UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
School of Mathematical Sciences

Optimal Design of Experiments for MNAR data

by

Adetola Adedamola Adediran
MTech, BTech
ORCiD: 0000-0003-3176-7872

A thesis for the degree of
Doctor of Philosophy

January 2025

University of Southampton

Abstract

Faculty of Social Sciences
School of Mathematical Sciences

Doctor of Philosophy

Optimal Design of Experiments for MNAR data

by Adetola Adedamola Adediran

The presence of missing data leads to biases in data analyses. To overcome these biases, it is crucial to understand the type of missing data that is present in the data. Amongst the three types of missing data (known as missing data mechanisms) that will be formally introduced in this thesis, the Missing Not At Random (MNAR) mechanism is the most complex. MNAR poses the most difficulties as it is an untestable assumption based on the current incomplete data. A recovery of some of the missing data is required to test its presence. In this research, we developed two statistical tests for testing the presence of MNAR in datasets and provide the theoretical framework of the tests. In the first test, the recovery design consists of a random sampling of the responses whose covariates lie within a particular region while the second test is based on an assignment of probabilities. We introduced techniques from Design of Experiments to improve the properties of these tests. The developed tests are compared with a random follow-up of missing responses, which will act as our benchmark design throughout. We formulate an easy and simple conjecture that uses the empirical density of the covariates to obtain the recovery region. Through simulations, the performance of the tests is evaluated.

Keywords: Missing data; Missing not at random; Selection model; Recovery region; Conjecture.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
    i Optimal follow up - an integrated approach to test for missing not at random - submitted to Journal of the American Statistical Association.
    ii An integrated approach to test for missing not at random - submitted to Journal of the American Statistical Association.
    iii An integrated approach to test for missing not at random - submitted to arXiv as arXiv:2208.07813 (15 Aug 2022).
    iv An efficient algorithm to detect MNAR missingness - International Society for Business and Industrial Statistics (ISBIS) 2023 Conference at Brock University, St. Catharines, Canada (July 13 - 14, 2023).
    v Designing follow-up samples: a comprehensive approach to detect MNAR missingness - Model-oriented Data Analysis and Optimum design (mODa) workshop at the University of Southampton, UK (July 9 - 14, 2023)

vi Designing follow up samples – A comprehensive approach to detect MNAR missingness efficiently - International Conference on Design of Experiments (ICODOE), Memphis, Tennessee, U.S.A (May 8 - 11, 2023).

vii Comparing recovery sample designs to test for the presence of MNAR - International Workshop on Statistical Modelling, Trieste, Italy (July 18-22, 2022).

Signed: Date:22/12/2024

Acknowledgements

Thanks to God Almighty for seeing me through this study. Special thanks to my Supervisors: Dr. Anthony Overstall, Prof Dankmar Bohning and Prof. Stefanie Biedermann for their amazing support and advice during this program. I would like to specially appreciate Prof. Stefanie Biedermann and Dr. Robin Mitra for their immense research guidance. Special appreciation to Dr. Jack Noonan, I am very grateful for your mentorship and collaboration during this research period.

My immense gratitude goes to the Commonwealth Scholarship Commission (CSC) for the fully funded scholarship I received. Without this financial support, it would have been nearly impossible for me to study at the University of Southampton, UK. I am grateful to my parents and siblings for their love and support in my academic pursuits. I would like to thank the Federal University of Technology, Akure (FUTA) for granting me a study leave to pursue my PhD. Also, to Prof. Femi Adebola, Prof. Olatunde Adeoti, Prof. Adegoke Ajiboye and Dr. Olusegun Ewemooje, I am grateful for their fatherly support.

To my darling husband Dr. Kehinde Ogunade, thank you for your love and support during the tough period encountered during this research period. I cannot but appreciate my friends on this doctoral journey Akinleye Folorunsho, Wole Ademola and Adedoyin Agunbiade for their encouragement.

Thank you all and God bless you.

To God Almighty and my parents, Elder Isaac and Mrs. Dolapo Adediran.

Chapter 1

Introduction

The importance of data for decision-making in today's world cannot be overemphasized. The majority of scientific and industrial processes involve data collection, analyses and interpretation of the collected data (Carpenter and Kenward, 2012). Data can occur in various forms, such as the overall population of a country, the number of illegal immigrants in a country, the maternal mortality rate in a community in a given year, the number of live births nationwide, the recovery rate of patients from a disease, and the effectiveness of a vaccine, among other illustrations. Data collection is essential in decision-making as conclusions can only be drawn from available data. Data can be collected or gathered from two sources: a primary source and a secondary source. The primary source of data involves the researcher or interviewer collecting data from the source i.e. originally getting the required data from respondents (Ajayi, 2017). Secondary data involves collecting previously gathered data (Ajayi, 2017). In the process of collecting the required data, there are chances that some of the required information would not be available or missing. This is known as missing data.

Missing data are defined as unavailable information that is required and useful for analysis if they were to be available (Little and Rubin, 2019). Missing data poses a problem in many research areas that involve data collection, either from primary source or secondary sources. It is often unlikely not to have missing data when dealing with data collection and its presence can introduce bias in the inferences drawn (Carpenter and Kenward, 2007). Missing data is referred to as non-response in surveys, missing or loss of results in experiments and attrition in longitudinal studies (Little and Rubin, 2002). Missing data occurs in surveys as non-response when respondents refuse to provide the information required by the interviewer or researcher, or when the respondent cannot be reached to get the required information (Cobben, 2009). Some of this information could be: investment, income, participants in a programme not revealing their age or pregnant women refusing to attend antenatal on some days among others. In surveys, there are two types of non-response: unit non-response and item non-response. Unit non-response occurs when the required information about a respondent is unavailable

(impossible to contact the respondent or the respondent refuses to provide all required information) while item non-response occurs when a respondent provides some, but not all, of the required information (Yan and Curtin, 2010). Asking questions about a sensitive attribute can lead to non-response or false responses. As an example, if asked the question "Are you a cultist?", many individuals would not provide a truthful answer due to many reasons such as fear of being exposed to the law or fear of stigmatization. This situation would most likely lead to the presence of evasive answers. When some responses are missing, this automatically reduces the sample size and hence, the accuracy of estimates is reduced. Warner (1965) developed the randomized response model used in estimating the proportion of people that belong to a sensitive attribute. Non-response in surveys has gained recognition and a lot of researchers have further expanded the work of Warner by reducing the non-response errors in surveys, see (Greenberg et al., 1969; Kim and Warde, 2004; Mangat, 1994; Adebola and Adepetun, 2011; Adebola et al., 2017; Adediran et al., 2020; Ewemooje et al., 2018) amongst others.

Aside from surveys, other processes that involve the use of data can also experience missing data. In agriculture, an experimenter could lose data on some units if they forget to record the results. Dropouts in clinical trials are also examples of how missing data can be present (Little and Rubin, 2019). Incorrect analysis of missing data can lead to bias and a loss in efficiency. It is not always possible to have all the required information and therefore analysing missing data correctly is important. Since the 1950s, a lot of useful statistical literature on missing data has been in existence because of its effect on analysis and inference (Carpenter and Kenward, 2012). Little and Rubin (2002) discussed diverse methods for handling missing data problems. The type of missing data problem present would inform which of the methods is appropriate to be used. To correctly analyse and make correct inferences in the presence of missing data, a classification of these problems known as Missing Data Mechanisms (MDMs) was introduced by Rubin (1976). The choice of method to analyse a particular missing data problem depends heavily on the type of missing mechanism present, hence, the importance of the classification (Little and Rubin, 2002). These mechanisms are Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) and will be discussed in detail in Section 2.1.

MCAR is the simplest among these mechanisms and can be easily analysed using standard statistical methods. In most cases, the mechanism is frequently assumed to be MAR; this assumption is frequently made out of necessity for simpler analysis rather than with a clear conviction that it is true. When the original incomplete sample is MNAR, the analysis results in biases in the study and is an untestable assumption (Little and Rubin, 2002). The complication of MNAR is that (without strong assumptions) it is an untestable assumption based on the original incomplete dataset. This is because the probability of missingness depends on the missing observations. Many researchers

have been able to work on MCAR and MAR, since these two MDMs are easier to detect and analyse. The complexity of MNAR and the need to develop more comprehensive approaches to address the problems it poses motivated this research. If some of the missing values can be recovered, then the presence of MNAR can be tested. Recovery could be in the form of a follow-up through surveys or telephone calls to patients to get the missing information. The majority of methods for handling missing data deal with this problem post hoc; this is usually due to their inability to be implemented or lack of planning. It is a recommended practice to account for missing values before data collection, according to a well-conducted scientific study and common sense in general, this allows the study design to be effectively planned appropriately. (Imhof et al., 2002; Lee et al., 2018a,b, 2019) are some of the studies that developed theories to construct optimal study designs that take into consideration the possibility of non-response and the possible benefits.

In this research, we investigate how the recovery of missing observations can facilitate tests for MNAR. Avoiding unverifiable assumptions and planning a follow-up sample to recover some of the missing values is the most effective way to learn about the existence of MNAR. Follow-up sampling, also known as double sampling in survey sampling (Elliott et al., 2000; Guan et al., 2018; Miao et al., 2021; Alho, 1990; Drew and Fuller, 1980; Qin and Follmann, 2014) or repeated attempt design in design of experiments (Jackson et al., 2010; Aronow et al., 2015; Daniels et al., 2015; Coppock et al., 2017), has been identified as a useful method to address missing data problems. In this work, we develop a framework that involves constructing follow-up sample designs to optimise the ability of a statistical test to detect the presence or absence of MNAR. The follow-up sample is designed such that it significantly improves the power of the test and will be compared to a random recovery of missing values which will often form our benchmark design. Additionally, we explore the efficiency and robustness of the designs through simulation studies. Finally, we formulate a conjecture design that is less computationally expensive, easier to understand and robust to finding efficient designs.

1.1 Aim and objectives

This research is aimed at developing an efficient framework that detects MNAR data.

The objectives of this research are to:

i develop a statistical test for MNAR with well understood theoretical properties;

ii incorporate Design of Experiments (DOE) techniques to improve the test's properties;

iii assess the performance of the optimised designs; and

iv assess the robustness of the designs.

1.2 Report Structure

The structure of this report is as follows. In Chapter 2, we present a literature review on missing data, missing data mechanisms, existing ways of handling missing data problems, follow-up sampling, experimental design and optimal designs. In Chapter 3, we provide a simulation study on some existing ways of handling the problem of missing data and highlight that determining the correct mechanism is crucial for correct inference, thus highlighting the importance of a test. In Chapter 4, we provide our novel research on dealing with MNAR, describing the tests for the presence of MAR vs MNAR and discuss a Type I error problem encountered when using an existing test present in the literature. In Chapter 5, we develop a consistent test for MNAR and develop an algorithm for improving its properties. Here, we also assess the robustness of this algorithm. In Chapter 6, we propose a new testing and design framework based on subsampling probabilities and assess the robustness of the designs. In Chapter 7, we consider higher-dimensional problems using conjectures formulated in Chapter 6. In Chapter 8, we demonstrate how incorporating a test for MNAR can improve a particular estimation problem, thus further motivating our research. In Chapter 9, we provide a summary of our findings and future research. The main theoretical advancements in the area of MNAR are shown in Chapters 5 and 6.

Chapter 2

Literature Review

One of the biggest challenges in data mining and data analysis projects is dealing with the presence of missing data (Silva and Zárate, 2014). This chapter can be divided into two parts. Part one provides a literature review of the field of missing data and the most popular approaches to tackle missing data issues; an area of research that has gained significant attention since the 1950s (Carpenter and Kenward, 2012; Little and Rubin, 2019; Alho, 1990; Rubin, 1976). Part two provides a literature review of experimental design. The two areas of research are inherently connected as will become clear throughout the chapter.

This chapter acknowledges and provides an investigation on some research connected to missing data and the optimal design of experiments.

2.1 Missing Data Mechanisms

Rubin (1976) categorised missing data into three mechanisms, which are based on the relationship between the missing and observed values. Understanding the concept of missing data mechanisms helps to identify the right analysis to be used (Fielding et al., 2009). Using the notation and definition in Little and Rubin (2019), let $X = (x_{ij})$ be an $n \times p$ matrix with no observation missing, such that the ith row $x_i = (x_{i1}, \ldots, x_{ip})$, where x_{ij} represents the value of X_j for unit i. Let $M = (m_{ij})$ represent the missing indicator matrix such that x_{ij} may be observed or missing and denote the rows of M by $m_i = (m_{i1}, \ldots, m_{ip})$. The missingness pattern is defined as:

$$m_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is missing} \\ 0 & \text{if } x_{ij} \text{ is not missing (observed).} \end{cases}$$

For simplicity, the rows (x_i, m_i) are assumed to be independently and identically distributed over i. How the conditional distribution of M given X, $f(M|X, \theta)$, where θ is

the unknown parameter of this distribution, can be expressed determines the classification of the MDMs. In classifying the MDMs in the following subsections, let X_{miss} be the missing elements of X and X_{obs} be the observed elements of X.

2.1.1 Missing Completely at Random (MCAR)

In the mechanism MCAR, the probability of a missing value is unrelated to the observed and unobserved data (Carpenter and Kenward, 2012). Under such a mechanism, the missing cases are a random subset of the full data set, i.e. there is no difference in the distribution of the complete cases and the missing cases in MCAR (Bhaskaran and Smeeth, 2014). A lost questionnaire in a survey and a damaged experimental unit in a field among others are examples of MCAR cases. According to (Little and Rubin, 2002), the mathematical representation of MCAR is:

$$f(M|X,\theta) = f(M|\theta) \quad \text{for all } X,\theta.$$

Determining the validity of the MCAR assumption can be obtained by comparing the distribution of the complete cases and missing cases on the observed values (Little and Schenker, 1995). A single test statistic for testing MCAR was developed by Little (1988). This method is widely acceptable in testing the presence of MCAR missingness. Little (1988) stated that when data is multivariate normal and the asymptotic null distribution is provided, the small-sample null distribution can be calculated. For bivariate data, the test simplifies to a conventional t-test (Little, 1988). The MCAR mechanism has the effect of allowing the missing mechanism to be disregarded and making the assumption that conclusions might be drawn without increasing the likelihood of complicating the missing data (Heitjan and Basu, 1996).

2.1.2 Missing at Random (MAR)

For the mechanism MAR, the probability of missingness is dependent on the observed values. This mechanism, like MCAR, assumes that missingness can be ignored (Rubin, 1973). Suppose in the random selection of a population where all elements have an equal chance of being selected in the sample, a survey on drug addiction is taken. If the gender is fully observed and the males are less likely to give the required information on drug addiction, then the missingness is MAR because the missingness depends on their gender and not on the severity of the addiction. This can be written mathematically as:

$$f(M|X,\theta) = f(M|X_{obs},\theta) \quad \text{for all } X_{miss},\theta.$$

MAR has been studied widely and several standard methods have been proposed to tackle this problem. Kenward and Molenberghs (1998) explained the frequentist approach to drawing conclusions based on likelihood under MAR, highlighting the elements of inference that necessitate taking the MDM into account. Lu (2004) extensively explained MAR using two definitions stating that "MAR is necessary and sufficient for Likelihood Ignorability (LIG)".

2.1.3 Missing not at Random (MNAR)

Under an MNAR mechanism, missingness is dependent on the missing values X_{miss} of X. Suppose in the survey of drug addiction, missingness occurs in those that are severely addicted, then we have missing not at random. This can be written mathematically as:

$$f(M|X,\theta) = f(M|X_{miss}, X_{obs}, \theta) \quad \text{for all } X_{miss}, \theta.$$

This mechanism is non-ignorable and must be correctly diagnosed before analysis. Analysing MNAR data is more complicated than other MDMs, as some important information is lost (unobserved) and as a result, some additional assumptions need to be tested before analysis. Hence, the MAR assumption is a common starting point for the analysis of clinical trials (Carpenter and Kenward, 2007). Without following up with the non-responders, it is impossible to determine the true mechanism for MNAR cases (Little and Rubin, 2002). Due to the untestable and restrictive assumptions of handling MNAR data, sensitivity analysis is often necessary (Briggs et al., 2003).

McPherson et al. (2015) compared three missing data strategies (MAR model, MNAR model and Wu–Carroll MNAR) in a clinical trial. It was concluded that an examination to see the connection of assumptions with the models and sensitivity analyses needs to be done in clinical trial research. In an attempt to reduce the bias of a regression analysis when data is MNAR, Tchetgen Tchetgen and Wirth (2017) proposed a simple CCA (see Section 2.2) with modification to the regression model of interest by including the instrumental variable design to account for selection bias meticulously. The approach was developed for the identity, log and logit link functions. Leurent et al. (2018) provided a tutorial for sensitivity analysis for MNAR using the pattern mixture framework with multiple imputation. A distinctive selection model-based method for analysing incomplete binary multilevel data with MNAR assumption was introduced by Hammon (2020). The proposed method performed better than existing methods specifically in terms of coverage and bias of the estimates of the parameters of interest.

2.2 Techniques For Handling Missing Data

The purpose of this subsection is to provide a flavour of the existing techniques in the literature to handle missing data when performing inference. We will not discuss each technique in great detail, but provide a high-level overview with suitable references.

Over the years, a lot of methods have been developed for handling missingness. Some of these are:

- Complete Case Analysis (CCA): this is also known as listwise deletion. In CCA, the missing cases are not taken into consideration and only the complete observed cases are analysed. This method leaves out the missing cases by analysing all the cases without missing values. Due to the simplicity of this method, it can be applied with standard statistical methods without any adjustments, which is an advantage (Little and Rubin, 2002). When this method is used, it is expected that the result should be similar to that of the complete data set i.e. it assumes that the complete cases represent a random subsample of the full data set (Hedges and Cooper, 2009). CCA performs satisfactorily for MCAR, however, for MAR and MNAR missingness, it may lead to bias because the observed cases are not representative of the full dataset (Hedges and Cooper, 2009). Its major disadvantage is bias and loss of precision as some of the data which are valid are discarded (Kang, 2013).
- Available case analysis: also known as pairwise deletion reduces the number of deleted cases by using all the available cases rather than discarding all cases with missing values, which makes it a better option than CCA (Baraldi and Enders, 2010). This method's limitation is that distinct sample subsets are obtained for various variables in the dataset, and it can only function satisfactorily for MCAR while bias may result when data are MAR and MNAR (Hedges and Cooper, 2009).

.

2.2.1 Single Imputation Methods

Single imputation methods involve the replacement of a missing value with an imputed value and analysing the data set as a complete data set. Rather than ignoring the missing values like CCA and available case analysis do, a replacement is done by imputation in single imputation methods. There are different types of single imputation methods, some of which are discussed as follows.

- The Unconditional Mean Imputation (UMI): this is the simplest imputation method. The means of the observed values are used in replacing the missing values (Little and Rubin, 2019). For example, if there are 300 observed values and 50 missing values, the mean of the 300 observed values is used to replace the 50 missing values. The mean of the variables are preserved in this method, however, the data set has less variability (Briggs et al., 2003). This method of imputation generally results in biased regression coefficients and invalid inferences (Enders, 2022). This also occurs in MCAR, which is the simplest missing mechanism because it imputes the same value for all missing values at the centre of the distribution (Enders, 2022).

- Conditional Mean Imputation (CMI): this imputation method is an improvement on the unconditional mean imputation. This method is conditional as it replaces the missing values with the conditional means given observed values (Briggs et al., 2003). A regression analysis is fitted based on the complete cases, and the missing values are replaced by the predictions from the regression (Enders, 2022). This method is better than CCA, available case analysis and unconditional mean because variation has been introduced in the distribution. The disadvantage of this method is that it increases the correlation between variables (Enders, 2022).

- Hot deck imputation: this is a single imputation method that handles missingness by replacing each missing value with a value from observed cases that are related to the missing value (Andridge and Little, 2010). A replacement of the missing values is done with "one or more variables for a non-respondent (called the recipient) with observed values from a respondent (the donor) that is similar to the non-respondent with respect to characteristics observed by both cases" (Andridge and Little, 2010), i.e the missing case would be replaced with the value of an observed case that falls in the same class with the missing case (Kalton and Kish, 1981). The restriction of the donors to fully complete variables results in the preservation of multivariate relationships (Marker et al., 2002).

- Substitution: this is another method of handling missing data problems commonly used in survey sampling. Suppose in a survey, some of the sampled unit provides the required information (respondents) while some do not give the required information (non-respondents). This method replaces a non-respondent with a unit or variable that was not initially sampled (Little and Rubin, 2019).

- Posterior Predictive Distribution Draw (PPDD): was developed as an improvement over the existing single imputation methods. It uses the Bayesian framework to draw from the posterior predictive distribution of the variable (Gelman et al., 2004). This is currently the best-performing single imputation method.

2.2.2 Principled Missing Data Methods

The principled missing data methods are methods of analysing missing data that use the information available from the observed data with statistical assumptions to assess the MDM and estimate the parameters of interest, rather than directly replacing the missing values (Dong and Peng, 2013). These methods when correctly applied, aid in recovering the underlying inferential model and the validity of a study is maximized (Lang and Little, 2018). Examples of principled missing data methods are Multiple Imputation and Maximum Likelihood.

2.2.2.1 Multiple Imputation

Multiple imputation addresses the overconfidence associated with single imputation by incorporating a degree of uncertainty into the imputed data set (Sterne et al., 2009). The Bayesian framework serves as motivation for multiple imputation and the general approach for this type of imputation is "to impute using the posterior predictive distribution of the missing data given the observed data and some estimate of the parameters" (Jamshidian and Mata, 2007). This method was proposed by Rubin (1978) and clarification was provided by Rubin (2004). This method replaces each missing value with a vector of length $M \geq 2$. The M vectors create M completed data sets from the vectors of imputations such that the first vector component replaces the missing value to obtain the first complete data set, the second data set is obtained when the second vector of imputation replaces the missing values until M complete data sets are obtained (Little and Rubin, 2019). Each imputed data set is analysed as complete data using statistical methods, and the resulting M complete data analyses can be combined to obtain an inference that reflects sampling variability because of the missing values (Little and Rubin, 2019). According to Little and Rubin (2019), generate M imputed datasets and analyse using standard statistical methods that are available to use in the absence of missingness. Let $\hat{\theta}_m$ be the parameter of estimation for the mth complete data set and W_m represents the sample variance estimate associated to $\hat{\theta}_m$ for $m = 1, 2, 3, ..., M$ for each of the M imputations i.e. the variance that would be present in the sample when there are no missing data. The combined estimate using Rubin's combining rule is:

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m.$$

The variance of $\bar{\theta}_M$ can be decomposed into two components: average within-imputation variation ($\overline{W}_M$) and between-imputation variation ($\bar{B}_M$),

$$\overline{W}_M = \frac{1}{M}\sum_{m=1}^{M} W_m,$$

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2.$$

The total variation associated with $\bar{\theta}_M$ is calculated as:

$$T_M = \overline{W}_M + \frac{M+1}{M}B_M.$$

2.2.2.2 Maximum-Likelihood (ML)

Using all of the available data points, the variance-covariance matrix for each variable in the model can be obtained using the ML approach (Soley-Bori, 2013). The model's regression parameters can then be estimated using the acquired variance-covariance matrix (Soley-Bori, 2013). The ML method is simple because it only requires the model specification, unlike the MI approach, which requires some decisions to be made beforehand such as the number of iterations to be utilised, the choice of prior distribution, among others. (Soley-Bori, 2013).

Full Information Maximum-Likelihood (FIML)

The FIML is also called the direct Maximum Likelihood. This method suggests that, under the assumed linear model, the multivariate normal likelihood function is directly maximised (Soley-Bori, 2013). Imputation does not occur in FIML unlike MI. It uses the maximization of the likelihood function of the observed values to obtain the parameter estimates (Dong and Peng, 2013). When the missing mechanism is MAR and the joint distribution of all the variables are multivariate normal, the FIML method estimates are unbiased (Dong and Peng, 2013). An advantage of this method is that it is more efficient than the existing methods of analysing missing data (Soley-Bori, 2013). The disadvantage is its computational complexity as it is not easy to compute (Brown, 1981).

Expectation Maximization (EM)

The EM is an ML method introduced by Dempster et al. (1977). This is an iterative method for obtaining MLE when data are missing (Little and Rubin, 2019). Every iteration in this method contains an expectation step and a maximization step, repeatedly until the maximum likelihood estimates are obtained (Dempster et al., 1977). The

expectation step obtains "the conditional expectation of the missing data given the observed data and current estimated parameters, and then substitutes these expectations for the missing data" (Little and Rubin, 2019) while the maximization step involves performing ML estimation of the parameter of interest like when there is no missing data present (Dempster et al., 1977). This method can be used when the sample size is large and MAR is assumed (Soley-Bori, 2013). EM is less demanding computationally than the FIML, however, a disadvantage of this method is that it cannot be used for other models except linear and log-linear models (Soley-Bori, 2013).

2.3 Follow-up Sampling

The MNAR missing mechanism requires a recovery of some missing values to be able to test its presence (Little and Rubin, 2019). Follow-up sampling has been used in various fields to address the problem of missing data. In order to obtain the required information and increase response rate, follow-up has been extensively used in survey sampling and observational studies (Miao et al., 2021).

In survey sampling, Elliott et al. (2000) studied the effect of subsampling callbacks on survey efficiency. In this research, recovery was done by randomly sampling from the callback units initially sampled, this is because a lot of cost is associated with the callbacks needed for a small recovery proportion. Hence, subsampling the callbacks shows that there is a reduction in interview costs and an increase in the efficiency of collected data. For nonignorable nonresponse with callbacks, Guan et al. (2018) proposed a semiparametric maximum likelihood estimator. The proposed method was applied to survey data with missing responses and results show that the estimator is more efficient than existing methods. Alho (1990) used the logistic regression model to reduce the nonresponse bias in sample surveys with an assumption that there is at least one callback. This paper provided inspiration for some of the results of Chapter 6.

For randomised experiments, a combined method that utilises double-sampling and worst-case bounds to address missing data problems was proposed by Aronow et al. (2015). This method makes use of little assumptions (allowing the presence of missing responses in the recovery sample) by relaxing the double-sampling assumption that all nonrespondents followed up would be recovered. Their method reveals that there is a significant reduction in uncertainty in controlled randomised experiments when double-sampling is used. An et al. (2009) worked on the need for double-sampling in survival studies and its application to the President's Emergency Plans for AIDS Relief (PEPFAR). They considered four methods in this study: one without double-sampling and three with double-sampling. The methods with double-sampling yielded higher estimates than the method without double-sampling. The research shows the importance of double-sampling for accurate data collection when data is missing. Daniels

et al. (2015) proposed the Repeated Attempt Pattern Mixture Model (RAM-PMM). This method uses the pattern mixture for modelling in repeated attempt designs. In comparison to previously existing models used for modelling repeated attempt data, this method provides flexibility and transparency in identifying parameters and performs better.

Carpenter and Kenward (2012) formulated two tests for MNAR vs MAR, but no further information on their respective properties were given. To the best of our knowledge, these are the only tests for MNAR vs MAR based on follow-up sampling in the literature. These tests will be formally introduced in Chapter 4.

2.4 Experimental Design

Atkinson et al. (2007) defined "a well-designed experiment as an efficient method of learning about the world". According to (Oehlert, 2010), the treatments, experimental units, assignment of treatments to the units, and observed responses are the elements that define an experiment. Experimentation is an indispensable part of scientific method. A set of guidelines that shows how experimental units are distributed among the treatments is known as experimental design (Dean and Voss, 1999) i.e. a rule that determines the allocation of treatments to units. Many industrial and medical processes involve experimentation. The basic designs of experiments are completely randomized design (CRD), randomized complete block design (RCBD) and Latin square design (LSD). Several researchers have developed other designs like the Incomplete Block Design (IBD), Factorial Design, and Balanced Incomplete Block Design (BIBD) among others. There are three basic principles of experimentation. A detailed explanation was provided by Montgomery (2017):

i Randomization: treatments should be applied to experimental units randomly such that each treatment has an equal chance of being allocated to any of the experimental units. It helps in providing a basis for inference and statistical testing among treatments. The absence of randomization in a process or an experiment may lead to the presence of systematic bias (Dean and Voss, 1999), hence, randomization helps in bias reduction.

ii Replication: this involves independently allocating treatments to different experimental units. To precisely measure the accompanying variability and the effects of interest in an experiment by repeating the experimental settings or condition is known as replication (Dean and Voss, 1999).

iii Blocking: this is also known as grouping or stratification. When the experimental units are heterogeneous in nature, blocking helps in grouping the population into smaller groups as homogeneous as possible. These smaller groups are called blocks

and the experimental units in each group are similar (Dean and Voss, 1999). Blocking reduces variability and increases precision.

2.4.1 Optimal Design

Rady et al. (2009) defined optimal designs as "experimental designs that are generated based on a particular optimality criterion and are generally optimal only for a specific statistical model". In statistical analysis, parameter estimation with minimal variance and lack of bias is possible with optimal designs (Fasoranbaku and Daramola, 2018). The fundamental principle underlying the theory of experimental design is that careful choice of the control variables can strengthen the statistical inference of the quantities of interest (Chaloner and Verdinelli, 1995). For example, a control variable in a chemical experiment could be the temperature at which the reaction of interest is run. A design would then be the set of temperatures for the different runs of the reaction. The information matrix can be used to assess the accuracy of an experimental design (Oladugba and Madukaife, 2009). It is assumed that the reader has some knowledge of Experimental design to prevent the replication of material. For an excellent introduction to the design of experiments, see Montgomery (2017) and Dean and Voss (1999).

2.4.1.1 Optimality Criteria

Oladugba and Madukaife (2009) defined optimality criterion as "a single-valued measure that determines how good a design is, and it is maximized or minimized by an optimal design". Optimality criteria can be defined as criteria that tell us about the quality of a design (Rady et al., 2009). There are several types of optimality criteria and the choice of optimality criterion depends on the purpose of the experiment, for example, precise estimation or prediction, or achieving high power of a test. Let ξ^* denote an optimal design with respect to the optimality criterion under consideration.

Consider the regression model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\mathbf{Y}$ is a vector of observed responses, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of unknown parameters and $\boldsymbol{\epsilon}$ is the vector of random errors. The Fisher information can be calculated by specifying the likelihood function based on the distribution of the errors $\boldsymbol{\epsilon}$. Next, construct the likelihood function $L(\boldsymbol{\beta}; \mathbf{Y})$ according to the specified distribution, then, the log-likelihood function $\log L(\boldsymbol{\beta}; \mathbf{Y})$ is derived. Calculate the first and second derivatives of the log-likelihood with. Lastly, the Fisher information can be computed using $I(\theta) = \mathbb{E}\left[\left(\frac{\partial \log L(\theta;\mathbf{Y})}{\partial \theta}\right)^2\right]$ or $I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log L(\theta;\mathbf{Y})}{\partial \theta^2}\right]$. Note that the theory outlined below also holds for non-linear and generalised linear models.

Let χ be the design region, and $\{x_1, x_2, \ldots, x_m\}$ be the support points of the design. Lee et al. (2018a) defined a continuous design ξ as a set of weights $w_1, w_2, \ldots, w_m$ assigned

to these support points, where $w_i \geq 0$ for all i, and the weights sum to 1. Mathematically defined as:

$$\xi = \left\{ \begin{array}{ccc} x_1 & \dots & x_m \\ w_1 & \dots & w_m \end{array} \right\} \tag{2.1}$$

subject to the conditions:

$$\sum_{i=1}^{m} w_i = 1 \quad \text{and} \quad w_i \geq 0 \quad \text{for all} \quad i.$$

The design points $x_1, \dots, x_m$ belong to the design space χ, and the weights w_i represent the proportion of total observations allocated to each design point.

Let Σ represent the class of all continuous designs, which consists of all possible choices of ξ that satisfy the design constraints and $F = F(\xi)$ be the Fisher information matrix of the regression model defined above. Let $\lambda_1, ..., \lambda_p$ be the eigenvalues of $F(\xi)$.

Rady et al. (2009) gave a detailed explanation on the various types of optimality criteria. Some of the discussed criteria are:

i A optimality: uses the Fisher's information matrix, the average variance of the parameter estimate is minimized by this optimality (Atkinson et al., 2007). This optimality can be defined as:

$$\min_{\xi \in \Sigma} \ trace(F)^{-1}.$$

The efficiency of the design ξ is:

$$A(\xi) = \frac{tr[F^{-1}(\xi_A^*)]}{tr[F^{-1}(\xi)]}.$$

In terms of eigenvalues, (Atkinson et al., 2007) defined A optimality as:

$$\min_{\xi \in \Sigma} \sum_{i=1}^{p} \frac{1}{\lambda_i}.$$

ii C Optimality: minimizes the variance of the linear combination of the model parameters $C^T\theta$, where C is a vector of known constants (Atkinson et al., 2007), i.e.

$$\min_{\xi \in \Sigma} \ var(C^T\hat{\theta}).$$

The C-efficiency is given as:

$$C(\xi) = \frac{C^T F^{-1}(\xi_c^*)C}{C^T F^{-1}(\xi)C}.$$

iii D Optimality: This is the most widely and commonly studied design criterion. In this design, the determinant of the information matrix is maximized.

$$\max_{\xi \in \Sigma} |F| = \min_{\xi} |(F)^{-1}|.$$

Maximising the determinant of the information matrix is equivalent to minimising the volume of the confidence ellipsoid for the parameter vector θ, thus ensuring that the estimates of the model parameters are as precise as possible. Atkinson et al. (2007) defined D optimality in terms of eigenvalues as:

$$\min_{\xi \in \Sigma} \prod_{i=1}^{p} \frac{1}{\lambda_i}.$$

iv E Optimality: this optimality criterion is concerned with maximizing the minimum eigenvalue of the information matrix (Rady et al., 2009):

$$\max_{\xi \in \Sigma} \lambda_{min}(F)^{-1} = min\ \lambda_{max}(F)^{-1}.$$

Atkinson et al. (2007) defined E optimality in terms of eigenvalues as:

$$\min_{\xi \in \Sigma} \max_{i} \frac{1}{\lambda_i}.$$

v G Optimality: this criterion is also known as a prediction criterion and it involves the minimization of the maximum variance of a predicted response value, $\hat{y}_x$, over the experimental space (Rady et al., 2009). This is defined as:

$$\min_{\xi \in \Sigma} \max_{x \in \chi} var(\hat{y}_x).$$

vi I Optimality: this is also known as the integrated variance, it minimizes the integrated prediction variance (Rady et al., 2009). Goos et al. (2016) defined this optimality criterion as a design that "minimizes the average prediction variance over the experimental region χ" and mathematically defined it as:

$$\text{Average variance} = \frac{1}{\int_\chi dx} \cdot tr[F^{-1}B],$$

where

$$B = \int_{\chi} t(x)t'(x)dx$$

is called the moments matrix, with x representing the mixture of proportions and $t(x)$ is the vector of model terms. Here, dx denotes an infinitesimal volume element in the domain χ over which integration is performed.

vii T Optimality: is an optimality criterion that is used in discriminating between two or more models where one of the models is true (Rady et al., 2009). When discriminating between two models, T optimality is defined as:

$$\max_{\xi \in \Sigma} \Delta_2(\xi) = \int [\eta_t(x) - \eta_2(x, \hat{\theta}_2(\xi))]^2 \xi(dx),$$

with $\eta_t(x)$ as the true model and $\hat{\theta}_2$ is the second model's parameter estimates. $\Delta(\xi)$ represents the non-centrality parameter and x is a set of independent variables. In Chapter 6, we show how T_E-optimality is used in our algorithm for design constructions that maximize the power of our test for MNAR. Waterhouse et al. (2008) formulated T_E-optimality, which is similar to T-optimality but has more appealing statistical properties, such as an asymptotic chi-square distribution under the null hypothesis. T_E-optimality selects a continuous design that maximizes the expected reduction in deviance. T_E optimality has specifically been constructed for comparing two generalised linear models, which makes it so useful for our work in Chapter 6. According to Waterhouse et al. (2008), a design $\xi^*_{T_E}$ is said to be T_E-optimal if:

$$\mathbb{E}\{R(\xi^*_{T_E}, X)\} = \max_{\xi \in \Xi} \mathbb{E}\{R(\xi, X)\},$$

where the reduction deviance R is defined as:

$$R(\xi, x) = D_1(\xi, x) - D_2(\xi, x) = 2\{l(\xi, \hat{\pi}_2, x) - l(\xi, \hat{\pi}_1, x)\},$$

with D_1 and D_2 as the deviance and $\hat{\pi}_1$ and $\hat{\pi}_2$ represent the maximum likelihood estimates of π_1 and π_2 from the unknown parameters of the compared models M_1 and M_2 respectively.

In a correspondence research between D and D_L optimal designs to see if a design that is D optimal is also D_L optimal, Oladugba and Madukaife (2009) introduced D_L optimal design where the determinant of the loss of information matrix is maximized. A numerical consideration was carried out in the regular geometric experimental region and irregular geometric experimental region. The result showed that for a bivariate linear response function, there is a correspondence between D and D_L optimality in a regular experimental region with or without blocking. In an irregular experimental region, correspondence doesn't always exist between D and D_L optimality.

While there is an extensive body of literature on optimal design, there are only very few papers that take missing data into account at the design stage. Baek et al. (2006) proposed the Bayesian optimal designs for a quantal dose-response study with potentially missing observations. Ghosh (1979) worked on the robustness of designs against incomplete data, Lee et al. (2018a) proposed an optimal design for experiments with possibly incomplete observations. Imhof et al. (2002) developed a framework applicable to linear and non-linear models to obtain optimal designs in the presence of missing observations using the expectation of the information matrix. Lee et al. (2018b) extended the work of Imhof et al. (2002) by considering missing not at random data.

2.4.2 Definitions

In this section, we introduce some statistics that will be used in Chapter 3 to examine the performance of some selected missing data analysis methods on the three missing data mechanisms. For a general estimator $\hat{\theta}$ of θ, let $\hat{\theta}_r$ be the estimated value of θ ($\hat{\theta}$) for the r^{th} replicate. Let $\bar{\theta}$ be the arithmetic mean of all $\hat{\theta}_r$. We then define the following:

- Bias: denoted as B can be defined as the difference between the expected value of the parameter and the true value of the parameter (Walther and Moore, 2005).

$$\textbf{Bias}(\bar{\theta}) = \bar{\theta} - \theta.$$

- Coverage: denoted as C Walther and Moore (2005) defined this as the proportion of times that the true parameter is contained in the 95% confidence interval. Casella and Berger (2024) defined coverage as:

$$\text{Coverage Probability} = \mathbb{P}(L(X) \leq \theta \leq U(X)),$$

  where θ is the true value of the parameter of interest, $(L(X)$ and $U(X))$ represent the lower and upper bounds of a confidence interval for θ based on a random sample Y respectively.

- Percentage Bias: denoted as PB measures how the bias of an estimator compares to the true parameter in percentage.

$$PB = 100 \times \frac{\text{Bias}}{\theta}.$$

- Estimated Variance: denoted as EV

$$EV = \frac{1}{R} \times \sum_{r=1}^{1000} (s.e.(\hat{\theta}_r))^2,$$

s.e is the standard error defined as $\sqrt{Var(\hat{\theta}_r)}$.

- True Variance denoted as TV

$$TV = \frac{1}{R-1} \times \sum_{r=1}^{1000} (\hat{\theta}_r - \bar{\theta})^2,$$

- Variance Ratio: denoted as VR is the ratio of the Estimated Variance to the True Variance.

$$VR = \frac{\text{Estimated Variance}}{\text{True Variance}}.$$

We considered VR to know how accurate our estimation is. The farther the VR value from 1, the less accurate our estimation.

- Mean Squared Error: denoted as MSE

$$MSE(\theta) = (\text{Bias})^2 + \text{EV}.$$

Chapter 3

Review of Some Existing Methods of Handling Missing Data Mechanisms

In this chapter, some existing methods of handling missing data mechanisms highlighted in Section 2.2 are studied. We review the performance of each method on the different types of missing mechanisms. We will demonstrate that while some methods can handle MCAR and MAR missingness, none of these methods are able to adequately deal with MNAR missing mechanism.

3.1 Simulation study

The following scenario is used to illustrate the current methods of handling missing data mechanisms. With 1000 replicates, 1000 observations from a normal distribution were simulated in the following way:

$$X_{ij} \sim N(\mu_X, \sigma_X), \tag{3.1}$$

$$Y_{ij}|(X_{ij} = x) \sim N(\beta_0 + \beta_1 x, \sigma_Y). \tag{3.2}$$

For each of the three Missing Data Mechanisms, we will introduce circa 30% missingness in the response variable y (the exact mechanism used will be discussed later). We will be interested in estimating the following four parameters: mean of X (μ_X), mean of Y (μ_Y) and the regression coefficients β_0 and β_1 using the samples X_{ij} where $i = 1, \ldots, R$ and $j = 1, \ldots, N$ generated randomly for this study, the corresponding response variable Y_{ij}, and the known missingness indicators m_{ij}. This estimation

will be done using the Complete Case Analysis (CCA), Unconditional Mean Imputation (UMI), Conditional Mean Imputation (CMI) and Posterior Predictive Distribution Draw (PPDD) methods discussed in Section 2.2. Let $N = 1000$, $R = 1000$, $\mu_X = 5$, $\sigma_X = 1$, $\beta_0 = 1$, $\beta_1 = 2$ and $\sigma_Y = 4$. These values are chosen arbitrarily to examine the performance of the existing methods of handling missing data problems on the three missing mechanisms. The performance of each imputation method on the summary statistics was studied and compared with the full data summary statistics.

3.2 Full data analysis

The summary statistics for the full data were obtained and compared to the summary statistics obtained using different missing data techniques on each of the three Missing Data Mechanisms that will be discussed shortly. In Table 3.1, the coverages for the means and regression parameters are all close to 0.95 and the estimated variances are close to the true variances. The variance ratios are approximately 1. The biases are low leading to low Percentage bias, and the MSE for the four parameters are low resulting from the small biases and variances.

TABLE 3.1: Summary Statistics for Full Data Analysis.

θ	μ_X	μ_Y	β_0	β_1
Coverage	0.943	0.950	0.958	0.959
Bias	-0.001	-0.001	0.008	-0.042
PB	-0.012	-0.012	0.849	-0.087
EV	0.001	0.008	0.112	0.004
TV	0.001	0.008	0.106	0.004
VR	0.959	0.995	1.056	1.035
MSE	0.001	0.008	0.106	0.004

3.3 Missing Completely at Random

The summary statistics for different methods of analysis when data is Missing Completely at Random (MCAR) is shown in Table 3.2.

- **Coverage:** For μ_X, all methods perform similarly with coverage close to 0.95. For μ_Y, full data, Complete Case Analysis (CCA), and Posterior Predictive Distribution Draw (PPDD) maintain coverage around 0.95, but Unconditional Mean Imputation (UMI) shows a significant reduction to 0.836, while Conditional Mean Imputation (CMI) shows moderate improvement at 0.903. For β_0 and β_1, the full

data and CCA methods show high coverage (close to 0.95), while UMI performs poorly with coverage of 0.000 for both. CMI and PPDD slightly recover coverage, but still fall short compared to full data and CCA.

- **Bias:** Bias for μ_X and μ_Y is very low across all methods. For β_0, UMI introduces a large positive bias of 2.986, while full data and CMI show much smaller biases. PPDD introduces a negative bias of -0.209. Similarly, for β_1, UMI has a substantial negative bias of -0.597, whereas the other methods show minimal bias, with full data and CCA having the smallest values.

- **Percentage Bias (PB):** For μ_X and μ_Y, percentage bias remains small across methods, except for PPDD, which introduces a higher PB for μ_Y (0.209). UMI displays a very large percentage bias for β_0 (298.637%) and a significant negative percentage bias for β_1 (-29.857%). Other methods have much smaller percentage biases.

- **Estimated Variance (EV):** EV for μ_X and μ_Y is consistent across methods, except for slight variations in CCA and UMI for μ_Y. For β_0, full data has the largest EV, followed by CCA and PPDD. UMI shows lower EV, while CMI has the smallest EV. Similarly, for β_1, the estimated variances are generally low, with CMI showing the smallest value.

- **True Variance (TV):** TV values for μ_X are stable across methods, while for μ_Y, UMI has a slightly larger TV compared to the others. For β_0, UMI again shows higher TV, while CMI presents the highest true variance. Other methods remain comparable to full data. TV for β_1 follows a similar pattern to EV.

- **Variance Ratio (VR):** The variance ratio for μ_X is near 1 for all methods, with UMI and PPDD having slightly higher values. For μ_Y, UMI displays a notably low VR (0.511), while CMI also shows a reduced ratio compared to the other methods. For the regression parameters β_0 and β_1, UMI and CMI result in lower VRs than other methods, indicating discrepancies between the estimated and true variances. Full data, CCA, and PPDD maintain VR values closer to 1.

- **Mean Squared Error (MSE):** The MSE for μ_X is stable and low across all methods. For μ_Y, UMI and CCA show slightly higher MSE values compared to full data. UMI dramatically increases the MSE for β_0 (9.036) and β_1 (0.361) due to large biases. The other methods maintain much lower MSE values, with CMI showing the smallest errors for β_0 and β_1 among the imputation methods.

