in $\bar{w}_M$ can have a large impact on $\lambda^{Z_{org}|\theta}$. $\bar{u}_M \approx \bar{w}_M$ is also

an issue raised by Reiter [2008], which sometimes leads to a negative T_M. Reiter [2008] handled this issue by using 0 instead of $\bar{u}_M - \bar{w}_M$ whenever $\bar{u}_M - \bar{w}_M < 0$. This solution however cannot be applied here when to computing the rates of missing information because we cannot have a denominator equal to 0. Other methods are needed to better deal with this problem.

Appendix A

A.1 Proof of lemma 1

Proof of lemma 1. Without loss of generality, we assume that the first n_1 entries in $\boldsymbol{t}$ are 1, and the rest $n_2 = n - n_1$ are 0. Let's use $\boldsymbol{I}$ to denote the $n \times n$ matrix, with diagnol elements equal to 1, $\boldsymbol{J}$ to denote the $n \times n$ matrix with all entries equal to 1. $\boldsymbol{J}_i$, $\boldsymbol{I}_i$ are $n_i \times n_i$ matrices defined as $\boldsymbol{J}$ and $\boldsymbol{I}$, i = 1, 2. $\mathbf{1}$ is an n dimensional vector with all entries equal to 1; $\mathbf{1}_i$ is an n_i dimentional vector defined as $\mathbf{1}$, i = 1,2. Therefore, we can easily see that, $\boldsymbol{t} = \begin{bmatrix} \boldsymbol{I_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{1}$.

Using the notations above, we have

$$(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \begin{bmatrix} \boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t} & \boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X} \\ \boldsymbol{X}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t} & \boldsymbol{X}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X} \end{bmatrix}^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{A.1}$$

With the knowledge of the inverse of block matrix, we have

$$
\begin{aligned}
D^{-1} &= \boldsymbol{X}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X} - \frac{\boldsymbol{X}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t}\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}}{(\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t})} \\
A &= (\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t})^{-1} \\
&\quad + (\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t})^{-1}\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}D\boldsymbol{X}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t}(\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t})^{-1} \\
B &= -(\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t})^{-1}\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}D \\
C &= -D\boldsymbol{X}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t}(\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t})^{-1}
\end{aligned}
\tag{A.2}
$$

Knowing $\boldsymbol{t} = \begin{bmatrix} \boldsymbol{I_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{1}$. , we have $\boldsymbol{t't} = \begin{bmatrix} \boldsymbol{J_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$, $(\boldsymbol{t}'(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{t})^{-1} = \frac{n}{n_1 n_2}$, we have,

$$
D^{-1} = \boldsymbol{X'}[(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}) - \frac{(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\begin{bmatrix} \boldsymbol{J_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})}{\frac{n_1 n_2}{n}}]\boldsymbol{X}
$$

$$
\begin{aligned}
\frac{(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\begin{bmatrix} \boldsymbol{J_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})}{\frac{n_1 n_2}{n}} &= \frac{n}{n_1 n_2}\begin{bmatrix} \boldsymbol{I_1} - \frac{n_1}{n}\boldsymbol{J_1} & \boldsymbol{0} \\ -\frac{1}{n_1}\boldsymbol{1_2'1_1} & \boldsymbol{0} \end{bmatrix}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}) \\
&= \frac{n}{n_1 n_2}\begin{bmatrix} (\frac{n_2}{n})^2\boldsymbol{J_1} & -\frac{n_1 n_2}{n^2}\boldsymbol{1_1 1_2'} \\ -\frac{n_1 n_2}{n^2}\boldsymbol{1_2 1_1'} & (\frac{n_1}{n})^2\boldsymbol{J_2} \end{bmatrix}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
D &= (\boldsymbol{X'}\begin{bmatrix} \boldsymbol{I_1} - \frac{1}{n_1}\boldsymbol{J_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I_2} - \frac{1}{n_2}\boldsymbol{J_2} \end{bmatrix}\boldsymbol{X})^{-1} \\
&= \boldsymbol{S_x^{-1}}
\end{aligned}
\tag{A.3}
$$