Full Data performs best across all statistics, as expected. CCA maintains solid performance with minimal bias, variance ratios close to 1, and stable MSE values. It is effective when data is MCAR, especially given that the 300 missing values have little impact on the results. UMI introduces significant bias and increases MSE, especially for the regression parameters β_0 and β_1, making it inefficient for restoring summary

statistics. CMI improves upon UMI with reduced bias and variance but still falls short of achieving optimal results, especially in variance ratio and MSE. PPDD slightly improves coverage and variance estimates for some parameters but still introduces bias, especially for μ_Y and β_0. CCA is the best approach when dealing with MCAR data in this analysis, outperforming imputation methods like UMI, CMI, and PPDD, which struggle to fully restore the summary statistics.

TABLE 3.2: MCAR Summary Statistics.

Statistic	θ	Full Data	CCA	UMI	CMI	PPDD
Coverage	μ_X	0.943	0.947	0.952	0.950	0.944
	μ_Y	0.950	0.956	0.836	0.903	0.950
	β_0	0.958	0.946	0.000	0.839	0.899
	β_1	0.959	0.949	0.000	0.834	0.878
Bias	μ_X	-0.001	0.000	-0.000	-0.001	0.001
	μ_Y	-0.001	0.000	0.000	-0.001	0.023
	β_0	0.008	-0.008	2.986	0.004	-0.209
	β_1	-0.042	0.001	-0.597	-0.001	0.046
PB	μ_X	-0.012	0.008	-0.006	-0.014	0.028
	μ_Y	-0.012	0.002	0.002	-0.012	0.209
	β_0	0.849	-0.796	298.637	0.351	-20.857
	β_1	-0.087	0.073	-29.857	-0.034	-2.287
EV	μ_X	0.001	0.001	0.001	0.001	0.001
	μ_Y	0.008	0.011	0.008	0.007	0.008
	β_0	0.112	0.149	0.095	0.073	0.102
	β_1	0.004	0.006	0.004	0.002	0.004
TV	μ_X	0.001	0.001	0.001	0.001	0.001
	μ_Y	0.008	0.011	0.011	0.010	0.008
	β_0	0.106	0.148	0.117	0.139	0.105
	β_1	0.004	0.006	0.004	0.005	0.004
VR	μ_X	0.959	0.959	1.039	1.039	1.060
	μ_Y	0.995	1.043	0.511	0.714	1.023
	β_0	1.056	1.004	0.807	0.523	0.968
	β_1	1.035	1.025	0.843	0.510	0.959
MSE	μ_X	0.001	0.001	0.001	0.001	0.001
	μ_Y	0.008	0.011	0.011	0.010	0.008
	β_0	0.106	0.148	9.036	0.139	0.149
	β_1	0.004	0.006	0.361	0.005	0.006

3.4 Missing at Random

We introduce 30% missingness using the following expit function. We consider a Missing at Random (MAR) missing data mechanism of the form:

$$g(x) = \frac{\exp(\alpha_0 + \alpha_1 x)}{1 + \exp(\alpha_0 + \alpha_1 x)}, \tag{3.3}$$

with $\alpha_0 = -3$ and $\alpha_1 = 0.42$ resulting in approximately 300 missing values of y. This is a MAR mechanism because the missingness depends only on x. Figure 3.1 shows the scatterplot for x and y when data is missing at random. In this figure, the red points correspond to the missing values and the blue points represent the observed values. This figure shows that the missing values are widely spread and not concentrated at a point. About 46% of highest 100 x values have missing y and 18% of lowest 100 x values have missing y.

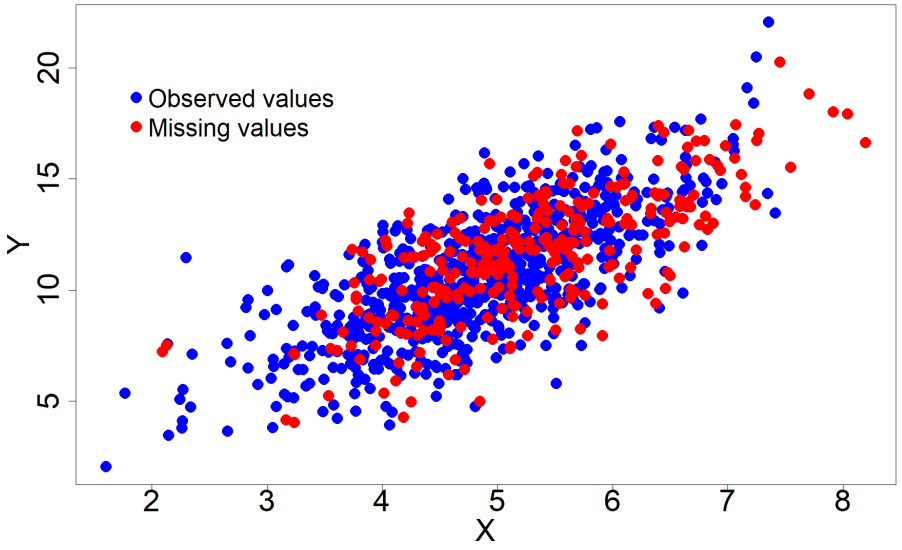

FIGURE 3.1: Scatterplot of y against x for Missing at Random.

Table 3.3 shows the summary statistics for MAR using some existing methods of handling missing data problems.

- **Coverage:** The Full Data, UMI, CMI, and PPDD methods exhibit high coverage values for μ_X around $0.94 - 0.96$. However, CCA's coverage for μ_X is significantly low at 0.110. For μ_Y, Full Data and PPDD maintain high coverage around 0.95, while CCA's coverage is moderate at 0.390, and UMI has the lowest at 0.195. In the case of β_0, Full Data and CCA show high coverage values of 0.958 and 0.965, respectively, whereas UMI reports zero coverage, with CMI and PPDD showing

slightly lower values at 0.860 and 0.850. For β_1, similar trends are observed, with Full Data and CCA reporting high coverage (0.959 and 0.966), while UMI again reports zero coverage, and CMI and PPDD have moderately lower values.

- **Bias:** The Full Data and UMI methods show nearly no bias for μ_X with values between -0.001 and 0.000, while CCA has a considerable negative bias of -0.120. For μ_Y, UMI has a high negative bias of -0.238, whereas the other methods show minimal bias. For β_0, UMI displays a significant positive bias of 2.953, while other methods show small biases ranging from -0.008 to 0.008. For β_1, UMI shows a notable negative bias of -0.638, with the other methods exhibiting small biases ranging from -0.042 to 0.046.

- **Percentage Bias (PB):** In this category, CCA and UMI have large negative percentage bias values for μ_X, reported as -2.406 and -0.001, respectively, while other methods show small percentage biases. For μ_Y, UMI and CCA exhibit significant negative percentage biases of -2.164, while other methods have smaller values. For β_0, UMI shows a substantial positive percentage bias of 295.281, while other methods range from -0.713 to 0.849. Lastly, for β_1, UMI has a significant negative percentage bias of -31.905, with the other methods displaying much smaller values, from -0.087 to 0.100.

- **Estimated Variance (EV):** In terms of estimated variance, all methods have similar small EV values for μ_X around 0.001. For μ_Y, EV ranges between 0.006 and 0.011 across all methods, with slight variability. For β_0, the highest EV is in Full Data at 0.112, while CMI has the lowest at 0.073, with other methods falling between 0.095 and 0.146. The values for β_1 across all methods remain small, ranging from 0.002 to 0.006.

- **True Variance (TV):** All methods show small true variance values for μ_X around 0.001. The values for μ_Y range from 0.008 to 0.011, with CMI and PPDD showing slightly larger values. For β_0, UMI exhibits the highest true variance at 0.132, while other methods vary from 0.105to 0.139. For β_1, true variance values are low across all methods, ranging from 0.004 to 0.008.

- **Variance Ratio (VR):** In terms of variance ratios, UMI and PPDD exhibit slightly inflated ratios at 1.039 and 1.060, while the other methods remain close to 1.0 for μ_X. For μ_Y, Full Data shows a high variance ratio at 0.995, whereas UMI is low at 0.491, with other methods ranging from 0.715 to 1.043. For β_0, CMI and PPDD demonstrate reduced variance ratios at approximately $0.537 - 0.551$, while Full Data displays higher values at 1.056. Lastly, for β_1, UMI has a low variance ratio of 0.685, while Full Data and CCA exhibit values closer to 1.0.

- **Mean Squared Error (MSE):** All methods report a small mean squared error for μ_X, around 0.001. The MSE for μ_Y is highest in UMI at 0.068, while other methods

range between 0.008 and 0.011. For β_0, UMI reports a significantly high MSE of 8.851, whereas other methods show MSE values ranging from 0.106 to 0.195. Finally, for β_1, UMI also displays the highest MSE at 0.413, while other methods have low MSE values between 0.004 and 0.008.

Overall, the Full Data method consistently performs the best, with low bias, high coverage, and low mean squared error across all statistics. The CCA method displays very low coverage for μ_X and μ_Y while maintaining good performance on β_0 and β_1. The UMI method reports significant biases and mean squared errors, particularly for β_0 and β_1, suggesting poor performance. Finally, the CMI and PPDD methods demonstrate moderate performance, balancing coverage and mean squared error, with some variability across the metrics, although they exhibit lower variance ratios.

TABLE 3.3: MAR Summary Statistics.

Statistic	θ	Full Data	CCA	UMI	CMI	PPDD
Coverage	μ_X	0.943	0.110	0.945	0.956	0.955
	μ_Y	0.950	0.390	0.195	0.903	0.920
	β_0	0.958	0.965	0.000	0.860	0.850
	β_1	0.959	0.966	0.000	0.853	0.826
Bias	μ_X	-0.001	-0.120	0.000	0.001	-0.001
	μ_Y	-0.001	-0.238	-0.238	0.001	0.001
	β_0	0.008	-0.007	2.953	0.002	0.005
	β_1	-0.042	0.002	-0.638	-0.001	-0.000
PB	μ_X	-0.012	-2.406	-0.001	0.020	-0.027
	μ_Y	-0.012	-2.164	-2.164	0.010	0.012
	β_0	0.849	-0.713	295.281	0.221	0.517
	β_1	-0.087	0.100	-31.905	-0.031	-0.009
EV	μ_X	0.001	0.001	0.001	0.001	0.001
	μ_Y	0.008	0.011	0.006	0.007	0.008
	β_0	0.112	0.146	0.096	0.074	0.105
	β_1	0.004	0.006	0.004	0.003	0.004
TV	μ_X	0.001	0.001	0.001	0.001	0.001
	μ_Y	0.008	0.011	0.011	0.010	0.011
	β_0	0.106	0.134	0.132	0.134	0.195
	β_1	0.004	0.005	0.005	0.005	0.008
VR	μ_X	0.959	0.995	0.991	1.044	1.086
	μ_Y	0.995	0.994	0.491	0.715	0.740
	β_0	1.056	1.092	0.729	0.551	0.537
	β_1	1.035	1.097	0.685	0.524	0.495
MSE	μ_X	0.001	0.016	0.001	0.001	0.001
	μ_Y	0.008	0.068	0.068	0.010	0.011
	β_0	0.106	0.134	8.851	0.133	0.195
	β_1	0.004	0.005	0.413	0.005	0.008

3.5 Missing Not at Random

We introduce 30% missingness using the following expit function. This is an example of a Missing Not at Random (MNAR) missing mechanism because the missing values depend on Y. This is sometimes called a self-censoring MNAR mechanism.

$$g(x,y) = \frac{\exp(\alpha_0 + \alpha_1 y)}{1 + \exp(\alpha_0 + \alpha_1 y)}. \tag{3.4}$$

The use of $\alpha_0 = -5$ and $\alpha_1 = 0.358$ results in approximately 300 missing values of Y. Figure 3.2 shows the scatterplot for x and y for MNAR, 51% of highest 100 x values have missing y and 19% of lowest 100 x values have missing y implying that the missing values are more at the upper end than the lower end.

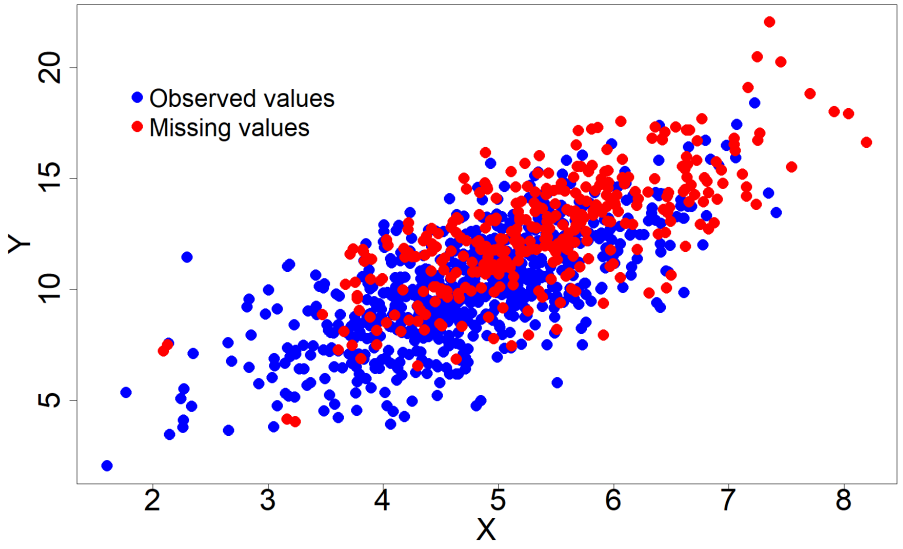

FIGURE 3.2: Scatterplot of y against x for Missing Not at Random.

Table 3.4 shows the comparison of the full data analysis and some methods of analysing missing data.

- **Coverage:** The Full Data method exhibits consistently high coverage across all parameters, with values ranging from 0.943 to 0.959. UMI also performs well for μ_X (0.950) but shows zero coverage for μ_Y, β_0, and β_1. CMI and PPDD demonstrate moderately high coverage for μ_X (0.961 and 0.950, respectively) and fair coverage for β_0 (0.616 and 0.711, respectively). However, both have low coverage for μ_Y and β_1. CCA performs poorly in terms of coverage, particularly with μ_X and μ_Y (0.001 and 0.000, respectively), although it has somewhat better coverage for β_0 (0.819) and β_1 (0.462).

- **Bias:** Full Data shows negligible bias for all statistics, with values close to 0. UMI has minimal bias for μ_X (-0.002) but large biases for μ_Y (-0.709), β_0 (3.226), and β_1 (-0.786), indicating poor performance. CCA demonstrates significant negative bias for μ_X (-0.180) and μ_Y (-0.709), and smaller but still notable biases for β_0 and β_1. CMI and PPDD have smaller biases across the parameters but show some issues with μ_Y and β_0.

- **Percentage Bias (PB):** Full Data and PPDD exhibit the smallest percentage bias across all statistics. UMI, on the other hand, demonstrates substantial percentage bias, particularly for β_0 (322.579%) and β_1 (-39.323%). CCA shows large percentage biases for μ_X (-3.592%) and μ_Y (-6.448%), as well as smaller percentage biases for the other parameters. CMI shows moderately high percentage bias for μ_Y and β_0.

- **Estimated Variance (EV):** All methods show similar small estimated variances for μ_X, μ_Y, and β_1, with values near 0.001. Full Data and PPDD exhibit the highest EV for β_0 (0.112 and 0.096, respectively), while CMI and PPDD have slightly lower EV values across the parameters.

- **True Variance (TV):** True variance values remain consistently small for μ_X (0.001) across all methods. For μ_Y, true variance values range from 0.008 to 0.010, with CMI and PPDD showing slightly larger values. True variance for β_0 is highest in CMI and PPDD (0.141 and 0.156, respectively), while Full Data has a lower value (0.106). True variance for β_1 is small across all methods.

- **Variance Ratio (VR):** UMI shows a slightly inflated variance ratio for μ_X (1.004), while other methods hover around 1.0. For μ_Y, Full Data and CCA exhibit higher variance ratios close to 1.0, while UMI shows a much lower ratio (0.490). Variance ratios for β_0 are highest in Full Data and CCA (1.056 and 1.074, respectively), and lowest in UMI (0.783). Variance ratios for β_1 are similar across the methods.

- **Mean Squared Error (MSE):** Full Data and PPDD demonstrate the lowest MSE values across all statistics. UMI, however, reports notably high MSE values, particularly for β_0 (10.518) and β_1 (0.623). CCA has a high MSE for μ_Y (0.513) but remains competitive with lower MSE values for other statistics. CMI and PPDD exhibit moderate MSE values, particularly for μ_Y and β_0.

As we have seen in this chapter, the existing methods of handling missing data problems do not perform satisfactorily for the MNAR missing mechanism. To correctly test for the presence of MNAR and analyse MNAR data, there is a need to recover some or all of the missing observations. Therefore, in the next chapter, we will look at the recovery approach to test the presence of MNAR vs MAR.

TABLE 3.4: MNAR Summary Statistics.

Statistic	θ	Full Data	CCA	UMI	CMI	PPDD
Coverage	μ_X	0.943	0.001	0.950	0.961	0.950
	μ_Y	0.950	0.000	0.000	0.006	0.010
	β_0	0.958	0.819	0.000	0.616	0.711
	β_1	0.959	0.462	0.000	0.240	0.357
Bias	μ_X	-0.001	-0.180	-0.002	0.000	-0.002
	μ_Y	-0.001	-0.709	-0.709	-0.379	0.382
	β_0	0.008	0.390	3.226	0.386	0.372
	β_1	-0.042	-0.154	-0.786	-0.153	-0.150
PB	μ_X	-0.012	-3.592	-0.043	0.006	0.034
	μ_Y	-0.012	-6.448	-6.448	-3.446	-3.470
	β_0	0.849	39.026	322.579	38.631	37.182
	β_1	-0.087	-7.678	-39.323	-7.659	-7.504
EV	μ_X	0.001	0.001	0.001	0.001	0.001
	μ_Y	0.008	0.010	0.005	0.006	0.007
	β_0	0.112	0.136	0.088	0.068	0.096
	β_1	0.004	0.006	0.003	0.003	0.004
TV	μ_X	0.001	0.001	0.001	0.001	0.001
	μ_Y	0.008	0.010	0.010	0.009	0.010
	β_0	0.106	0.127	0.132	0.141	0.156
	β_1	0.004	0.005	0.005	0.006	0.007
VR	μ_X	0.959	0.936	1.004	1.076	0.994
	μ_Y	0.950	0.995	0.490	0.682	0.742
	β_0	1.056	1.074	0.783	0.503	0.617
	β_1	1.035	1.098	0.752	0.464	0.563
MSE	μ_X	0.001	0.034	0.001	0.001	0.001
	μ_Y	0.008	0.513	0.513	0.153	0.155
	β_0	0.106	0.279	10.518	0.284	0.294
	β_1	0.004	0.029	0.623	0.029	0.029

Chapter 4

Testing for MNAR using a recovery sample

This chapter focuses on how a recovery of missing responses can be used to test the presence of MNAR. There are two tests for MAR vs MNAR in the literature and these tests can produce erroneous type 1 error rates. We will attempt to explain why in this chapter. We will also explore the use of different recovery designs and recovering different proportions of the missing responses. For simplicity, we will consider scenarios with one covariate in this chapter. Generalisations to an arbitrary number of covariates will be presented in Chapters 5 and 6.

4.1 Hypothesis Testing

Consider a scenario with a univariate response, Y and one covariate X, where

$$Y|(X = x) \sim N(\beta_0 + \beta_1 x, \sigma_y^2)\,. \tag{4.1}$$

For now, we will assume β_0, β_1 and σ_y^2 are known. It is assumed that the independent variable X is fully observed. In contrast, the response variable Y may contain missing values with M being the missing indicator such that it is 1 when Y is missing and 0 if observed. Let $y_1, \ldots, y_n$, $x_1, \ldots, x_n$ and $m_1, \ldots, m_n$ denote samples of size n from the model in (4.1). Let n_{obs} be the number of observed cases, n_{miss} represents the number of missing observations such that $n_{obs} + n_{miss} = n$. Without loss of generality, we assume the first n_{obs} of $y_1, \ldots, y_n$ are observed, meaning $y_{n_{obs}+1}, \ldots, y_n$ are all initially missing. Let $\mathbf{M} := \{n_{obs+1}, n_{obs+2} \ldots, n\}$ be the set of indices whose y values are missing. Assume resources permit follow up of a number of experimental units with missing responses to obtain (recover) their responses, e.g. through home visits to patients in

a clinical trial or follow-up telephone calls in a survey. We denote the number of recovered responses by n^*, where $n^* \leq n_{miss}$. We assume the choice of which responses to recover is in the practitioner's control, and thus this gives rise to the concept of a recovery design.

Definition 4.1. A recovery design $\mathbf{D} = \mathbf{D}(n^*)$ is a subset of size n^* from $\mathbf{M} := \{n_{obs+1}, n_{obs+2} \ldots, n\}$.

The recovery design will instruct the experimenter what missing values to recover as follows. For notation, relabel the observations such that the recovery design becomes $\mathbf{D} = \{k_1, k_2, \ldots, k_{n^*}\}$. The recovered responses from the initial missing responses become the elements of the vector $\mathbf{Y_R} := (Y_{k_1}, \ldots, Y_{k_{n^*}})$ and $\mathbf{Y_O} := (Y_1, Y_2, \ldots, Y_{n_{obs}})$ represent the observed responses. The augmented data is a combination of the observed cases and the recovered cases $\mathbf{Y_A} := (\mathbf{Y_O}, \mathbf{Y_R})^T$. In the same vein, the covariates and the missing indicator are: $\mathbf{X_A} := (\mathbf{X_O}, \mathbf{X_R})^T$ with $\mathbf{X_R} := (X_{k_1}, \ldots, X_{k_{n^*}})$ and $\mathbf{X_O} := (X_1, \ldots, X_{n_{obs}})$; $\mathbf{M_A} := (\mathbf{M_O}, \mathbf{M_R})^T$ with $\mathbf{M_O} = (0, \ldots, 0)$ and $\mathbf{M_R} = (1, \ldots, 1)$.

In the popular book by Carpenter and Kenward (2012), two tests for MNAR vs MAR were formulated for testing if the missing data mechanism is MAR vs MNAR. These tests could also be used in making inferences on the model parameters. The two tests are:

1. Selection Model framework (SMF): Fit a logistic regression of the missing indicator on the independent and response variables:

$$logitPr(M_A = 1) = \alpha_0 + \alpha_1 Y_A + \alpha_2 X_A + \alpha_3 X_A Y_A. \tag{4.2}$$

   Under the null hypothesis of MAR, we have $\alpha_1 = \alpha_3 = 0$, MAR is present if the hypothesis is true else otherwise.

2. Pattern mixture framework (PMF): This model fits a linear regression of the form:

$$\mathbb{E}(Y_A) = \beta_0 + \beta_1 X_A + \beta_2 M_A + \beta_3 X_A M_A. \tag{4.3}$$

   Under the null hypothesis of MAR, we have $\beta_2 = \beta_3 = 0$, MAR is present if the hypothesis is true else otherwise.

In addition to these hypotheses tests, we propose two tests based on just one parameter (excluding the interaction term), respectively.

For SMF, the model is:

$$logitPr(M_A = 1) = \alpha_0 + \alpha_1 Y_A + \alpha_2 X_A, \tag{4.4}$$

with $\alpha_1 = 0$ if MAR is true, otherwise MNAR. For PMF:

$$\mathbb{E}(Y_A) = \beta_0 + \beta_1 X_A + \beta_2 M_A, \tag{4.5}$$

with $\beta_2 = 0$ if MAR is true, otherwise MNAR.

4.2 Recovery Designs

In this section, we will consider four recovery designs. These designs serve as a preliminary study to assess the effect of different designs on the Type I error rate and power of the tests above. More principled ways of selecting recovery designs optimally will be discussed in detail in Chapters 5 and 6.

1. Random Sample Design: In this recovery design, a random sample of n^* is selected from the missing cases and added to the complete cases.
2. Highest Values Design: This design selects n^* highest values based on x as the sample.
3. Smallest Values Design: This design selects n^* smallest values based on x as the sample.
4. Half Highest/Half Smallest Values Design: This design selects the n^* sample such that $\frac{n^*}{2}$ is the highest values based on x and the other $\frac{n^*}{2}$ is the smallest values based on x.

Suppose we have 1000 observations of Y and X and Y contains approximately 300 missing values. Suppose further we can recover some responses from the approximately 300 missing values and Y_A is the augmented data comprising of the complete cases and the recovered cases. The recovered samples will be selected using the different designs above.

4.3 Simulation Studies

4.3.1 One-Parameter Model

We simulated the data with equations (3.1) and (3.2) with the missing mechanisms simulated according to equation (3.3) with $\alpha_0 = -3$ and $\alpha_1 = 0.42$ for MAR and equation (3.4) with $\alpha_0 = -3$ and $\alpha_1 = 0.19$ for MNAR. These α values are chosen such that there are approximately 300 missing values in the response variable Y. The Type I error and power analysis were computed for MAR and MNAR respectively using the models in

equations (4.4) and (4.5) under the different recovery designs and different recovered sample sizes. The purpose was to study how different recovery designs impact type 1 error and power.

The hypothesis for SMF is:

$$H_0 : \alpha_1 = 0$$
$$H_1 : \alpha_1 \neq 0.$$

The hypothesis for PMF is as follows:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0.$$

Figures 4.1 − 4.4 and Table A.1 shows the Type I error for MAR and power for MNAR using different recovery designs and frameworks. Using the Selection Model Framework for MAR, recovering the highest values of x is the worst design because it has Type I error values above 0.05 when less than 50% of the missing values are recovered. The smallest values design also gives Type I error values slightly above 0.05 when 10% of the missing values are recovered and approximately 0.05 at other sample sizes. The Random design and the half highest/half smallest designs are the best designs as all Type I error values are approximately 0.05 at all sample sizes. As the recovered sample size increases, the Type I error for the highest values design reduces. In Pattern Mixture Framework for MAR, the random design and the half highest/half smallest designs perform similarly as in SMF and all designs have Type I error values of 0.05 at all recovery sample sizes. Figure 4.1 shows the plot of the Type I error against the recovered sample sizes using SMF and Figure 4.2 for PMF, an increase in sample size from 30 to 50 leads to a reduction in the Type I error for all designs. For MNAR, under both frameworks, the random design has the highest power while the highest design has the least power among all the designs. The power increases and approaches 1 as n^* increases. This is shown graphically in Figures 4.3 and 4.4. These plots show that both SMF and PMF perform similarly for all the designs.

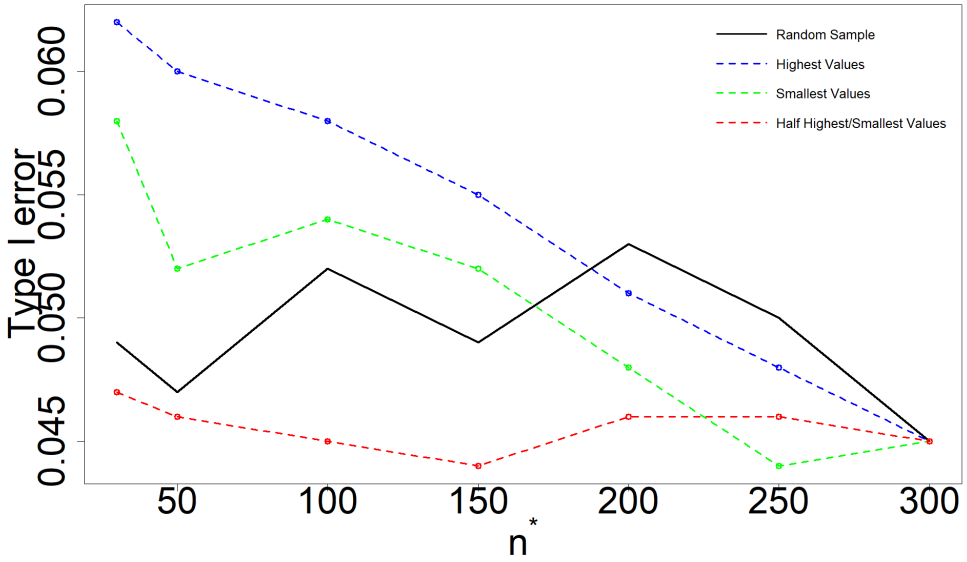

FIGURE 4.1: MAR Type I error plot using SMF for different recovery scenarios.

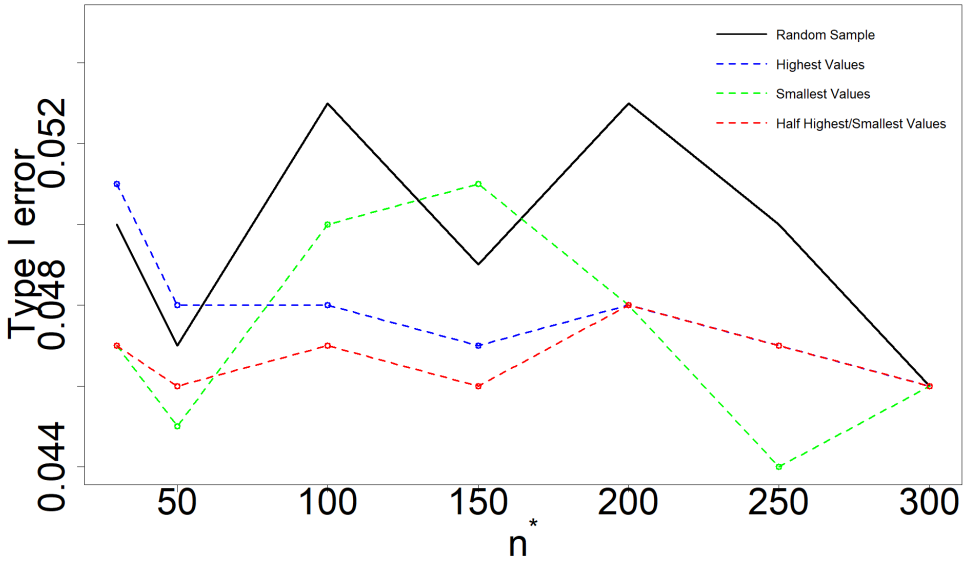

FIGURE 4.2: MAR Type I error plot using PMF for different recovery scenarios.

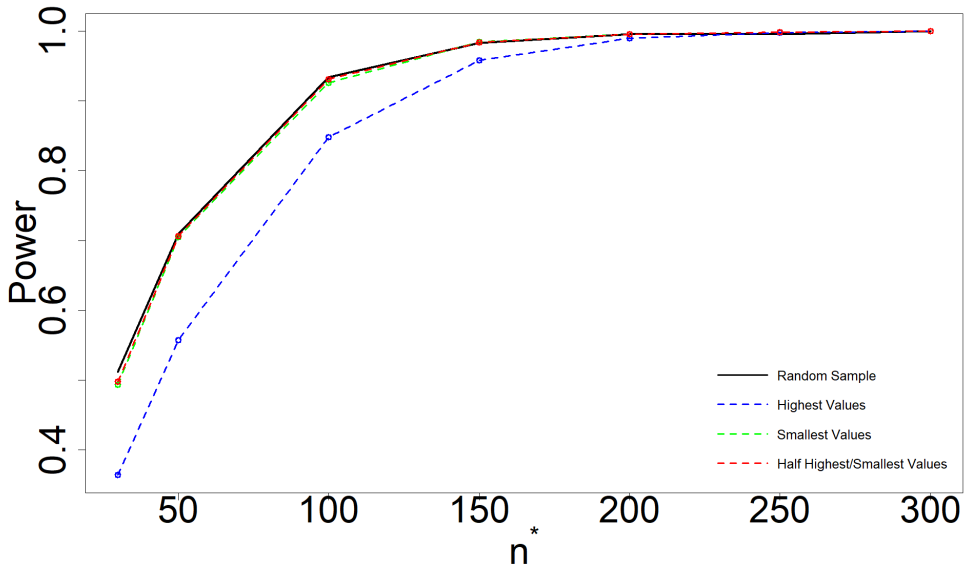

FIGURE 4.3: MNAR power plot using SMF for different recovery scenarios.

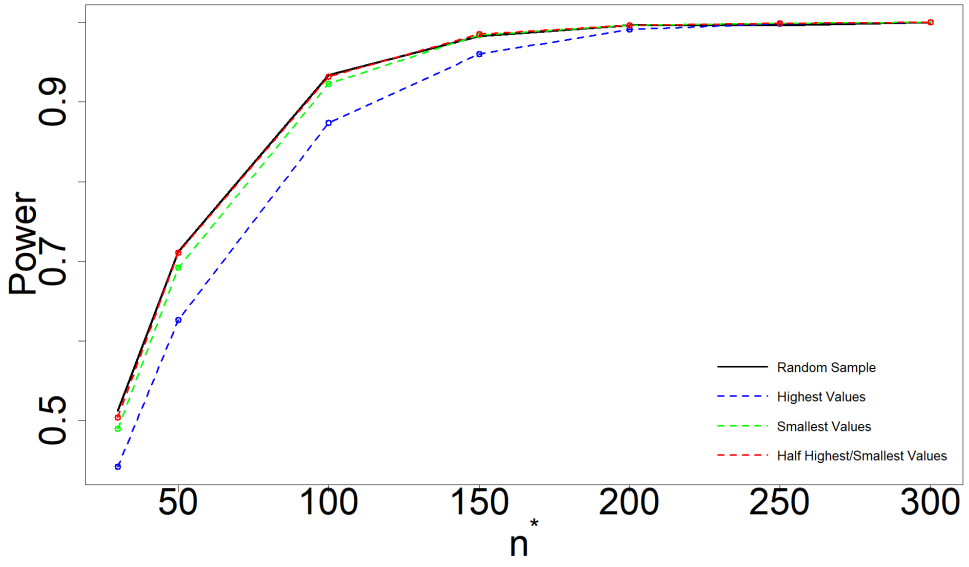

FIGURE 4.4: MNAR power plot using PMF for different recovery scenarios.

The power analysis for three different MNAR cases (which are detailed in Table 4.1) and different designs for varying sample sizes is shown in Table 4.1. The values of α_0 and α_1 in the three cases are chosen to preserve the same amount of missing data (approximately 30%) in the data. The values are varied so there would be different degrees of association to the variable y with equal amount of missing cases. The third case is the least associated and the first case is the most associated.

The first case has its lowest power under the highest design, the smallest design has the highest power and the power is 1 at $n^* = 100$ and above for all designs for both SMF and PMF. In the second case, the highest value design has the least power for both frameworks, both frameworks have power above 0.99 for all designs at $n^* = 100$ and above. For the third case, the random design has the highest power followed by the half highest/half smallest design using both frameworks. The power is less than 0.50 when n^* is 30 for the smallest value design. The half highest/half smallest design at $n^* = 30$ has approximately 0.5 power value, approximately 0.7 when $n^* = 50$ and tends towards 1 as the sample sizes increase. For all designs, the power increases as we recover more missing values.

The SMF and PMF perform similarly for the random design at all recovered sample sizes for the three cases of MNAR except for $n^* = 30$ for the first case where SMF performed slightly better than PMF. For the highest values design and half highest/smallest values design, the PMF performs better than the SMF for all the cases studied. For the smallest values design, the SMF performs better than the PMF.

TABLE 4.1: Power analysis for different MNAR cases and different recovery designs in 10000 replicates.

Design	n^*	$\alpha_0 = -5, \alpha_1 = 0.358$		$\alpha_0 = -4, \alpha_1 = 0.270$		$\alpha_0 = -3, \alpha_1 = 0.190$	
		SMF	PMF	SMF	PMF	SMF	PMF
Random	30	0.954	0.951	0.798	0.798	0.512	0.512
	50	0.996	0.996	0.945	0.945	0.709	0.712
	100	1.00	1.00	0.997	0.997	0.934	0.934
	150	1.00	1.00	1.00	1.00	0.983	0.938
	200	1.00	1.00	1.00	1.00	0.996	0.996
	250	1.00	1.00	1.00	1.00	0.996	0.996
	300	1.00	1.00	1.00	1.00	1.00	1.00
Highest	30	0.690	0.876	0.560	0.691	0.364	0.442
	50	0.913	0.972	0.798	0.872	0.557	0.626
	100	0.998	0.999	0.982	0.990	0.848	0.874
	150	1.00	1.00	0.999	0.999	0.958	0.960
	200	1.00	1.00	1.00	1.00	0.990	0.991
	250	1.00	1.00	1.00	1.00	0.998	0.998
	300	1.00	1.00	1.00	1.00	1.00	1.00
Smallest	30	0.956	0.953	0.786	0.782	0.493	0.490
	50	0.997	0.997	0.944	0.941	0.705	0.692
	100	1.00	1.00	0.999	0.999	0.926	0.923
	150	1.00	1.00	1.00	1.00	0.985	0.984
	200	1.00	1.00	1.00	1.00	0.996	0.996
	250	1.00	1.00	1.00	1.00	0.999	0.999
	300	1.00	1.00	1.00	1.00	1.00	1.00
Half highest/half smallest	30	0.938	0.941	0.778	0.786	0.498	0.504
	50	0.995	0.995	0.938	0.941	0.706	0.711
	100	1.00	1.00	0.998	0.998	0.931	0.932
	150	1.00	1.00	0.999	0.999	0.984	0.985
	200	1.00	1.00	1.00	1.00	0.996	0.996
	250	1.00	1.00	1.00	1.00	0.999	0.999
	300	1.00	1.00	1.00	1.00	1.00	1.00

4.3.2 Two-Parameter Model

Here, we used the two-parameter model for both SMF and PMF to test the type of missingness present using the same data generated in Subsection 4.3.1.

The hypothesis for SMF is given as:

$$H_0 : \alpha_1 = \alpha_3 = 0$$
$$H_1 : \text{at least one } \alpha_i \neq 0.$$

The hypothesis for PMF is as follows:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \text{at least one } \beta_i \neq 0.$$

The MAR Type I error and MNAR power analysis for pattern mixture two-parameter model are shown in Figures 4.5 and 4.6 respectively below and in Table A.2 in the appendix. For MAR, irrespective of the recovery design used, the Type I error is the same when all the missing values are recovered. All designs have Type I error values of approximately 0.05 at all values of n^*. For MNAR, the highest design has the least power followed by the half highest/half smallest design. The smallest design has the highest power while the random design has better power than the random design and the half highest/half smallest design.

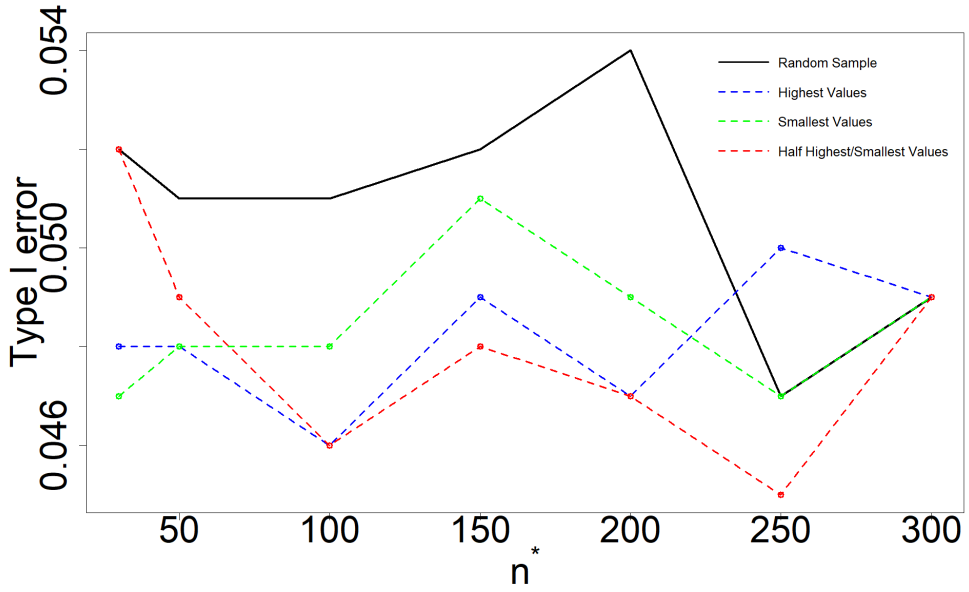

FIGURE 4.5: MAR Type I error plot using PMF for different recovery scenarios.

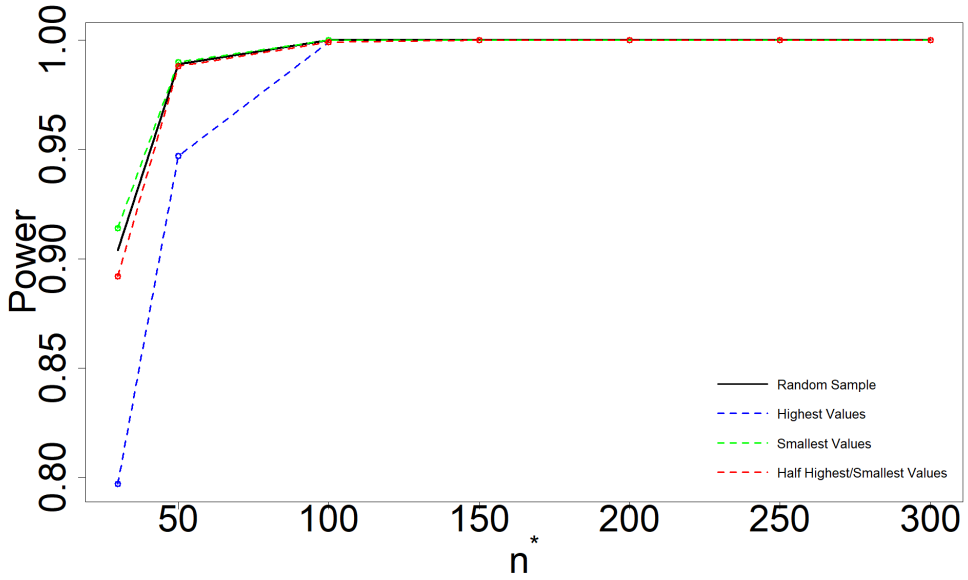

FIGURE 4.6: MNAR Power plot using PMF for different recovery scenarios.

In Table 4.2, the pattern mixture framework for the two-parameter model was used for three different MNAR cases which are detailed in the table. The first case has all powers greater than 0.9 for the random design and smallest design. At $n^* = 30$, the highest value and half highest/half smallest design give power values 0.797 and 0.892 respectively. For the second case of MNAR, the random design has approximately 0.7 power value at $n^* = 30$, the highest value design, smallest value design and half highest/half smallest design have values lesser than 0.7 when $n^* = 30$. At $n^* = 150$ and above, the power is approximately 1. For the third case, the random design has the highest power at $n^* = 30$ followed by the half highest/half smallest design. The highest design has the least power at all recovered sample sizes except at $n^* = 300$ which is the same as other designs. The power increases as n^* increases for all designs. Conclusively, the random design has the largest power while the highest design has the least power.

TABLE 4.2: Power analysis for a two-parameter model for different MNAR cases and different recovery designs in 10000 replicates.

Design	n^*	Pattern Mixture		
		$\alpha_0 = -5, \alpha_1 = 0.358$	$\alpha_0 = -4, \alpha_1 = 0.270$	$\alpha_0 = -3, \alpha_1 = 0.190$
Random	30	0.904	0.701	0.414
	50	0.989	0.893	0.609
	100	1.00	0.994	0.881
	150	1.00	0.999	0.965
	200	1.00	1.00	0.989
	250	1.00	1.00	0.990
	300	1.00	1.00	1.00
Highest	30	0.797	0.585	0.351
	50	0.947	0.798	0.515
	100	0.999	0.974	0.798
	150	1.00	0.998	0.925
	200	1.00	1.00	0.978
	250	1.00	1.00	0.995
	300	1.00	1.00	1.00
Smallest	30	0.914	0.690	0.392
	50	0.990	0.893	0.591
	100	1.00	0.996	0.869
	150	1.00	1.00	0.965
	200	1.00	1.00	0.989
	250	1.00	1.00	0.997
	300	1.00	1.00	1.00
Half highest/half smallest	30	0.892	0.688	0.400
	50	0.988	0.887	0.608
	100	0.999	0.993	0.882
	150	1.00	0.999	0.965
	200	1.00	1.00	0.989
	250	1.00	1.00	0.996
	300	1.00	1.00	1.00

Comparing the PMF one-parameter model for MNAR in Table 4.1 and the PMF two-parameter model in Table 4.2, the one-parameter model produced better powers than the two-parameter model for all the designs at all recovered sample sizes.

In Table 4.3 below, the two-parameter model under SMF is used. The result shows Type I error values above 0.05 for all designs except the random design. Some investigation on the two-parameter model for SMF is provided in the next section.

TABLE 4.3: MAR Type I error for two-parameter using the SMF in 10000 replicates.

Design	n^*	Type I error
Random	30	0.054
	50	0.057
	100	0.054
	150	0.053
	200	0.055
	250	0.054
	300	0.047
Highest	30	0.528
	50	0.778
	100	0.979
	150	0.995
	200	0.991
	250	0.845
	300	0.047
Smallest	30	0.899
	50	0.984
	100	0.999
	150	0.999
	200	0.994
	250	0.787
	300	0.047
Half highest/half Smallest	30	1.00
	50	1.00
	100	0.999
	150	0.981
	200	0.724
	250	0.209
	300	0.047

4.4 Problem with the Type I error

The two parameters model for the selection model framework does not work as the presence of the interaction term seems to affect the Type I error rate. As shown in Table 4.3, the random sample design has values close to 0.05. The highest values design gives values ranging from 0.528 to 0.845, the values increase as n^* increases to 150 and then decrease from 150. The smallest values design and half highest/smallest values design also have values above the expected Type I error. Since the one-parameter model works well under the SMF, the addition of the interaction term seems to affect the model performance in the two-parameter case.

This problem exists because we fit the wrong model on the augmented data. Taking the half highest/smallest recovery design as a case study, there are about 700 zeros out of the 1000 observations scattered across the space, corresponding to the observed data. A selection of ones is obtained with some at the lower end of the space and some at the higher end. The fitted model tries to ascertain whether there is a relationship between the probability of being a one and the x variable. The half highest/smallest design shows that there seem to be two groups of ones, the first group with higher values of x and the other group with smaller values of x.

We tried to fit a functional relationship between the two variables (the binary variable M_A and x) with a choice of a polynomial function and a linear function to see what happens.

There is a connection between the interaction of x with y obtained from:

$$y = \beta_0 + \beta_1 x + e, \tag{4.6}$$

where we can obtain xy as:

$$xy = \beta_0 x + \beta_1 x^2 + ex. \tag{4.7}$$

The inclusion of the interaction between x and y implicitly includes a squared term of x, which may contribute to explaining why we have problematic Type I error rate when the interaction term is included. Moreover, from equation (4.7), the error term ex makes the variance depend on x. Multiplying the error term e by x results in heteroscedasticity making the variance depend on x. This violates the homoscedasticity assumption in regression model. This variance dependency can result in erroneous standard error estimates, which could result in erroneously rejecting the null hypothesis when it is true. The exclusion of the interaction term avoids the problem of heteroscedasticity resulting in a better Type I error.

The result for comparing the models below is shown in Figure 4.7.

$$\text{Full Model is logit } Pr(M_i = 1) = \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2, \tag{4.8}$$

$$\text{Reduced Model is logit } Pr(M_i = 1) = \alpha_0 + \alpha_1 X_i. \tag{4.9}$$

The Type I errors obtained for all the recovery designs are all greater than 0.05. The random sample gave the smallest Type I errors compared to other designs, followed by the smallest values design. The Type I error decreases as the recovered sample increases. While the true model is the one in equation (4.9) the test far too often rejects this model in favour of the (incorrect) quadratic model for the reasons outlined in equation (4.7).

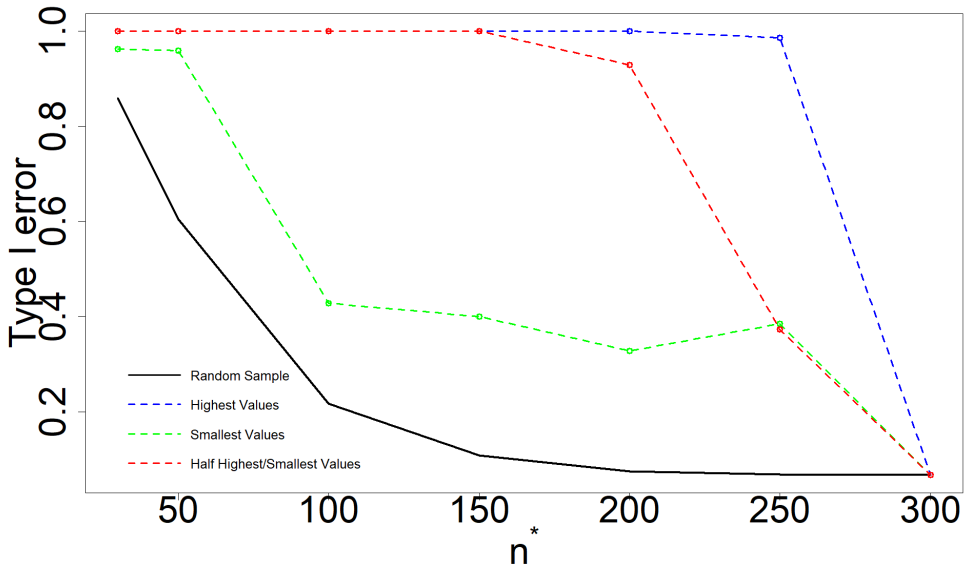

FIGURE 4.7: MAR Type I error plot using SMF for different recovery scenarios.

Figure 4.8 shows the result for comparing the models below.

$$\text{Full Model is logit } Pr(M_i = 1) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{1i}^3, \tag{4.10}$$

$$\text{Reduced Model is logit } Pr(M_i = 1) = \alpha_0 + \alpha_1 X_{1i}. \tag{4.11}$$

We considered a cubic model to see if there would be a good Type I error rate compared to the quadratic model. Just as the quadratic of x, the cubic of x performed similarly. The Type I errors are all above 0.05 and the random sample designs and the smallest values design gave values smaller than the highest values and half highest/smallest designs. There is a decrease in the Type I error as the recovered sample size increases. For all the designs, recovering all the missing values gives the same Type I error of 0.066.

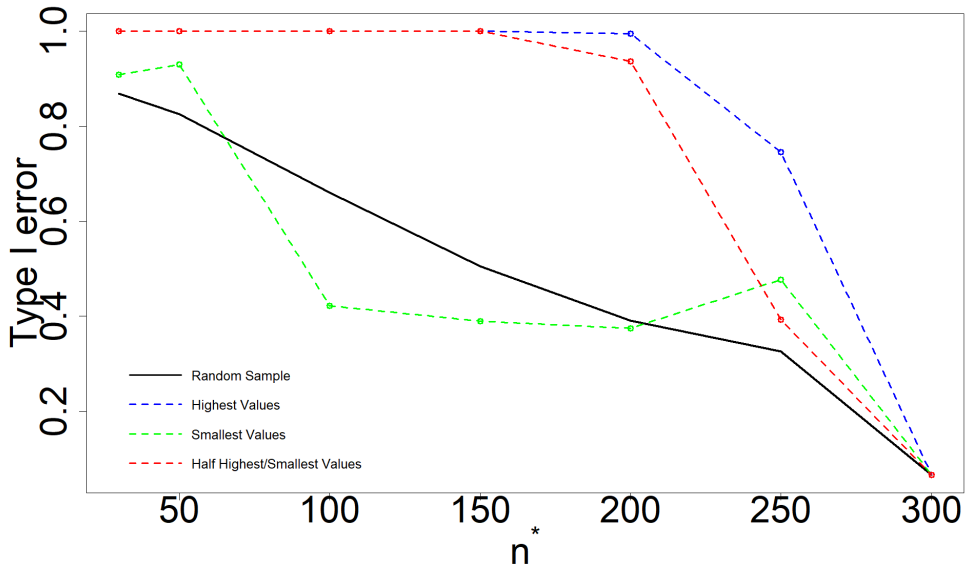

FIGURE 4.8: MAR Type I error plot using SMF for different recovery scenarios.

The above simulations show that the types of recovery designs discussed do not perform better than the random design, hence, a need to develop a better recovery approach that would produce better power with correct Type I error. Two such approaches will be developed in Chapters 5 and 6, respectively.

Chapter 5

Selection Model Framework

In this chapter, we formulate a consistent test for MAR vs MNAR using the recovery designs discussed in the previous chapter. We explore ideas from the design of experiments to obtain an optimal design within a region that increases the power of the test. We will focus solely on the SMF test because of its natural relationship with the MDM.

5.1 A consistent test for MAR vs MNAR

We consider the use of a design region, where the missing values in this region would be augmented with the observed values in this region. The theoretical framework for this test and simulation study will be discussed in this section.

Here, we consider a scenario similar to Chapter 4 with a univariate response, Y, and a p-dimensional covariate vector $X = (X_1, \ldots, X_p)^T$, where

$$Y|(X = x) \sim N(\beta_0 + \beta^T x, \sigma_y^2)\,, \tag{5.1}$$

with $x = (x_1, \ldots, x_p)^T$, $\beta_0 \in \mathbb{R}$ and $\beta = (\beta_1 \ldots, \beta_p)^T \in \mathbb{R}^p$. Let X be a random vector with any arbitrary joint density function $f(x)$. For the construction of the design region R_A, we assume that β, β_0 and σ_y^2 are all known. This assumption may be seen as a restrictive assumption, however, it follows the principle of locally optimal designs (Chernoff, 1953).

It is assumed that the realizations of X are always fully observed and only the response variable Y contains missing responses/values. Let M indicate the missing data predictor that equals one if Y is missing and zero if observed. The missing data mechanism according to Rubin (1976) is determined by the conditional distribution of M, $\Pr(M = 1|X = x, Y = y)$. Under MAR, we have $\Pr(M = 1|X = x, Y = y) = \Pr(M = 1|X = x)$;

which implies that, given the covariates, M does not depend on Y. Under MNAR, this property does not hold as M conditionally depends on X and Y.

Let's assume the missing mechanism has the form

$$g(\Pr(M = 1|X = x, Y = y)) = w^T\lambda + z^T\psi\,, \tag{5.2}$$

where g is any arbitrary link function for a generalized linear model (GLM) with a binary response, w is a q-dimensional vector whose components could depend on functions of x but not y, and z is an s-dimensional vector whose components additionally depend on functions of y, for example an interaction $x_i y$, and (or) a function of just y. Without loss of generality, we assume the first component of w equals one; this corresponds to assuming an intercept term is present in the missing mechanism. Here λ and ψ are the unknown vectors of coefficients.

Examples of g include the logit, probit, and complementary log-log link functions. Taking the inverse of the link function results in an equivalent form of (5.2) that models the conditional distribution directly.

$$\Pr(M = 1|X = x, Y = y) = g^{-1}(w^T\lambda + z^T\psi)\,. \tag{5.3}$$

For the logit link function, the inverse is given by expit $= \frac{\exp(t)}{1+\exp(t)}$. The inverse of the probit link function is $\Phi(t)$, where $\Phi(t)$ is the standard normal distribution function. The inverse of the complementary log-log model is expressed as $1 - \exp(-\exp(t))$.

To determine the presence of MNAR thus involves determining the value of ψ, the missing mechanism is MAR when $\psi = 0$. However, based on the original sample (incomplete dataset), this parameter is inestimable as Y contains missing values.

In order to address this, the use of a two-stage experiment for the data collection would be deployed. The first stage consists of a sample size n of the response variable, covariates and corresponding missing indicator. Here, we shall recall the notation of Section 4.1. $y_1, \ldots, y_n$, $x_1, \ldots, x_n$ and $m_1, \ldots, m_n$ are samples of size n from the model in (5.1). The total number of observed cases and missing cases in the dataset are denoted as n_{obs} and n_{miss} respectively. This implies that $n_{miss} + n_{obs} = n$. Without loss of generality, suppose the dataset is sorted such that the first n_{obs} of $y_1, \ldots, y_{n_{obs}}$ are observed and $y_{n_{obs}+1}, \ldots, y_n$ are missing. Let $\mathbf{D_O} = \{D_{i_1}, D_{i_2}, \ldots, D_{i_{n_{obs}}}\} = \{0, 0, \ldots, 0\}$ be the indicator for the observed cases and $\mathbf{M} := \{m_1, m_2, \ldots, m_{n_{miss}}\}$ indicate the set of indices m whose Y_m are missing at the first stage.

Using the notations in Section 4.1, the number of recovered responses is a proportion of the missing observations, n^* can be expressed as $n^* = \lceil c_1 \cdot n_{miss} \rceil$ with $0 < c_1 \leq 1$.

The choice of which responses to be recovered depends on the experimenter, resulting in the concept of a recovery design.

Assuming a recovery design takes a random sample within a specified p-dimensional region of the covariate space (more detail will be given in section 5.1.2), the key developments in this section are as follows. Firstly, for a missing mechanism of the form in (5.3) with logit link, we show that the entire sample of augmented data (i.e. observed plus recovered) restricted to this space maintains the *expit* functional form, but with a modified value of λ. Secondly, for any other link function, with the same recovery design, we determine a randomly sampled pre-specified proportion of the observed data restricted to this space must be taken in order to preserve the mechanism's functional form. From these, we can establish the necessary properties of the statistical test, such as the Type I error rate, enabling it to be used with confidence. Furthermore, the inherent benefit of using a logit link (in the absence of evidence to support an alternative link) is evident. The logit link permits all augmented data within the design region to be used for inference while other link functions may require subsampling of the observed record pool to ensure an appropriate statistical test. Deriving the relevant theory leads us to the third key development, which provides a framework for optimizing the power of the test by considering power as a function of the recovery design.

5.1.1 A test for MNAR with logistic regression

From the notation in Section 4.1, recall that $\mathbf{Y_A} := (\mathbf{Y_O}, \mathbf{Y_R})^T$, $\mathbf{X_A} := (\mathbf{X_O}, \mathbf{X_R})^T$ and $\mathbf{M_A} := (\mathbf{M_O}, \mathbf{M_R})^T$. For $Y_i^* \in \mathbf{Y_A}$, let $\mathbf{X}^*_{\mathbf{A},i}$ be the corresponding i^{th} row in the matrix $\mathbf{X_A}$ and let $M_i^* \in \mathbf{M_A}$ be the corresponding indicator value. The superscript * denotes the augmented data. Let w_i and z_i be the values of w and z at observation i of the augmented data. A test for MNAR using the SMF fits the model below:

$$\Pr(M_i^* = 1 | \mathbf{X}^*_{\mathbf{A},i} = \mathbf{x}, Y_i^* = y) = g^{-1}(w_i^T \lambda_A + z_i^T \psi_A)\,. \tag{5.4}$$

The parameters λ_A and ψ_A are unknown regression coefficients based on the augmented data. The relation to the parameters λ_A and ψ_A must be determined to perform inference or estimate on the original parameters λ and ψ. In Section 5.1.2, it is shown that if the augmented data is fashioned in a particular way, for the logit link function, we obtain the relation $\psi_A = \psi$ and $\lambda_A = \lambda + (constant, 0, \ldots, 0)^T \in \mathbb{R}^q$; recall that without loss of generality it is assumed that the first element of w_i is equal to one and corresponds to the intercept term in the GLM. To determine the presence of MNAR or not involves testing the hypothesis $\psi_A = 0$. For any other link function, it is ensured that λ and ψ are maintained at the loss of some information.

To perform this test, we use the likelihood ratio test. The log-likelihood ratio test statistic for testing a general hypothesis $\psi = \psi_0$ is given by $2[l(\hat{\psi}, \hat{\lambda}_A) - l(\psi_0, \hat{\lambda}_0)]$, where l

denotes the log-likelihood function based on a sample of size n_A (the number of observations in the augmented data) and $(\hat{\psi}, \hat{\lambda}_A)$ and $\hat{\lambda}_0$ denotes the maximum likelihood estimators under the alternative and null models, respectively. Under the null hypothesis, the distribution of the likelihood ratio statistic is approximated by a central chi-square distribution with s degrees of freedom, i.e. the classical statistical approximation.

5.1.2 A mixture distribution for the augmented data

There is a possibility that the distribution of the observed cases and missing cases would be different, hence, the augmented data is a mixture of distributions. These mixture distributions comprise a weighted combination of the distribution of the observed data and the distribution of the missing data. Carefully constructing the marginal distributions provides an expression for the missing data mechanism in the augmented data. Let $\mathcal{R}_A \subseteq \mathbb{R}^p$ be a p-dimensional region constructed by the cartesian product of intervals in $\mathbb{R}$ of positive length. It is possible to have $\mathcal{R}_A = \prod_{j=1}^{p}[p_j, q_j]$ with $p_j < q_j$ and $p_j, q_j \in \mathbb{R}$; $\mathcal{R}_A$ becomes a p-dimensional hypercuboid if p_j and q_j are finite. However, we could also permit more general sets such as e.g. $\mathcal{R}_A = \prod_{j=1}^{p}[p_j, q_j] \cup [r_j, s_j]$ with $r_j < s_j$ and $r_j, s_j \in \mathbb{R}$. The region $\mathcal{R}_A$ will be used to instruct the recovery design **D**. In particular, we will only recover n^* missing responses $Y_{n_{obs}+1}, \ldots, Y_n$ whose corresponding p-dimensional covariate vectors $X_{n_{obs}+1}, \ldots, X_n$ lie within $\mathcal{R}_A$. For the random vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ introduced in Section 5.1, the intersection of events is defined as:

$$\begin{aligned} \mathcal{M}_O &:= \{M = 0\} \cap \{\mathbf{X} \in \mathcal{R}_A\} \\ \mathcal{M}_R &:= \{M = 1\} \cap \{\mathbf{X} \in \mathcal{R}_A\}, \end{aligned}$$

Here, M is the missing indicator that takes value 1 when Y is missing and 0 if observed.

It is assumed that $\mathcal{R}_A$ can be chosen such that $\Pr(\mathcal{M}_O) > 0$ and the region $\mathcal{R}_A$ could also be chosen such that:

$$\Pr(\mathcal{M}_R) = c_1 \cdot \Pr(M = 1). \tag{5.5}$$

The above helps to have sufficient missing cases and observed cases within the region where the augmented cases fall and also makes sure that the recovered sample is a fixed proportion c_1 of the missing cases but if the recovery design involves a random sampling without replacement, $\mathcal{R}_A$ is chosen such that:

$$\Pr(\mathcal{M}_R) \geq c_1 \cdot \Pr(M = 1). \tag{5.6}$$

If $\mathcal{R}_A = \mathbb{R}^p$ is set such that $\Pr(\mathcal{M}_R) = \Pr(M = 1)$, then $\mathbf{D}$ would be a random sample of size n^* from all the missing responses. This design is a random recovery design and will serve as a benchmark design.

From the observed data whose covariates lie within $\mathcal{R}_A$, we introduce the capability of bringing a random sample of a particular proportion into the augmented data. As a result, let $0 < c_2 \leq 1$ be the proportion of observed data whose covariates lie within $\mathcal{R}_A$ that are used within the augmented data. If $c_2 = 1$, then all of the observed observations that lie in $\mathcal{R}_A$ will be used. If $c_2 = 0.6$, then 60% of observed data lying within $\mathcal{R}_A$ is included in the augmented data, provided they are sampled randomly from all the observed covariates that are contained in $\mathcal{R}_A$. If c_2 is chosen carefully, the original missing mechanism can be maintained in the augmented data for any GLM link function. With this framework, we derive the following:

Definition 5.1. Let M_A be the missing indicator variable in the augmented data that takes values 0 when Y is observed and 1 when Y is missing. M_A satisfies:

$$M_A = \begin{cases} 1 & \text{with probability } \dfrac{c_1 \cdot \Pr(M = 1)}{c_1 \cdot \Pr(M = 1) + c_2 \cdot \Pr(\mathcal{M}_O)} \\ 0 & \text{with probability } \dfrac{c_2 \cdot \Pr(\mathcal{M}_O)}{c_1 \cdot \Pr(M = 1) + c_2 \cdot \Pr(\mathcal{M}_O)} \end{cases}. \tag{5.7}$$

Proof of 5.1

If (5.5) holds and $c_2 = 1$, $M_A = M|\{\mathcal{M}_O \cup \mathcal{M}_R\}$.

$$\begin{aligned} \Pr(M_A = 1) &= \Pr(M = 1|\{\mathcal{M}_O \cup \mathcal{M}_R\}) \\ &= \frac{\Pr(\{M = 1\} \cap \{\mathcal{M}_O \cup \mathcal{M}_R\})}{\Pr(\{\mathcal{M}_O \cup \mathcal{M}_R\})} \\ &= \frac{\Pr(\{M = 1\} \cap \{\mathcal{M}_O \cup \mathcal{M}_R\})}{\Pr(\{M = 1\} \cap \{\mathcal{M}_O \cup \mathcal{M}_R\}) + \Pr(\{M = 0\} \cap \{\mathcal{M}_O \cup \mathcal{M}_R\}))} \\ &= \frac{\Pr(\{M = 1\} \cap \mathcal{M}_R)}{\Pr(\{M = 1\} \cap \mathcal{M}_R) + \Pr(\{M = 0\} \cap \mathcal{M}_O))} \\ &= \frac{c_1 \cdot \Pr(M = 1)}{c_1 \cdot \Pr(M = 1) + \Pr(\mathcal{M}_O)}. \end{aligned}$$

If condition (5.6) holds rather than condition (5.5) and c_2 is not necessarily equal to one, then we need to modify the claim $M_A = M|\{\mathcal{M}_O \cup \mathcal{M}_R\}$ to account for the random subsample in the observed and recovered data. Let U be a uniform random variable on $[0, 1]$ and define the event $\mathcal{B} = \{U \leq c_1 \cdot \Pr(M = 1)/\Pr(\mathcal{M}_R)\}$. Then $M_A =$

$M|\{\{\mathcal{M}_O \cap \{U \leq c_2\}\} \cup \{\mathcal{M}_R \cap \mathcal{B}\}\}$.

$$\begin{aligned}
\Pr(M_A = 1) &= \Pr(M = 1|\{\{\mathcal{M}_O \cap \{U \leq c_2\}\} \cup \{\mathcal{M}_R \cap \mathcal{B}\}\}) \\
&= \frac{\Pr(\{M = 1\} \cap (\{\mathcal{M}_O \cap \{U \leq c_2\}\} \cup \{\mathcal{M}_R \cap B\}\}))}{\Pr(\mathcal{M}_O \cap \{U \leq c_2\} \cup \{\mathcal{M}_R \cap B\})} \\
&= \frac{\Pr(\{M = 1\} \cap \{\mathcal{M}_R \cap B\})}{\Pr(\mathcal{M}_O \cap \{U \leq c_2\} + \Pr\{\mathcal{M}_R \cap B\})} \\
&= \frac{\Pr(B)\Pr(\{M = 1\} \cap \mathcal{M}_R\})}{c_2 \cdot \Pr(\mathcal{M}_O) + c_1 \cdot \Pr(M = 1)}
\end{aligned}$$

since $\Pr(B) = \dfrac{c_1 \cdot \Pr(\mathrm{M} = 1)}{\Pr(\mathcal{M}_\mathrm{R})}$ and $\Pr(\{M = 1\} \cap \mathcal{M}_R\}) = \Pr(\mathcal{M}_\mathrm{R})$

$$\begin{aligned}
\Pr(B)\Pr(\{M = 1\} \cap \mathcal{M}_R\}) &= \frac{c_1 \cdot \Pr(\mathrm{M} = 1)}{\Pr(\mathcal{M}_\mathrm{R})} \cdot \Pr(\mathcal{M}_\mathrm{R}) \\
&= c_1 \cdot \Pr(\mathrm{M} = 1) \\
\Pr(M_A = 1) &= \frac{c_1 \cdot \Pr(M = 1)}{c_1 \cdot \Pr(M = 1) + c_2 \cdot \Pr(\mathcal{M}_O)} .
\end{aligned}$$

Accordingly, define the random variables and random vectors

$$Y_O := Y \,|\, \mathcal{M}_O; \quad Y_R := Y \,|\, \mathcal{M}_R; \quad X_O := X \,|\, \mathcal{M}_O; \quad X_R := X \,|\, \mathcal{M}_R \,.$$

Definition 5.2. The augmented response denoted as Y_A and augmented covariate X_A are realizations from random variable/vector:

$$Y_A := (1 - M_A)Y_O + M_A Y_R \tag{5.8}$$

$$X_A := (1 - M_A)X_O + M_A X_R \,. \tag{5.9}$$

These expressions are constructed according to how the augmented data are constructed, being a combination of observed responses and recovered responses.

Using the above definitions, we state the key theorem, followed by two corollaries, below.

Theorem 5.3. *For $0 < c_1 \leq 1$ and $0 < c_2 \leq 1$, provided $\mathcal{R}_A$ satisfies* (5.5) *or* (5.6)*, then for $x := (x_1, \ldots, x_p)$ the missing data mechanism in the augmented data has the following form. For $x \in \mathcal{R}_A$ and any link function g:*

$$\Pr(M_A = 1 \,|\, X_A = x, Y_A = y) = \frac{c^* \Pr(M = 1|X = x, Y = y)}{c^* \Pr(M = 1|X = x, Y = y) + \Pr(M = 0|X = x, Y = y)} \,,$$

where $c^* = \frac{c_1 \cdot \Pr(M=1)}{c_2 \cdot \Pr(M=1, X \in \mathcal{R}_A)}$. Otherwise, the probability is zero.

Proof of Theorem

For $\mathbf{x} \in \mathcal{R}_A$:

We start by considering the probability of observing X_R and Y_R in a small region dx and dy respectively, restricted to the region $\mathcal{R}_A$:

$$\begin{aligned}\Pr(X_R \in dx, Y_R \in dy) &= \frac{\Pr(X \in dx, Y \in dy, \mathcal{M}_R)}{\Pr(\mathcal{M}_R)} \\ &= \frac{\Pr(X \in dx, Y \in dy, M = 1)}{\Pr(\mathcal{M}_R)}.\end{aligned}$$

Mathematically, $\Pr(X_R \in dx) = f(x)dx$, where $f(x)$ represents the probability density function of X_R, dx and dy are infinitesimally small intervals around the variables X and Y, respectively.

Similarly, the probability for the observed cases is:

$$\Pr(X_O \in dx, Y_O \in dy) = \frac{\Pr(X \in dx, Y \in dy, M = 0)}{\Pr(\mathcal{M}_O)}.$$

The conditional probability $\Pr(M_A = 1 \mid Y_A = y, X_A = x)$ using Bayes' theorem is given by:

$$\begin{aligned}\Pr(M_A = 1 \mid Y_A = y, X_A = x) &= \frac{\Pr(M_A = 1, X_A \in dx, Y_A \in dy)}{\Pr(X_A \in dx, Y_A \in dy)} \\ &= \frac{\Pr(M_A = 1)\Pr(X_R \in dx, Y_R \in dy)}{\Pr(X_A \in dx, Y_A \in dy)} \\ &= \Pr(M_A = 1) \times \frac{\Pr(X_R \in dx, Y_R \in dy)}{\Pr(X_A \in dx, Y_A \in dy)}.\end{aligned}$$

Define:

$$\begin{aligned}A &= \frac{\Pr(M_A = 1)\Pr(M = 1 \mid X = x, Y = y)\Pr(Y \in dy \mid X = x)\Pr(X \in dx)}{\Pr(\mathcal{M}_R)} \\ B &= \frac{\Pr(M_A = 0)\Pr(M = 0 \mid X = x, Y = y)\Pr(Y \in dy \mid X = x)\Pr(X \in dx)}{\Pr(\mathcal{M}_O)}.\end{aligned}$$

Thus:

$$\Pr(M_A = 1 \mid Y_A = y, X_A) = \frac{\frac{\Pr(M_A = 1)\Pr(M = 1 \mid X = x, Y = y)\Pr(X \in dx, Y \in dy)}{\Pr(\mathcal{M}_R)}}{A + B}$$

Multiply all through by $\Pr(\mathcal{M}_R)$ to get:

$$\Pr(M_A = 1 | Y_A = y, X_A) = \frac{\Pr(M_A = 1)\Pr(M = 1 | X = x, Y = y)\Pr(X \in dx, Y \in dy)}{C + D}$$

where,

$$\begin{aligned} C &= \Pr(M_A = 1)\Pr(M = 1|X = x, Y = y)\Pr(X \in dx, Y \in dy) \\ D &= \frac{\Pr(\mathcal{M}_R)}{\Pr(\mathcal{M}_O)}\Pr(M_A = 0)\Pr(M = 0|X = x, Y = y)\Pr(X \in dx, Y \in dy)\,. \end{aligned}$$

Divide all through by $\Pr(X \in dx, Y \in dy)$ and $\Pr(M_A = 1)$,

Therefore we have:

$$\frac{\Pr(M = 1|X \in dx, Y = y)\Pr(M = 0, X \in \mathcal{R}_A)}{\Pr(M = 1|X = x, Y = y)\Pr(\mathcal{M}_O) + \frac{\Pr(M_A=0)}{\Pr(M_A=1)} \cdot \Pr(\mathcal{M}_R) \cdot \Pr(M = 0|X = x, Y = y)}$$

which simplifies to:

$$\frac{\Pr(M{=}1|X{=}x,Y{=}y)\cdot\Pr(\mathcal{M}_O)}{\Pr(M{=}1|X{=}x,Y{=}y)\cdot\Pr(\mathcal{M}_O)+\frac{\Pr(M_A=0)}{\Pr(M_A=1)}\cdot\Pr(\mathcal{M}_R)\cdot\Pr(M{=}0|X{=}x,Y{=}y)} \tag{5.10}$$

From Definition 5.1, the following relation can be obtained:

$$\frac{\Pr(M_A = 0)}{\Pr(M_A = 1)} = \frac{\dfrac{c_2 \cdot \Pr(\mathcal{M}_O)}{c_1 \cdot \Pr(M = 1) + c_2 \cdot \Pr(\mathcal{M}_O)}}{\dfrac{c_1 \cdot \Pr(M = 1)}{c_1 \cdot \Pr(M = 1) + c_2 \cdot \Pr(\mathcal{M}_O)}}$$

$$\frac{\Pr(M_A = 0)}{\Pr(M_A = 1)} = \frac{c_2 \cdot \Pr(\mathcal{M}_O)}{c_1 \cdot \Pr(M = 1)}\,.$$

Substitute this in 5.10 to obtain:

$$\frac{\Pr(M = 1|X = x, Y = y)\Pr(\mathcal{M}_O)}{\Pr(M = 1|X = x, Y = y)\Pr(\mathcal{M}_O) + \frac{c_2\cdot\Pr(\mathcal{M}_O)}{c_1\cdot\Pr(M=1)} \cdot \Pr(\mathcal{M}_R) \cdot \Pr(M = 0|X = x, Y = y)}\,,$$

Divide through by $\Pr(\mathcal{M}_O)$ to obtain:

$$\frac{\Pr(M = 1|X = x, Y = y)}{\Pr(M = 1|X = x, Y = y) + \frac{c_2\cdot\Pr(\mathcal{M}_R)}{c_1\cdot\Pr(M=1)} \cdot \Pr(M = 0|X = x, Y = y)}\,,$$

Multiply through by $\frac{c_1\cdot\Pr(M=1)}{c_2\cdot\Pr(\mathcal{M}_R)}$,

$$\Pr(M_A = 1|X_A = x, Y_A = y) = \frac{c^*\Pr(M = 1|X = x, Y = y)}{c^*\Pr(M = 1|X = x, Y = y) + \Pr(M = 0|X = x, Y = y)}$$

where,

$$c^* = \frac{c_1 \cdot \Pr(M = 1)}{c_2 \cdot \Pr(\mathcal{M}_R)}\,. \tag{5.11}$$

Which concludes the proof of Theorem 5.3.

An important corollary of this theorem is:

Corollary 5.4. *Under the conditions of Theorem* (5.3)*, provided*

$$c_2 = \frac{c_1 \cdot \Pr(M = 1)}{\Pr(M = 1, X \in \mathcal{R}_A)},$$

then for $x \in \mathcal{R}_A$*, we have* $\Pr(M_A = 1 \mid X_A = x, Y_A = y) = \Pr(M = 1 \mid X = x, Y = y)$*, otherwise zero.*

Given a recovery proportion c_1, Corollary 5.4 provides the value of c_2, the proportion of observed data lying in $\mathcal{R}_A$ to randomly sample when constructing the augmented data, in order to maintain the original missing mechanism. If c_2 does not satisfy the requirements of Corollary 5.4, then $c^* \neq 1$ in Theorem 5.3 and there is no cancellation. Consequently, the true missing mechanism in the augmented data will not be of any well-known GLM form and therefore obtaining estimates for the parameters and testing the presence of MNAR becomes more complicated. One could build a custom link function from Theorem 5.3, however easy implementation and the loss of optimized procedures that are present in statistical software packages for well-known link functions will likely make analysis cumbersome. An exception to this is the logit link function which maintains the same form regardless of the choice of c_2. This is a consequence of the following corollary.

Corollary 5.5. *If the original missing data mechanism utilizes the logit link function:*

$$\Pr(M = 1 \mid Y = y, X = x) = \frac{\exp(w^T\lambda + z^T\psi)}{1 + \exp(w^T\lambda + z^T\psi)},$$

then for all $x \in \mathcal{R}_A$ *and any* c_2*, the missing mechanism in the augmented data has the form*

$$\Pr(M_A = 1 \mid Y_A = y, X_A = x) = \frac{\exp(w^T\lambda_A + z^T\psi)}{1 + \exp(w^T\lambda_A + z^T\psi)},$$

with $\lambda_A = \lambda + (\log(c^*), 0, \ldots, 0)^T \in \mathbb{R}^q$*.*

While Corollary 5.5 assumes an MNAR mechanism, it also holds under MAR as ψ can equal 0. The interpretation of Corollary 5.5 is as follows. If the recovery design is a random sample of the missing responses $Y_{n_{obs}+1}, \ldots, Y_n$ whose covariate vectors $X_{n_{obs}+1}, \ldots, X_n$ lie within $\mathcal{R}_A$, and we augment our recovered data with the observed data whose p-dimensional covariate vectors also lie within $\mathcal{R}_A$, then only the intercept in the missing mechanism's linear predictor changes regardless of whether the mechanism is MAR or MNAR. In particular, the coefficients in front of terms involving y in the augmented data are the same as its counterpart based on the full sample, permitting MNAR's presence to be reliably inferred using the augmented sample.

Since in Corollary 5.5, c_2 can take any value in the interval $(0, 1]$, when considering the logit link function we will set $c_2 = 1$. This results in more observations in the augmented data and the variance of estimators is reduced. For any other link function, we will assume c_2 is obtained from Corollary 5.4.

5.1.3 A special case of Corollary 5.5

In this section, we consider a simple example of Corollary 5.5 to help illustrate the result. We will set $c_2 = 1$ with $p = 1$ covariate and recover the missing values that fall within an interval; the recovery region $\mathcal{R}_A$ becomes $[a, b]$. This section is also used to highlight the potential for increasing the power of the test for MNAR. If changing the interval impacts the power of the test (hopefully improving), then we can explore methods for selecting the interval, rather than just arbitrarily. In this section, we redefine the designs as follows:

1. Highest Design: select c_1 highest x values and augment with the observed cases that fall in the recovery interval.
2. Smallest Design: select c_1 smallest x values and augment with the observed cases that fall in the recovery interval.
3. Half Highest/Half Smallest Design: select $\frac{c_1}{2}$ highest x values and $\frac{c_1}{2}$ smallest x values. The c_1 recovered cases and the observed cases that fall in the recovery interval form the augmented data.

When considering the highest, smallest and half highest/half smallest designs, by construction we have $\Pr(\mathcal{M}_{\mathrm{R}}) = \mathrm{c}_1 \cdot \Pr(\mathrm{M} = 1)$. Using (5.11), this results in $c^* = 1$ and $\log(c^*) = 0$, therefore, no change or shift is seen in the intercept. This will be demonstrated in the example below.

Generate n observations from a normal distribution in the following way:

$$X \sim N(5, 1),$$

$$Y|X \sim N(1 + 2x, 4).$$

Approximately 30% missing values were introduced using equation (3.3) with $\alpha_0 = -3$ and $\alpha_1 = 0.42$ for MAR and equation (3.4) with $\alpha_0 = -5$ and $\alpha_1 = 0.358$ for MNAR. The interval $a < X < b$ was used to recover some missing values in the data. a and b are real values, in this case, the quantiles of the missing values were used. The augmented data consists of the missing values and observed values that fall within this interval.

Tables 5.1 and 5.2 show the model coefficients for MNAR and MAR respectively. A sample size of 1000000 was used with approximately 30% missing values, the recovered samples were selected from the missing data $X's$ that falls within $a < X < b$, the observed values based on x that falls within this range were augmented with the recovered samples. A logistic regression model was fitted on the augmented data and the effect on power and Type I error was studied.

In Table 5.1, the model coefficients for MNAR $(-5 + 0.358y)$ and the expected value of intercept $+ log(c^*)$ are shown. For random sample design, $log(c^*)$ changes for different recovered proportions. The intercepts approximate the expected value and showed that a shift of $log(c^*)$ was seen. The coefficient of x is approximately 0 for all the proportions and the coefficient of y reduces and tends to the true value of 0.358 as the recovered proportion increases. The highest design shows that there is no shift in the intercept, and the coefficients approximate the true values. The smallest design and half highest/smallest design also show the same result as the highest design. The coefficients of x are approximately 0 and the coefficients of y approximate 0.358.

TABLE 5.1: MNAR Model Coefficients

Design	Recovered Proportion	$\alpha_0 + \alpha_1 x + \alpha_2 y$	-5 + log(c*)
Random	0.1	-7.357-0.0028x+0.364y	-7.303
	0.2	-6.621-0.0081x+0.362y	-6.609
	0.3	-6.205-0.0084x+0.362y	-6.204
	0.4	-5.909-0.0062x+0.360y	-5.916
	0.5	-5.682-0.0067x+0.360y	-5.693
	0.6	-5.500-0.0074x+0.360y	-5.511
	0.7	-5.348-0.0046x+0.359y	-5.357
	0.8	-5.212-0.0041x+0.359y	-5.223
	0.9	-5.094-0.0039x+0.359y	-5.105
	1.0	-4.994-0.0023x+0.359y	-5.00
Highest	0.1	-4.987-0.0012x+0.358y	-4.999
	0.2	-5.030-0.0002x+0.360y	-4.999
	0.3	-4.992-0.0033x+0.359y	-4.999
	0.4	-5.008-0.0021x+0.358y	-4.999
	0.5	-4.973-0.0073x+0.359y	-4.999
	0.6	-4.969-0.0056x+0.358y	-4.999
	0.7	-4.980-0.0043x+0.359y	-5.00
	0.8	-4.986-0.0030x+0.358y	-4.999
	0.9	-5.006-0.0005x+0.359y	-4.999
	1.0	-4.994-0.0023x+0.359y	-5.00
Smallest	0.1	-4.972-0.0043x+0.358y	-4.999
	0.2	-4.921-0.0225x+0.359y	-4.999
	0.3	-4.957-0.0119x+0.359y	-4.999
	0.4	-4.974-0.0077x+0.359y	-4.999
	0.5	-4.987-0.0025x+0.358y	-4.999
	0.6	-5.002-0.0015x+0.359y	-4.999
	0.7	-4.995-0.0017x+0.359y	-4.999
	0.8	-4.997-0.0011x+0.358y	-4.999
	0.9	-4.993-0.0025x+0.359y	-4.999
	1.0	-4.994-0.0023x+0.359y	-5.00
Half Highest/Smallest	0.1	-4.977+0.0007x+0.357y	-5.00
	0.2	-4.982-0.0016x+0.358y	-5.00
	0.3	-4.983-0.0056x+0.359y	-5.00
	0.4	-4.994-0.0044x+0.359y	-5.00
	0.5	-4.994-0.0031x+0.359y	-5.00
	0.6	-4.994-0.0029x+0.359y	-5.00
	0.7	-4.991-0.0005x+0.358y	-5.00
	0.8	-4.992-0.0022x+0.358y	-5.00
	0.9	-4.995-0.0032x+0.359y	-5.00
	1.0	-4.994-0.0023x+0.359y	-5.00

The model coefficients for MAR $(-3 + 0.42x)$ are shown in Table 5.2. For the random design, the coefficients are close to the true and expected values, the intercepts are close to the expected value of intercept $+ log(c^*)$. The coefficient of y is approximately 0 and the coefficient of x is very close to 0.42, as the sample size increases, the coefficients of x

get closer to the true value. The highest design, smallest design and half highest/smallest design all performed similarly with no change in the intercept and the coefficients of x and y approximate their true values.

TABLE 5.2: MAR Model Coefficients

Design	Recovered Proportion	$\alpha_0 + \alpha_1 x + \alpha_2 y$	-3 + log(c*)
Random	0.1	-5.252+0.402x+0.0039y	-5.303
	0.2	-4.586+0.414x+0.0008y	-4.609
	0.3	-4.178+0.411x+0.0018y	-4.204
	0.4	-3.885+0.412x+0.0010y	-3.916
	0.5	-3.671+0.413x+0.0011y	-3.693
	0.6	-3.496+0.416x+0.0005y	-3.511
	0.7	-3.341+0.416x+0.0002y	-3.357
	0.8	-3.210+0.415x+0.0011y	-3.223
	0.9	-3.088+0.414x+0.0010y	-3.105
	1.0	-2.984+0.415x+0.0009y	-3.00
Highest	0.1	-2.901+0.400x+0.0023y	-2.999
	0.2	-2.888+0.406x-0.0016y	-2.999
	0.3	-2.901+0.409x-0.0021y	-2.999
	0.4	-2.991+0.419x-0.0007y	-2.999
	0.5	-2.964+0.415x-0.0005y	-3.00
	0.6	-2.974+0.416x-0.0004y	-2.999
	0.7	-2.966+0.413x+0.0004y	-2.999
	0.8	-2.981+0.416x+0.0003y	-2.999
	0.9	-2.984+0.417x+0.0001y	-2.999
	1.0	-2.984+0.415x+0.0009y	-3.00
Smallest	0.1	-3.060+0.424x+0.0064y	-2.999
	0.2	-3.007+0.417x+0.0028y	-2.999
	0.3	-2.985+0.413x+0.0018y	-2.999
	0.4	-2.999+0.415x+0.0024y	-2.999
	0.5	-2.992+0.414x+0.0020y	-3.00
	0.6	-2.999+0.417x+0.0017y	-2.999
	0.7	-2.981+0.412x+0.0019y	-2.999
	0.8	-2.987+0.415x+0.0014y	-2.999
	0.9	-2.990+0.416x+0.0008y	-2.999
	1.0	-2.984+0.415x+0.0009y	-3.00
Half Highest/Smallest	0.1	-2.989+0.401x+0.0072y	-2.999
	0.2	-2.983+0.406x+0.0048y	-3.00
	0.3	-2.983+0.410x+0.0031y	-3.00
	0.4	-2.984+0.415x+0.0010y	-3.00
	0.5	-2.987+0.416x+0.0006y	-3.00
	0.6	-2.987+0.417x+0.0002y	-2.999
	0.7	-2.985+0.415x+0.0092y	-2.999
	0.8	-2.984+0.415x+0.0011y	-2.999
	0.9	-2.984+0.415x+0.0007y	-2.999
	1.0	-2.984+0.415x+0.0009y	-3.00

The MAR Type I error for four different designs is shown in Figure 5.1, the random design has a Type I error of less than or approximately 0.05 for all the recovered proportions. All other designs have approximately 0.05 Type I error values. Figure 5.3

shows the power for the different designs. The random and smallest designs give the highest power values while the highest design has the least power.

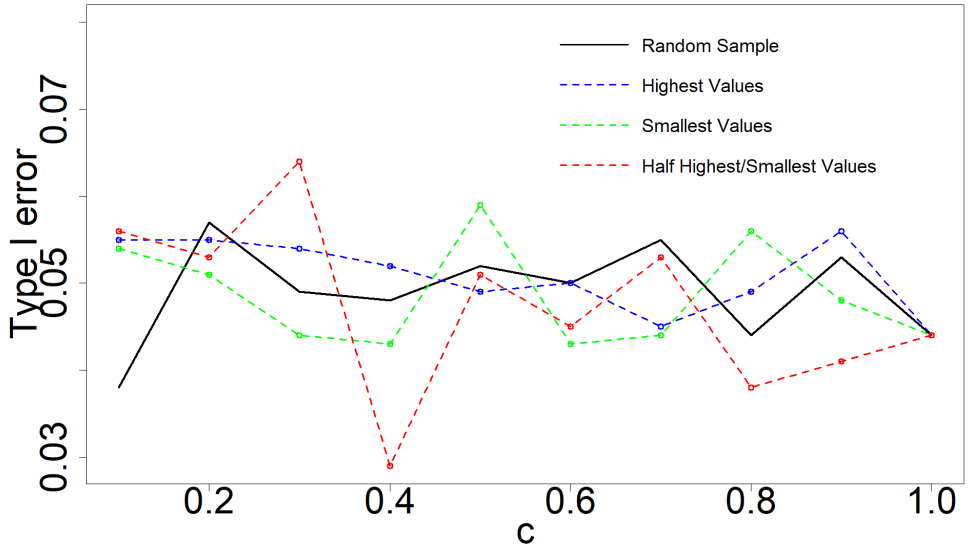

FIGURE 5.1: MAR Type I error plot using SMF for different recovery scenarios.