and

$$B = -\frac{n}{n_1 n_2}(\mathbf{1_1'}, \mathbf{0'}) \begin{bmatrix} \boldsymbol{I_1} - \frac{1}{n}\boldsymbol{J_1} & -\frac{1}{n}\mathbf{1_1 1_2'} \\ -\frac{1}{n}\mathbf{1_2 1_1'} & \boldsymbol{I_2} - \frac{1}{n}\boldsymbol{J_2} \end{bmatrix} \boldsymbol{X} D$$

$$= -\Delta_x' \boldsymbol{S_x}^{-1} \tag{A.4}$$

$$C = -\boldsymbol{S_x}^{-1}\Delta_x'$$

$$A = \frac{n}{n_1 n_2} + \Delta_x' \boldsymbol{S_x}^{-1}\Delta_x'$$

■

A.2 Chapter 3: more power tables

Power = 0.5

δ	0.2			0.5			0.8		
p_{mis}	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6
$\rho^2_{\boldsymbol{x}y\|t}$= 0.3	458	574	812	76	96	136	32	42	58
	(486)	(648)	(970)	(80)	(108)	(160)	(36)	(48)	(70)
$\rho^2_{\boldsymbol{x}y\|t}$= 0.6	428	494	628	72	84	106	30	36	46
	(486)	(648)	(970)	(80)	(108)	(160)	(36)	(48)	(70)
$\rho^2_{\boldsymbol{x}y\|t}$ = 0.9	398	414	448	66	70	76	28	30	32
	(486)	(648)	(970)	(80)	(108)	(160)	(36)	(48)	(70)

Power = 0.6

δ	0.2			0.5			0.8		
p_{mis}	0.2	0.4	0.6	0.2	0.4	0.6	0.2	0.4	0.6
$\rho^2_{\boldsymbol{x}y\|t}$=0.3	582	732	1032	96	122	172	40	50	72
	(616)	(820)	(1230)	(104)	(138)	(206)	(44)	(58)	(86)
$\rho^2_{\boldsymbol{x}y\|t}$=0.6	544	628	800	90	104	134	38	44	56
	(616)	(820)	(1230)	(104)	(138)	(206)	(44)	(58)	(86)
$\rho^2_{\boldsymbol{x}y\|t}$ = 0.9	506	526	570	84	88	94	34	36	40
	(616)	(820)	(1230)	(104)	(138)	(206)	(44)	(58)	(86)

Table A.1: Some sample size tables

These are the sample size tables for required power equals to 0.5 and 0.6. The values in the tables are the required sample sizes (n) when MI is used; the values in the brackets are the corresponding required sample size under the same conditions if CCA is used instead. For this table, $m = 50$, the two samples are of the same size $n/2$.

Bibliography

Rebecca Andridge and Katherine Jenny Thompson. Using the fraction of missing information to identify auxiliary variables for imputation procedures via proxy pattern-mixture models. *International Statistical Review*, 83(3):472–492, 2015. ISSN 1751-5823. doi: 10.1111/insr.12091. URL http://dx.doi.org/10.1111/insr.12091. 10.1111/insr.12091.

Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011. doi: 10.1002/mpr.329. URL http://dx.doi.org/10.1002/mpr.329.

T. Baguley. Understanding statistical power in the context of applied research. *Applied Ergonomics*, 2004.

Richard S. Balkin and Carl J. Sheperis. Evaluating and reporting statistical power in counseling research. *Journal of Counseling Development*, 89(3):268–272, 2011. ISSN 1556-6676. doi: 10.1002/j.1556-6678.2011.tb00088.x. URL http://dx.doi.org/10.1002/j.1556-6678.2011.tb00088.x.