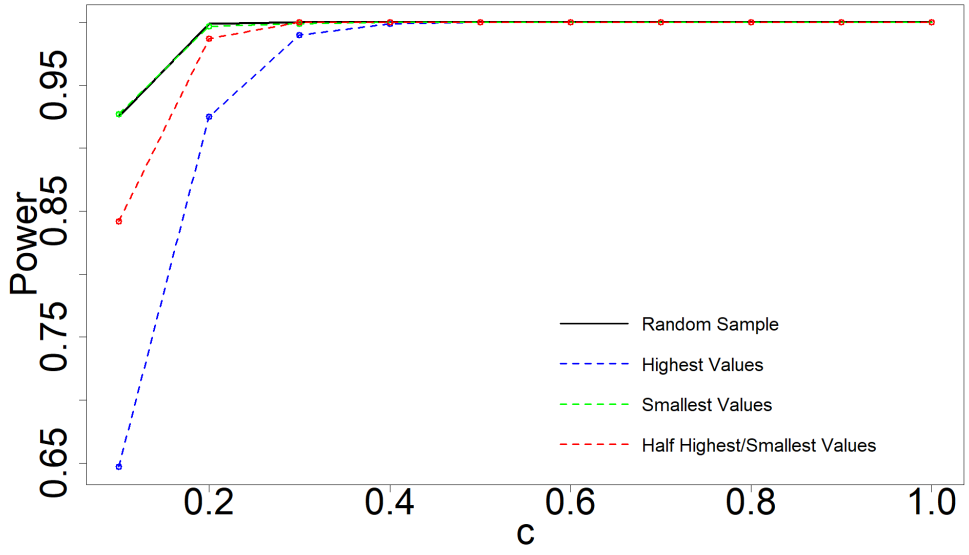

FIGURE 5.2: MNAR power plot using SMF for different recovery scenarios.

Figure 5.3 shows the power analysis for MNAR of type $-3 + 0.19y$, the random design gives the best power values at all recovered proportions. The smallest design performs better than the highest and the half highest/smallest designs. The highest design performs least of all the designs.

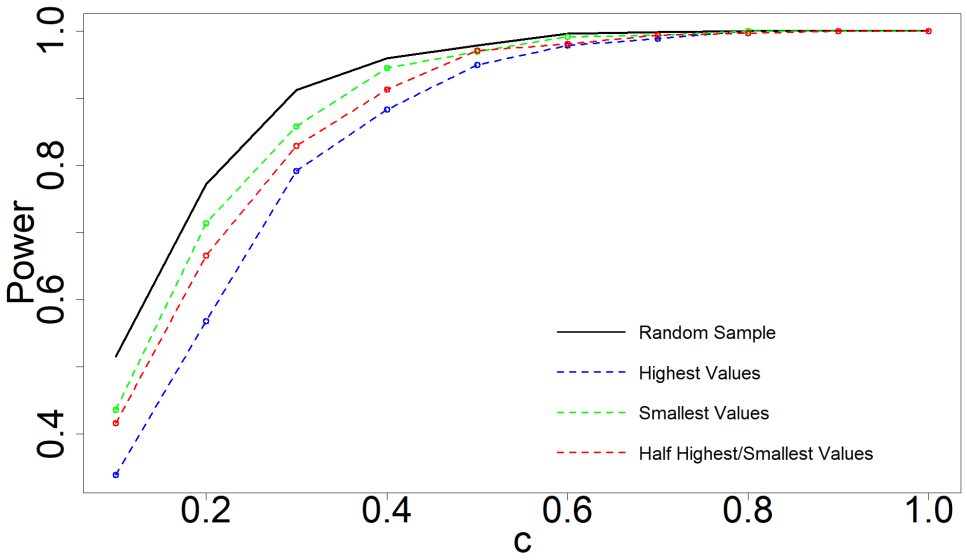

FIGURE 5.3: MNAR power plot using SMF for different recovery scenarios.

Results from simulations above showed that sampling within a restricted region corrects the Type I error problem seen in Chapter 4. In the following Sections, we find an optimal design that leads to better power than the random design.

5.2 Optimal design

In this section, we explore ideas from optimal experimental design to increase the power of the test. The main idea is to choose the region $\mathcal{R}_A$ that minimises the (asymptotic) variance of the maximum likelihood estimator for the parameter of interest.

5.2.1 Designing region R_A

5.2.1.1 Approximating the deviance by a non-central χ^2 distribution

Using the conditions of Theorem 5.3, Corollary 5.4 and Corollary 5.5. Assuming the logit link function, then $c_2 = 1$ and have $\lambda_A = \lambda + (\log(c^*), 0, \ldots, 0)^T$ and $\psi_A = \psi$. For any other link function, c_2 is determined from Corollary 5.4 and have $\lambda_A = \lambda$ and $\psi_A = \psi$. To maximize the power of the likelihood ratio test for $\psi = 0$ as described in Section 5.1.1, is equivalent to testing between MAR (Null) and MNAR (Alternative), we thereby integrate the results of Self et al. (1992). The main focus of this section is to approximate the power of the likelihood ratio test by approximating the distribution of the log-likelihood ratio statistic with a non-central chi-square distribution with

s degrees of freedom. The non-centrality parameter used in the approximation is computed by equating the expected value of a non-central chi-square random variable to an approximation of the expected value of the likelihood ratio statistic, which involves taking the expected value of lead terms in an asymptotic expansion of the likelihood ratio statistic. The technique of Self et al. (1992) is adopted to decompose the likelihood ratio statistic into the following three components:

$$\begin{aligned} 2[l(\hat{\psi},\hat{\lambda}_A) - l(\psi_0,\hat{\lambda}_0)] &= 2[l(\hat{\psi},\hat{\lambda}_A) - l(\psi,\lambda_A)] - 2[l(\psi_0,\hat{\lambda}_0) - l(\psi_0,\lambda_0^*)] \\ &+ 2[l(\psi,\lambda_A) - l(\psi_0,\lambda_0^*)]\,, \end{aligned} \tag{5.12}$$

where λ_0^* is the limiting value of $\hat{\lambda}_0$ and is given in Definition 5.6 below. The asymptotic expansion of the first component in (5.12) was considered in Cordeiro (1983). Taking only the lead term in this expansion results in an approximate expected value for the first component in (5.12) of $q+s$. The expected value of the lead term in this expansion of the second component in (5.12) is equal to the trace of the matrix U given by

$$U = \left\{ \mathbb{E}\left[-\frac{\partial^2 l(\psi,\lambda_A)}{\partial \lambda_A^2}\bigg|_{(\psi_0,\lambda_0^*)} \right] \right\}^{-1} \mathbb{E}\left[\left(\frac{\partial l(\psi,\lambda_A)}{\partial \lambda_A}\bigg|_{(\psi_0,\lambda_0^*)} \right) \left(\frac{\partial l(\psi,\lambda_A)}{\partial \lambda_A}\bigg|_{(\psi_0,\lambda_0^*)} \right)^T \right].$$

In Self et al. (1992), formulae for computing U are provided for generalized linear models. Here, we will derive the results for the specific case of a logit model, but other link functions can easily be considered. We obtain from Self et al. (1992, p. 33):

$$\mathbb{E}\left[-\frac{\partial^2 l(\psi,\lambda_A)}{\partial \lambda_A^2}\bigg|_{(\psi_0,\lambda_0^*)} \right] = \sum_{i=1}^{n_A} \frac{\exp(\theta_i^*)}{(1+\exp(\theta_i^*))^2} w_i {w_i}^T$$

and

$$\mathbb{E}\left[\left(\frac{\partial l(\psi,\lambda_A)}{\partial \lambda_A}\bigg|_{(\psi_0,\lambda_0^*)} \right) \left(\frac{\partial l(\psi,\lambda_A)}{\partial \lambda_A}\bigg|_{(\psi_0,\lambda_0^*)} \right)^T \right] = \sum_{i=1}^{n_A} \frac{\exp(\theta_i)}{(1+\exp(\theta_i))^2} w_i {w_i}^T\,,$$

where $\theta_i^* = w_i^T \lambda_0^*$ and $\theta_i = w_i^T \lambda_A + z_i^T \psi$.
The expected value of the third component in (5.12), denoted by Δ, can be calculated exactly. For the logit link function, it is given by

$$\Delta = 2\sum_{i=1}^{n_A} \left\{ \frac{\exp(\theta_i)}{1+\exp(\theta_i)}(\theta_i - \theta_i^*) - \log\left(\frac{1+\exp(\theta_i)}{1+\exp(\theta_i^*)} \right) \right\}.$$

Since the expected value of a non-central chi-square random variable with s degrees of freedom and non-centrality parameter γ is $s+\gamma$, using the recommendation of Self et al. (1992), we approximate the distribution of the likelihood ratio statistic by a

non-central chi-square distribution with non-centrality parameter $q + \Delta - tr(U)$.

To compute U and Δ, one has to know the limiting value of $\hat{\lambda}_0$. For an arbitrary link function g, we can appeal to the following definition.

Definition 5.6. The limiting value of $\hat{\lambda}_0$, that is λ_0^*, minimizes the Kullback-Leibler divergence between the alternative and null models assuming the alternative model is true. Consequently, it satisfies:

$$\arg\min_{\lambda_0^*} \mathbb{E}_1 \Bigg[g^{-1}(W^T\lambda_A + Z^T\psi) \log\left(\frac{g^{-1}(W^T\lambda_A + Z^T\psi)}{g^{-1}(W^T\lambda_0^*)}\right) + (1 - g^{-1}(W^T\lambda_A + Z^T\psi)) \log\left(\frac{1 - g^{-1}(W^T\lambda_A + Z^T\psi)}{1 - g^{-1}(W^T\lambda_0^*)}\right) \Bigg],$$

where the capitalization W and Z emphasizes that the elements of w and z are now functions of the random variables defined in Definition 5.2, $\mathbb{E}_1$ denotes expectation under the alternative model and is taken with respect to X_A and Y_A.

To compute the expectation with respect to X_A and Y_A, one can exploit the forms of (5.8) and (5.9). After conditioning on $M_A = 1$ and $M_A = 0$, one needs to know the joint distributions of (X_R, Y_R) and (X_O, Y_O) respectively to compute the required expectation.

Lemma 5.7. *For $x \in \mathcal{R}_A$:*

$$\begin{aligned}
\Pr(X_R \in dx, Y_R \in dy) &= \frac{\Pr(M = 1 \mid X = x, Y = y)\Pr(Y \in dy \mid X = x)\Pr(X \in dx)}{\Pr(\mathcal{M}_R)} \\
\Pr(X_O \in dx, Y_O \in dy) &= \frac{\Pr(M = 0 \mid X = x, Y = y)\Pr(Y \in dy \mid X = x)\Pr(X \in dx)}{\Pr(\mathcal{M}_O)},
\end{aligned}$$

otherwise zero. Here, $dx = (dx_1, \ldots, dx_p)$ is a vector of infinitesimals and the relation $\Pr(Y \in dy \mid X = x)$ can be obtained from (5.1).

The matrix A and scalar Δ are random quantities since they depend on w_i. However, by the law of large numbers, for the logit model as $n_A \to \infty$:

$$\begin{aligned}
\frac{1}{n_A}\sum_{i=1}^{n_A} \frac{\exp(\theta_i^*)}{(1+\exp(\theta_i^*))^2} w_i {w_i}^T &\to \mathbb{E}\left[\frac{\exp(\theta^*)}{(1+\exp(\theta^*))^2} WW^T\right] \\
\frac{1}{n_A}\sum_{i=1}^{n_A} \frac{\exp(\theta_i)}{(1+\exp(\theta_i))^2} w_i {w_i}^T &\to \mathbb{E}\left[\frac{\exp(\theta)}{(1+\exp(\theta))^2} WW^T\right],
\end{aligned}$$

where $\theta^* = W^T\lambda_0^*$ and $\theta = W^T\lambda_A + Z^T\psi$.

Therefore, for n_A suitably large, we propose replacing each term in A with their limiting forms. We also suggest replacing Δ with $\mathbb{E}\Delta$ (although this expectation will be

dropped from notation). The expectations can be computed by averaging over X_A using Lemma 5.7.

We now propose two similar algorithms for choosing our recovery design. The first algorithm will be aimed at GLMs with an arbitrary link function whereas the second will focus only on the case of the logit link, which has the appealing property alluded to in Corollary 5.5. For simplicity, we only consider $\mathcal{R}_A$ taking the form of a p-dimensional cuboid.

Algorithm 1 Algorithm 1a (General Link Function)

Input: $0 < c_1 \leq 1$ and Corollary 5.4.
Output: Recovery design **D**.
Initialization:
  Determine c_2 from Corollary 5.4.
Steps:
Select the p-dimensional cuboid $\mathcal{R}_A$ such that the noncentrality parameter $q + \Delta - \text{tr}(U)$ is maximized, subject to the constraints:
- $\Pr(\mathcal{M}_R) \geq c_1 \cdot \Pr(M = 1)$
- $\Pr(\mathcal{M}_O) > 0$

Construct the recovery design **D** consisting of:
- A random sample of n^* points within $\mathcal{R}_A$.
- A random sample containing $c_2 \times 100\%$ of the observed data within $\mathcal{R}_A$.

Return: Recovery design **D**.

Algorithm 2 Algorithm 1b (Logit Link Function)

Input: Value $0 < c_1 \leq 1$.
Output: Recovery design **D**.
Initialization:
  Set $c_2 = 1$.
Steps:
Select the p-dimensional cuboid $\mathcal{R}_A$ such that the noncentrality parameter $q + \Delta - \text{tr}(U)$ is maximized, subject to the constraints:
- $\Pr(\mathcal{M}_R) \geq c_1 \cdot \Pr(M = 1)$
- $\Pr(\mathcal{M}_O) > 0$

Construct the recovery design **D** consisting of:
- A random sample of n^* points within $\mathcal{R}_A$.
- All of the observed data within $\mathcal{R}_A$.

Return: Recovery design **D**.

In practice, implementing Algorithm 1*a* and 1*b* (in both cases) will likely lead to locally optimal solutions. Furthermore, in a finite regime, there will likely be scenarios where $\mathcal{R}_A$ includes slightly fewer points than n^*. In this case, we recommend uniform enlargement until n^* points lie within $\mathcal{R}_A$.

5.2.2 Minimizing asymptotic variance

An alternative approach to design the region $\mathcal{R}_A$ is explored here and optimizes the power of the SMF test (5.4). This method is more in line with classical design theory, in particular D_1-optimality. The derivations here only permit scalar ψ. Accordingly, we only consider w and z of the form $w = (1, x_1, \ldots, x_p)^T$ and $z = (y)$. Note, under regularity conditions, the MLE of $\alpha := (\lambda_A, \psi)^T$ satisfies

$$(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, I^{-1}(\alpha)) \quad \text{as} \quad n_A \to \infty\,, \tag{5.13}$$

where n_A is the number of observations in the augmented data and $I(\alpha)$ is the observed Fisher Information matrix of α. A design that provides the most precise information about α_{p+1}, the coefficient of y, leads to higher power (asymptotically) for the SMF test in (5.4) among MLEs, is one that minimizes the (asymptotic) variance of $\hat{\alpha}_{p+1}$ or, equivalently, $I^{-1}(\alpha)[p+1, p+1]$, the $(p+1)$th diagonal element of $I^{-1}(\alpha)$. Atkinson and Fedorov (1975) showed that, for completely observed data, this approach and maximizing the asymptotic power of the test that the coefficient of interest is zero are equivalent.

Inverting $I(\alpha)$ can be computationally expensive and unnecessary since only the last diagonal element is of interest.

by appealing to Cramer's rule to obtain

$$I^{-1}(\alpha)[p+1, p+1] = \det(I_p(\alpha))/\det(I(\alpha))\,, \tag{5.14}$$

where $I_p(\alpha)$ is the $p \times p$ submatrix of $I(\alpha)$ comprising its first p rows and columns.

$I(\alpha)$ satisfies $I(\alpha) = -\mathbb{E}[H(l(\alpha))]$ where H is the Hessian and l is the log-likelihood function of the augmented data. By considering missing mechanisms of the form (5.5), its properties (continuity, differentiability, and being bounded on $(0, 1)$) mean we can take the expectation within the Hessian. Exploiting independent and identically distributed observations, we obtain:

$$\begin{aligned} I(\alpha) \;=\; & -H\big(\mathbb{E}[n_A] \cdot \mathbb{E}\big(M_A \log\Big[g^{-1}(W^T\lambda_A + Z^T\psi_A)\Big] \\ & + (1 - M_A)\log\Big[1 - g^{-1}(W^T\lambda_A + Z^T\psi_A)\Big]\big)\big)\,. \end{aligned} \tag{5.15}$$

One can also show that $\mathbb{E}[n_A] = n \cdot (c_1 \Pr(M = 1) + c_2 \Pr(\mathcal{M}_O))$.

For the logit link function, we have $\lambda_A = \lambda + (\log(c^*), 0, \ldots, 0)^T$ and $\psi_A = \psi$. Otherwise, provided we operate under the conditions of Corollary 5.4 and choose c_2 accordingly, we have for any other link function, $\lambda_A = \lambda$ and $\psi_A = \psi$. To compute expectations in (5.15), we appeal to Lemma 5.7.

We now propose an alternative algorithm to Algorithm 1*a* and 1*b*. Again, for simplicity, we only consider $\mathcal{R}_A$ taking the form of a p-dimensional cuboid and divide the new Algorithm into two parts depending on what link function is used. The simulation study in this research is focused on the logit link function.

Algorithm 3 Algorithm 2a (Arbitrary Link Function)

Input: $0 < c_1 \leq 1$ and Corollary 5.4.
Output: Recovery design **D**.
Initialization:
  Determine c_2 from Corollary 5.4.
Steps:
Select the p-dimensional cuboid $\mathcal{R}_A$ such that the right-hand side of (5.14) is minimized, subject to the constraints:
- $\Pr(\mathcal{M}_R) \geq c_1 \cdot \Pr(M = 1)$
- $\Pr(\mathcal{M}_O) > 0$

Construct the recovery design **D** consisting of:
- A random sample of n^* points within $\mathcal{R}_A$.
- A random sample containing $c_2 \times 100\%$ of the observed data within $\mathcal{R}_A$.

Return: Recovery design **D**.

Algorithm 4 Algorithm 2b (Logit Link Function)

Input: Value $0 < c_1 \leq 1$.
Output: Recovery design **D**.
Initialization:
  Set $c_2 = 1$.
Steps:
Select the p-dimensional cuboid $\mathcal{R}_A$ such that the right-hand side of (5.14) is minimized, subject to:
- $\Pr(\mathcal{M}_R) \geq c_1 \cdot \Pr(M = 1)$
- $\Pr(\mathcal{M}_O) > 0$

Construct the design **D** consisting of:
- A random sample of n^* points within $\mathcal{R}_A$.
- All the observed data within $\mathcal{R}_A$.

Return: Recovery design **D**.

5.2.2.1 Single Covariate

For a single covariate ($p = 1$), consider the following different MNAR cases. For each example, all values are chosen to introduce circa 30% missingness in Y.

Case 1. Generate 1000 points following a simple linear regression model in 10000 replicates:

$$Y|(X = x) \sim N(2 - 2x, 4)\,,$$

with $X \sim N(0, 16)$. Introduce MNAR missingness into y using:

$$P(M = 1|Y = y, X = x) = \frac{\exp(-2 + 0.4x - 0.13y)}{(1 + \exp(-2 + 0.4x - 0.13y))}.$$

We apply test (5.4) and test the hypothesis $H_0 : \psi = 0$ to obtain the MSE and power for the random design and optimal design. Figure 5.4 shows the mean squared error (MSE) of the MLE, $\hat{\psi}$, and the power of test (5.4) is shown in Figure 5.5, for different recovery proportions, c_1 (called c in what follows for simplicity), under designs constructed from Algorithm 2b (red dashed line). We use Algorithm 2b because the link function is logistic. Algorithm 2 is slightly cheaper and simpler to execute than Algorithm 1 for this example. We also present equivalent results from randomly recovered observations, i.e. $\mathcal{R}_A = \mathbb{R}$ (solid black). As c increases, the MSE reduces as shown in Figure 5.4. The optimal design has a smaller MSE than the random design. In Figure 5.5, the power values are shown. The power increases as c increases and the optimal design has powers slightly above the random design.

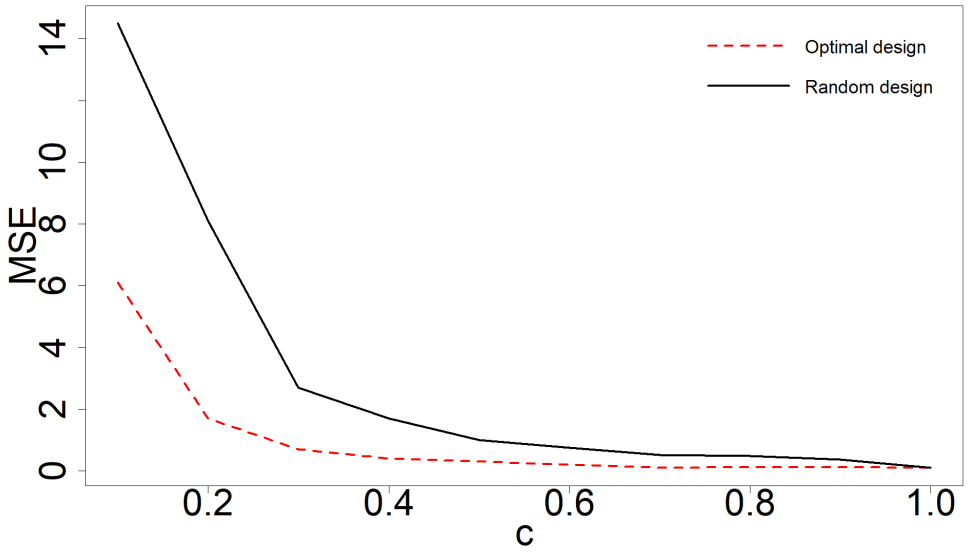

FIGURE 5.4: MSE comparison between random design and optimal design for $p = 1$ case 1.

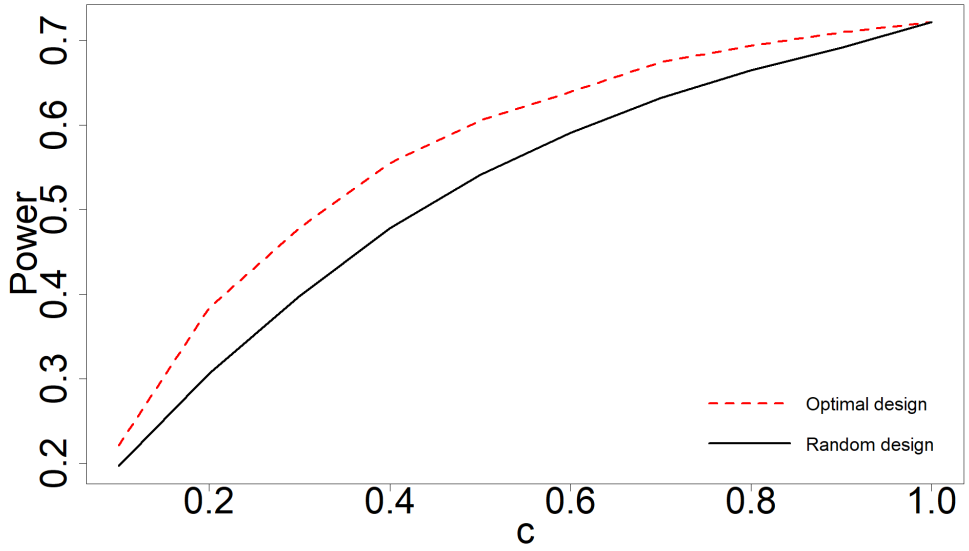

FIGURE 5.5: Power comparison between random design and optimal design for $p = 1$ case 1.

Case 2. In the example below, the variance of the covariate X is reduced from 16 to 1, and the mean of the response variable $Y|X$ changes from $2 - 2x$ to $1 + 2x$, in order to examine whether there will be a difference in the results when comparing the MSE and power of the two designs. These values are used to ensure that there are approximately 30% missing values in Y.

$n = 1000$ points were generated as follows:

$$Y|(X = x) \sim N(1 + 2x, 4)\,,$$

with $X \sim N(5, 1)$. Introduce MNAR missingness into the model using:

$$P(M = 1|Y = y, X = x) = \frac{\exp(0.89 - 0.15y)}{1 + \exp(0.89 - 0.15y)}.$$

The SMF test was applied and the hypothesis $H_0 : \psi = 0$ was tested to obtain the MSE and power for the random design and optimal design.

For this case, the MSE and Power for random design and optimal design for different values of c are shown in Figures 5.6 and 5.7 respectively. As c increases, the MSE reduces as shown in Figure 5.6. The random design has a higher MSE than the optimal design. In Figure 5.7, the optimal design has higher power than the random design. The power increases as c increases.

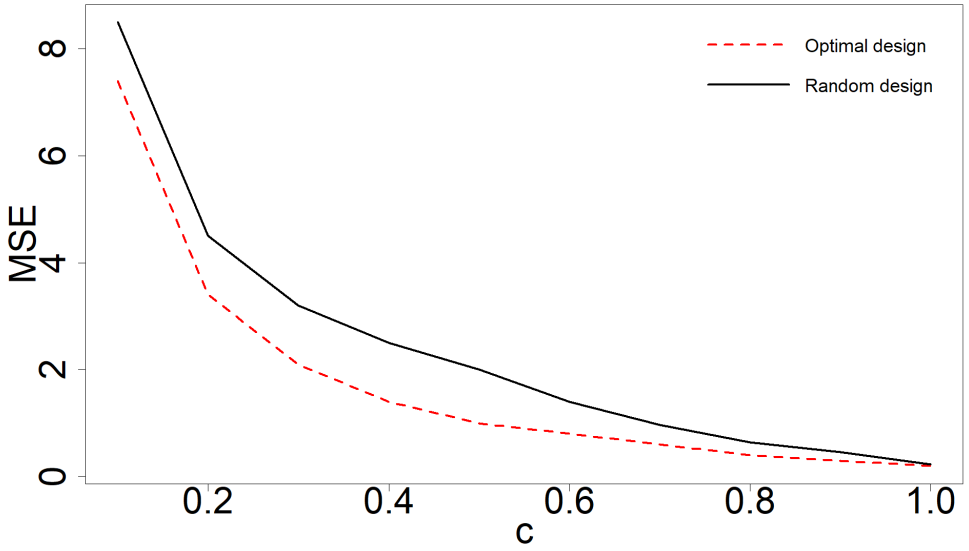

FIGURE 5.6: MSE comparison between random design and optimal design for $p = 1$ case 2.

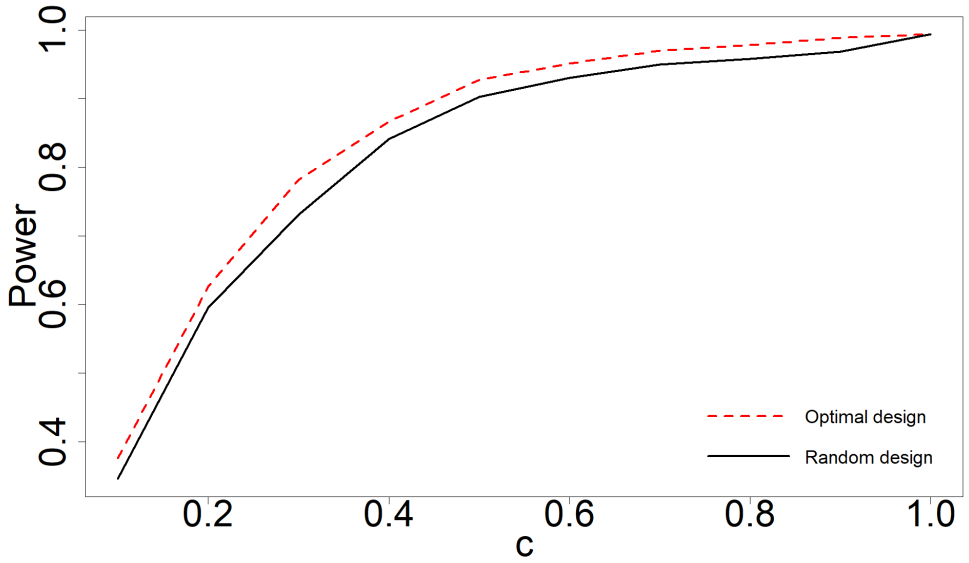

FIGURE 5.7: Power comparison between random design and optimal design for $p = 1$ case 2.

5.2.2.2 Multiple Covariates

For $p = 2$, generate 1000 points following a simple linear regression model in 10000 replicates:

$$Y|(X_1 = x_1, X_2 = x_2) \sim N(2 - 2x_1 + 2x_2, 4)\,,$$

with $X_1 \sim N(0,16)$ and $X_2 \sim N(2,4)$. Introduce MNAR missingness into the model using:

$$P(M=1|Y=y,X=x) = \frac{\exp(-2+0.4x_1+0.2x_2-0.15y)}{1+\exp(-2+0.4x_1+0.2x_2-0.15y)}.$$

For this case, Using these parameters above introduce about 30% missing values in Y. The optimal design performs better than the random design. In Figure 5.8, as c increases, the MSE reduces for both designs with the optimal design having the smaller MSE at all values of c. As shown in Figure 5.9, the power increases as c increases, the random design has smaller power than the optimal design. An upward trend can be seen for the power for both designs and a downward trend for the MSE.

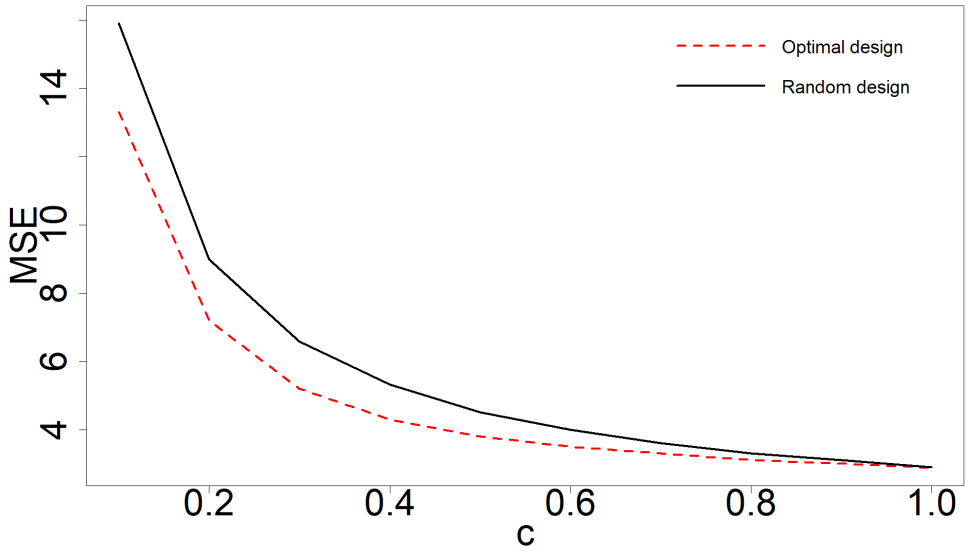

FIGURE 5.8: MSE comparison between random design and optimal design when $p = 2$.

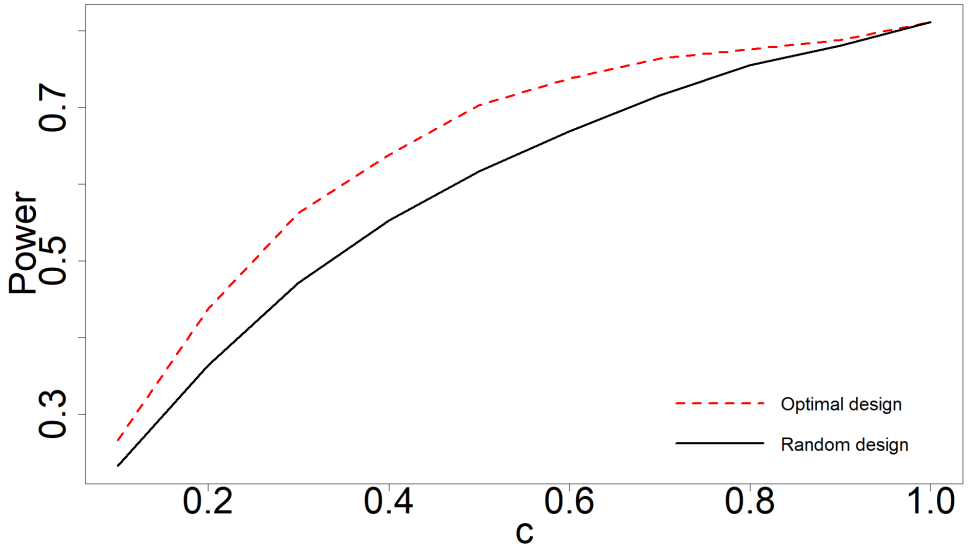

FIGURE 5.9: Power comparison between random design and optimal design when $p = 2$.

Figures 5.10 and 5.11 show the missing, observed and recovered data for $p = 2$ when $c = 0.2$ and 0.9 respectively. The green rectangles in the figures represent the regions $\mathcal{R}_A$, the red and blue points represent (x_1, x_2) that are observed and missing respectively. From both figures below, increasing c from 0.2 to 0.9 extends the recovery region to the right but not the left. This leads to an increase in the dimension on the X_1 axis while the dimension does not change on the X_2 axis. It leaves out the extremes of the distributions.

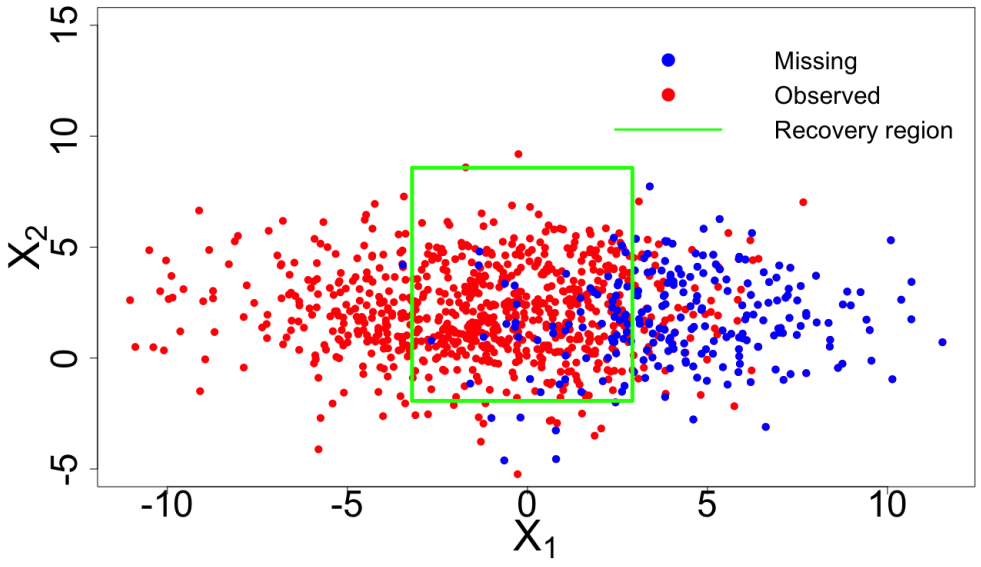

FIGURE 5.10: Region $\mathcal{R}_A$ for $c = 0.2$.

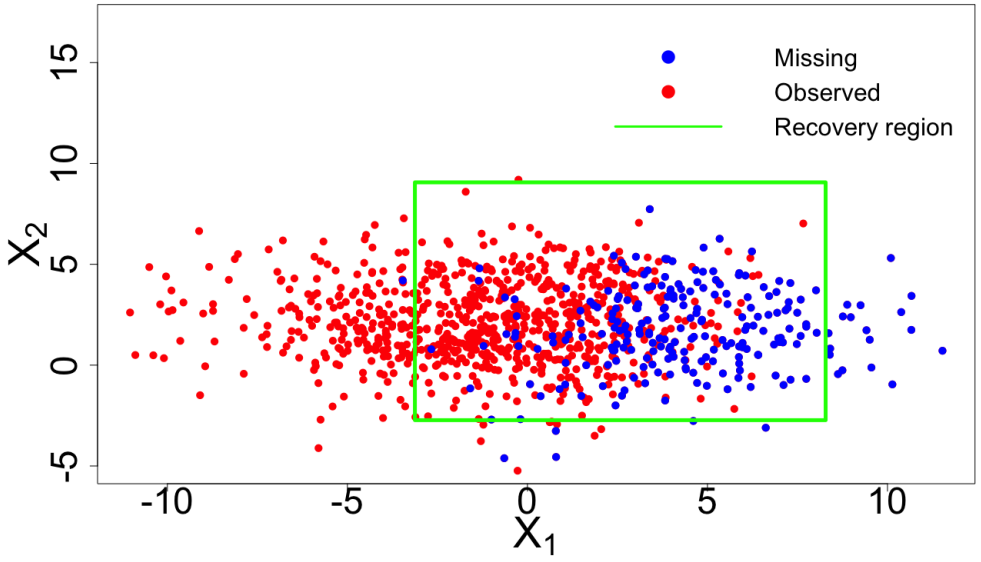

FIGURE 5.11: Region $\mathcal{R}_A$ for $c = 0.9$.

5.3 Assessing the robustness

This Section examines the robustness of the proposed design methodology. In the derived methodology, we assume that the potential MNAR mechanism, the regression model and its parameters are known. We investigate the design's robustness from various misspecifications, such as:

- Getting the MNAR mechanism wrong. This could be having an incorrect value for the intercept or the wrong coefficient on either x or y. In particular, we will investigate the effect of a change in sign on the coefficient of x and y. Will there be a significant change when the sign changes?
- Misspecifying the regression relation. What happens when the regression relation is wrong? Does a change in the regression coefficients or a change in sign affect the power of the test?

In order to assess the robustness of the proposed methodology, we misspecify one parameter and find the optimal recovery regions. For each parameter, we consider scenarios where we respectively, add and subtract 10% of its true value, to see how the performance of the design is affected if the misspecification occurs in a neighbourhood of the true values. To assess more severe misspecifications, we also consider scenarios where the parameters change signs and, for the coefficient of y in the linear predictor of the missing mechanism and the coefficient of x in the linear regression model, we also assess the effect of doubling or multiplying the value of the coefficients.

Generate 1000 points following a simple linear regression model in 100000 replicates:

$$Y|(X = x) \sim N(2 - 2x, 4)\,,$$

with $X \sim N(0, 16)$. Introduce MNAR missingness into the model using:

$$P(M = 1|Y = y, X = x) = \frac{\exp(-2 + 0.4x - 0.15y)}{(1 + \exp(-2 + 0.4x - 0.15y))}.$$

These values are chosen such that approximately 30% missing cases are introduced in Y. Applying the SMF test and testing the hypothesis to obtain the MSE and power for the random design, optimal design and different misspecifications in the mechanism and regression model gives the following result presented in tables below.

Tables 5.3 and 5.4 show the power and MSE for the different designs considered in 100000 replicates respectively. The optimal design has the highest power among all other designs at all values of recovery proportion. For the missing mechanism; at 10% increase and decrease in the intercept, the designs $(-2.2, 0.4, -0.15)$ and $(-1.8, 0.4, -0.15)$ performed better than the random design. For a 10% increase and decrease on the coefficient of x, the designs $(-2, 0.44, -0.15)$ and $(-2, 0.36, -0.15)$ have better power at all values than the random design. A 10% increase and decrease on the coefficient of y also result in better power than the random design at all values of recovery proportion. A change in sign on the missing mechanism does not negatively affect the power of the test as it results in better power than the random design, however, the sign change on the coefficient on x and y gives the least power among all other misspecified designs. For the regression relation $2 - 2x$, an increase or decrease in the coefficient does not affect the power of the test. The power obtained for all misspecifications outperformed that of the random design at all recovery proportions. A change in the sign on the coefficient also performs well irrespective of the coefficient with the sign change. A change in sign on the coefficient of x gives the least power among the misspecified values for the regression relation. For the MSE, the random design has the largest MSE values among all other designs.

In Tables A.3 and A.4 in the Appendix, different recovery proportions were considered for different misspecified designs in 10000 replicates for power and MSE respectively. As the sample size increases, the power increases. At all recovery proportions and sample sizes, the random selection performed worse than the true optimal design and misspecified designs while the misspecified designs performed less than the optimal design. Among all the misspecified designs, the one with the change in sign on the coefficient of y performs worse but better than the random design. The MSE decreases as the sample size and recovery proportion increases, the random design has the largest MSE values amongst all other designs.

TABLE 5.3: Power for different designs in 100000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
true optimal	0.287	0.463	0.595	0.672	0.724	0.758	0.767	0.796	0.809	0.830
random	0.243	0.385	0.493	0.576	0.642	0.693	0.740	0.774	0.803	0.830
Missing mechanism										
(-2,0.4,-0.30)	0.287	0.461	0.592	0.672	0.702	0.711	0.756	0.777	0.804	0.830
(-2.2,0.4,-0.15)	0.280	0.459	0.584	0.668	0.721	0.757	0.765	0.791	0.809	0.830
(-1.8,0.4,-0.15)	0.287	0.464	0.568	0.654	0.680	0.711	0.744	0.781	0.803	0.830
(2,0.4,-0.15)	0.283	0.393	0.578	0.589	0.694	0.750	0.759	0.790	0.804	0.830
(-2,0.44,-0.15)	0.286	0.461	0.574	0.671	0.722	0.738	0.757	0.772	0.804	0.830
(-2,0.36,-0.15)	0.286	0.462	0.594	0.672	0.723	0.757	0.762	0.789	0.806	0.830
(-2,-0.4,-0.15)	0.252	0.390	0.495	0.591	0.670	0.720	0.757	0.777	0.804	0.830
(-2,0.4,-0.165)	0.287	0.460	0.584	0.671	0.726	0.737	0.764	0.786	0.805	0.830
(-2,0.4,-0.135)	0.285	0.459	0.586	0.672	0.723	0.738	0.766	0.781	0.803	0.830
(-2,0.4,0.15)	0.252	0.391	0.501	0.597	0.665	0.719	0.755	0.781	0.807	0.830
Regression Coefficients										
(2.2-2x)	0.284	0.462	0.589	0.671	0.722	0.738	0.765	0.778	0.804	0.830
(1.8-2x)	0.285	0.463	0.594	0.648	0.703	0.730	0.755	0.783	0.805	0.830
(-2-2x)	0.285	0.462	0.581	0.671	0.712	0.747	0.764	0.785	0.807	0.830
(2-2.2x)	0.287	0.463	0.575	0.629	0.703	0.709	0.745	0.776	0.802	0.830
(2-1.8x)	0.285	0.461	0.594	0.650	0.700	0.730	0.755	0.780	0.803	0.830
(2+2x)	0.252	0.398	0.507	0.594	0.660	0.714	0.755	0.775	0.803	0.830

TABLE 5.4: MSE for different designs in 100000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
true optimal	0.0118	0.0066	0.0048	0.0040	0.0036	0.0033	0.0032	0.0030	0.0029	0.0027
random	0.0148	0.0084	0.0062	0.0050	0.0043	0.0038	0.0034	0.0032	0.0029	0.0027
Missing mechanism										
(-2,-0.4,-0.30)	0.0118	0.0066	0.0048	0.0040	0.0037	0.0036	0.0033	0.0031	0.0029	0.0227
(-2.2,0.4,-0.15)	0.0144	0.0075	0.0056	0.0044	0.0039	0.0035	0.0033	0.0031	0.0030	0.0027
(-1.8,0.4,-0.15)	0.0119	0.0066	0.0051	0.0042	0.0039	0.0037	0.0034	0.0031	0.0029	0.0027
(2,0.4,-0.15)	0.0120	0.0084	0.0050	0.0049	0.0038	0.0033	0.0032	0.0030	0.0029	0.0027
(-2,0.44,-0.15)	0.0119	0.0066	0.0051	0.0040	0.0038	0.0037	0.0033	0.0031	0.0029	0.0027
(-2,0.36,-0.15)	0.0119	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0030	0.0029	0.0027
(-2,-0.4,-0.15)	0.0141	0.0084	0.0061	0.0048	0.0041	0.0036	0.0033	0.0031	0.0029	0.0227
(-2,0.4,-0.165)	0.0118	0.0065	0.0049	0.0040	0.0036	0.0034	0.0032	0.0031	0.0029	0.0027
(-2,0.4,-0.135)	0.0119	0.0066	0.0049	0.0040	0.0036	0.0034	0.0032	0.0031	0.0029	0.0027
(-2,0.4,0.15)	0.0141	0.0083	0.0061	0.0048	0.0041	0.0036	0.0033	0.0031	0.0029	0.0027
Regression Coefficients										
(2.2-2x)	0.0119	0.0066	0.0049	0.0040	0.0036	0.0034	0.0032	0.0031	0.0029	0.0029
(1.8-2x)	0.0118	0.0066	0.0048	0.0042	0.0038	0.0035	0.0033	0.0031	0.0029	0.0027
(-2-2x)	0.0118	0.0067	0.0050	0.0040	0.0037	0.0034	0.0032	0.0031	0.0029	0.0027
(2-2.2x)	0.0118	0.0066	0.0051	0.0045	0.0038	0.0037	0.0034	0.0031	0.0029	0.0027
(2-1.8x)	0.0118	0.0066	0.0048	0.0042	0.0038	0.0035	0.0033	0.0031	0.0029	0.0027
(2+2x)	0.0142	0.0081	0.0059	0.0048	0.0041	0.0036	0.0033	0.0031	0.0029	0.0027

Tables A.5 and A.6 in the appendix show the Power and MSE for extreme cases of misspecification respectively. The extreme cases considered here involve a change in sign

and at least 100% increase in the coefficients of x and y. For the missing mechanisms: for $(-2, 0.4, 0.15)$, at all recovered proportions, the power exceeds that of random and we get smaller MSE compared to random. Implying that, even when the sign changes, the design performs better than the random design. $(-2, 0.4, 0.30)$ performs better than the random design both for power and MSE. For $(-2, 0.4, 0.60)$ and $(-2, 0.4, 0.30)$, a 400% and 600% increase on the coefficient of y respectively, the designs performed slightly below the random design at the recovery proportions 0.1 and 0.2, and better than random design as proportion increases. For the regression coefficients, a change in sign and increment in the coefficient of x were considered. At all values of recovery proportion, the misspecified designs outperformed the random design both in power and MSE.

Extreme cases of at least 100% increase in the coefficients of x and y without change in the sign were considered and the power and MSE results are shown in Tables 5.5 and 5.6 respectively. For the missing mechanisms: for $(-2, 0.4, -0.3)$, at all recovered proportions, the power exceeds that of random and smaller MSE compared to random. For $(-2, 0.4, 0.6)$ and $(-2, 0.4, 0.3)$, a 400% and 600% increase on the coefficient of y respectively, the designs also performed better than the random design at all recovery proportions. For the regression coefficients, an increment in the coefficient of x was considered. At all values of recovery proportion, the misspecified designs outperformed the random design both in power and MSE.

TABLE 5.5: Power for extreme designs for $n = 1000$ in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Missing Mechanism	true optimal	0.296	0.469	0.597	0.677	0.729	0.766	0.788	0.797	0.818	0.828
	random	0.243	0.395	0.487	0.574	0.636	0.690	0.741	0.777	0.802	0.828
	(-2,0.4,-0.3)	0.289	0.468	0.586	0.670	0.729	0.739	0.769	0.795	0.810	0.828
	(-2,0.4,-0.6)	0.249	0.399	0.496	0.665	0.710	0.727	0.764	0.794	0.809	0.828
	(-2,0.4,-0.9)	0.245	0.396	0.494	0.590	0.728	0.761	0.772	0.792	0.810	0.828
Regression Coefficient	(2-4x)	0.285	0.462	0.597	0.682	0.723	0.750	0.781	0.787	0.811	0.828
	(2-6x)	0.282	0.478	0.594	0.657	0.716	0.744	0.774	0.786	0.810	0.828
	(2-8x)	0.253	0.476	0.593	0.646	0.709	0.744	0.773	0.782	0.807	0.828

TABLE 5.6: MSE for extreme designs for $n = 1000$ in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Missing Mechanism	true optimal	0.0118	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0029	0.0028	0.0027
	random	0.0149	0.0085	0.0063	0.0051	0.0044	0.0039	0.0035	0.0032	0.0030	0.0027
	(-2,0.4,-0.3)	0.0118	0.0066	0.0049	0.0040	0.0036	0.0034	0.0032	0.0030	0.0029	0.0027
	(-2,0.4,-0.6)	0.0145	0.0082	0.0061	0.0040	0.0037	0.0035	0.0032	0.0030	0.0029	0.0027
	(-2,0.4,-0.9)	0.0118	0.0084	0.0063	0.0049	0.0035	0.0033	0.0031	0.0030	0.0029	0.0027
Regression Coefficient	(2-4x)	0.0119	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0030	0.0029	0.0027
	(2-6x)	0.0118	0.0066	0.0048	0.0042	0.0037	0.0034	0.0031	0.0030	0.0029	0.0027
	(2-8x)	0.0156	0.0066	0.0048	0.0043	0.0037	0.0034	0.0031	0.0030	0.0029	0.0027

Table A.7 in the Appendix shows the power and MSE for optimal design, random design and highest values design. The highest values here imply that the recovered values

are the highest of all the missing values for different recovery proportions. At 0.1, the 10% highest values from the missing values would recovered. The highest design is the worst design here as it has the least power and highest MSE at all values recovery proportion.

All in all, as the sample size increases, the power increases and the MSE decreases. The performance of the misspecified design in terms of power and MSE is in the direction of the sign. A change in sign affects the power more than an increment in the coefficient. From the results above, the design is robust to misspecifications because at all points of increment and change in sign, all the misspecified designs outperformed the random design.

TABLE 5.7: Power and MSE for different designs for $n = 1000$ in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Power	true optimal	0.296	0.469	0.597	0.677	0.729	0.766	0.788	0.797	0.818	0.828
	random	0.243	0.395	0.487	0.574	0.636	0.690	0.741	0.777	0.802	0.828
	Highest values	0.088	0.114	0.120	0.165	0.244	0.333	0.463	0.594	0.725	0.826
MSE	true optimal	0.0118	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0029	0.0028	0.0027
	random	0.0149	0.0085	0.0063	0.0051	0.0044	0.0039	0.0035	0.0032	0.0030	0.0027
	Highest values	0.1021	0.0414	0.0243	0.0160	0.0114	0.0084	0.0063	0.0048	0.0037	0.0028

Chapter 6

Subsampling based on probabilities

In Chapter 5, the recovery design consisted of a random recovery of responses whose covariates lie within a particular region. The recovered responses are combined with the observed variables restricted to this region to get the augmented data. In this chapter, we define the recovery design as an assignment of probabilities leading to a likelihood ratio test for testing MAR vs MNAR. This recovery design makes use of all the observed cases as opposed to the recovery design in Chapter 5 where only the observed cases that fall within the design region can be used.

Note that in this chapter, we have changed the missing mechanism indicator to $M = 1$ when Y is observed and 0 if missing. This is to make sure there is an alignment with the notation present in the literature. In Chapter 5, c_1 represents the recovery proportion and c_2 represents the proportion of observed data that lies in the recovery region which is added to the recovered units to get the augmented data. However, in this chapter, we represent the recovery proportion as c because there is no restriction on the recovery region and all the observed cases are used to augment the recovered cases. This chapter provides a generalization that can be used to obtain the results of Chapter 5.

In Section 6.1, we develop the testing framework for MNAR. Section 6.2 investigates optimal design using inspiration from T_E-optimality. In Section 6.3, we provide the results from a simulation study, which also investigates the robustness of our designs. Section 6.4 delves more deeply into the area of robustness, developing non-parametric alternatives to design construction.

6.1 Testing for MNAR

6.1.1 Problem formulation

To motivate our methodology, consider a setting involving a univariate response, Y, and a p-dimensional covariate vector $X = (X_1, \ldots, X_p)^T$. We assume we can express the conditional distribution, $p(Y|X = x, \theta)$ in closed form, where θ corresponds to the parameters characterising the distribution. We will initially assume $p(Y|X = x, \theta)$ is known before exploring robust approaches that will not require this knowledge. We also assume that the covariates X are fully observed so that only the response Y has missing values. This assumption is very common in the field, see (Kim and Yu, 2011). Let M be an indicator random variable that takes values 0 and 1. When Y is observed, M equals one and zero otherwise.

Under MAR, we have

$$\Pr(M|X = x, Y = y) = \Pr(M|X = x) \tag{6.1}$$

i.e. $M \perp\!\!\!\perp Y|X$, the absence of this conditional independence implies the presence of MNAR. This makes the inability to detect this relationship based on the original (incomplete) sample to be clearly evident. Y is only observed when $M = 1$ (by definition) and for the case $M = 0$, there are no Y values. This shows the need for a follow-up sample to recover a proportion of the missing Y values is necessary to be able to construct a statistical test to detect this property.

Let $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ be realisations of size n from the continuous random vector/variables (X, Y) with some distribution function partly characterised by $p(Y|X = x, \theta)$. Let $I = \{1, \ldots, n\}$ be the set of indices for n individuals. We assume that we have the capacity to make one more additional attempt to recover the response variable of a unit if it is missing at the first attempt. Let $I_1 \subset I$ be the set of individuals whose responses are captured at the first attempt, $I_2 \subset I \backslash I_1$ be the set of individuals whose responses are captured at the second attempt and let I_3 be the individuals not captured at all. Accordingly, we let $n_i = |I_i|$, for $i = 1, 2, 3$, denote the cardinality of each set.

We further assume the missing mechanism has the form (5.2).

Typical choices of g include the logit, probit, and complementary log-log link functions. By taking the inverse of the link function, an equivalent form of (5.2) that models the conditional distribution directly is (5.3).

In order to determine the type of missing mechanism present, the value of ψ needs to be determined, with MAR present when $\psi = 0$ and MNAR if otherwise.

Let $p_{i,1}$ be the probability that unit i is captured at the first attempt; that is

$$p_{i,1} := \Pr(M_i = 1 | X_i = x_i, Y_i = y_i)\,. \tag{6.2}$$

Typical forms for $p_{i,1}$ are constructed around the logit link function (Alho, 1990; Kim and Yu, 2011). Consequently, we will typically assume $p_{i,1}$ have the form:

$$p_{i,1} = p_1(\lambda, \psi; w_i, z_i) = \text{expit}(w_i^T \lambda + z_i^T \psi)\,. \tag{6.3}$$

where w_i and z_i denote the values of w and z at observation i.

After observing missing responses in the model, assume resources permit the follow-up of a number of experimental units with missing responses to obtain (recover) their responses. Since the covariates are assumed to be fully observed always, we then design the follow-up sample around the covariates with the missing responses. Define $p_{i,2}$ to be the probability that the response of unit i is captured at the second attempt given it was not captured at the first attempt. In this Chapter, there is a change in the concept of recovery design that was first introduced in Chapter 4. We then define the design as below.

Definition 6.1. A recovery design **D** is an assignment of the probabilities $p_{i,2}$.

Definition 6.1 can be linked to traditional Design of Experiments. In a continuous design, as defined in equation (2.1), we have a number of support points with corresponding weights, which reflect the proportion of observations to be made in each support point. When we need a design that can be run in practice, we usually round the weights such that the products of each weight with the sample size are integers that sum to the sample size. These will be the number of observations to be taken in each support point. For our scenario of a recovery design, we can view the covariate values of units with missing responses as 'potential' support points, and the allocation of probabilities $p_{i,2}$ to these potential support points then forms the design. This is somewhat similar in nature to a continuous design. However, when we need to run a recovery design in practice, we draw randomly (with probabilities $p_{i,2}$) from the set of potential support points. Both types of design, traditional and recovery, can be embedded in an Optimal Design of Experiments framework. In the traditional case, usually both support points and weights are optimised with respect to some criterion, whereas for the recovery design only the probabilities need to be optimised.

From hypothesis testing considerations, **D** should be designed in a way that maximises the power of a test for MNAR. This is of course subject to the constraint that we are unlikely to have the resources or capability to recover all of the missing information; that is, we are unlikely to have $p_{i,2} = 1$ for all i. To compare the effect of different designs and different tests for MNAR, we will consider the power as a function of the

proportion of recovered responses. That is, we will consider the power of the tests as a function of c with $n_2 = \lceil c \cdot (n - n_1) \rceil$ and $0 < c \leq 1$. How to select the probabilities $p_{i,2}$ under this constraint is one of the main considerations of this Chapter.

We will assume $p_{i,2} = p_2(x_i)$ where $p_2 : \mathbb{R}^p \to [0,1]$ is a continuous function. A convenient form for p_2 that we will typically take is

$$p_2(x) = p_2(x; \gamma_0, \gamma) := \text{expit}(\gamma_0 + \gamma^T x)\,, \tag{6.4}$$

γ_0 and $\gamma = (\gamma_1, \ldots, \gamma_p)^T$ are the parameters to be optimised. This model for $p_{i,2}$ provides adequate flexibility when trying to locate beneficial regions to recover missing responses although in theory any legitimate form for p_2 could be used. To ensure that, on average, we recover the correct proportion of missing responses given a specified value of c, we require for a given γ, the intercept γ_0 in (6.4) to be the solution of the equation

$$\int p_2(x; \gamma_0, \gamma) \Pr(X \in dx | M = 0) = c\,. \tag{6.5}$$

Henceforth, we shall assume γ_0 satisfies this condition. The problem of designing an 'optimal' follow-up design (given this expit form of p_2) is transformed to finding the 'optimal' value of γ.

6.1.2 Constructing the likelihood function

In this section, we construct the likelihood function for our data which is crucial for formulating a likelihood ratio test for MAR vs MNAR. The likelihood function can be obtained from the arguments of Alho (1990).

Define the unconditional probabilities

$$\mu_{i,1} := p_{i,1}, \quad \mu_{i,2} := p_{i,2}(1 - p_{i,1}), \quad \mu_{i,3} := 1 - \mu_{i,1} - \mu_{i,2}\,, \tag{6.6}$$

and let $U_i := (U_{i,1}, U_{i,2}, U_{i,3})^T$, where $U_{i,1} = 1$ if the response for unit i is captured at the first attempt, $U_{i,2} = 1$ if the response for unit i is captured at the second attempt and $U_{i,3} = 1$ if the response for unit i is not captured at all. Note the random variable $U_{i,1}$ has the same distribution as M. We assume

$$U_i \sim \text{Multinomial}(1, \mu_{i,1}, \mu_{i,2}, \mu_{i,3})\,. \tag{6.7}$$

Let $R_i = 1 - U_{i,3}$ where $R_i = 1$ if the response for unit i is captured, otherwise zero. Construct the vectors

$$U = (U_1, \ldots, U_n)^T, \;\; R = (R_1, \ldots, R_n)^T, \tag{6.8}$$

and define the probabilities

$$v_{i,1} = v_1(\lambda, \psi; w_i, z_i) = \frac{\mu_{i,1}}{\mu_{i,1} + \mu_{i,2}}, \qquad v_{i,2} = v_2(\lambda, \psi; w_i, z_i) = 1 - v_{i,1} = \frac{\mu_{i,2}}{\mu_{i,1} + \mu_{i,2}}. \tag{6.9}$$

Then the conditional likelihood of U given R is (up to a constant) given by

$$L(U|R) = \prod_{R_i=1} v_{i,1}^{U_{i,1}} v_{i,2}^{U_{i,2}},$$

and consequently, the log-likelihood is given by

$$\begin{aligned} \ell(\lambda, \psi) &= \sum_{R_i=1} \sum_{j=1}^{2} U_{i,j} \log(v_{i,j}) \\ &= \sum_{i \in I_1 \cup I_2} U_{i,1} \log(v_{i,1}) + (1 - U_{i,1}) \log(1 - v_{i,1}). \end{aligned} \tag{6.10}$$

Here we have used the fact that if an observation is observed then $U_{i,2} = 1 - U_{i,1}$. For $p_{i,2} = p_2(x; \gamma_0, \gamma)$ with $p_2(x; \gamma_0, \gamma)$ given in (6.4), we obtain

$$\begin{aligned} v_{i,1} = v_1(\lambda, \psi; w_i, z_i, \gamma) &= \frac{p_{i,1}}{p_{i,1} + p_{i,2}(1 - p_{i,1})} \\ &= \frac{g^{-1}(w_i^T \lambda + z_i^T \psi)}{g^{-1}(w_i^T \lambda + z_i^T \psi) + p_2(x_i; \gamma_0, \gamma) \cdot (1 - g^{-1}(w_i^T \lambda + z_i^T \psi))}, \end{aligned}$$

where we have introduced γ into the notation to highlight the dependence on the recovery design.

6.1.3 A mixture distribution for the augmented data

The observed data combined with the recovered data, that is all observations $i \in I_1 \cup I_2$, will be referred to as the augmented data. The augmented data, being a combination of observed and recovered data, has a natural mixture distribution which we will formulate in the following lemma.

Definition 6.2. The indicator variable M_A follows a Bernoulli distribution, denoted $M_A \sim \text{Ber}(p)$, where the probability parameter p is given by:

$$p = \frac{\Pr(M = 1)}{c \cdot \Pr(M = 0) + \Pr(M = 1)}.$$

Therefore, M_A is defined by:

$$M_A := \begin{cases} 0 & \text{with probability } \dfrac{c \cdot \Pr(M = 0)}{c \cdot \Pr(M = 0) + \Pr(M = 1)} \\ 1 & \text{with probability } \dfrac{\Pr(M = 1)}{c \cdot \Pr(M = 0) + \Pr(M = 1)} \end{cases}. \tag{6.11}$$

Then the augmented response/covariates are realisations from random variable/vector:

$$Y_A := M_A Y_O + (1 - M_A) Y_R \tag{6.12}$$
$$X_A := M_A X_O + (1 - M_A) X_R \,, \tag{6.13}$$

where $Y_O := Y \mid M = 1$; $X_O := X \mid M = 1$, and X_R and Y_R have distribution functions

$$\begin{aligned}
\Pr(X_R \in dx) &= \Pr(X \in dx | M = 0, U_{i,2} = 1) = p_2(x) \Pr(X \in dx | M = 0)/c \,, \\
\Pr(Y_R \in dy) &= \Pr(Y \in dy | M = 0, U_{i,2} = 1) = \int \Pr(Y \in dy | X = x, M = 0) \Pr(X_R \in dx) \,.
\end{aligned}$$

The probabilities $v_1(\lambda, \psi; w, z, \gamma)$ and $v_2(\lambda, \psi; w, z, \gamma)$ correspond to the missing data mechanism in the augmented data. More precisely, they satisfy

$$\begin{aligned}
v_1(\lambda, \psi; w, z, \gamma) &= \Pr(M_A = 1 \mid X_A = x, Y_A = y), \\
v_2(\lambda, \psi; w_i, z_i, \gamma) &= \Pr(M_A = 0 \mid X_A = x, Y_A = y) \,.
\end{aligned}$$

6.1.4 SMF tests for MNAR

6.1.4.1 The likelihood ratio test

The ability to construct the log-likelihood in (6.10) means the likelihood ratio test can provide a test for MAR vs MNAR. The likelihood ratio test statistic for testing a general hypothesis $\psi = \psi_0$ is given by $2[\ell(\hat{\lambda}, \hat{\psi}) - \ell(\hat{\lambda}_0, \psi_0)]$, where $(\hat{\lambda}, \hat{\psi})$ and $\hat{\lambda}_0$ denotes the maximum likelihood estimators under the alternative and null models, respectively. The quantity $2[\ell(\hat{\lambda}, \hat{\psi}) - \ell(\hat{\lambda}_0, \psi_0)]$, often referred to as the reduction in deviance, is given by

$$\begin{aligned}
&2[\ell(\hat{\lambda}, \hat{\psi}) - \ell(\hat{\lambda}_0, \psi_0)] \\
&= \sum_{i \in I_1 \cup I_2} U_{i,1} \log \left(\frac{v_1(\hat{\lambda}, \hat{\psi}; w_i, z_i)}{v_1(\hat{\lambda}_0, \psi_0; w_i, z_i)} \right) + (1 - U_{i,1}) \log \left(\frac{1 - v_1(\hat{\lambda}, \hat{\psi}; w_i, z_i)}{1 - v_1(\hat{\lambda}_0, \psi_0; w_i, z_i)} \right) \,.
\end{aligned}$$

When testing for MAR vs MNAR, we set $\psi_0 = 0$. For large $n_1 + n_2$, under the null hypothesis the statistic $2[\ell(\hat{\lambda}, \hat{\psi}) - \ell(\hat{\lambda}_0, \psi_0)]$ will be approximately chi-square distributed with s degrees of freedom, i.e. a classical statistical approximation. Unfortunately, the maximum likelihood estimators $\hat{\lambda}$, $\hat{\psi}$ and $\hat{\lambda}_0$ cannot be explicitly written and must be found numerically.

6.1.4.2 A benchmark recovery design

A random recovery of a proportion of the missing responses is perhaps the simplest and most intuitive initial design; this corresponds to selecting $p_{i,2} = p_2(x_i; \gamma_0, \gamma) = c$ for all i and therefore $\gamma = 0$. We shall denote this design by $\mathbf{D_B}$ as it will form the benchmark design for numerical comparisons in later sections. For this choice of p_2 and for any link function g, we have

$$v_1(\lambda, \psi; w, z, \gamma) = \frac{g^{-1}(w^T\lambda + z^T\psi)}{g^{-1}(w^T\lambda + z^T\psi) + c \cdot (1 - g^{-1}(w^T\lambda + z^T\psi))}\,, \tag{6.14}$$

which, for the logit model, simplifies to

$$v_1(\lambda, \psi; w, z, \gamma) = \frac{\exp(w^T\lambda + z^T\psi)}{c + \exp(w^T\lambda + z^T\psi)} = \text{expit}(w^T\lambda_A + z^T\psi)\,, \tag{6.15}$$

where $\lambda_A := \lambda - (\log(c), 0, \ldots, 0)^T \in \mathbb{R}^q$. Recall, λ and ψ are the unknown vectors of coefficients. w is a q-dimensional vector whose components could depend on functions of x but not y, and z is an s-dimensional vector whose components additionally depend on functions of y, for example an interaction $x_i y$, and (or) a function of just y. c is the recovery proportion and γ is the parameter to be optimised.

In this scenario, $v_{i,1}$ corresponds to the link function of a logit model and therefore (6.10) will correspond to a likelihood function of a logistic regression model with a shifted intercept. This is equivalent to saying for the augmented data

$$\Pr(M_A = 1 \mid Y_A = y, X_A = x) = \text{expit}(w^T\lambda_A + z^T\psi)\,.$$

This is a particularly appealing property since one can take advantage of existing glm software. For example, in R we can directly use the function *glm* with a logit link function and can avoid having to manually find all maximum likelihood estimators. Importantly, the coefficient ψ remains unchanged and the estimators will be consistent with this approach.

6.1.4.3 A random recovery within a region

Here, we will prove that the results of Chapter 5 can be obtained from the results of this Chapter. If instead of selecting $p_{i,2} = c$ for all i, one could set $p_{i,2} = \textit{constant}$ for particular i and zero for the remaining i. For example, let $\mathcal{R}_A \subseteq \mathbb{R}^p$ be a p-dimensional region chosen large enough so that

$$\Pr(M = 0, X \in \mathcal{R}_A) \geq c \cdot \Pr(M = 0)\,. \tag{6.16}$$

Then for

$$p_{i,2} = \begin{cases} c \cdot \Pr(M = 0)/\Pr(M = 0, X \in \mathcal{R}_A) & \text{if } x_i \in \mathcal{R}_A \\ 0 & \text{otherwise,} \end{cases} \tag{6.17}$$

the log-likelihood in (6.10) becomes:

$$\ell(\lambda, \psi) = \sum_{\substack{i \in I_1 \cup I_2 \\ i: x_i \in \mathcal{R}_A}} U_{i,1} \log(v_{i,1}) + (1 - U_{i,1}) \log(1 - v_{i,1}), \tag{6.18}$$

where $v_{i,1}$ has the same form as (6.14), but with c replaced with $c^* = c \cdot \Pr(M = 0)/\Pr(M = 0, X \in \mathcal{R}_A)$.

A careful choice of $\mathcal{R}_A$ can improve the power of the LRT over $\mathbf{D_B}$. This has been shown in Chapter 5.

6.1.4.4 Tests for MAR vs MNAR

We will now explicitly formulate two tests for MAR vs MNAR that come under the umbrella of the likelihood ratio test mentioned at the beginning of this section. The first test will cover more general choices of $p_{i,2}$ and does not require any restrictions on the augmented data. The main drawback is pre-existing software packages cannot be used and estimators have to be manually found.

The second test focuses on the logit model with $p_{i,2}$ chosen according to (6.17) which, as shown, has the advantage that a logistic regression model can be fitted directly to the augmented data and correct inferences about ψ can be made. It does, however, require restricting the augmented data to lie within $\mathcal{R}_A$ which could result in a loss of power by ignoring some observations.

Test T_1**.** *For* $0 < c \leq 1$ *and a given* γ*, select* $p_{i,2}$ *according to* (6.4) *where* γ_0 *is the solution of* (6.5)*. Perform the likelihood ratio test for the hypothesis* $\psi = 0$ *using the likelihood function* (6.10)*.*

Test T_2**.** *For* $0 < c \leq 1$*, provided* $p_{i,1}$ *have the form of* (6.3)*,* $\mathcal{R}_A$ *satisfies* (6.16) *and* $p_{i,2}$ *are selected according to* (6.17)*, perform the likelihood ratio test for the hypothesis* $\psi = 0$ *using the likelihood function* (6.18)*. This is equivalent to fitting a logistic regression model to the augmented data within* $\mathcal{R}_A$ *with linear predictor* $w^T\lambda_A + z^T\psi$ *and testing whether* $\psi = 0$*.*

6.2 Designing the recovery

6.2.1 T-optimality criteria

In experimental design, T-optimality (and some of its analogues) is a criterion used for discriminating between competing forms of a model (Atkinson and Fedorov, 1975; de Leon and Atkinson, 1992; Waterhouse et al., 2008; Tommasi and López-Fidalgo, 2010). The construction of designs for model discrimination experiments for linear or nonlinear models was first studied in Atkinson and Fedorov (1975) where the idea of T-optimality was first formulated. For linear or non-linear models, the main idea behind T-optimality is to select a design that maximises the non-centrality parameter in the $F-$test. For discrimination designs for glm's, de Leon and Atkinson (1992) formulated the T-optimality criterion in terms of the deviance. In Waterhouse et al. (2008), a criterion similar to the T-optimality was formulated as T_E-optimality. This criterion, which has more appealing statistical properties like an asymptotic chi-square distribution under the null hypothesis, selects a continuous design that maximises the expected reduction in deviance.

There are two main issues that prevent any direct application of previous results in the design literature on T-optimality. The first and obvious issue is the presence of missing observations. The second is that the covariates are not selected in advance; instead, we are only given covariates after the outcome from an experiment with missing data. Nevertheless, we can obtain insight for formulating the optimality criterion for the problem considered.

Using inspiration from T_E-optimality, we will look to maximise the expected reduction in deviance by tuning γ, where the expectation is taken with respect to X_A and Y_A under the alternative hypothesis of MNAR. Taking expectation of the likelihood ratio statistic given in Section 6.1.4.1 provides:

$$\begin{aligned}&\mathbb{E}\{2\left[\ell(\hat{\lambda},\hat{\psi})-\ell(\hat{\lambda}_0,0)\right]\}=2\mathbb{E}(n_1+n_2)\\&\quad\times\ \mathbb{E}\left[M_A\log\left(\frac{v_1(\hat{\lambda},\hat{\psi};W_A,Z_A,\gamma)}{v_1(\hat{\lambda}_0,0;W_A,Z_A,\gamma)}\right)+(1-M_A)\log\left(\frac{1-v_1(\hat{\lambda},\hat{\psi};W_A,Z_A,\gamma)}{1-v_1(\hat{\lambda}_0,0;W_A,Z_A,\gamma)}\right)\right],\end{aligned}$$

The capitalisation W_A and Z_A and the subscript A emphasises that the elements of w and z are now functions of the random variables X_A and Y_A. Since $\mathbb{E}(n_1+n_2)$ remains constant for each design, one could select γ on the basis of the following criterion:

$$\begin{aligned}&T(n,\gamma^*)=\\&\max_{\gamma}\mathbb{E}\left[M_A\log\left(\frac{v_1(\hat{\lambda},\hat{\psi};W_A,Z_A,\gamma)}{v_1(\hat{\lambda}_0,0;W_A,Z_A,\gamma)}\right)+(1-M_A)\log\left(\frac{1-v_1(\hat{\lambda},\hat{\psi};W_A,Z_A,\gamma)}{1-v_1(\hat{\lambda}_0,0;W_A,Z_A,\gamma)}\right)\right]\end{aligned}\tag{6.19}$$

However, computing this expectation theoretically using Lemma 6.2 or even estimating via Monte Carlo methods is impractical. For large values of $n_1 + n_2$, it will be more convenient to consider an asymptotic form of (6.19). By doing so, we will make the assumption we know the missing data mechanism including the values of λ and ψ. Therefore any claimed optimality will only be classed as local. We select γ according to:

$$T(\gamma^*) := \lim_{n\to\infty} T(n,\gamma^*) =$$
$$\max_{\gamma}\min_{\lambda_0} \mathbb{E}\left[M_A \log\left(\frac{v_1(\lambda,\psi;W_A,Z_A,\gamma)}{v_1(\lambda_0,0;W_A,Z_A,\gamma)}\right) + (1-M_A)\log\left(\frac{1-v_1(\lambda,\psi;W_A,Z_A,\gamma)}{1-v_1(\lambda_0,0;W_A,Z_A,\gamma)}\right)\right]$$

To compute the expectation in $T(\gamma^*)$, one can exploit the forms of (6.12) and (6.13) in Lemma 6.2. After conditioning on $M_A = 1$ and $M_A = 0$, one needs to know the joint distributions of (X_R, Y_R) and (X_O, Y_O) respectively to compute the required expectation.

Lemma 6.3. *For $dx = (dx_1, \ldots, dx_p)$ a vector of infinitesimals we have*

$$\begin{aligned}\Pr(X_R \in dx, Y_R \in dy) &= \Pr(Y \in dy \,|\, X = x, M = 0)\Pr(X_R \in dx)\\ \Pr(X_O \in dx, Y_O \in dy) &= \Pr(Y \in dy \,|\, X = x, M = 1)\Pr(X_O \in dx)\,.\end{aligned}$$

One could then use Lemma 6.2 to compute the distribution of X_R and X_O assuming one knows the distribution of $X, p(Y|X = x, \theta)$ and the missing data mechanism. We will show in Section 6.4 that some of these fairly strong assumptions can be avoided. In practice, whilst one could evaluate the expectation theoretically, we would recommend using Monte Carlo methods for simplicity. Using the fact

$$\Pr(M_A = 1 | X_A = x, Y_A = y) = v_1(\lambda,\psi;w,z,\gamma)\,,$$

we can equivalently express our criterion as

$$\begin{aligned}T(\gamma^*) &= \max_{\gamma}\min_{\lambda_0}\mathbb{E}\left[v_1(\lambda,\psi;W_A,Z_A,\gamma)\log\left(\frac{v_1(\lambda,\psi;W_A,Z_A,\gamma)}{v_1(\lambda_0,0;W_A,Z_A,\gamma)}\right)\right.\\ &+ \left.(1-v_1(\lambda,\psi;W_A,Z_A,\gamma))\log\left(\frac{1-v_1(\lambda,\psi;W_A,Z_A,\gamma)}{1-v_1(\lambda_0,0;W_A,Z_A,\gamma)}\right)\right],\end{aligned}$$

which is related to the expected Kullback-Leibler divergence. Criteria similar to this have been studied in Tommasi and López-Fidalgo (2010). The main algorithm for determining the recovery design in test T_1 is as follows.

If we are using test T_2, instead of optimising with respect to γ, we seek to optimise over the choice of the region $\mathcal{R}_A$. We will only consider scenarios where $\mathcal{R}_A$ is in the form

Algorithm 5 Algorithm 1 for T_1

1: **Input:** Value $0 < c \leq 1$, function $T(\gamma^*)$.
2: **Output:** Value of γ.
3: **Steps:**
4: Approximate $T(\gamma^*)$ using the provided function $T(\gamma^*)$.
5: Choose γ based on the approximation of $T(\gamma^*)$.
6: **Return:** γ.

of a hypercuboid as mentioned in Section 6.1.4.4. An equivalent criterion is

$$C\left(\mathcal{R}_A^*\right) := \max_{\mathcal{R}_A} \min_{\lambda_0} \mathbb{E}\left[M_A \log\left(\frac{v_1(\lambda,\psi;W_A,Z_A,\gamma)}{v_1(\lambda_0,0;W_A,Z_A,\gamma)}\right) + (1-M_A)\log\left(\frac{1-v_1(\lambda,\psi;W_A,Z_A,\gamma)}{1-v_1(\lambda_0,0;W_A,Z_A,\gamma)}\right)\right].$$

With this criterion, we formulate the algorithm for determining the recovery design when using test T_2:

Algorithm 6 Algorithm 1 for T_2

1: **Input:** Value $0 < c \leq 1$, function $C(\mathcal{R}_A^*)$.
2: **Output:** Recovery region $\mathcal{R}_A$.
3: **Steps:**
4: Choose $\mathcal{R}_A$ according to $C(\mathcal{R}_A^*)$.
5: **Return:** $\mathcal{R}_A$.

6.3 Simulation studies

In this section, we perform a simulation study assessing the benefits of Algorithm 1 for tests T_1 and T_2. We will consider two cases: $p = 1$ and $p = 2$. For $p = 1$, we generate n points as follows from $Y|(X = x) \sim N(\beta_0 + \beta_1 x, \sigma_y^2)$ with $X \sim N(\mu_x, \sigma_x^2)$ and MNAR missingness is introduced into y values using $\Pr(M = 1|Y = y, X = x) = \text{expit}(\alpha_0 + \alpha_1 x + \alpha_2 y)$. The parameters are chosen such that approximately 30% of points are missing their y value. We will repeat this process 10,000 times and in each replication apply the tests, with $H_0 : \psi = 0$, to the generated sample.

(a) $n = 400$, $(\beta_0, \beta_1) = (1, 0.7)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (2, 16)$, $(\alpha_0, \alpha_1, \alpha_2) = (-0.2, 0.8, 0.6)$.
(b) $n = 1000$, $(\beta_0, \beta_1) = (2, -2)$, $\sigma_y^2 = 4$, $(\mu_x, \sigma_x^2) = (0, 16)$, $(\alpha_0, \alpha_1, \alpha_2) = (2, 0.4, -0.15)$.

In the examples above, we increased the sample size from 400 in example (a) to 1000 in example (b). The value of β_0 increases from 1 to 2, and β_1 decreases from 0.7 to -2. The variance of Y, σ_y^2, increases from 1 to 4. The mean of x decreases from 2 to 0, while the variance remains constant. We increased the value of α_0 from -0.2 to 2, and reduced α_1 and α_2 from 0.8 to 0.4, and from 0.6 to -0.15, respectively. Despite the changes in

parameter values, there is no change in the results: Algorithm 1 for T_1 has the largest power, followed by Algorithm 1 for T_2, with the random design having the least power.

In Figures 6.1 and 6.2, as a function of the recovery proportion c, we depict the power of three tests for detecting the MNAR scenarios described in (a) and (b) respectively. The dashed red line represents the power of test T_1 using Algorithm 1 for T_1 to design the recovery design. With a dotted blue line, we plot the power of the test T_2 using the recovery design constructed from Algorithm 1 for T_2. With a solid black line, we plot the power of the test T_1 or T_2 using a random recovery; the test T_1 with $\gamma = 0$ and the test T_2 with $\mathcal{R}_A = \mathbb{R}^d$ are equivalent and both situations correspond to a random recovery. From this figure, we see considerable gains over random sampling for both tests. The test T_1 utilising Algorithm 1 has more power than test T_2 with Algorithm 1 for the scenarios considered here. This is due to the fact that when using test T_2, we have to restrict the augmented data to lie within $\mathcal{R}_A$. This loss of observations is the cause of a slight loss in power. The two designs perform better than the random design.

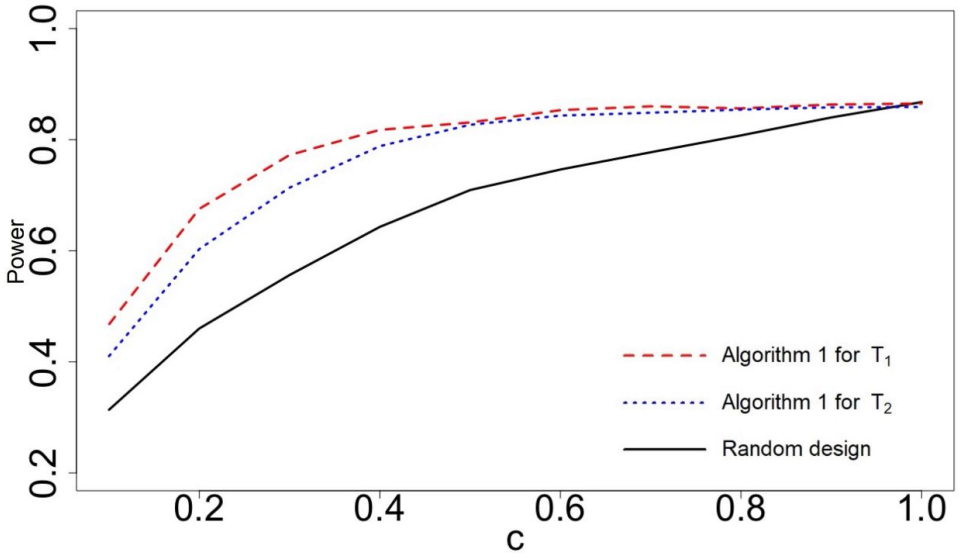

FIGURE 6.1: Power for different recovery proportions for example (a): Red.

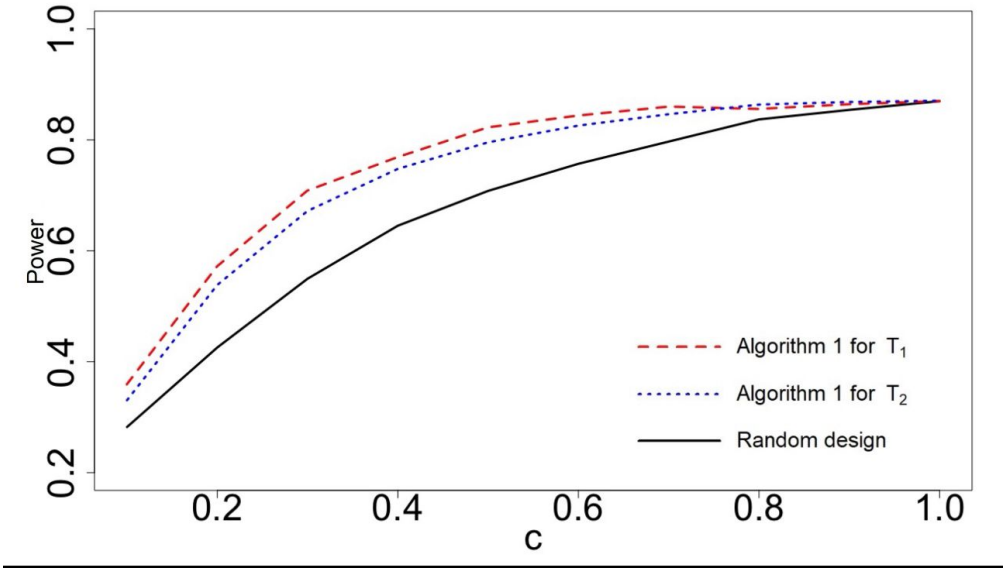

FIGURE 6.2: Power for different recovery proportions for example (b): Red.

In the following examples, we will take $p = 2$ and generate n points as follows from $Y|(X = x) \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma_y^2)$ with $X_1 \sim N(\mu_{1,x}, \sigma_{1,x}^2)$ and $X_2 \sim N(\mu_{2,x}, \sigma_{2,x}^2)$. MNAR missingness is introduced into y values using

$$\Pr(M = 1|Y = y, X = x) = \text{expit}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 y)\,.$$

We consider the following scenarios.

(c) $n = 1000$, $(\beta_0, \beta_1, \beta_2) = (2, -2, 2)$, $\sigma_y^2 = 4$, $(\mu_{1,x}, \mu_{2,x}, \sigma_{1,x}^2, \sigma_{2,x}^2) = (2, 2, 16, 16)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-2.7, 0.4, 0.2, -0.15)$.
(d) $n = 400$, $(\beta_0, \beta_1, \beta_2) = (0, 2, -2)$, $\sigma_y^2 = 1$, $(\mu_{1,x}, \mu_{2,x}, \sigma_{1,x}^2, \sigma_{2,x}^2) = (-2, 0, 4, 16)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (2.9, -0.4, 0.4, 0.5)$.

In the examples above, we varied some of the parameters to observe if there would be any changes in the performance of the designs. In example (d), the sample size was reduced from 1000 to 400. The value of β_0 was reduced to 0, β_1 was increased to 2, and β_2 was decreased to -2. The variance of Y was reduced from 4 in example (c) to 1 in example (d). The mean and variance of X_1 were reduced to -2 and 4, respectively. The mean of X_2 was reduced to 0, while its variance remained the same as in example (c). Additionally, α_0 increased from -2.7 to 2.9, α_1 changed from 0.4 to -0.4, and α_2 increased from -0.15 to 0.5.

Despite these changes in parameter values, the performance of the designs remained unaffected across the examples. In both examples, the random design performed

worse than Algorithm 1 for T_1.

Figure 6.3 and Figure 6.5 show the power for T_1 with Algorithm 1 against the random design for (c) and (d) respectively. Figure 6.4 and Figure 6.6 show the optimal values of $\gamma = (\gamma_1, \gamma_2)$ as a function of c obtained from using this algorithm for (c) and (d) respectively. The dot-dashed orange line shows the value of γ_1 whereas the dashed green line shows the value of γ_2. The benefit of both algorithms over the random recovery becomes even more evident under the scenarios considered here, especially for smaller values of c where the power of the optimised tests is significantly higher.

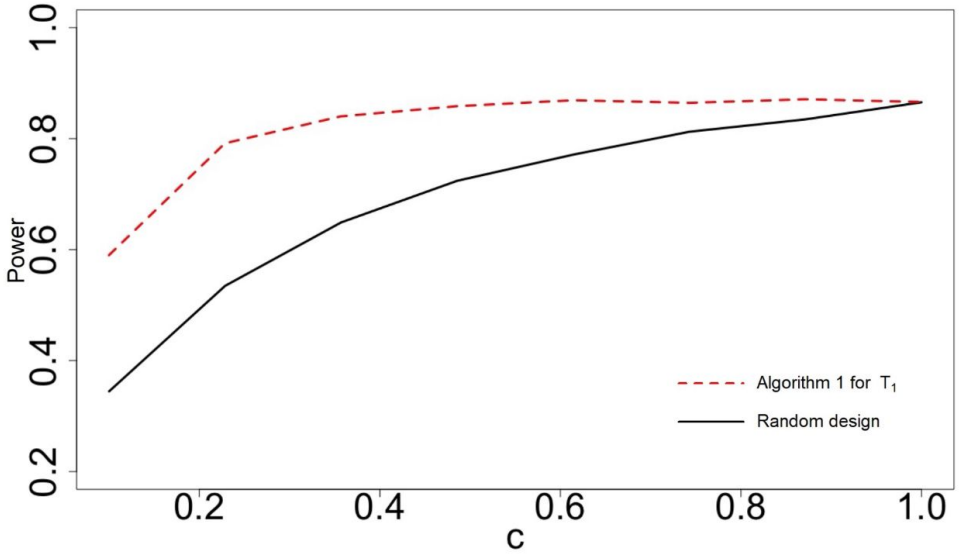

FIGURE 6.3: Power for different recovery proportions for example (c).

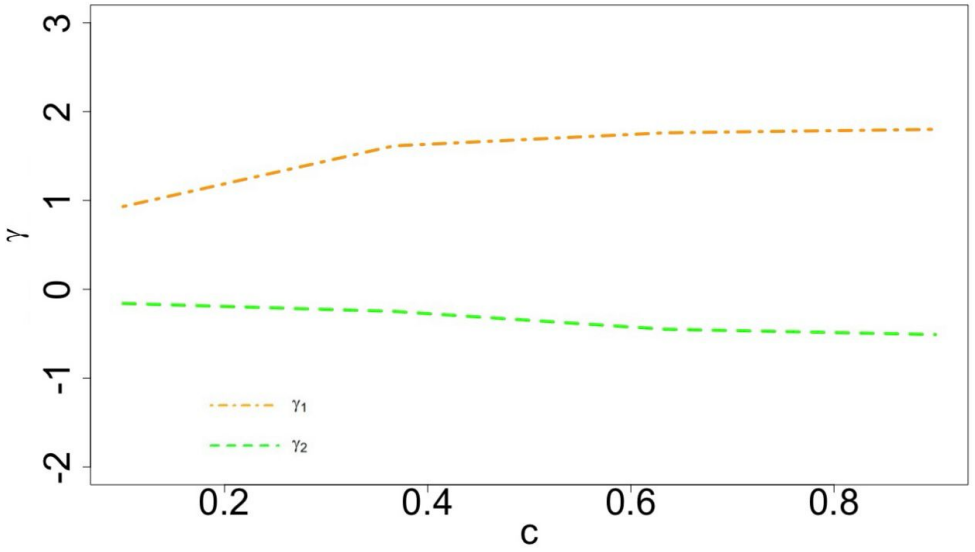

FIGURE 6.4: Optimal values of γ for different recovery proportions with example (c).