Jonathan W. Bartlett, Ofer Harel, and James R. Carpenter. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American Journal of Epidemiology*, 182(8):730–736, 2015. doi: 10.1093/aje/kwv114. URL http://dx.doi.org/10.1093/aje/kwv114.

A. A. Beaujean. Sample size determination for regression models using monte carlo methods in r. *Practical Assessment, Research and Evaluation*, 19(12), 2014.

T. E. Bodner. What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4):651–675, 2008. doi: 10.1080/10705510802339072. URL https://doi.org/10.1080/10705510802339072.

British Medical Journal. Resources for readers: 7. the t tests. URL http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/7-t-tests.

J. R. Carpenter and M.G. Kenward. *Multiple Imputation and its application*. Wiley, 2013.

S. Champely, C. Ekstrom, P. Dalgaard, J. Gill, J. Wunder, and H. D. Rosario. *Basic Functions for Power Analysis*, 8 2015.

J. Cohen. *Statistical Power Analysis for Behavioral Science*. Routledge, 2nd edition edition, 1988.

Linda M. Collins, Joseph L. Schafer, and Chi-Ming Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001.

A. Davey and J. Savla. *Statistical Power Analysis with Missing Data*. Routledge, 1st edition, 2010.

Adam Davey and Jyoti Savla. Estimating statistical power with incomplete data. *Organizational Research Methods*, 12(2):320–346, 2009. doi: 10.1177/1094428107300366. URL `https://doi.org/10.1177/1094428107300366`.

M Delacre, D Lakens, and C. Leys. Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Phychology*, 30(1): 92–101, 2017. doi: http://doi.org/10.5334/irsp.82.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

Van der Sluis S, Dolan CV, and Posthuma D. Neale MC. Power calculations using exact data simulation: A useful tool for genetic study designs. *Behavior Genetics*, 38:202–211, 2008.

M. Desai, D. A. Esserman, M. D. Gammon, and M. B. Terry. The use of complete-case and multiple imputation-based analyses in molecular epidemiology studies that assess interaction effects. *Epidemiol Perspectives and Innovations*, 8, 2011.

R. B. Donald. Multiple imputations in sample surveys - a phenomenological bayesian approach to nonresponse. In *Survey Research Methods Section of the American Statistical Association*, pages 20–28, 01 1978.

J. D. Elashoff. nquery advisor® version 7.0 user's guide. 2007.

C.K. Enders. *Applied Missing Data Analysis*. Methodology in the Social Sciences. Guilford Publications, 2010. ISBN 9781606236406. URL https://books.google.com/books?id=5o-yNRJTQ1EC.

F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39:175–191, 2007a.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007b. ISSN 1554-3528. doi: 10.3758/BF03193146. URL https://doi.org/10.3758/BF03193146.

Chris Fraley. On computing the largest fraction of missing information for the em algorithm and the worst linear function for data augmentation. *Computational Statistics Data Analysis*, 31(1):13 – 26, 1999. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/S0167-9473(99)00003-1. URL http://www.sciencedirect.com/science/article/pii/S0167947399000031.

G.S. Ginsburg, K.L. Drake, JY Tein, R Teetse, and M.A. Riddle. Preventing onset of anxiety disorders in offspring of anxious parents: A randomized controlled trial of a family-based intervention. *The American Journal of Psychiatry*, 172:1207–1214, December 2015.

John Graham and Stewart Donaldson. Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. 78: 119–28, 03 1993.

John W Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

J.W. Graham, A.E. Olchowski, and T.D. Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8:206–213, 2007.

M. H. Hansen, W. N. Hurwitz, and W. G Madow. *Sample Survey Methods and Survey*. Wiley, 1st edition, 1953.

O. Harel. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, 4:75–89, January 2007a.

O. Harel. *Strategies for Data Analysis with Two Types of Missing Values: From Theory to Application*. LAP Lambert Academic Publishing, 2009.

O. Harel and XH Zhou. Multiple imputation: review of theory, implementation and software. *Statistical in Medicine*, 26, July 2007.