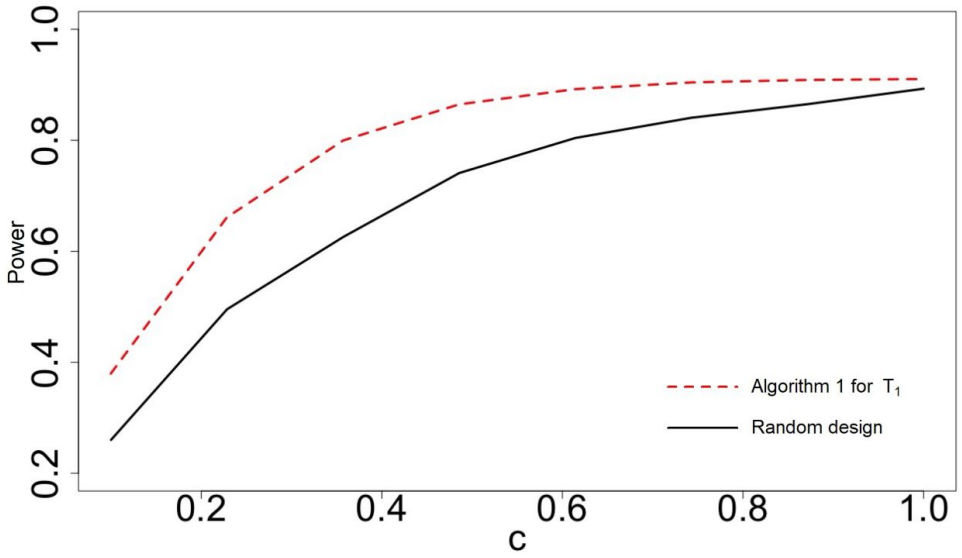

FIGURE 6.5: Power for different recovery proportions for example (d).

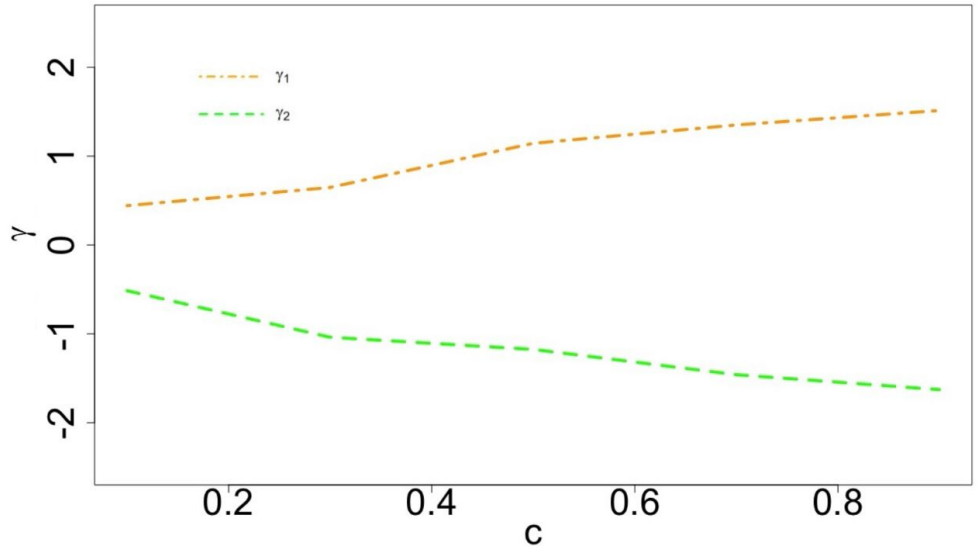

FIGURE 6.6: Optimal values of γ for different recovery proportions with example (d).

For an example with interaction between the covariates and the response variable, take $p = 1$ and include an interaction term into the missing data mechanism.
Generate $n = 1000$ points as follows:

$$Y|(X = x) \sim N(2 - 0.5x, 4)\,,$$

with $X \sim N(1, 4)$. We study the performance of the algorithm 1 for T_1 and random design for two MNAR mechanisms:

(e) $P(M = 1|Y = y, X_1 = x_1)$ =expit$(1 + 0.5x_1 + 0.05y + 0.05x_1y)$
(f) $P(M = 1|Y = y, X_1 = x_1)$ =expit$(1.8 - 0.5x_1 + 0.05y - 0.04x_1y)$.

In the examples above, using the same sample size and dataset, we introduced missingness in Y using the expit function with different parameters when the interaction term was included. In comparison to example (e), example (f) has a larger α_0, the same α_2, and smaller α_1 and α_3.

As shown in Figure 6.7, we see for Example (e) that the optimised recovery design produces significant benefits over the pure random recovery. This is not the case for Example (f) in Figure 6.8, where for this MNAR mechanism the optimised recovery appears to produce identical results, in terms of power, to pure random recovery. This is important, as it highlights an instance where the optimal recovery design is not qualitatively better than random recovery, but nevertheless we see that the optimal design does not perform any worse than random recovery.

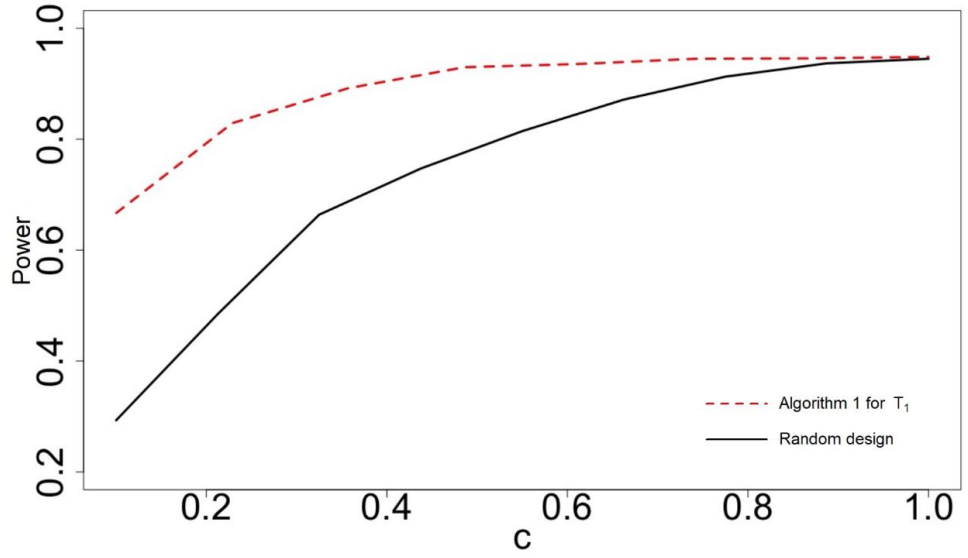

FIGURE 6.7: Power for different recovery proportions and designs with Example (e).

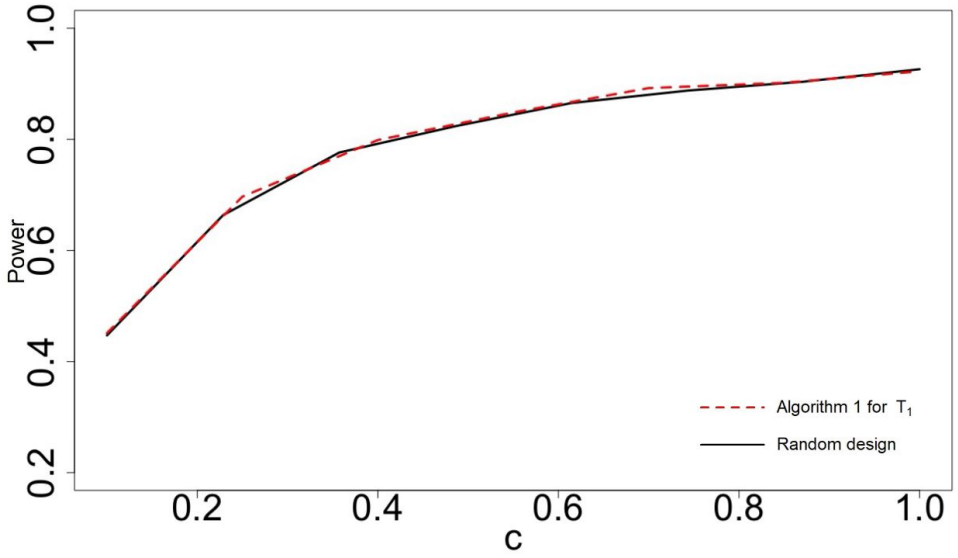

FIGURE 6.8: Power for different recovery proportions and designs with Example (f).

6.3.1 Assessing the robustness

In this section, we assess the robustness of Algorithm 1 for test T_1 as the robustness of Algorithm 1 for test T_2 has been discussed in Chapter 5. For a univariate and multivariate cases, we misspecify some of the parameters to examine the power of the test.

For $p = 1$, generate n points from $Y|(X = x) \sim N(\beta_0 + \beta_1 x, \sigma_y^2)$ with $X \sim N(\mu_x, \sigma_x^2)$.

Introduce MNAR missingness into y values using $\Pr(M = 1|Y = y, X = x) = \text{expit}(\alpha_0 + \alpha_1 x + \alpha_2 y)$. This selection introduces approximately 30% of points missing their y value. We will repeat this process 10,000 times and in each replication, test the hypothesis $H_0 : \psi = 0$, to the generated sample.

(a) $n = 400$, $(\beta_0, \beta_1) = (1, 0.7)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (2, 16)$, $(\alpha_0, \alpha_1, \alpha_2) = (-0.2, 0.8, 0.6)$.

Table 6.1 shows the power of algorithm 1 for test T_1 (True optimal), random design and some misspecified designs. At all values of recovery proportion c, the test T_1 performs better than the other designs. All other misspecified designs outperformed the random design at all values of c except at $(-0.2, -0.8, 0.6)$ where there is a change in sign for the coefficient of x in the missing mechanism model. This leads us to vary the values of the α_1 and α_2 from the true value to different values to see how this affects the power of the test. Tables 6.2 and 6.3 show the γ_1 values and the corresponding power for different values of α_1 and α_2. The designs with misspecified α_2 performs better than the random design while the designs with $(-0.2, -0.8, 0.6)$, $(-0.2, -0.7, 0.6)$ and $(-0.2, -0.6, 0.6)$ performed slightly worse than the random design. This worse performance than the random occurs when α_1, the coefficient of x, whose true value is $+0.8$ is

severely misspecified. This worse in performance occurs when the misspecified α_1 lies between -0.5 and -0.4. Neglecting the error in the regression model and substituting the regression relation for y into the missing mechanism will lead to a sign change in the coefficient of x when the misspecified α_1 is -0.42. This may in turn cause a sign change of the optimal γ_1, and a design with the incorrect sign for γ_1 is likely to perform worse than the random design (which corresponds to $\gamma_1 = 0$). This shows that a severe misspecification on the coefficient of x is more likely to affect the test than that of y. Therefore, a change in the sign of α_1 affects the power of the test.

TABLE 6.1: Power for different designs in 10000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.484	0.700	0.800	0.837	0.850	0.858	0.860	0.862	0.863
Random	0.304	0.443	0.546	0.643	0.700	0.753	0.799	0.817	0.840
Missing mechanism									
(0.2,0.8,0.6)	0.478	0.681	0.797	0.834	0.845	0.855	0.857	0.860	0.862
(-0.2,-0.8,0.6)	0.280	0.373	0.419	0.444	0.448	0.538	0.595	0.685	0.773
(-0.2,0.8,-0.6)	0.426	0.628	0.727	0.783	0.828	0.842	0.851	0.857	0.858
(-0.4,0.8,0.6)	0.477	0.690	0.797	0.834	0.845	0.856	0.858	0.860	0.862
(-0.2,0.6,0.6)	0.485	0.684	0.778	0.818	0.838	0.849	0.854	0.858	0.859
(-0.2,0.8,0.9)	0.478	0.692	0.794	0.828	0.847	0.853	0.858	0.859	0.862
(-0.1,0.8,0.6)	0.480	0.689	0.795	0.828	0.847	0.855	0.859	0.860	0.863
(-0.2,0.4,0.6)	0.474	0.688	0.767	0.818	0.836	0.850	0.854	0.858	0.860
(-0.2,0.8,0.3)	0.479	0.687	0.800	0.825	0.845	0.857	0.860	0.862	0.863
Regression Coefficients									
(1-0.7x)	0.420	0.621	0.729	0.788	0.838	0.850	0.859	0.860	0.861
(-1+0.7x)	0.480	0.692	0.794	0.828	0.847	0.856	0.858	0.860	0.862
(1+0.4x)	0.469	0.686	0.797	0.817	0.845	0.857	0.859	0.860	0.863
(2+0.7x)	0.474	0.689	0.798	0.825	0.848	0.857	0.858	0.860	0.862
(-1-0.7x)	0.418	0.617	0.730	0.788	0.838	0.844	0.859	0.860	0.862

TABLE 6.2: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.8768	0.9934	0.9167	0.9532	0.9544	0.8251	0.9494	0.9605	0.6712
(-0.2,-0.8,0.6)	-0.4122	-0.4599	-0.5279	-0.5279	-0.5778	-0.6478	-0.8864	-0.6943	-0.8055
(-0.2,-0.7,0.6)	-0.3465	-0.3678	-0.3982	-0.4868	-0.4629	-0.4895	-0.6206	-0.6600	-0.5279
(-0.2,-0.6,0.6)	-0.2354	-0.2375	-0.2458	-0.3512	-0.3565	-0.3882	-0.4888	-0.5014	-0.5278
(-0.2,-0.5,0.6)	-0.0954	-0.0243	-0.1003	-0.1819	-0.2361	-0.2361	-0.2185	-0.2589	-0.3981
(-0.2,-0.4,0.6)	0.0688	0.0657	0.0315	-0.0106	0.0447	0.0698	0.1132	0.0372	0.2522
(-0.2,-0.3,0.6)	0.2393	0.2367	0.2713	0.2360	0.2146	0.2854	0.3560	0.3694	0.6365
(-0.2,-0.2,0.6)	0.3222	0.2361	0.3620	0.4347	0.5347	0.4488	0.5197	0.6177	0.8079
(-0.2,-0.1,0.6)	0.8197	0.5010	0.5279	0.4155	0.4979	0.5056	0.6157	0.6101	0.8125
(-0.2,0.1,0.6)	0.5279	0.5278	0.8463	0.6872	0.7379	0.8025	0.7794	0.8911	0.7768
(-0.2,0.2,0.6)	0.7151	0.8450	0.7285	0.7746	0.8340	0.9285	0.8365	0.7773	0.8059
(-0.2,0.3,0.6)	0.6102	0.8491	0.7570	0.9320	0.9130	0.9301	0.9775	0.8481	0.9128
(-0.2,0.4,0.6)	0.7583	0.8952	0.9294	0.9791	0.9571	0.9213	0.8597	0.9775	0.8587
(-0.2,0.5,0.6)	0.6161	0.7420	0.8880	0.9413	0.9508	0.9665	0.9783	0.9066	0.7715
(-0.2,0.6,0.6)	0.8861	0.8309	0.8871	0.9161	0.9595	0.8994	0.9518	0.8796	0.7608
(-0.2,0.7,0.6)	0.8088	0.9832	0.9547	0.9894	0.9523	0.9684	0.9123	0.8328	0.7041
(-0.2,0.8,0.5)	0.7791	0.9445	0.8847	0.9829	0.8814	0.9691	0.9365	0.7645	0.6987
(-0.2,0.8,0.4)	0.8982	0.9311	0.9183	0.8885	0.9840	0.9675	0.8531	0.7687	0.9588
(-0.2,0.8,0.3)	0.9168	0.8983	0.9415	0.9147	0.9600	0.9775	0.8968	0.9428	0.7012
(-0.2,0.8,0.2)	0.8894	0.7739	0.8399	0.7999	0.9320	0.9736	0.9799	0.8971	0.9516
(-0.2,0.8,0.1)	0.7796	0.8118	0.8193	0.8886	0.9889	0.9736	0.9065	0.8372	0.7266
(-0.2,0.8,-0.1)	0.8096	0.8572	0.9665	0.9562	0.9511	0.9042	0.8411	0.8359	0.7541
(-0.2,0.8,-0.2)	0.6927	0.7790	0.9213	0.8573	0.9344	0.8939	0.9275	0.7076	0.9156
(-0.2,0.8,-0.3)	0.5434	0.6964	0.8192	0.8730	0.7690	0.9532	0.8436	0.7766	0.8820
(-0.2,0.8,-0.4)	0.6229	0.8194	0.6385	0.6651	0.8986	0.8436	0.7494	0.7310	0.5294
(-0.2,0.8,-0.5)	0.4416	0.5645	0.5293	0.8921	0.7907	0.8514	0.7495	0.6113	0.5279
(-0.2,0.8,-0.6)	0.4590	0.5160	0.5167	0.5422	0.7283	0.7005	0.6665	0.8058	0.6033
Regression Coefficients									
(1.0657+0.6875x)	0.9338	0.9855	0.9737	0.9832	0.9543	0.9645	0.8676	0.9572	0.6019

TABLE 6.3: Power for different designs in 10000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.484	0.700	0.800	0.837	0.850	0.858	0.860	0.862	0.863
Random	0.304	0.443	0.546	0.643	0.700	0.753	0.799	0.817	0.840
Missing mechanism									
(-0.2,-0.8,0.6)	0.280	0.373	0.419	0.444	0.448	0.538	0.595	0.685	0.773
(-0.2,-0.7,0.6)	0.291	0.302	0.381	0.432	0.524	0.591	0.637	0.703	0.798
(-0.2,-0.6,0.6)	0.299	0.349	0.442	0.487	0.559	0.625	0.668	0.727	0.798
(-0.2,-0.5,0.6)	0.305	0.443	0.513	0.559	0.607	0.671	0.727	0.772	0.810
(-0.2,-0.4,0.6)	0.331	0.486	0.572	0.635	0.710	0.763	0.803	0.820	0.853
(-0.2,-0.3,0.6)	0.376	0.549	0.664	0.720	0.760	0.809	0.832	0.846	0.859
(-0.2,-0.2,0.6)	0.398	0.549	0.690	0.769	0.814	0.825	0.842	0.854	0.859
(-0.2,-0.1,0.6)	0.450	0.623	0.728	0.766	0.813	0.830	0.848	0.854	0.859
(-0.2,0.1,0.6)	0.438	0.631	0.775	0.801	0.829	0.845	0.852	0.858	0.859
(-0.2,0.2,0.6)	0.468	0.688	0.762	0.808	0.831	0.850	0.853	0.858	0.859
(-0.2,0.3,0.6)	0.451	0.688	0.767	0.818	0.836	0.850	0.854	0.858	0.859
(-0.2,0.4,0.6)	0.474	0.688	0.767	0.818	0.836	0.850	0.854	0.858	0.860
(-0.2,0.5,0.6)	0.474	0.694	0.782	0.820	0.838	0.850	0.853	0.858	0.860
(-0.2,0.6,0.6)	0.485	0.684	0.778	0.818	0.838	0.849	0.854	0.858	0.859
(-0.2,0.7,0.6)	0.479	0.701	0.784	0.821	0.838	0.851	0.854	0.858	0.859
(-0.2,0.8,0.5)	0.474	0.693	0.790	0.834	0.848	0.853	0.857	0.858	0.863
(-0.2,0.8,0.4)	0.480	0.691	0.794	0.823	0.845	0.857	0.858	0.859	0.862
(-0.2,0.8,0.3)	0.479	0.687	0.800	0.825	0.845	0.857	0.860	0.862	0.863
(-0.2,0.8,0.2)	0.478	0.664	0.786	0.811	0.851	0.857	0.860	0.862	0.862
(-0.2,0.8,0.1)	0.476	0.680	0.772	0.816	0.840	0.851	0.854	0.858	0.862
(-0.2,0.8,-0.1)	0.479	0.688	0.785	0.819	0.838	0.849	0.853	0.858	0.859
(-0.2,0.8,-0.2)	0.463	0.674	0.780	0.815	0.837	0.848	0.854	0.857	0.860
(-0.2,0.8,-0.3)	0.440	0.663	0.772	0.816	0.832	0.851	0.853	0.858	0.859
(-0.2,0.8,-0.4)	0.452	0.682	0.749	0.799	0.836	0.846	0.852	0.857	0.858
(-0.2,0.8,-0.5)	0.420	0.636	0.728	0.816	0.832	0.847	0.852	0.853	0.858
(-0.2,0.8,-0.6)	0.426	0.628	0.727	0.783	0.828	0.842	0.851	0.857	0.858
Regression Coefficients									
(1.0657+0.6875x)	0.490	0.702	0.785	0.820	0.838	0.850	0.854	0.857	0.858

In example (b) below, compared to example (a), the sample size was increased from 400 to 1000. The value of β_0 increased from 1 to 2, while β_1 decreased from 0.7 to -2. The mean of X, μ_x, was reduced to 0, and the variance, σ_x^2, remained constant. Additionally, α_0 increased from -0.2 to 2, α_1 decreased from 0.8 to -0.4, and α_2 decreased from 0.6 to -0.15. Despite these changes in parameters, the designs performed similarly to those in example (a). These parameters are chosen such that there are approximately 30% missingness in Y.

(b) $n = 1000$, $(\beta_0, \beta_1) = (2, -2)$, $\sigma_y^2 = 4$, $(\mu_x, \sigma_x^2) = (0, 16)$, $(\alpha_0, \alpha_1, \alpha_2) = (2, 0.4, -0.15)$.

The results for this example are shown in the appendix. In Table A.8, the power of algorithm 1 for test T_1, random design and some misspecified designs are shown. At all values of recovery proportion c, the test T_1 performs better than the other designs. All other misspecified designs outperformed the random design at all values of c except at $(2, -0.4, -0.15)$ where there is a change in sign for the coefficient of x in the missing mechanism model. Tables A.9 and A.10 show the γ_1 values and the resulting power for the different designs. All designs performed better than the random design except the design where the sign of α_1 changes. As in example (a) above, this shows that a change in the sign of α_1 affects the test.

In example (c) below, compared to example (b), the sample size remains the same at 1000. The value of β_0 is unchanged at 2, but β_1 increases from -2 to 1 in example (c). The mean of X, μ_x, increases from 0 to 2, while the variance of X, σ_x^2, decreases from 16 to 1. Additionally, α_0 decreases from 2 in example (b) to -2 in example (c), α_1 increases from 0.4 to 1.3, and α_2 increases from -0.15 to 0.15. The choice of parameters ensures there are approximately 30% missing values in Y. Despite these changes in parameters, the designs performed similarly in both examples.

(c) $n = 1000$, $(\beta_0, \beta_1) = (2, 1)$, $\sigma_y^2 = 4$, $(\mu_x, \sigma_x^2) = (2, 1)$, $(\alpha_0, \alpha_1, \alpha_2) = (-2, 1.3, 0.15)$.

Tables A.11 and A.12 in the appendix show the γ_1 values and power of algorithm 1 for test T_1, random design and some misspecified designs respectively. The test T_1 has the best power at all values of c, as the coefficient of x changes in the missing mechanism, the misspecified designs also outperformed the random design at all values of c except when the coefficient of x falls between -0.15 and -1.3. As explained in example (a) above, when the regression relation for y is substituted in the missing mechanism, it would lead to a change in sign in the γ_1 which in turns perform worse than the random design. It could be seen on table A.11 that the sign of γ_1 changes from $+$ to $-$ when there is a change in sign on γ_1 and changes from $-$ to $+$ as soon as the γ_1 changes to positive in the missing mechanism. This further shows that a change in sign on the coefficient on x affects the performance of the test and it is therefore important to have the correct sign.

For $p = 2$, generate n points from $Y|(X = x) \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma_y^2)$ with $X_1 \sim N(\mu_{1,x}, \sigma_{1,x}^2)$ and $X_2 \sim N(\mu_{2,x}, \sigma_{2,x}^2)$.

Introduce MNAR missingness into y values using $\Pr(M = 1|Y = y, X = x) = \text{expit}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_2 y)$. There are approximately 30% of points missing their y value using the parameter values in the example below. We will repeat this process $10,000$ times and in each replication and test the hypothesis $H_0 : \psi = 0$, to the generated sample.

(d) $n = 400$, $(\beta_0, \beta_1, \beta_2) = (0, 2, -2)$, $\sigma_y^2 = 1$, $(\mu_{1,x}, \mu_{2,x}, \sigma_{1,x}^2, \sigma_{2,x}^2) = (-2, 0, 4, 16)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (2.9, -0.4, 0.4, 0.5)$.

Tables 6.5 and 6.6 show the γ_1 and γ_2 values respectively. The power for the random design, optimal design and different misspecifications in the model are shown in Table 6.4. Irrespective of the misspecification, all designs outperform the random design. The random design has the least power among all other designs. The misspecified design $(2.9, 0.4, 0.4, 0.5)$ performs slightly lower than the optimal design but better than the random design. The $(2.9, -0.4, -0.4, 0.5)$ and $(2.9, 0.4, -0.4, 0.5)$ designs have power values above the optimal design at $c < 0.5$. Showing that a change in the sign of the coefficients of the covariates does not affect the power of the test.

TABLE 6.4: Power for different designs in 10000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.388	0.641	0.786	0.827	0.877	0.892	0.894	0.905	0.912
Random	0.304	0.473	0.565	0.664	0.744	0.795	0.855	0.883	0.911
Missing mechanism									
(2.9,-0.3,0.4,0.5)	0.386	0.627	0.782	0.831	0.871	0.878	0.898	0.901	0.912
(2.9,-0.2,0.4,0.5)	0.396	0.640	0.760	0.831	0.868	0.885	0.897	0.908	0.913
(2.9,-0.1,0.4,0.5)	0.370	0.639	0.790	0.831	0.881	0.882	0.899	0.907	0.913
(2.9,0.1,0.4,0.5)	0.379	0.635	0.750	0.820	0.869	0.884	0.896	0.908	0.912
(2.9,0.2,0.4,0.5)	0.398	0.634	0.725	0.813	0.875	0.888	0.897	0.906	0.912
(2.9,0.3,0.4,0.5)	0.394	0.616	0.771	0.840	0.873	0.888	0.895	0.907	0.912
(2.9,0.4,0.4,0.5)	0.384	0.612	0.785	0.822	0.883	0.884	0.891	0.904	0.911
(2.9,-0.4,0.3,0.5)	0.404	0.650	0.776	0.842	0.870	0.885	0.898	0.907	0.911
(2.9,-0.4,0.2,0.5)	0.388	0.639	0.763	0.849	0.871	0.887	0.897	0.901	0.912
(2.9,-0.4,0.1,0.5)	0.393	0.681	0.773	0.837	0.874	0.893	0.902	0.907	0.914
(2.9,-0.4,-0.1,0.5)	0.422	0.658	0.777	0.821	0.867	0.884	0.900	0.906	0.912
(2.9,-0.4,-0.2,0.5)	0.411	0.686	0.785	0.858	0.875	0.885	0.893	0.908	0.912
(2.9,-0.4,-0.3,0.5)	0.427	0.652	0.790	0.832	0.873	0.891	0.899	0.906	0.911
(2.9,-0.4,-0.4,0.5)	0.421	0.683	0.789	0.840	0.871	0.885	0.895	0.907	0.911
(2.9,0.4,-0.4,0.5)	0.430	0.678	0.793	0.858	0.877	0.885	0.894	0.903	0.911
Regression Coefficients									
$(0.07+1.96x_1 - 1.98x_2)$	0.383	0.649	0.756	0.828	0.870	0.889	0.899	0.908	0.911

TABLE 6.5: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.3883	0.7683	1.2086	0.9198	1.2699	1.0780	0.5982	0.6391	2.1061
(2.9,-0.3,0.4,0.5)	0.3500	0.4309	1.2217	0.7250	0.8125	0.8112	0.9772	0.5906	1.4896
(2.9,-0.2,0.4,0.5)	0.6500	0.9118	0.5428	0.7512	0.1997	0.9559	1.0476	1.5378	2.1250
(2.9,-0.1,0.4,0.5)	0.2617	0.4583	1.5745	1.1745	1.4444	1.1061	1.1891	0.9106	1.2619
(2.9,0.1,0.4,0.5)	0.1330	0.8532	1.6948	2.9932	0.9903	0.9297	0.7618	1.1438	0.7628
(2.9,0.2,0.4,0.5)	0.8541	1.2904	2.1188	2.5703	2.8186	1.7148	0.9930	0.9051	0.7430
(2.9,0.3,0.4,0.5)	0.4583	1.3278	0.8004	1.1060	1.7387	1.1036	2.0707	1.2636	0.3439
(2.9,0.4,0.4,0.5)	0.4277	2.0845	2.7870	1.1419	1.3015	1.1568	1.0850	1.4200	0.4813
(2.9,-0.4,0.3,0.5)	0.3428	0.3720	0.7825	0.9972	0.8372	0.8839	1.1198	1.7977	1.5968
(2.9,-0.4,0.2,0.5)	0.3629	0.5750	0.4441	1.2838	0.9261	1.3422	1.5105	0.5219	2.1125
(2.9,-0.4,0.1,0.5)	0.3896	1.1634	0.7031	1.1334	1.1164	1.3727	2.1243	1.4478	0.6000
(2.9,-0.4,-0.1,0.5)	0.6822	0.7511	0.7369	0.7090	0.7495	1.2539	1.3563	0.7313	0.5219
(2.9,-0.4,-0.2,0.5)	0.9144	1.8711	1.2206	1.9333	1.3964	1.3245	0.5750	1.0865	0.8156
(2.9,-0.4,-0.3,0.5)	1.0836	1.0250	1.7325	1.045	1.5826	1.2401	0.8798	0.9436	0.4531
(2.9,-0.4,-0.4,0.5)	1.4474	1.1750	1.3378	1.2881	1.3441	2.0750	2.8750	0.8403	1.3500
(2.9,0.4,-0.4,0.5)	2.1231	2.5130	2.1520	2.5750	2.3159	0.7987	1.1641	0.4282	0.6469
Regression Coefficients									
$(0.07+1.96x_1 - 1.98x_2)$	0.4500	0.7173	0.5281	0.6935	0.8255	1.1688	1.6344	0.9598	1.3975

TABLE 6.6: γ_2 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	-0.6016	-0.8612	-1.2371	-1.0366	-1.3711	-1.1516	-1.3023	-1.3813	-1.9468
(2.9,-0.3,0.4,0.5)	-0.8000	-0.6900	-1.1742	-1.2001	-0.8500	-0.8452	-1.5016	-1.1250	-1.5330
(2.9,-0.2,0.4,0.5)	-0.6000	-0.8485	-0.9852	-0.9956	-0.9020	-0.8359	-1.2184	-1.6448	-2.5000
(2.9,-0.1,0.4,0.5)	-0.5438	-0.7995	-1.5419	-1.1400	-1.3162	-1.5498	-1.2172	-1.2089	-1.3664
(2.9,0.1,0.4,0.5)	-0.8133	-0.6745	-0.9713	-1.4914	-1.3578	-1.3125	-1.2946	-1.1750	-0.9654
(2.9,0.2,0.4,0.5)	-0.8745	-0.8619	-0.8750	-1.1630	-1.7351	-1.0323	-0.9733	-1.2578	-0.8281
(2.9,0.3,0.4,0.5)	-0.6193	-0.7597	-0.7037	-0.9080	-0.9999	-1.1276	-1.1733	-1.2960	-0.9148
(2.9,0.4,0.4,0.5)	-0.5837	-0.9375	-2.0068	-0.6989	-1.2956	-1.0821	-1.0067	-1.4326	-0.7502
(2.9,-0.4,0.3,0.5)	-1.3181	-1.2973	-1.0026	-1.1661	-0.9547	-1.3278	-1.3448	-2.4844	-1.9325
(2.9,-0.4,0.2,0.5)	-0.9297	-0.8000	-1.1899	-1.3621	-1.3672	-1.6188	-2.0059	-1.0375	-2.8500
(2.9,-0.4,0.1,0.5)	-0.7742	-2.1236	-1.5250	-1.9373	-1.7938	-2.1444	-2.4993	-1.8379	-1.2000
(2.9,-0.4,-0.1,0.5)	-1.6981	-1.4039	-1.4989	-2.0557	-1.3991	-2.3125	-1.6250	-1.1500	-0.8063
(2.9,-0.4,-0.2,0.5)	-1.5932	-2.6094	-2.2251	-2.2993	-2.7080	-2.3991	-1.2000	-1.2418	1.1375
(2.9,-0.4,-0.3,0.5)	-2.9798	-1.3000	-2.6983	-2.3115	-1.9101	-2.0654	-1.4314	-2.3497	-0.9375
(2.9,-0.4,-0.4,0.5)	-2.1181	-1.9001	-2.6006	-3.7214	-2.8103	-4.2000	-6.8000	-1.1749	-1.8000
(2.9,0.4,-0.4,0.5)	-2.1286	-2.5702	-2.5435	-2.7000	-2.7171	-1.8544	-1.1625	-0.6875	
Regression Coefficients									
$(0.07+1.96x_1 - 1.98x_2)$	-0.7125	-1.0127	-0.7625	-1.0754	-1.0508	-1.1568	-1.8375	-1.0403	-2.7070

In section 7.2, we introduce a heuristic method to find robust designs.

6.4 Non-parametric alternatives

In the arguments of this section, we adopt the findings of Kim and Yu (2011) in order to increase the robustness of our MNAR test to model misspecifications. As shown in Lemma 6.3, a key component of computing expectations present in the T-optimality criteria developed in this work is the knowledge of $\Pr(Y \in dy \,|\, X = x, M = 0)$ and $\Pr(Y \in dy \,|\, X = x, M = 1)$. It seems reasonable to suggest the probability $\Pr(Y \in dy \,|\, X = x, M = 1)$ can be estimated using the observed data, and it indeed can. However, estimating $\Pr(Y \in dy \,|\, X = x, M = 0)$ is not trivial due to the unobserved responses. A key result that connects the properties of the observed and unobserved responses is provided in the following lemma.

Lemma 6.4.

$$\Pr(Y \in dy \,|\, X = x, M = 0) = \Pr(Y \in dy \,|\, X = x, M = 1) \times \frac{O(x, y)}{\mathbb{E}(O(x, Y)|X = x, M = 1)},$$

where

$$O(x, y) = \frac{\Pr(M = 0 \,|\, X = x, Y = y)}{\Pr(M = 1 \,|\, X = x, Y = y)}.$$

The quantity $O(x, y)$ is known as the conditional odds of non-response. By specialising Lemma 6.4 for the logit model given in (6.3), we obtain

$$\Pr(Y \in dy \,|\, X = x, M = 0) = \Pr(Y \in dy \,|\, X = x, M = 1) \times \frac{\exp(-\psi y)}{\mathbb{E}(\exp(-\psi Y)|X = x, M = 1)}.$$

This result states the density of the non-responding responses is an exponential tilting of the density for the responding responses (Kim and Yu, 2011). A consistent estimate of $\Pr(Y \in dy \,|\, X = x, M = 1)$ can be nonparametrically obtained using kernel density estimation. If $\Pr(M = 1 \,|\, X = x, Y = y)$ is known, then ψ does not require estimation. Otherwise, $-\psi$ will have to be estimated using a follow-up of some non-respondents.

Lemma 6.5. *From Kim and Yu (2011), a non-parametric consistent estimator of* $\Pr(Y \in dy \,|\, X = x, M = 0)$ *is given by* $m_0(x, y; \psi) \cdot dy$ *with*

$$m_0(x, y; \psi) \;=\; \sum_{i=1}^{n} \frac{M_i K_h(x, x_i) K_h(y, y_i) \exp(-\psi y_i)}{\sum_{j=1}^{n} M_j K_h(x, x_j) \exp(-\psi y_j)},$$

where $K_h(u, x) = h^{-1} K\{(u - x)/h\}$, $K(\cdot)$ *is a symmetric density function (kernel) on* $\mathbb{R}$ *and* $h = h_n$ *is the bandwidth such that* $h_n \to 0$ *and* $nh_n \to \infty$ *as* $n \to \infty$.

Using Lemma 6.5 and Lemma 6.3, we can estimate the expected value of a function $g(X_A, Y_A)$ as follows:

$$\begin{aligned}\mathbb{E}g(X_A, Y_A) &= \mathbb{E}g(X_O, Y_O)\Pr(M_A = 1) + \mathbb{E}g(X_R, Y_R)\Pr(M_A = 0) \\ &\simeq \frac{1}{n_1 + c \cdot (n - n_1)} \sum_{i \in I_1} g(x_i, y_i) \\ &+ \frac{1}{n_1 + c \cdot (n - n_1)} \sum_{i \in S} \int g(x_i, y) m_0(x_i, y; \psi) dy\,, \end{aligned} \tag{6.20}$$

where S is a set obtained from sampling $c \cdot (n - n_1)$ elements without replacement from $I \backslash I_1$ with each item having probability $p_2(x_i)$ of being selected. A particularly important form of g is

$$\begin{aligned} g(X_A, Y_A) &= \Bigg[v_1(\lambda, \psi; W_A, Z_A, \gamma) \log\left(\frac{v_1(\lambda, \psi; W_A, Z_A, \gamma)}{v_1(\lambda_0, 0; W_A, Z_A, \gamma)}\right) \\ &+ (1 - v_1(\lambda, \psi; W_A, Z_A, \gamma)) \log\left(\frac{1 - v_1(\lambda, \psi; W_A, Z_A, \gamma)}{1 - v_1(\lambda_0, 0; W_A, Z_A, \gamma)}\right) \Bigg] \end{aligned} \tag{6.21}$$

since we require estimating the mean of g for the optimal design criteria $T(\gamma^*)$.

We will now formulate the second algorithm for test T_1 which can be seen as a relaxed form of Algorithm 1 for T_1.

Algorithm 7 Algorithm 2 for T_1

1: **Input:** Value $0 < c \leq 1$, approximation formula (6.20), and function g given in (6.21).
2: **Output:** Value of γ.
3: **Steps:**
4: Approximate $T(\gamma^*)$ using the approximation formula (6.20) with function g from (6.21).
5: Choose γ based on the approximation of $T(\gamma^*)$.
6: **Return:** γ.

To use Lemma 6.5, we need to specify a kernel and a bandwidth. Whilst this choice is fairly arbitrary, we will align with Kim and Yu (2011) and choose the Gaussian kernel and set $h = \hat{\sigma}_x n^{-0.2}$, where $\hat{\sigma}_x$ is an estimator for the standard deviation of X.

In Figures 6.9 and 6.10, as a function of the recovery proportion c, using a dashed blue line we plot the power of Algorithm 2 for T_1 and compare that to Algorithm 1 for T_1 using the same dashed red line as previous figures. For scenario (a) where $n = 400$ it is easier to distinguish between the two algorithms as is shown in Figure 6.9. If we reduced σ_x from 4 to 2, Algorithm 2 produces designs with similar power to the ideal (but more restrictive) Algorithm 1. This is shown in Figure 6.10.

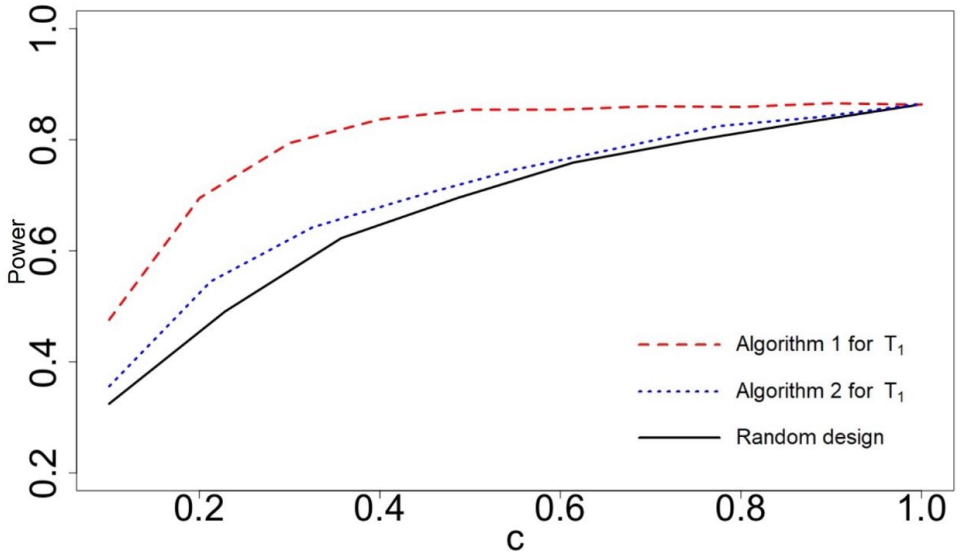

FIGURE 6.9: Power for different recovery proportions comparing Algorithm 1 and Algorithm 2 for T_1 versus a random recovery.

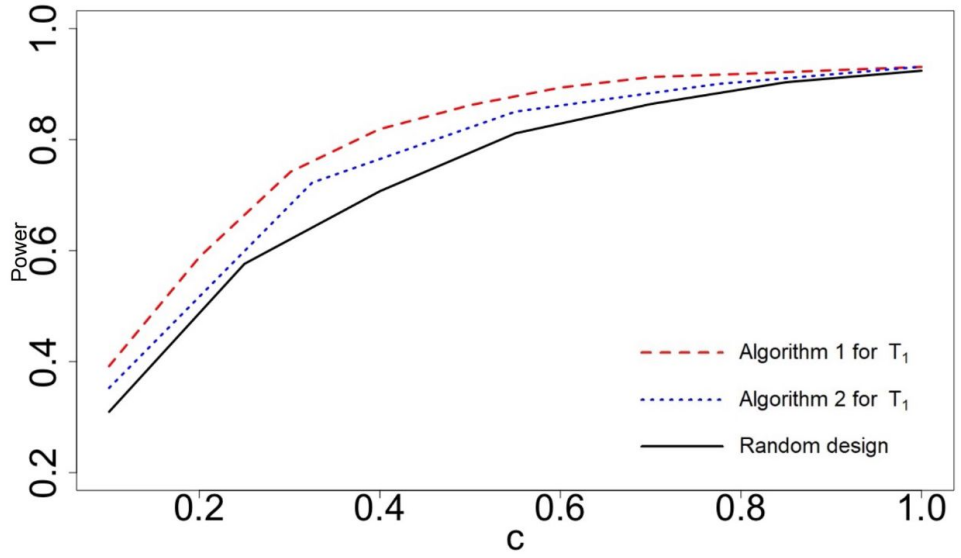

FIGURE 6.10: Power for different recovery proportions comparing Algorithm 1 and Algorithm 2 for T_1 versus a random recovery.

Chapter 7

Multivariate Case

This chapter focuses on how to approach the design problem in a multivariate scenario. It is computationally expensive to find optimal designs based on regions or probabilities when there are three or more covariates using Algorithm 1 for T_1 or Algorithm 1 for T_2. Therefore, in this chapter we explore two different approaches for addressing these difficulties. Section 7.1 uses the correlation coefficient between covariates and response variable in designing the optimal region based on the concept of an optimal design region developed in Chapter 5. Section 7.2 focuses on a conjecture that uses the empirical densities of the covariates to produce a robust and efficient design which is based on the framework developed in Chapter 6. The framework in Section 7.2 will also allow us to make our MNAR test more robust to model misspecifications.

7.1 Using the Correlation Coefficient

This section focuses on how to approach the design region in a multivariate scenario using the correlation coefficient between covariates and response variables in designing the optimal region for a multivariate scenario. For the multivariate case, if optimisation with a large number of covariates becomes too complicated, which combination of covariates should be included in computations to obtain the best design region? Can the covariate with the highest correlation value with the response variable be used for the design region? What happens when the covariates have almost the same correlation coefficients? Results in this chapter show that the higher the correlation with the response variable, the higher the power of the optimal design when the covariate is used in constructing the design region.

7.1.1 Simulation studies

Firstly, we present an example with 2 covariates where the missing data mechanism is MAR. This example compares the MLE for α_{p+1} and the Type I error of the test using random and optimal design with all covariates. All parameters used in this section ensures there are approximately 30% missing cases in the response variable Y.

MAR: Generate 1000 points following a multiple linear regression model in 10000 replicates:

$$Y|(X_1 = x_1, X_2 = x_2) \sim N(1 - x_1 + 2x_2, 25)\,,$$

with $X_1 \sim N(3, 36)$ and $X_2 \sim N(1, 16)$. Introduce MAR missingness into the model using:

$$P(M = 1|X = x) = \frac{\exp(-3 + 0.42x_1 + 0.25x_2)}{1 + \exp(-3 + 0.42x_1 + 0.25x_2)}.$$

Figures 7.1 and 7.2 below show the Type I error values and MSE for both designs respectively. Figure 7.1 shows that both designs have approximately 0.05 Type I error values. Figure 7.2 shows that the optimal design has a smaller MSE compared to the random design.

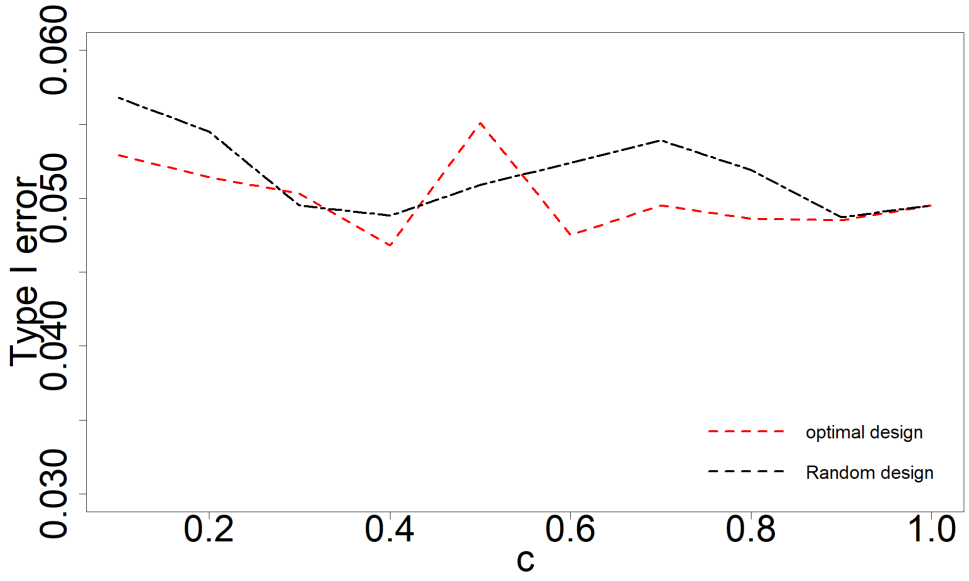

FIGURE 7.1: Type I error comparison between random design and optimal design.

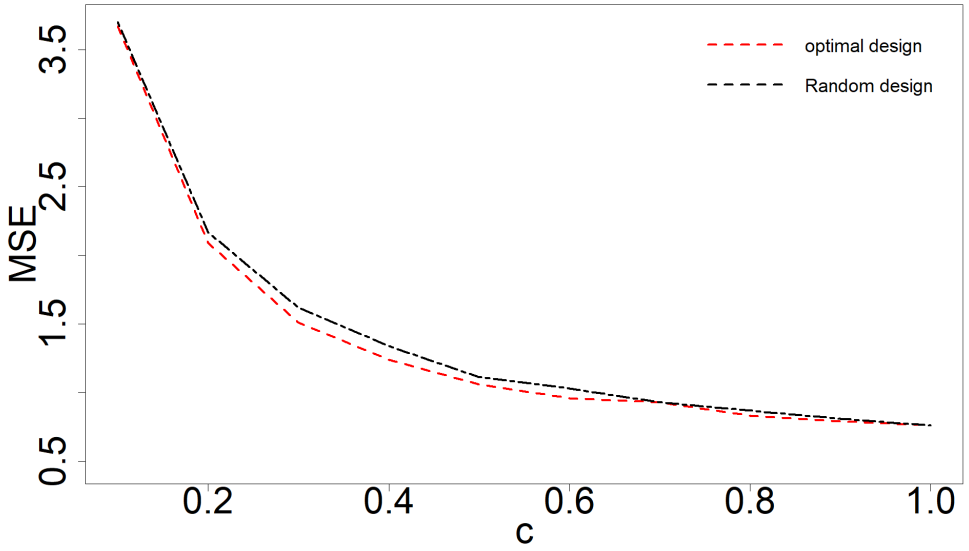

FIGURE 7.2: MSE comparison between random design and optimal design.

In the example below, we find out if a design based on only one variable can outperform the random design. We considered the random design, optimal design region formed using each covariate, both covariates and a linear combination (z) of both covariates.

MNAR 1: Generate 1000 points following a multiple linear regression model in 10000 replicates:

$$Y|(X_1 = x_1, X_2 = x_2) \sim N(2 - 2x_1 + 2x_2, 4)\,,$$

with $X_1 \sim N(0, 16)$ and $X_2 \sim N(2, 4)$. Introduce MNAR missingness into the model using:

$$P(M = 1|Y = y, X = x) = \frac{\exp(-2 + 0.4x_1 + 0.2x_2 - 0.15y)}{1 + \exp(-2 + 0.4x_1 + 0.2x_2 - 0.15y)}.$$

In the model above, let z be a linear combination of the covariates such that: $z = 0.4x_1 + 0.2x_2$. The variable z is used to form the design region for the recovery. z is taken from the missing mechanism model. Figure 7.3 shows the Power comparison for different designs considered. The optimal design using x_1 gives the highest power at all values of recovery proportion followed by the optimal design when both covariates are used in constructing the design region. The optimal design using the linear combination z and the optimal design using x_2 performs similarly to the random design. The power increases as the recovery proportion increases for all the designs. Using X_1 gives higher power than X_2 because the correlation between X_1 and Y is greater than that of X_2 and Y. $Cor(X_1, Y) = -0.8727201$ and $Cor(X_2, Y) = 0.4362057$.

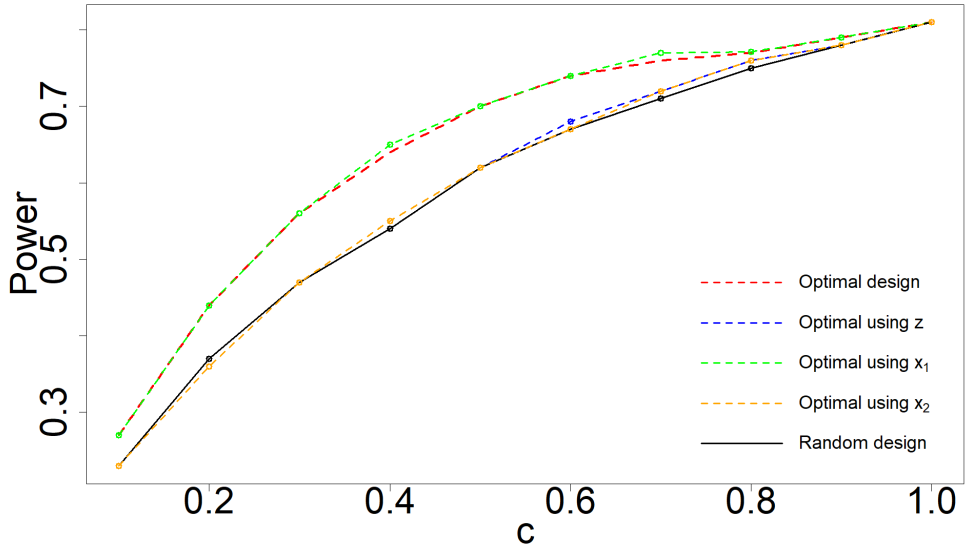

FIGURE 7.3: Power comparison for random design, optimal design using all covariates, optimal design using z and optimal design using each covariate for different recovery proportions.

From the above example, the optimal design with one interval (z) did not perform well but similarly as the optimal design with x_2 (the least correlated variable with y). In the following examples, we focus on designing the region using all the covariates, one covariate at a time and compare these optimal designs with the random design.

MNAR 2: Generate 1000 points following a multiple linear regression model in 10000 replicates:

$$Y|(X_1 = x_1, X_2 = x_2) \sim N(2 - 2x_1 + 2x_2, 4)\,,$$

with $X_1 \sim N(0,4)$ and $X_2 \sim N(2,4)$. Introduce MNAR missingness into the model using:

$$P(M = 1|Y = y, X = x) = \frac{\exp(-2.5 + 0.4x_1 + 0.4x_2 - 0.11y)}{1 + \exp(-2.5 + 0.4x_1 + 0.4x_2 - 0.11y)}.$$

In the above model, X_1 and X_2 have the same variance, the correlation coefficient for X_1 and Y is -0.6653363 and that of X_2 and Y is 0.6661896. The results for this simulation using the random design, optimal design for both covariates, optimal design using X_1 and optimal design using X_2 are shown in Figure 7.4 below. With the absolute correlation values being approximately 0.67, the optimal design using X_1 and X_2 performs similarly to the random design. At all values of recovery proportion, the optimal design using both covariates outperforms the other designs.

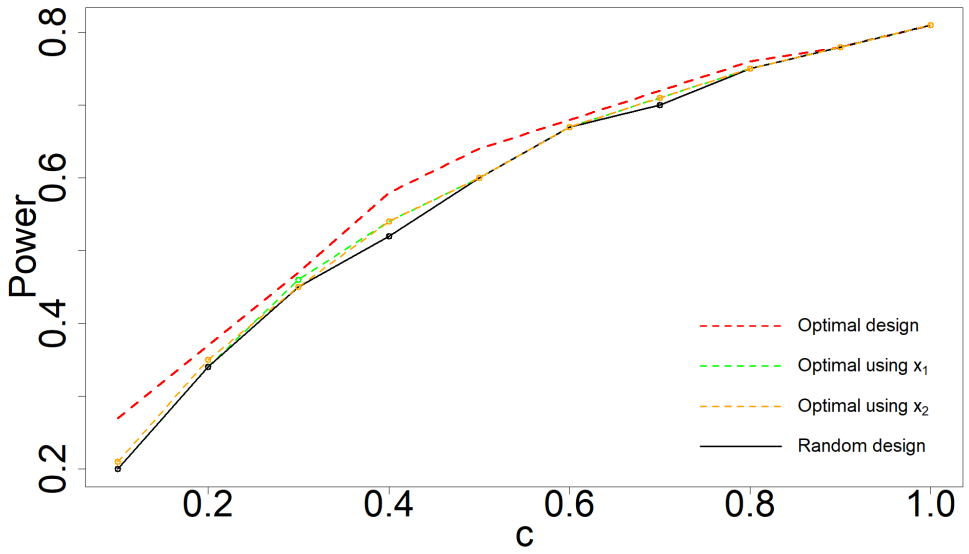

FIGURE 7.4: MSE comparison between random design and optimal design using covariates: Red: random, Black: optimal using both covariates, Blue: optimal using X_1, Green: optimal using X_2.

There are three covariates in the example below. Here, we focus on designing the regions using each covariate and compare the result of each optimal design with the random design.

MNAR 3: Generate 1000 points following a multivariate linear regression model in 10000 replicates:

$$Y|(X_i = x_i) \sim N(2 - 2x_1 + 2x_2 - 0.5x_3, 4),$$

such that $i = 1, \ldots, 3$, with $X_1 \sim N(2,4)$, $X_2 \sim N(1,9)$, X_3 and $\sim N(4,1)$. The model below is used to introduce MNAR missingness into y:

$$P(M = 1|Y = y, X = x) = \frac{\exp(2 + 0.16x_1 + 0.6x_2 - 1.8x_3 - 0.15y)}{1 + \exp(2 + 0.16x_1 + 0.6x_2 - 1.8x_3 - 0.15y)}.$$

Figure 7.5 shows the results for different designs. Using X_2 gives the highest power than other covariates because it has the highest correlation value with Y. $Cor(X_1, Y) = -0.5456$, $Cor(X_2, Y) = 0.8026$ and $Cor(X_3, Y) = -0.0653$. Using a covariate with a higher correlation coefficient indicates greater power. Using X_3 results in a smaller power as that of random design because it has the least correlation coefficient among all the covariates. Building the design around X_1 produces designs that are slightly better than the random design.

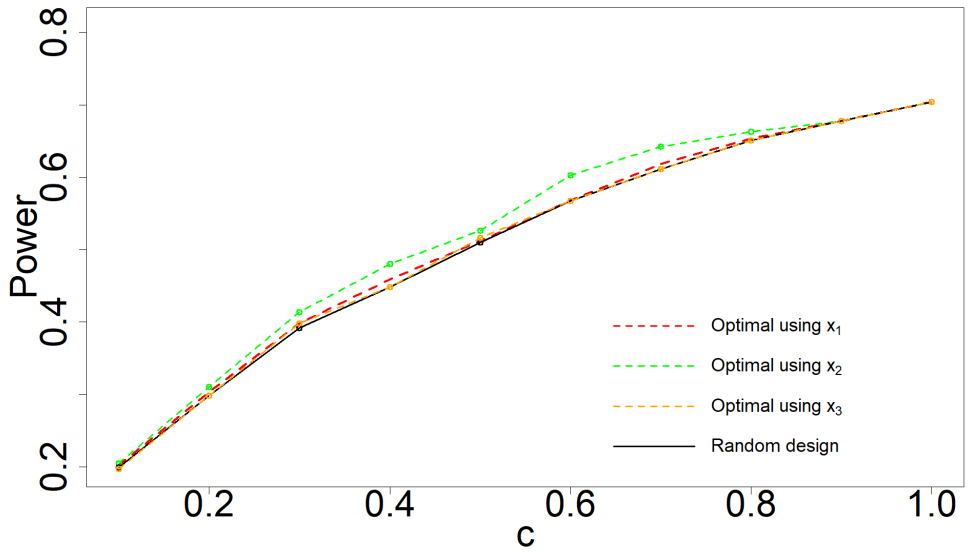

FIGURE 7.5: Power comparison for random design and optimal designs each covariate for different recovery proportions.

Conclusively, due to the complexities that can arise in designing the optimal region for a multivariate case, using the covariate with the highest correlation with the response variable to design the optimal region would result in better power than the random design. The covariate with the least correlation with the response variable would perform similarly to the random design.

7.2 Conjecture

In this section, we consider designs that are based on probabilities as in Chapter 6. We propose a simple and robust method for finding efficient designs, including for the multivariate case. We will demonstrate how the empirical densities of the covariate(s) can be used in obtaining an efficient design.

7.2.1 A simple and robust method to find efficient designs

In a single covariate or multivariate problem, instead of computationally obtaining the values of γ_1 for a single covariate problem and γ_1, ..., γ_p in a multivariate scenario (where p is the number of covariates), it is possible to obtain the empirical densities of the covariates with missing response and observed response. For different examples below, we depict three empirical densities of covariate values: covariate values for the optimal design computed from Algorithm 1 for T_1 (green), the covariate values where the response is missing (red) and the covariate values where the response is observed

(black). Using the parameters in examples (a) and (d) in Section 6.3.1, the density plots for the covariate in example (a) are shown in Figures 7.6 using missing mechanism $(-0.2, 0.8, 0.6)$ and 7.7 using missing mechanism $(-0.2, 0.8, -0.6)$. Figures 7.8 and 7.9 show the density plots for the covariates (x_1 and x_2) respectively with missing mechanism $(2.9, -0.4, 0.4, 0.5)$.

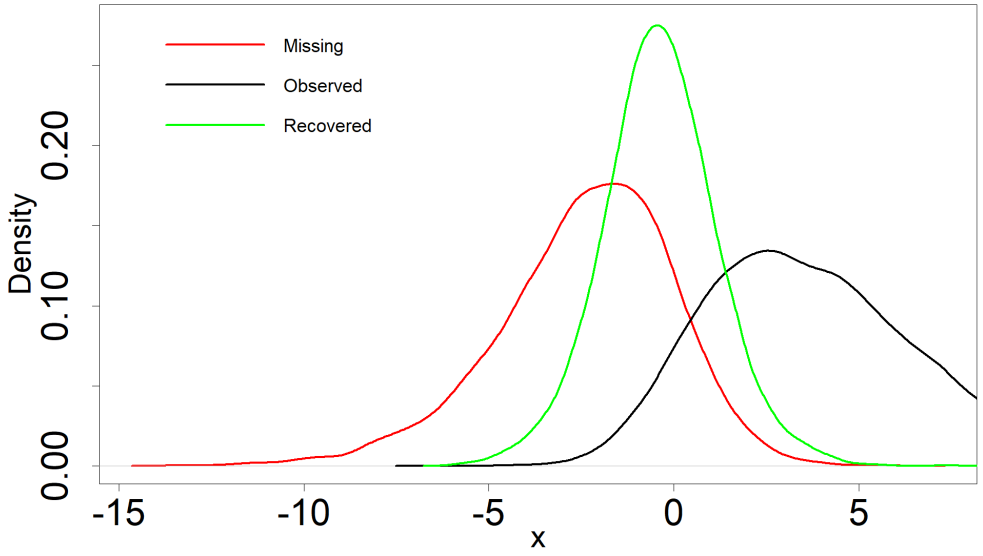

FIGURE 7.6: Covariate density plot for $(-0.2, 0.8, 0.6)$.

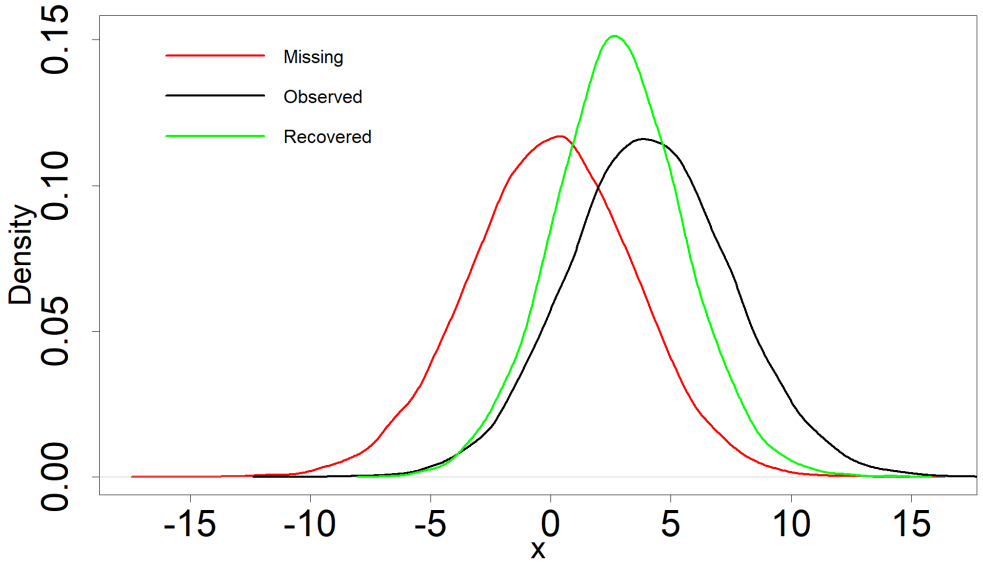

FIGURE 7.7: Covariate density plot for $(-0.2, 0.8, -0.6)$.

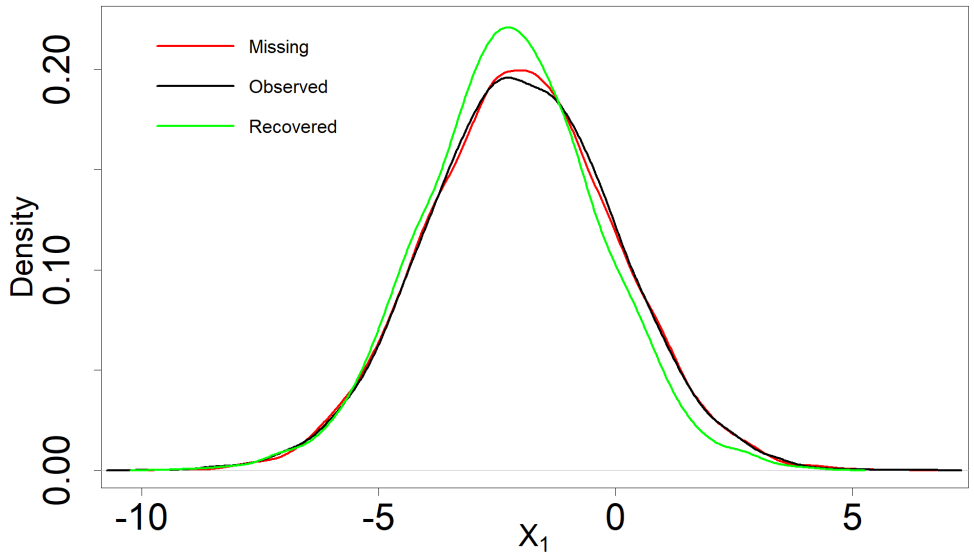

FIGURE 7.8: Covariate density plot for $(2.9, -0.4, 0.4, 0.5)x_1$.

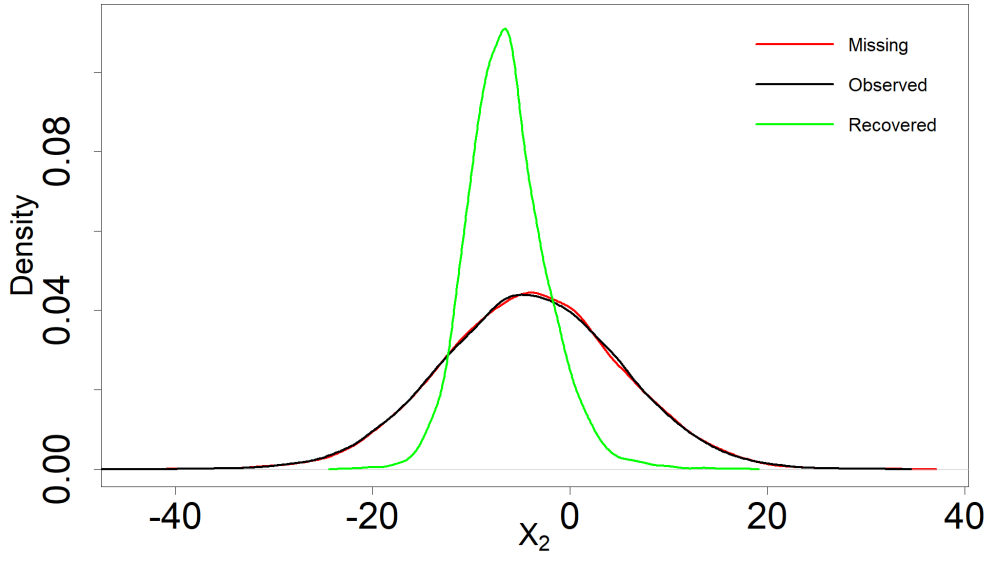

FIGURE 7.9: Covariate density plot for $(2.9, -0.4, 0.4, 0.5)x_2$.

It could be seen from the figures above that the peak of the optimal recovery (green) occurs roughly where the red and the black densities intersect. This suggests that recovering the responses whose covariate values fall in the intersection of covariate values with missing and observed responses would result in a very close optimal design. This leads to the formulation of the conjecture below:

Conjecture 1. Recovery designs whose empirical densities peak at the intersection of the densities formed by the covariate values with missing and observed response, respectively, will be close to optimal.

Conjecture 1 provides a simple method to find efficient designs. The following is the algorithm:

Algorithm 8 Designing Efficient Recovery

1: **Input:** Covariate values with missing and observed responses, given value of c.
2: **Output:** Optimal γ values.
3: **Steps:**
4: Plot the densities formed by the covariate values with missing response.
5: Plot the densities formed by the covariate values with observed response.
6: Through trial and error, determine a value of γ that places the density of the recovery design at the intersection of the missing response and observed response densities for the given value of c.
7: **Return:** γ values.

This is a very easy method, as it is easy to implement in practice, without performing a complicated optimisation procedure. Also, it is not necessary to specify any values for the model parameters making the design highly robust to parameter misspecifications. In the examples below, we find the power of the optimal design (red dashed line), conjectured design (blue dashed line) and random design (black solid line) for MAR and MNAR missing mechanisms. β_0, β_1 and γ_i used in this section are chosen to introduce circa 30% missing data into the response variable.

Table 7.1 shows the γ_1 values for one covariate case used in obtaining the power values for the optimal design and conjectured design. The power of the optimal, random and conjectured designs are shown in Figure 7.10, the conjectured design is very close to the optimal design while the random design performed worse at all values of c.

(a) $n = 400$, $(\beta_0, \beta_1) = (1, 0.7)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (2, 16)$, $(\alpha_0, \alpha_1, \alpha_2) = (-0.2, 0.8, 0.6)$.

TABLE 7.1: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(-0.2,0.8,0.6)	0.698	0.864	0.892	0.893	0.952	0.876	0.798	0.829	0.839
Conjecture	1.20	1.40	1.56	1.61	1.73	1.77	1.81	1.84	1.89

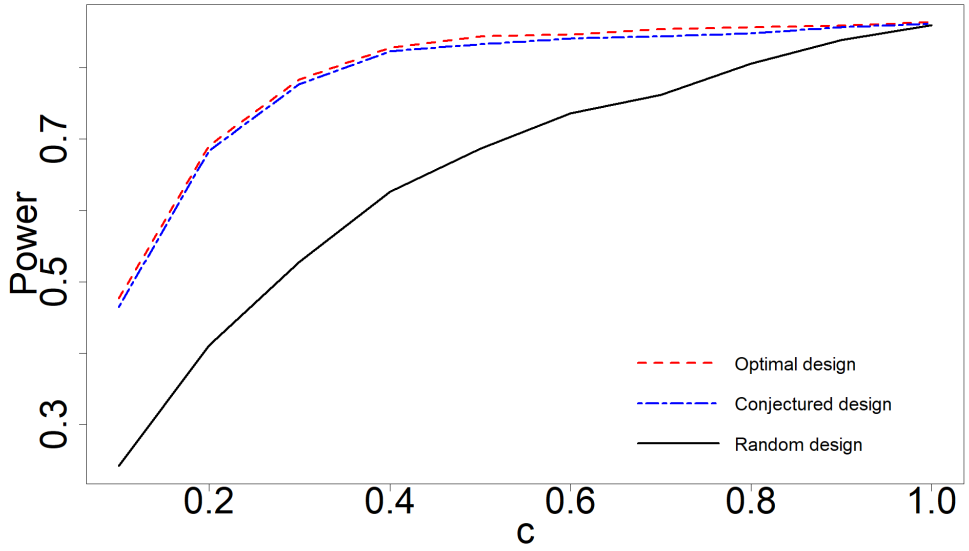

FIGURE 7.10: Power plot using different designs.

The two covariates case is shown in example (b) below. Tables 7.2 and 7.3 show the γ_1 and γ_2 values respectively. Figure 7.11 shows the power plot for the different designs, the optimal design has the highest power, followed by the conjectured design which is close to the optimal design. The random design has the least power among all the designs for all values of c.

(b) $n = 1000$, $(\beta_0, \beta_1, \beta_2) = (2, -2, 2)$, $\sigma_y^2 = 4$, $(\mu_{1,x}, \mu_{2,x}, \sigma_{1,x}^2, \sigma_{2,x}^2) = (2, 2, 16, 16)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-2.7, 0.4, 0.2, -0.15)$.

TABLE 7.2: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(-2.7,0.4,0.2,-0.15)	0.913	0.835	1.246	1.088	1.227	1.088	0.782	0.793	1.100
Conjecture	1.23	1.35	1.42	1.46	1.50	1.52	1.55	1.57	1.59

TABLE 7.3: γ_2 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(-2.7,0.4,0.2,-0.15)	-0.084	-0.130	-0.287	-0.342	-0.489	-0.136	-0.462	-0.477	-0.575
Conjecture	-0.30	-0.21	-0.20	-0.17	-0.16	-0.12	-0.10	-0.09	-0.08

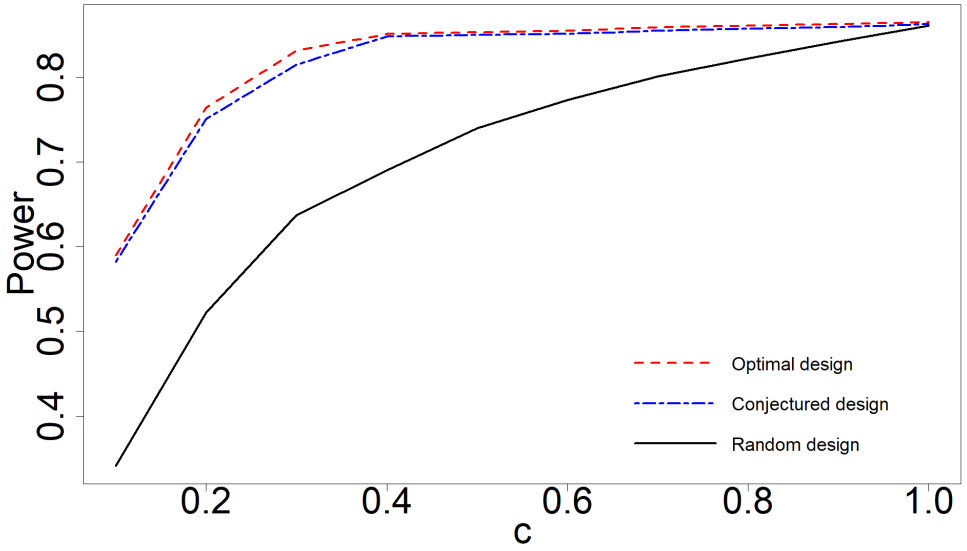

FIGURE 7.11: Power plot using different designs.

In order to have the same number of missing values in a different dataset as in example (c) and (d), the sample size increased from 400 in example (a) to 1000. The coefficient β_0 increases from 1 to 2, while β_1 decreases to -2. In this example, X has a smaller variance of 2 and a larger mean of 3. The model for introducing the missing values also changes, with α_0 changing from positive to negative, α_1 from negative to positive, and α_2 taking a value of 0 for MAR. For the MNAR example in (d), all parameters remain the same, except for α_2, which takes a value of 0.3.

The example below shows the Type I error and power for the three different designs with a quadratic term in the regression model and the missing mechanism model. The inclusion of the quadratic term is to see if this affects the Type I error of the test or not. The Type I errors for the three designs are shown in Table 7.4. Table 7.5 shows the γ_1 values for the optimal and conjectured designs. All three designs have Type I error values of approximately 0.05 showing that the inclusion of the quadratic term in the regression model does not affect the test.

(c) $n = 1000$, $(\beta_0, \beta_1) = (2, -2)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (3, 2)$, $(\alpha_0, \alpha_1, \alpha_2) = (2.9, -0.13, 0)$ with regression $\beta_0 + \beta_1 x^2$ and missing mechanism $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

TABLE 7.4: Type I error for different designs in 2000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
(2.9,-0.13,0)	0.051	0.048	0.054	0.056	0.048	0.050	0.050	0.050	0.050	0.052
Random	0.055	0.040	0.045	0.045	0.041	0.055	0.053	0.050	0.065	0.054
Conjecture	0.051	0.051	0.049	0.049	0.052	0.054	0.050	0.052	0.056	0.045

TABLE 7.5: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(2.9,-0.13,0)	-0.236	-0.469	-0.420	-0.162	-0.215	-0.226	-0.695	-0.072	0.037
Conjecture	-0.10	-0.30	-0.35	-0.38	-0.40	-0.42	-0.45	-0.47	-0.50

In the example below, we consider the MNAR mechanism and introduce a quadratic term in the regression and missing mechanism models. Figure 7.12 shows the power of the three designs. The γ_1 values are shown in Table 7.6 for optimal and conjectured designs. The optimal design has the highest power while the random design has the smallest power at values of c. The conjectured design is better than the random design and performed slightly worse than the optimal design (lies between the optimal design and random design).

(d) $n = 1000$, $(\beta_0, \beta_1) = (2, -2)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (3, 2)$, $(\alpha_0, \alpha_1, \alpha_2) = (2.9, -0.13, 0.3)$ with regression $\beta_0 + \beta_1 x^2$ and missing mechanism $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

TABLE 7.6: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(2.9,-0.13,0.03)	-0.927	-0.972	-0.957	-0.952	-0.937	-0.976	-0.974	-0.895	-0.579
Conjecture	-0.45	-0.42	-0.40	-0.38	-0.36	-0.35	-0.32	-0.13	-0.11

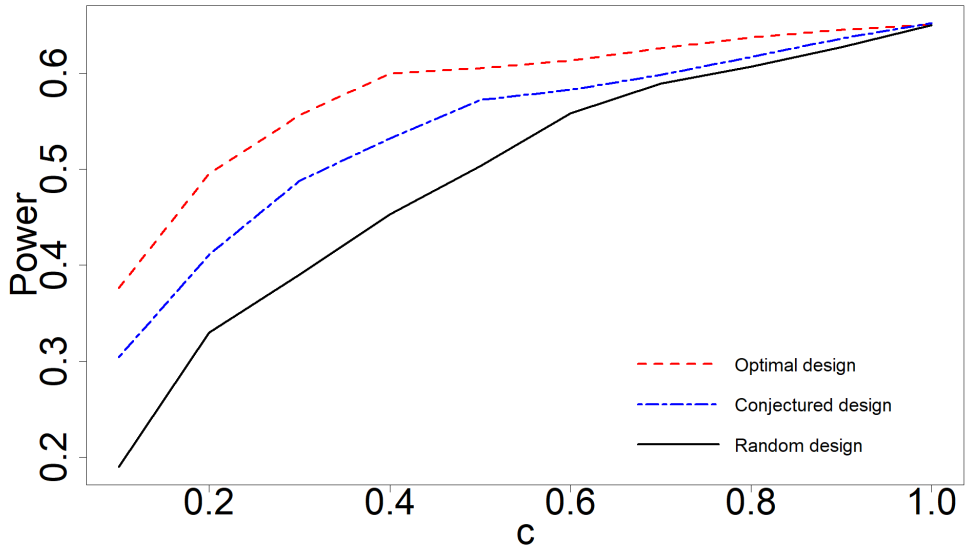

FIGURE 7.12: Power plot using different designs.

We consider an example with five covariates to examine the performance of the conjectured design. Figure 7.13 shows the power for both designs and the γ values for the

conjectured design are shown in Table 7.7. As seen in Figure 7.13, the random design has a smaller power than the conjectured design.

(e) Generate 1000 points following a multiple linear regression model in 10000 replicates:

$$Y|(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5) \sim N(2 - x_1 + 1.2x_2 + x_3 - 0.8x_4 + 1.5x_5, 1),$$

with $X_1 \sim N(3,4)$,$X_2 \sim N(2,1)$, $X_3 \sim N(1,4)$, $X_4 \sim N(2,4)$ and $X_5 \sim N(1,1)$. Introduce MNAR missingness into Y using:

$$P(M = 1|Y = y, X = x) = \frac{\exp(0.58 - 0.25x_1 + 0.13x_2 + 0.23x_3 + 0.1x_4 - 0.3x_5 + 0.23y)}{1 + \exp(0.58 - 0.25x_1 + 0.13x_2 + 0.23x_3 + 0.1x_4 - 0.3x_5 + 0.23y)}.$$

Using the random and conjectured designs, obtain the power of test.

TABLE 7.7: γ values

γ_i	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
γ_1	-0.1	-0.18	-0.2	-0.23	-0.26	-0.3	-0.32	-0.35	-0.38
γ_2	0.14	0.22	0.27	0.29	0.32	0.34	0.36	0.39	0.45
γ_3	0.24	0.25	0.28	0.32	0.35	0.37	0.39	0.42	0.46
γ_4	-0.05	-0.1	-0.16	-0.18	-0.2	-0.22	-0.25	-0.27	-0.3
γ_5	0.05	0.06	0.11	0.13	0.23	0.25	0.27	0.29	0.33

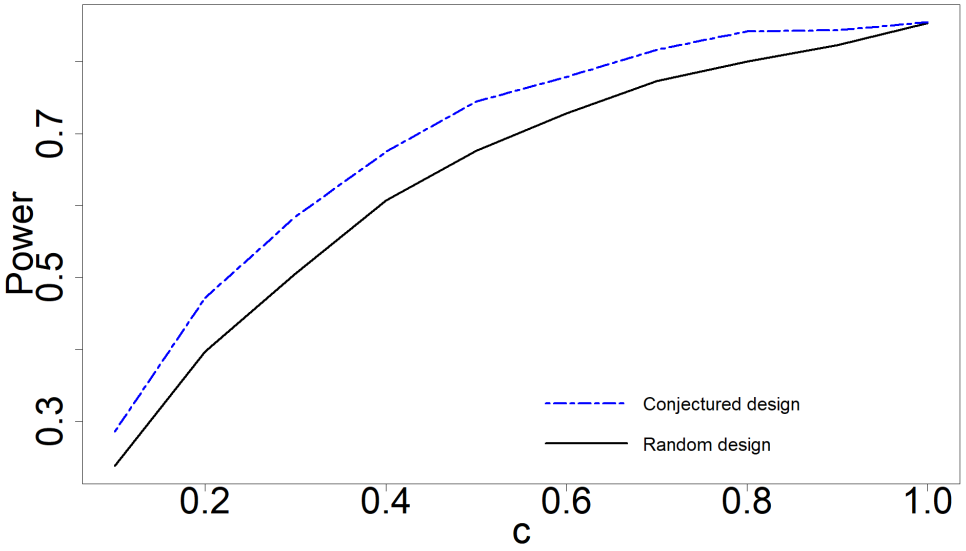

FIGURE 7.13: Power plot using different designs.

In this very extreme example, we consider the application of the conjecture for a scenario with ten covariates. The use of numerical optimisation when using Algorithm 1

for T_1 would not be possible for this instance. It is important to have a conjecture which can give us designs that are better than random for scenarios for which it would be numerically difficult to get an optimal design. The γ values for the conjectured design are shown in Table 7.8. Figure 7.14 shows that the random design has smaller power values compared to the conjectured design.

(f) Generate 1000 points following a multiple linear regression model in 10000 replicates:

$$Y|(X = x) \sim N(-2.3x_1 + 0.78x_2 + 0.5x_3 - 1.8x_4 + 0.65x_5 + 0.12x_6 - x_7 - 0.8x_8 + 1.5x_9 - 0.8x_{10}, 4),$$

for $i = 1, \ldots, 10$ with $X_1 \sim N(3,4)$, $X_2 \sim N(2,9)$, $X_3 \sim N(2,4)$, $X_4 \sim N(2,4)$ and $X_5 \sim N(1,9)$, $X_6 \sim N(3,4)$, $X_7 \sim N(2,4)$, $X_8 \sim N(3,9)$, $X_9 \sim N(1,4)$ and $X_{10} \sim N(2,1)$. Introduce MNAR missingness into Y using a logistic regression model with linear predictor: $0.8 - 1.55x_1 + 0.13x_2 + 2.3x_3 - 1.1x_4 - 0.35x_5 + 0.87x_6 + 0.48x_7 + 0.23x_8 + 0.41x_9 + 0.3x_{10} + 0.25y$.

Using the random and conjectured designs, obtain the power of test.

TABLE 7.8: γ values

γ_i	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
γ_1	-0.18	-0.25	-0.27	-0.30	-0.33	-0.36	-0.38	-0.40	-0.43
γ_2	0.03	0.05	0.07	0.08	0.10	0.13	0.15	0.18	0.20
γ_3	0.20	0.23	0.26	0.29	0.31	0.33	0.35	0.38	0.40
γ_4	-0.20	-0.22	-0.25	-0.28	-0.30	-0.32	-0.35	-0.38	-0.40
γ_5	-0.02	-0.04	-0.07	-0.09	-1.10	-1.30	-1.70	-1.85	-2.00
γ_6	0.10	0.13	-0.15	0.18	0.20	0.23	0.25	0.28	0.30
γ_7	-0.02	-0.05	-0.07	-0.08	-0.12	-0.15	-0.17	-0.22	-0.27
γ_8	-0.17	-0.20	-0.22	-0.24	-0.27	-0.28	-0.30	-0.32	-0.35
γ_9	0.13	0.15	0.17	0.19	0.20	0.22	0.24	0.25	0.30
γ_{10}	-0.20	-0.22	-0.23	-0.25	-0.27	-0.30	-0.33	-0.35	-0.38

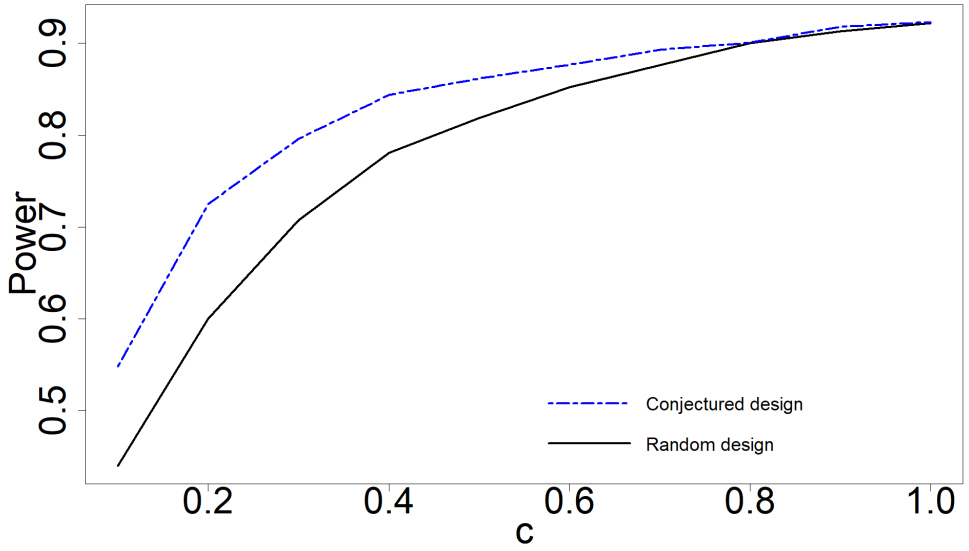

FIGURE 7.14: Power plot using different designs.

Tables A.13 - A.17 and Figures A.1 - A.3 in the Appendix show result for different cases using the optimal design, conjectured design and random design. These results show similar trends to those shown in this section. For all the cases, the optimal design outperformed the other design. The random design performed the least among all designs. Overall, these examples show that the conjecture design outperforms the random design and is often close to the optimal design.

7.2.2 Saturated Model Fitting

This subsection shows another use for Conjecture 1 that will not only make the recovery design but also Test T_1 more robust to model misspecifications. With all parametric tests, when the model terms are misspecified, the Type I error rate may be off. For example, in a one-covariate model, if we assume a linear relationship with the covariate in the linear predictor of the missing mechanism, but in reality it is quadratic, the test may not be level α. In a situation where there is very little knowledge of the true relationship, the conjectured design (which is purely based on the data and thus free from any model assumptions) can be used and then Test T_1 with a conservatively specified model can be used. For example, if we are not sure if there may be a quadratic term in x in the linear predictor, we could add quadratic and perhaps even cubic terms when performing the test. As long as this 'larger' model approximately contains the 'true' model there should be no problem with the Type I error, although we may encounter a slight reduction in power as a trade-off. These results are shown in the scenarios below.

Using the parameters and models in scenario 1, the Type I error of the test that tests a linear predictor with $\alpha_0 + \alpha_1 x^2$ against $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$ are shown in Tables 7.9 using the

conjecture and random designs. This scenario is to show that despite the inclusion of quadratic term in the the missing mechanism, the Type I error of the test does not fail. For MAR, the Type I error rate for both designs are close to 0.05. For MNAR, Figure 7.15 shows the power of the test using both designs. The power of the test using the random design is smaller than that of the conjectured design. As shown in Figure 7.15, the conjecture (blue) has higher power than the random design (black).

Scenario 1: $n = 1000$, $(\beta_0, \beta_1) = (0.5, -1.2)$, $\sigma_y^2 = 4$, $(\mu_x, \sigma_x^2) = (2, 3)$, $(\alpha_0, \alpha_1, \alpha_2) = (4, -0.45, -0.18)$ with regression $\beta_0 + \beta_1 x^2$ and missing mechanism $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

MAR: the Type I error rate of the test that tests a linear predictor with $\alpha_0 + \alpha_1 x^2$ against $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

TABLE 7.9: Type I error for different designs in 2000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Random	0.049	0.051	0.048	0.050	0.052	0.056	0.052	0.056	0.059	0.051
Conjecture	0.056	0.052	0.050	0.057	0.053	0.058	0.054	0.052	0.053	0.054

MNAR: power of the test that tests a linear predictor with $\alpha_0 + \alpha_1 x^2$ against $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

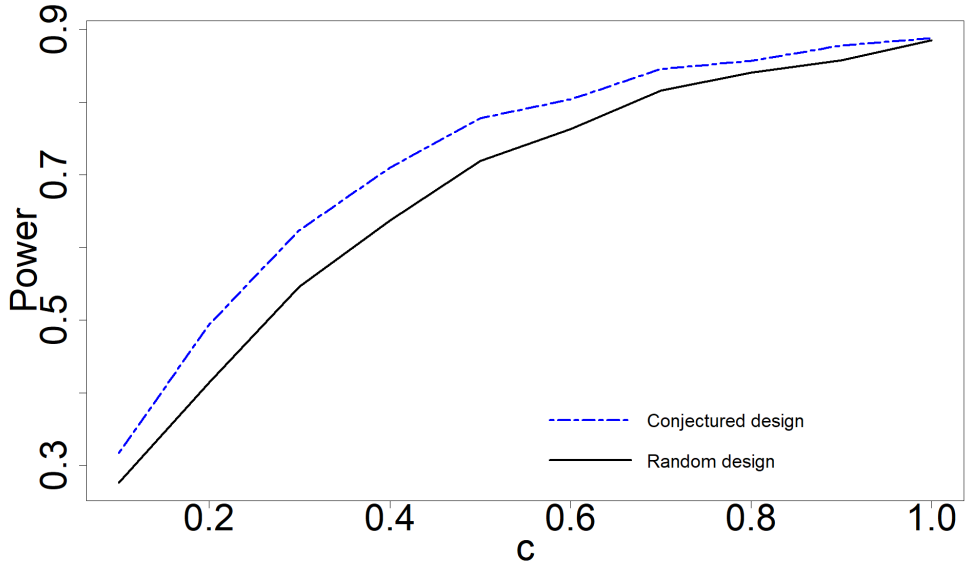

FIGURE 7.15: Power plot using different designs.

Scenario 2: $n = 1000$, $(\beta_0, \beta_1) = (2, -2)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (3, 2)$, $(\alpha_0, \alpha_1, \alpha_2) = (2.9, -0.13, 0.3)$ with regression $\beta_0 + \beta_1 x^2$ and missing mechanism $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

In this scenario, the Type I error and power of the test that tests $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$ against $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 y$ are respectively shown in Table 7.10 and Figure

7.16 respectively. The Type I error for both the random design and conjectured design are all close to 0.05 for all values of c. The conjectured design performs better than the random design. To see if there is a drop in the power of the test when the model is overfitted, Figure 7.17 shows the power of the test of $\alpha_0 + \alpha_1 x^2$ against $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$. In this case, the conjectured design performs better than the random design. Comparing Figures 7.16 and 7.17 shows that there is not much drop in the power obtained from optimising the recovery design using the true model or when the model is overfitted.