Ofer Harel. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, 4(1):75 – 89, 2007b. ISSN 1572-3127. doi: https://doi.org/10.1016/j.stamet.2006.03.002. URL `http://www.sciencedirect.com/science/article/pii/S1572312706000153`.

IBM Corp. *IBM SPSS Statistics for Windows, Version 22.0*. IBM Corp., Armonk, NY, 2013.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2nd edition edition, 2002.

Roderick J. A. Little. Missing-data adjustments in large surveys. *Journal of Business Economic Statistics*, 6(3):287–296, 1988. ISSN 07350015. URL http://www.jstor.org/stable/1391878.

Andrea Marshall, Douglas G Altman, Roger L Holder, and Patrick Royston. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC medical research methodology*, 9(1):1, 2009.

J. McGinniss and O. Harel. Multiple imputation in three or more stages. *Journal of Statistical Planning and Inference*, 176:33 – 51, 2016. ISSN 0378-3758. doi: http://dx.doi.org/10.1016/j.jspi.2016.04.001. URL http://www.sciencedirect.com/science/article/pii/S0378375815300720.

Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input (disc: p558-573). *Statistical Science*, 9:538–558, 1994.

Xiao-Li Meng and Martin Romero. Discussion: Efficiency and self-efficiency with multiple imputation inference. *International Statistical Review / Revue Internationale de Statistique*, 71(3):607–618, 2003. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403832.

D Moher, CS Dulberg, and GA Wells. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA*, 272(2):122–124, 1994. doi: 10.1001/jama.1994.03520020048013. URL +http://dx.doi.org/10.1001/jama.1994.03520020048013.

Geert Molenberghs, Caroline Beunckens, Cristina Sotto, and Michael G. Kenward. Every missingness not at random model has a missingness at random counterpart with

equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371–388, 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2007.00640.x. URL http://dx.doi.org/10.1111/j.1467-9868.2007.00640.x.

Tim P. Morris, Ian R. White, and Patrick Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1):75, Jun 2014. ISSN 1471-2288. doi: 10.1186/1471-2288-14-75. URL https://doi.org/10.1186/1471-2288-14-75.

K.R. Murphy, B. Myor, and A. Wolach. *Statistical Power Analysis:A Simple and General Model for Traditional and Modern Hypothesis Tests*. Routledge, 1st edition edition, 1998.

Linda K. Muthén and Bengt O. Muthén. How to use a monte carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4):599–620, 2002. doi: 10.1207/S15328007SEM0904_8. URL http://dx.doi.org/10.1207/S15328007SEM0904_8.

Utah USA NCSS, LLC. Kaysville. *PASS 15 Power Analysis and Sample Size Software (2017)*. 2017.

Søren Feodor Nielsen. Proper and improper multiple imputation. *International Statistical Review / Revue Internationale de Statistique*, 71(3):593–607, 2003a. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403831.

Søren Feodor Nielsen. [proper and improper multiple imputation]: Rejoinder. *International Statistical Review / Revue Internationale de Statistique*, 71(3):625–627, 2003b. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403834.

Randall M. Peterman. The importance of reporting statistical power: The forest decline and acidic deposition example. *Ecology*, 71(5):2024–2027, 1990. ISSN 00129658, 19399170. URL `http://www.jstor.org/stable/1937612`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models, 2001.

J. P. Reiter. Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95:933–946, 2008.

C.H. Rhoads. Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy*, 3(1), 2012.

D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1st edition, 1987.

D.B. Rubin. An overview of multiple imputation. *In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association.*, 1988.

Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

Donald B. Rubin. Discussion on multiple imputation. *International Statistical Review / Revue Internationale de Statistique*, 71(3):619–625, 2003. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403833.

Donald B. Rubin and Nathaniel Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394):366–374, 1986. doi: 10.1080/01621459.1986.10478280. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478280.

Graeme D. Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mannâ"whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006. doi: 10.1093/beheco/ark016. URL +http://dx.doi.org/10.1093/beheco/ark016.

SAS. *SAS/STAT 9.2 User's Guide*. SAS, Cary, NC, 2008.

SAS Institute Inc. *SAS/STAT Software, Version 9.3*. Cary, NC, 2011. URL http://www.sas.com/.

Victoria Savalei and Mijke Rhemtulla. On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(3):477–494, 2012. doi: 10.1080/10705511.2012.687669. URL http://dx.doi.org/10.1080/10705511.2012.687669.

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall.CRC, 1st edition, 1997.

J. L. Schafer and J. W. Graham. Multiple imutation: our view of the state of art. *Psychological Methods*, 7(2):147 – 177, 2002.

Joseph L Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15, 1999.

Joseph L Schafer and Maren K Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4): 545–571, 1998.

Z.J. Shen. *Nested multiple imputation.* PhD thesis, Department of Statistics, Harvard University, 2000.

J. Siddique, J. P. Reiter, A. Brincks, R. D. Gibbons, C. M. Crespi, and C. H. Brown. Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. *Statistics in Medicine*, 34:3399–3414, 2015.

StataCorp. *STATA POWER AND SAMPLE-SIZE REFERENCE MANUAL RELEASE 13*, 2013.

R.J. Steidl, J.P. Hayes, and E. Schauber. Statistical power analysis in wildlife research. *The Journal of Wildlife Management*, 61(2):270–279, 1997.

Jonathan Sterne, Ian R. White, John B Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. 338, 2009. doi: https://doi.org/10.1136/bmj.b2393.

ML Tang and NS Tang. Exact tests for comparing two paired proportions with incomplete data. *Biometrical Journal*, 46(1):72–82, 2002.

Yongqiang Tang. Closed-form reml estimators and sample size determination for mixed effects models for repeated measures under monotone missingness. *Statistics*

in Medicine, 36(13):2135–2147, 2017. ISSN 1097-0258. doi: 10.1002/sim.7270. URL http://dx.doi.org/10.1002/sim.7270. sim.7270.

M. Templ and P. Filzmoser. Visualization of missing values using the r-package vim. *Research report cs-2008-1, Department of Statistics and Probability Theory, Vienna University of Technology.*, 2008. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-1complete.pdf.

S. van Burren. *Flexible Imputation of Missing Data*. Chapman and Hall.CRC, 1st edition, 2012.

S. van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, 2007. doi: 10.1177/0962280206074463.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles*, 45(3):1–67, 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03. URL https://www.jstatsoft.org/v045/i03.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. URL http://www.jstatsoft.org/v45/i03/.

G Verbeke and G Molenberghs. chapter 21. Springer New York, New York, NY, 2000.

Gerko Vink, Laurence E. Frank, Jeroen Pannekoek, and Stef van Buuren. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68

(1):61–90, 2014. ISSN 1467-9574. doi: 10.1111/stan.12023. URL http://dx.doi.org/10.1111/stan.12023.

P.T. von Hippel. How many imputations do you need? a two-stage calculation using a quadratic rule. *Sociological Methods and Research*, 2018, in press.

James Wagner. The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly*, 74(2):223, 2010. doi: 10.1093/poq/nfq007. URL +http://dx.doi.org/10.1093/poq/nfq007.

I. R. White and J. B. Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate valuese size for planned missing designs. *Statistics in Medicine*, 29:2929–2931, December 2010.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011a.

Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011b.

Erika J. Wolf, Kelly M. Harrington, Shaunna L. Clark, and Mark W. Miller. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73(6):913–934, 2013. doi: 10.1177/0013164413495237. URL https://doi.org/10.1177/0013164413495237.

X. Xie and X. Meng. Dissecting multiple imputation from a multi-phase inference perspective: what happens when gods, imputers and analysts models are uncongenial? *Statistica Sinica*, 2014. doi: 10.5705/ss.2014.067.