MAR: the Type I error rate of the test that tests a linear predictor with $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$ against $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 y$.

TABLE 7.10: Type I error for different designs in 2000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Random	0.045	0.045	0.054	0.048	0.049	0.048	0.052	0.049	0.052	0.049
Conjecture	0.040	0.046	0.043	0.044	0.042	0.040	0.040	0.041	0.035	0.041

MNAR: the power of the test that tests a linear predictor with $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$ against $\alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 y$.

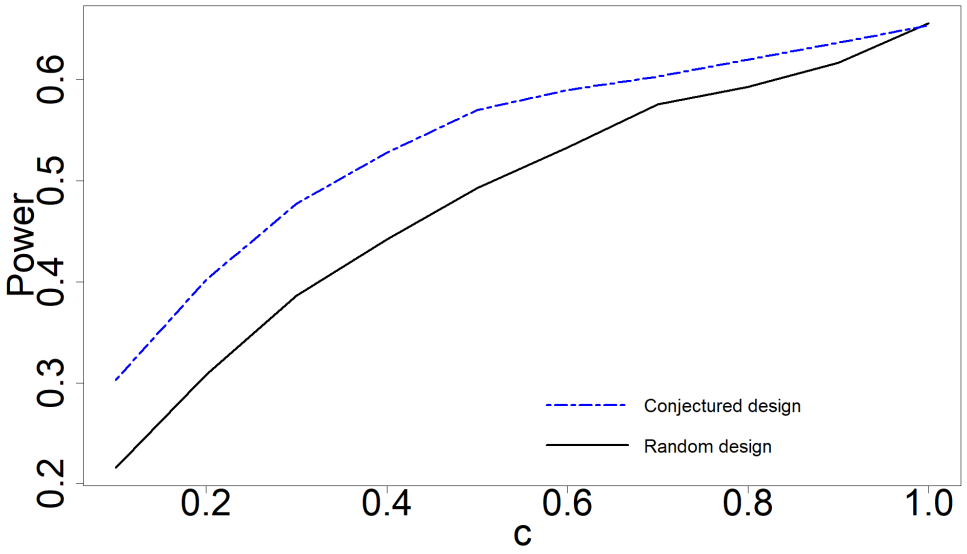

FIGURE 7.16: Power plot using different designs.

MNAR: power of the test that tests a linear predictor with $\alpha_0 + \alpha_1 x^2$ against $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$

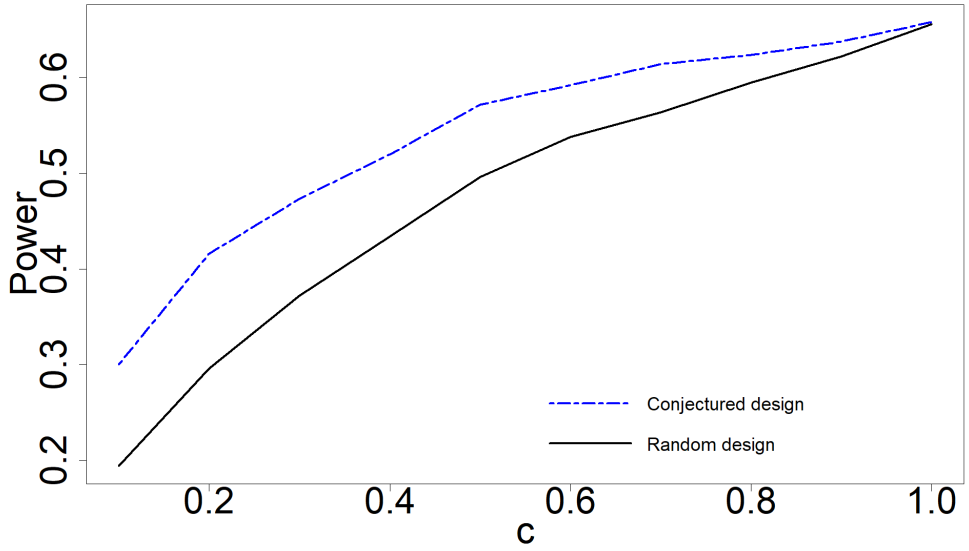

FIGURE 7.17: Power plot using different designs.

7.3 Application of methods to a real data example

We apply our methods to a scenario where our observed data values are not simulated from a known statistical distribution. In order to do this we use a study derived from the 1979 National Longitudinal Survey of Youth, commonly referred to as the NLSY79. This longitudinal survey, begun in 1979, interviewed a nationally representative sample of 12,686 young men and women in the US. From 1986, information on children born to women in the survey was also collected. For more information about the survey see Mitra and Reiter (2011, 2016). We note here that our sole purpose is to establish potential gains in detecting the presence of MNAR in this setting where the data generating mechanism is unknown. We do not seek to infer true causal effects, with the standard problem of unmeasured confounders being potentially relevant here.

Following Mitra and Reiter (2011, 2016) we subset on first born children only to avoid complications with family nesting. The resulting data set comprises 4888 observations. For our analysis model we consider a linear regression model of the form,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i\,, \quad \epsilon_i \sim N(0, \sigma^2),$$

where y_i corresponds to the ith child's Peabody Maths score (PIATM) administered at 5 or 6 years of age (taken as a proxy for cognitive development), and x_i corresponds to (the logarithm of) family income at birth, $i = 1, \ldots, 4888$. Since the developments of the paper assume covariates are fully observed, we focus on the subset on the 4888 observations with observed family income, resulting in 3596 observations. We note

alternative analysis models could be chosen but settled on this with evidence in the literature suggesting a relationship between these two variables (Cooper and Stewart, 2021). Of the 3596 observations, PIATM is missing in 1640 cases (a 45.6% missing rate). Figure 7.18 plots a scatterplot of the 1956 observed PIATM scores against the log of family income. The parameters that determine the distribution of $y|x$, $(\beta_0, \beta_1, \sigma)$, are unknown to us in this real data setting and estimated from the available data. This gives estimates $\beta_0 \approx 69.06, \beta_1 \approx 3.14$ and $\sigma \approx 13.14$. When optimizing our recovery designs, we will assume these are the true values. This seems reasonable from findings in Section 6.3.1 which indicate robustness to slight misspecification of these parameters. In Figure 7.19, a solid black line depicts the density estimate of log(income). A candidate density is the skewed normal distribution with $\xi = 11.3, \omega = 1.4, \alpha = -3$, indicated by the dashed blue line overlaid on the density estimate.

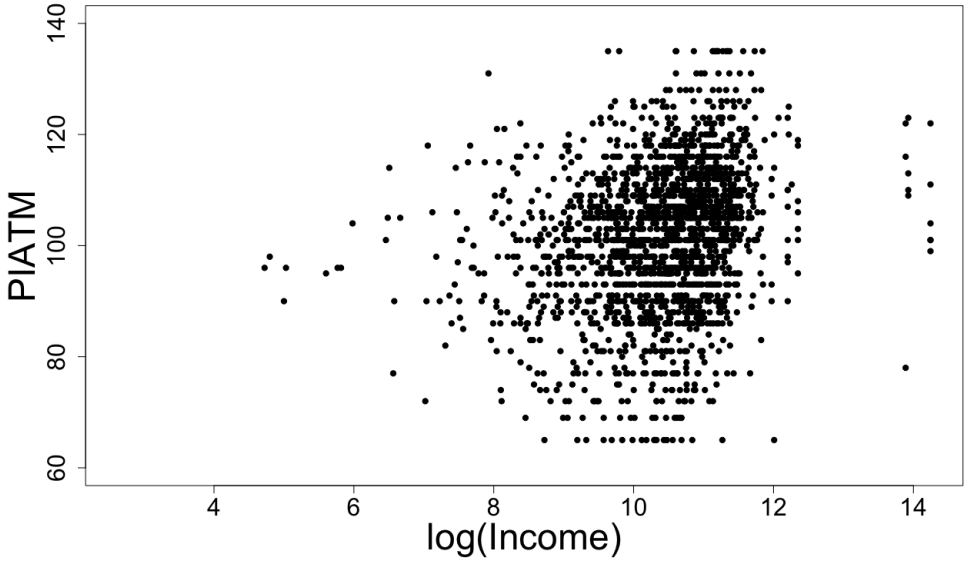

FIGURE 7.18: Peabody score vs log(income) scatterplot.

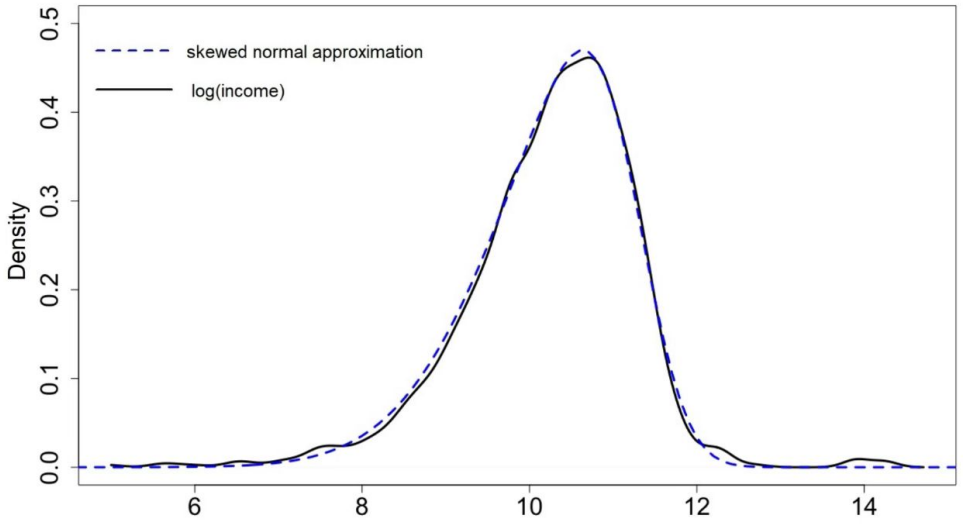

FIGURE 7.19: Density of log(income) and skewed normal approximation.

7.3.1 Simulation based on the complete case subsample

In this section, we subset our data on the complete case subsample, which comprise 1956 units. We re-introduce missing values into the outcome, y, using the following mechanism:

$$\Pr(M = 1|y, x) = \exp(\alpha_0 + \alpha_1 x + \alpha_2 y)/(1 + \exp(\alpha_0 + \alpha_1 x + \alpha_2 y)) . \tag{7.1}$$

We specifically consider two scenarios for the missing data mechanism:

A) $(\alpha_0, \alpha_1, \alpha_2) = (52, -5, 0.016)^T$,

B) $(\alpha_0, \alpha_1, \alpha_2) = (-50, 5, -0.016)^T$

The parameter values α_i are chosen to introduce approximately 40% missing data into the outcome (similar to the amount present in the original data). As before, we assume we are able to recover a proportion of the missing y_i values. In scenario A, α_0 and α_2 are positive and α_1 is negative while in B, α_0 and α_2 are negative and α_1 is positive. The absolute values for α_1 and α_2 are the same in both scenarios except for α_0 which reduces from 52 in A to 50 in B.

We apply Algorithm 1 for T_1 for optimizing the power of the test described in Section 6.2 above. We do this by repeatedly generate new samples from the complete cases by resampling rows of the data with replacement. The number of samples is set to 1956 to match the number of complete cases. In each scenario, we then introduce

missing values into the outcome using the mechanisms specified above. We compare the performance of Algorithm 1 for T_1 against a random recovery as well as designs formed from Conjecture provided in Section 7.2.1.

Our results are presented in Figures 7.20 and 7.21 below, where we use the scheme of dashed red, dot-dashed blue and solid black for Algorithm 1 for T_1, Conjecture 1 and the random recovery, respectively. We firstly remark that the power for Algorithm 1 and Conjecture 1 are almost identical, with an ever so slight advantage to Algorithm 1 for T_1 as we would anticipate. We see that in Scenario A as shown in Figure 7.20, there is a substantial increase in power with small recovery proportions, for example for a recovery proportion $c = 0.3$ the random recovery has a power of 0.45 approximately, compared to a power of 0.6 using the optimal recovery design. The significant increase in power is more noticeable in Figure 7.21 for Scenario B. For this scenario we can see that for $c = 0.4$ the power using the optimal recovery design is approximately 0.65, and to achieve a similar power using the random recovery design, a recovery proportion of almost 1 would be required, which may be impractical in many settings.

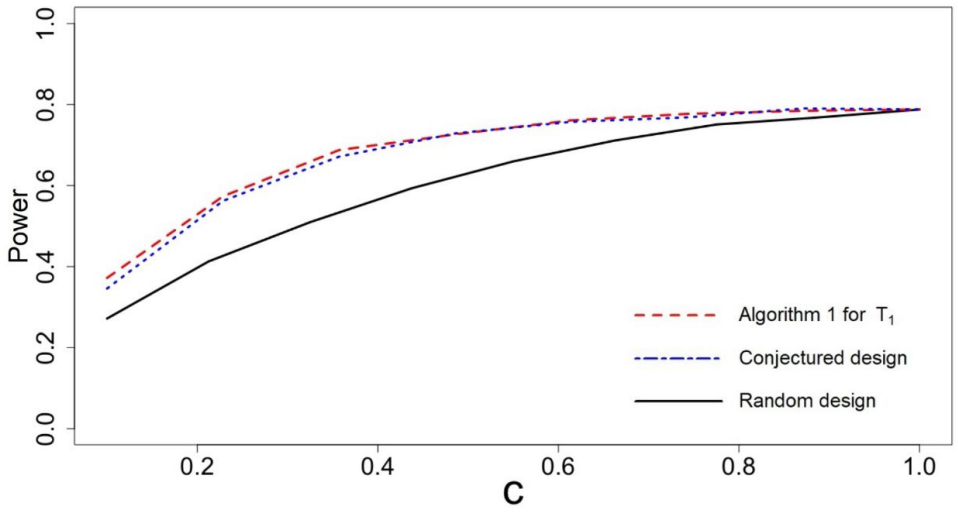

FIGURE 7.20: A comparison of power using the optimal recovery design using Algorithm 1 for T_1, the recovering design from Conjecture 1 and a random recovery design

.

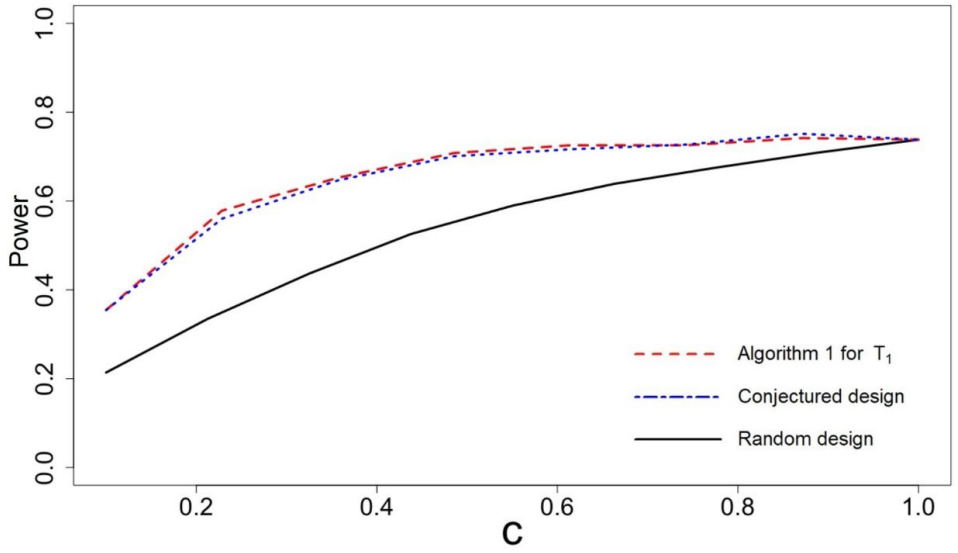

FIGURE 7.21: A comparison of power using the optimal recovery design using Algorithm 1 for T_1, the recovering design from Conjecture 1 and a random recovery design

.

Chapter 8

Using a test for MAR vs MNAR to improve estimation

In previous chapters, we have discussed how a follow up sample of missing responses allows us to perform tests for MAR vs MNAR. We have also demonstrated how assuming an incorrect missing mechanism can lead to considerable bias in estimation problems. The purpose of this chapter is to demonstrate that utilizing the tests for MAR vs MNAR developed in this thesis can help identify the correct MDM, and therefore improve the estimation of parameters.

8.1 Estimating $E(Y)$

In this section, we review the non-parametric estimators in Kim and Yu (2011) and develop a new estimator that utilises a test for MAR vs MNAR before estimation. Cheng (1994) proposed a non-parametric estimator for estimating functional mean of a response variable when the missing mechanism is MAR. In this section, this estimator will be called Cheng's estimator and is formally introduced in Definition 8.1. Kim and Yu (2011) discussed a non-parametric estimator used to estimate the mean of the response variable Y in the presence of missing values and compared with Cheng's estimator. When the missing mechanism is ignorable (i.e. MAR), the Cheng's estimator can be used while the semi-parametric estimator can be used when the missing mechanism is non-ignorable (MNAR).

Using the problem formulation in Subsection 6.1.1 with one covariate, suppose we are interested in estimating the mean of Y. Recall M_i is the missing indicator that takes value 1 when y_i is observed and 0 when y_i is missing.

Definition 8.1. Kim and Yu (2011) defined a consistent estimator of $\Pr(Y \in dy \mid X = x, M = 0)$ under missing at random mechanism, called the Chengs estimator and denoted as $\hat{\theta}_1$ as:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} M_i y_i + (1 - M_i)\hat{m}(x_i),$$

where $\hat{m}(x_i)$, the consistent estimator of $m(x_i) = E(y_i|x_i)$, is defined as:

$$\hat{m}(x_i) = \sum_{i=1}^{n} \frac{M_i K_h(x_i, x) y_i}{\sum_{i=1}^{n} M_i K_h(x_i, x)}.$$

Definition 8.2. For the non-ignorable missing mechanism, a consistent estimator of $\Pr(Y \in dy \mid X = x, M = 0)$ called the semi-parametric estimator (Kim and Yu, 2011) and denoted as $\hat{\theta}_{SE}$ is:

$$\hat{\theta}_{SE} = \frac{1}{n}\sum_{i=1}^{n} M_i y_i + (1 - M_i)\hat{m}(x_i; \hat{\psi}),$$

where $\hat{m}(x_i; \hat{\psi})$ is defined as:

$$\hat{m}(x; \hat{\psi}) = \sum_{i=1}^{n} \frac{M_i K_h(x, x_i) \exp(-\hat{\psi} y_i) y_i}{\sum_{i=1}^{n} M_i K_h(x, x_i) \exp(-\hat{\psi} y_i)},$$

where $K_h(u, x) = h^{-1}K\{(u - x)/h\}$, $K(\cdot)$ is a symmetric density function (kernel) on $\mathbb{R}$ and $h = h_n$ is the bandwidth such that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. A consistent estimator $\hat{\psi}$ of ψ can be obtained by solving:

$$\sum_{i=1}^{n} (1 - M_i)\delta_i\{y_i - m(x_i; \psi)\} = 0.$$

δ_i is an indicator variable that takes value 1 if the missing y_i is recovered in the follow-up sample and 0 if not recovered.

Definition 8.3. Under missing at random mechanism, introduce recovery or follow-up to Cheng's estimator in Definition 8.1 to obtain Cheng with follow-up denoted as $\hat{\theta}_{CE}$, a consistent estimator of $\Pr(Y \in dy \mid X = x, M = 0)$ is:

$$\hat{\theta}_{CE} = \frac{1}{n}\sum_{i=1}^{n} M_i y_i + (1 - M_i)(\delta_i y_i) + (1 - M_i)\hat{m}(x_i; 0),$$

where δ_i is the same as defined in Definition 8.2 and $\hat{m}(x_i)$ as defined in Definition 8.1

Combined estimator $\hat{\theta}_{TE}$: We introduce a new estimator called $\hat{\theta}_{TE}$. This is a combined estimator as it first tests the presence of MAR vs. MNAR and uses one of the estimators

earlier discussed. Let's consider a scenario with combinations of MAR and MNAR datasets. If the dataset is MAR, use $\hat{\theta}_{CE}$ for estimation and if otherwise, use $\hat{\theta}_{SE}$. The algorithm for this estimator is as follows:

Algorithm 9 Estimation Procedure for $\hat{\theta}_{TE}$

1: **Input:** A combination of datasets where each dataset may be MAR (Missing at Random) or MNAR (Missing Not at Random).
2: **Output:** Statistical properties of the estimator $\hat{\theta}_{TE}$ for $\mathbb{E}(Y)$.
3: **Initialization:**
4: Recover c of the missing values of y and augment them with the observed cases.
5: **Steps:**
6: Test the hypothesis $H_0 : \psi = 0$.
7: **if** MAR is present **then**
8: Use the estimator $\hat{\theta}_{CE}$.
9: **else**
10: Use the estimator $\hat{\theta}_{SE}$.
11: **end if**
12: Obtain the bias, variance, and MSE of the chosen estimator for $\mathbb{E}(Y)$.
13: **Return:** The estimate of the bias, variance, and MSE.

8.2 Simulation studies

For two regression models and eight different missing data mechanisms, we will compute the bias, variance and MSE for each estimator $\hat{\theta}_1$, $\hat{\theta}_{SE}$ and $\hat{\theta}_{CE}$. In mechanism 1, M1, we will consider a MAR scenario. In all other scenarios, we will consider MNAR. These examples are taken from the examples used in Kim and Yu (2011) to cover a large variety of different MNAR scenarios.

(a) Generate 200 observations with 2000 replications and obtain the bias, variance and MSE of Y from $X_i \sim N(2,1)$ and $e_i \sim N(0,1)$ for Model A: $y_i = 1 + 0.7x_i + e_i$ and model B: $y_i = 1 + 0.5(x_i - 2.5) + e_i$ with 15% recovery and $h = \hat{\sigma}_x n^{-0.2}$ ($\hat{\sigma}_x$ represents the estimated standard deviation of x_i in the sample) using the following missing mechanisms in both models:

$$M1 = \frac{\exp(\alpha_0 + \alpha_1 x_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i)},$$

where $(\alpha_0, \alpha_1) = (-1.5, 1.0)$ for both models.

$$M2 = \frac{\exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i)},$$

where $(\alpha_0, \alpha_1, \alpha_2) = (-0.85, 0.3, 0.3)$ for model A and $(\alpha_0, \alpha_1, \alpha_2) = (-1.58, 0.5, 0.7)$ for model B.

$$M3 = \frac{\exp(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 y_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 y_i)},$$

where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-2.0, 0.3, 0.3, 0.3)$ for model A and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-2.72, 2.72, -0.68, 0.7)$ for model B.

$$\begin{aligned} M4 &= 0.5 \; if y_i \leq p \\ &= 1 \; if y_i > p, \end{aligned}$$

where $p = 3.4$ for model A and $p = 2.5$ for model B. In context, the value of the response beyond a threshold, p, will always be observed, whereas, the value below p will only be observed with probability 0.5.

$$M5 = \frac{\exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i + \alpha_3 y_i^2)}{1 + \exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i + \alpha_3 y_i^2)},$$

where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-0.65, 0.1, 0.1, 0.1)$ for model A and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-0.85, 0.1, 0.1, 0.3)$ for model B.

$$M6 = \alpha(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i),$$

where $\alpha(\cdot)$ is the cdf of the standard normal distribution, $(\alpha_0, \alpha_1, \alpha_2) = (-0.64, 0.1, 0.3)$ for model A and $(\alpha_0, \alpha_1, \alpha_2) = (-0.53, 0.1, 0.4)$ for model B.

$$M7 = 1 - exp\{-exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i)\},$$

where $(\alpha_0, \alpha_1, \alpha_2) = (-1.4, 0.3, 0.3)$ for model A and $(\alpha_0, \alpha_1, \alpha_2) = (-1.15, 0.3, 0.3)$ for model B.

$$M8 = \frac{\exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i + \alpha_3 x_i y_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i + \alpha_3 x_i y_i)},$$

where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-1.4, 0.1, 0.1, 0.3)$ for model A and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-0.15, 0.1, 0.1, 0.1)$ for model B.

Table 8.1 shows the bias, variance and MSE values for the three different estimators. $\hat{\theta}_1$ has the highest bias and $\hat{\theta}_{SE}$ has the highest variance while $\hat{\theta}_{CE}$ has the least variance among the estimators for both models and missing mechanisms.

(b) Generate 200 observations with 2000 replications using $y_i = 1 + 0.7x_i + e_i$ such that, $X_i \sim N(2, 2)$ and $e_i \sim N(0, 1)$. Introduce missingness using the model below and 15% recovery.

$$M1 = \frac{\exp(\alpha_0 + \alpha_1 x_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i)},$$

where $(\alpha_0, \alpha_1) = (-1.5, 1.3)$.

TABLE 8.1: Monte Carlo biases, variances and mean squared errors for different estimators.

Missing Mechanism	Model	Estimates	$\hat{\theta}_1$	$\hat{\theta}_{SE}$	$\hat{\theta}_{CE}$
M1	A	Bias	0.0363	-0.0046	0.0274
		Var	0.0120	0.0208	0.0108
		MSE	0.0134	0.0208	0.0115
	B	Bias	-0.0701	-0.0007	-0.0607
		Var	0.0125	0.0243	0.0118
		MSE	0.0174	0.0243	0.0155
M2	A	Bias	0.1294	0.0020	0.1131
		Var	0.0102	0.0188	0.0098
		MSE	0.0269	0.0188	0.0225
	B	Bias	0.2409	-0.0001	0.2049
		Var	0.0107	0.0195	0.0106
		MSE	0.0688	0.0195	0.0525
M3	A	Bias	0.1366	0.0006	0.1307
		Var	0.0112	0.8011	0.0108
		MSE	0.0299	0.8011	0.0279
	B	Bias	0.1939	0.0026	0.1682
		Var	0.0116	0.0213	0.0104
		MSE	0.0492	0.0213	0.0388
M4	A	Bias	0.0037	0.0015	0.0006
		Var	0.0114	0.0389	0.0097
		MSE	0.0114	0.0389	0.0097
	B	Bias	-0.0046	0.0100	-0.0024
		Var	0.0181	0.0410	0.0101
		MSE	0.0181	0.0410	0.0102
M5	A	Bias	0.2077	-0.0013	0.1781
		Var	0.0104	0.0185	0.0102
		MSE	0.0536	0.0185	0.0419
	B	Bias	0.2829	0.0019	0.0240
		Var	0.01222	0.0171	0.0113
		MSE	0.0922	0.0171	0.0689
M6	A	Bias	-0.3104	0.0007	-0.2586
		Var	0.0149	0.0256	0.0134
		MSE	0.1112	0.0256	0.0802
	B	Bias	-0.4576	0.0004	-0.3904
		Var	0.0148	0.1304	0.0132
		MSE	0.2242	0.1304	0.1656
M7	A	Bias	-0.1617	-0.0024	-0.1407
		Var	0.0124	0.0226	0.0117
		MSE	0.0385	0.0226	0.0315
	B	Bias	-0.1870	0.0033	-0.1538
		Var	0.0136	0.0253	0.0125
		MSE	0.0486	0.0253	0.0361
M8	A	Bias	0.2369	0.0017	0.2026
		Var	0.0105	0.0188	0.0098
		MSE	0.0667	0.0188	0.0509
	B	Bias	0.0896	-0.0038	0.076
		Var	0.0110	0.0207	0.0106
		MSE	0.0190	0.0207	0.0164

In Table 8.2, the estimates using the different estimators are shown. The best estimator is $\hat{\theta}_{CE}$ as it has the least variance and MSE among the estimators. $\hat{\theta}_{SE}$ has the least bias but the highest variance among the estimators. As the recovery proportion c increases, the bias, variance and MSE decrease for $\hat{\theta}_{CE}$ and $\hat{\theta}_{SE}$. The estimates of $\hat{\theta}_1$ remain constant at all values of c because it ignores the missing observations.

(c) Using estimator $\hat{\theta}_{TE}$, generate 500 observations and 10000 replications such that each dataset is MAR or MNAR based on different percentages with $y_i = 1 + 0.7x_i + e_i$, $X_i \sim N(2,1)$ and $e_i \sim N(0,1)$ for a combination of MAR and MNAR using the model below and recovery proportion 0.15:

$$\pi_i = \frac{\exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i + \alpha_2 y_i)}.$$

where $(\alpha_0, \alpha_1, \alpha_2)$ is $(-1.5, 1, 0)$ for MAR and $(-0.85, 0.3, 0.3)$ for MNAR.

Tables A.18 and A.19 in the appendix show the estimates obtained for different combinations of MAR and MNAR datasets by first testing the type of missing mechanism present and using the corresponding estimator. P denotes the number of times H_0 is rejected. As c increases, P increases for each combination while other estimates reduce. For each value of c, P decreases as the percentage of MNAR in each combination decreases. At 100% MNAR, the value of P is the power of the test and at 100% MAR, P approximates the Type I error.

(d) For 1000 datasets, generate 1000 observations each such that the dataset is a combination of 90% MAR and 10% MNAR. Using $\hat{\theta}_{TE}$, $\hat{\theta}_{SE}$ and $\hat{\theta}_{CE}$. Obtain the bias, variance and MSE of Y.

Table A.20 shows the bias, variance and MSE for the 90% MAR and 10% MNAR combination. $\hat{\theta}_{SE}$ has the least bias at all values of c. $\hat{\theta}_{TE}$ has the least variance leading to the least MSE at all values of c, thereby implying that testing improves estimation. At higher values of c, all the estimators have similar MSE values.

TABLE 8.2: Monte Carlo biases, variances and mean squared errors for different estimators for different recovery proportions.

c	Estimates	$\hat{\theta}_1$	$\hat{\theta}_{SE}$	$\hat{\theta}_{CE}$
0.1	Bias	0.1563	-0.0047	0.1368
	Var	0.0212	0.0844	0.0206
	MSE	0.0457	0.0844	0.0393
0.2	Bias	0.1563	-0.0006	0.1254
	Var	0.0212	0.0493	0.0209
	MSE	0.0457	0.0493	0.0366
0.3	Bias	0.1563	-0.0044	0.1046
	Var	0.0212	0.0291	0.0180
	MSE	0.0457	0.0291	0.0289
0.4	Bias	0.1563	-0.0012	0.0941
	Var	0.0212	0.0284	0.0185
	MSE	0.0457	0.0284	0.0274
0.5	Bias	0.1563	-0.0014	0.0771
	Var	0.0212	0.0253	0.0164
	MSE	0.0457	0.0253	0.0224
0.6	Bias	0.1563	-0.0020	0.0617
	Var	0.0212	0.0205	0.0161
	MSE	0.0457	0.0205	0.0199
0.7	Bias	0.1563	-0.0001	0.0463
	Var	0.0212	0.0184	0.0157
	MSE	0.0457	0.0184	0.0179
0.8	Bias	0.1563	-0.0001	0.0315
	Var	0.0212	0.0169	0.0158
	MSE	0.0457	0.0169	0.0168
0.9	Bias	0.1563	-0.0005	0.0163
	Var	0.0212	0.0144	0.0141
	MSE	0.0457	0.0144	0.0144
1.0	Bias	0.1563	-2.60e-7	-1.11e-19
	Var	0.0212	0.0152	0.0149
	MSE	0.0457	0.0152	0.0149

(e) For different sample sizes and 0.15 recovery proportion, generate n observations each such that 75% of the dataset is MAR and 25% is MNAR. Using $\hat{\theta}_{TE}$, $\hat{\theta}_{SE}$ and $\hat{\theta}_{CE}$. Obtain the bias, variance and MSE of Y.

The result of this analysis is can be found in Table A.21 in the appendix. Figure 8.1 shows the Root Mean Square Error of the estimators. $\hat{\theta}_{TE}$ has the least RMSE among all estimators. The RMSE for $\hat{\theta}_{CE}$ is less than that of $\hat{\theta}_{SE}$ at sample size 2000 and lesser.

At $n > 2000$, $\hat{\theta}_{CE}$ has the highest RMSE among all estimators. As the sample size increases, the RMSE decreases for all estimators. Unlike the other two estimators that estimate $E(Y)$ without knowing the type of mechanism present, $\hat{\theta}_{TE}$ first tests for the type of missing mechanism present in the data before using the right estimation for each missing mechanism. This results in the least RMSE for $\hat{\theta}_{TE}$, thereby showing that testing before estimation improves estimation.

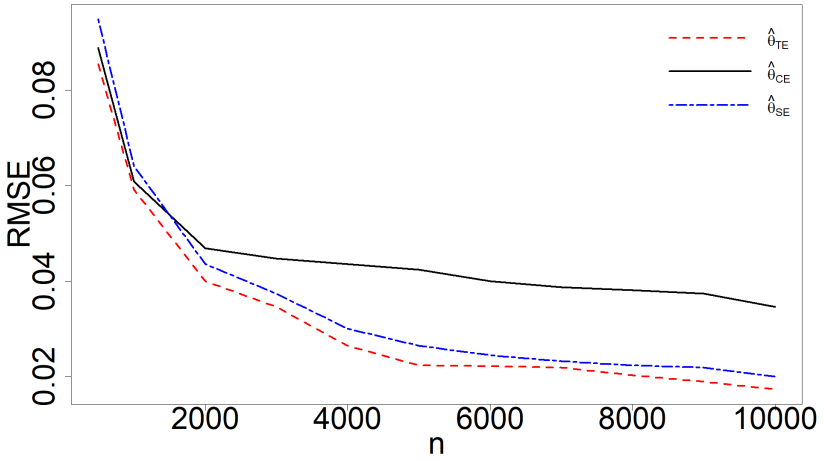

FIGURE 8.1: Root mean squared error (RMSE) for different estimators and sample sizes.

Chapter 9

Conclusion

9.1 Summary

This research started with exploring the existing methods of analysing missing data mechanisms. Some of the methods were sufficient to improve the inference and summary statistics for MCAR and MAR mechanisms. The existing methods performed poorly under a MNAR mechanism because of its complexity. The MNAR mechanism could only be tested with the presence of recovered missing values. In order to improve the inference on MNAR, the distribution of the complete cases and missing cases for MNAR were studied and the result showed that the distributional form does not change but the location does. This showed that when the model is fit on the observed data, the distribution changes, thereby leading to biased and inefficient estimates. The concept of recovery was introduced to MNAR data and studied for the improvement of inference. Initially, four different recovery designs were used and the effect of each design on power and summary statistics was studied. The PMF and SMF models were used on the four recovery scenarios to test for MAR and MNAR. The PMF model performs well for all the designs but the SMF model gives erroneous Type I errors.

We explored a comprehensive insight into testing for the presence of MNAR utilizing a recovery sample. Firstly, we conducted a theoretical study of the SMF test (5.4), where we established that the augmented data used for determining whether MNAR is present represents a sample from various mixture distributions. A careful construction of the marginal distributions then allowed us to express a formula for the missing data mechanism in the augmented data. The mathematical expressions for the missing data mechanism based on the subsample of observed plus recovered data permit principled inferences to be made in this setting. Notably, for the commonly used expit model, when the random design is used, we determine that there is a shift in the intercept of the model by $\log(c^*)$ while the other coefficients in the linear predictor, and indeed the form of the model, remain unchanged. For other link functions, to preserve

the original mechanism, the model must be fit to the recovered plus only a randomly sampled pre-specified proportion of the observed data. Consequently, the SMF test (5.4) with $H_0 : \psi = 0$ based on the augmented data will reliably test for the presence of MNAR in the original data, giving analysts confidence in their results. Secondly, using experimental design methods to construct the recovery sample allows inferences to be optimised. We considered how the power of the SMF test (5.4) can be increased through methods from the design of experiments. We propose constructing recovery designs that minimize the variance of the MLE (D_1-optimality) to increase efficiency and thereby improve the power of the SMF test (5.4). We also provide the equivalence to T-optimality (Atkinson and Fedorov, 1975) based on assigned subsampling probabilities to the covariates and how it considerably outperforms the random recovery sampling. The test based on subsampling probabilities uses all the observed cases as opposed to the restricted design that uses the observed cases that fall in the recovery region. This is a generalized design that can be used in obtaining the design in Chapter 5. This design leads to a better power than the design in Chapter 5 because it uses all the observed cases in addition to the recovered cases rather than restricting the observed cases to the cases that fall in the design region. The robustness of this design based on different misspecifications was considered and the optimal design outperformed the random design.

Multivariate cases were considered using the correlation of the covariates with the response variable to decide which of the covariates should be used in forming the design region. Results show that the covariate with the highest correlation with the response variable provides higher power than the other designs considered. We proposed a conjecture that forms the recovery design using the empirical densities of the covariates. The peak at the intersection of the densities formed by the missing and observed responses informs the recovery region. All observed cases are also used in this design. We compared results from the optimal, conjecture and random designs for more than two covariates. The optimal design and the conjecture design are similar in performance as they both outperformed the random design. The optimal design is slightly better than the conjecture design. However, the conjecture design is simple and can be found with a less complicated optimisation procedure.

This research provides a comprehensive approach to detecting MNAR missingness in an incomplete data set in practice, combining a design strategy with a corresponding reliable and well-understood test for MNAR. A well-chosen recovery design can achieve higher power to detect MNAR compared with random recovery sampling for fixed recovery proportions. Similarly, if the power is fixed in advance, an efficient recovery design can result in a smaller proportion of missing responses needing to be followed up to achieve this power, thus reducing costs.

In summary, the novel contributions of this Thesis are:

i a novel statistical test for MNAR. While the tests in (4.2) and (4.3) were originally proposed in Carpenter and Kenward (2012), to the best of our knowledge there has been no attempt in the literature to investigate their properties, or indeed the properties of any test for MNAR. Our research is thus the first comprehensive development and investigation of an MNAR test, which can subsequently be used in practice;

ii robust and efficient designs for recovering missing observations. An efficient design can considerably increase the power of the test compared with random sampling. As the recovery sample will often be small in practice, due to logistics and costs, this is a vital contribution to ensure the uptake of our test in practice;

iii a simple conjecture to find efficient designs without specific modelling assumptions, which could be particularly useful in multivariate settings. Where finding an optimal design would computationally be very expensive. Hence this conjecture, which makes it easy to find efficient designs, may further help to increase the uptake of our methodology in practice.

9.2 Future Work

It is assumed in this research that a follow-up unit would respond. However, in situations where this assumption is not realistic, a partial recovery during follow-up is a potential problem to look into in the future. One way to do this could be to extend the framework developed in Chapter 6 from a two-stage design to a three-stage design. This involves including a third attempt to obtain a response.

In the Thesis, we have seen some ways to make the MNAR test more robust to model misspecifications. It would be interesting to investigate this area further. An increase in the robustness of the MNAR test to model misspecifications could be achieved in two ways. Firstly, a semiparametric method could be generalized and tailored to fit the binary response model for the missing data mechanism. Secondly, a two-stage recovery design could be implemented, where the first stage is concerned purely with model exploration and selection. I.e. the augmented sample is utilized at the first stage to decide on the most appropriate parametric models for the data generating and missing data mechanisms. Then, in the second stage, use the formulated parametric models to optimize the remaining follow-up. This approach could be embedded in the repeated callbacks framework of Alho (1990). The approach raises two interesting follow-up questions: Firstly, must a uniform random recovery be used in the first stage to ensure consistent parameter estimates from the different models are obtained, or can a non-uniform recovery sample, based on an a priori conjecture, be implemented, thus leading to a more efficient overall procedure? Secondly, what would be the optimal division of follow-up sampling into the first and second stages?

Depending on the application, there may be more than one variable with missing values, resulting in a multivariate testing problem. There are two ways in which this problem could be addressed within our framework. Firstly, a multivariate response version of the likelihood ratio test could be developed, analogously to Test T_1 in Subsection 6.1.4.4. In this case, the T_E-optimality criterion would only need minor adjustment, i.e. in the calculation/approximation of the objective function $T(\gamma)$. The second approach would use multiple testing (test T_1 applied to each incomplete response) with Bonferroni correction. While this approach provides a simpler solution to the testing problem and would also inform the experimenter in which variable(s) MNAR missingness occurs, the design problem will become more complicated as there would be a need to deal with multi-objective optimization. To address this, a compound criterion (which maximizes a (weighted) average of the individual T_E-objective functions) or take a Pareto front approach could be considered. Substantial further investigation would be needed to find out which strategy is most beneficial.

In this research we focussed on the SMF, and hence provided an optimal design for the SMF test only. However, it is possible to obtain an optimal design for the PMF. The two tests can then be compared when both are run with their respective optimal designs and recommendations as to which of them should be used in which situation could be provided.

Another area of exploration is the use of the conjecture when the covariates are categorical variables. The marginal distributions of the categorical covariates can be used to find efficient designs rather than the empirical densities used in the continuous covariates. The following idea may be worth trying: when there is one categorical covariate in the model. Suppose the categorical covariate takes values: $1, 2, \ldots, p$. This data can be divided into the observed and missing datasets using the response variable. The marginal distribution of both the missing dataset and observed dataset can be obtained. The recovery distribution can then be constructed for each case as follows: For each category, we could assign a probability as the average of the probabilities for this category in the observed and the missing datasets, respectively. This can be expanded to more than one categorical covariate. Also, this research can be explored when there is a mixture of categorical and continuous covariates in the model.

Appendix A

Additional results

A.1 Tables relating to Chapter 4

In this section, the MAR Type I error and MNAR power analysis are shown using the different recovery designs.

TABLE A.1: MAR Type I error and MNAR power analysis with different recovery design and sample sizes in 10000 replicates.

Design	n^*	Selection Model		Pattern Mixture	
		MAR	MNAR	MAR	MNAR
Random	30	0.049	0.512	0.050	0.512
	50	0.047	0.709	0.047	0.712
	100	0.052	0.934	0.053	0.934
	150	0.049	0.983	0.049	0.983
	200	0.053	0.996	0.053	0.996
	250	0.050	0.996	0.050	0.996
	300	0.045	1.00	0.046	1.00
Highest	30	0.062	0.364	0.051	0.442
	50	0.060	0.557	0.048	0.626
	100	0.058	0.848	0.048	0.874
	150	0.055	0.958	0.047	0.960
	200	0.051	0.990	0.048	0.991
	250	0.048	0.998	0.047	0.998
	300	0.045	1.00	0.046	1.00
Smallest	30	0.058	0.493	0.047	0.490
	50	0.052	0.705	0.045	0.692
	100	0.054	0.926	0.050	0.923
	150	0.052	0.985	0.051	0.984
	200	0.048	0.996	0.048	0.996
	250	0.044	0.999	0.044	0.999
	300	0.045	1.00	0.046	1.00
Half highest/half smallest	30	0.047	0.498	0.047	0.504
	50	0.046	0.706	0.046	0.711
	100	0.045	0.931	0.047	0.932
	150	0.044	0.984	0.046	0.985
	200	0.046	0.996	0.048	0.996
	250	0.046	0.999	0.047	0.999
	300	0.045	1.00	0.046	1.00

TABLE A.2: MAR Type I error and MNAR power analysis for pattern mixture two-parameter model with different recovery design and sample sizes in 10000 replicates.

Design	n^*	Pattern Mixture	
		MAR Type I error	MNAR Power
Random Sample Selection	30	0.052	0.904
	50	0.051	0.989
	100	0.051	1.00
	150	0.052	1.00
	200	0.054	1.00
	250	0.047	1.00
	300	0.049	1.00
Highest x	30	0.048	0.797
	50	0.048	0.947
	100	0.046	0.999
	150	0.049	1.00
	200	0.047	1.00
	250	0.050	1.00
	300	0.049	1.00
Smallest x	30	0.047	0.914
	50	0.048	0.990
	100	0.048	1.00
	150	0.051	1.00
	200	0.049	1.00
	250	0.047	1.00
	300	0.049	1.00
Half Highest and Half smallest x	30	0.052	0.892
	50	0.049	0.988
	100	0.046	0.999
	150	0.048	1.00
	200	0.047	1.00
	250	0.045	1.00
	300	0.049	1.00

A.2 Tables relating to Chapter 5

The tables in this section shows the power or (and) MSE for the true optimal design, random design and different misspecified designs as discussed in Chapter 5.

TABLE A.3: Power for different designs in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
200	true optimal	0.115	0.148	0.176	0.192	0.214	0.220	0.229	0.248	0.254	0.266
	random	0.102	0.126	0.146	0.164	0.190	0.212	0.221	0.230	0.247	0.266
	Missing mechanism										
	(-2.2,0.4,-0.15)	0.110	0.140	0.167	0.186	0.206	0.214	0.229	0.232	0.253	0.266
	(-1.8,0.4,-0.15)	0.105	0.138	0.160	0.179	0.205	0.217	0.228	0.241	0.248	0.266
	(2,0.4,-0.15)	0.111	0.142	0.165	0.188	0.200	0.210	0.235	0.226	0.241	0.266
	(-2,0.44,-0.15)	0.103	0.145	0.160	0.188	0.207	0.213	0.223	0.241	0.249	0.266
	(-2,0.36,-0.15)	0.111	0.135	0.166	0.186	0.204	0.211	0.227	0.232	0.245	0.266
	(-2,-0.4,-0.15)	0.108	0.127	0.152	0.163	0.189	0.213	0.218	0.236	0.239	0.266
	(-2,0.4,-0.165)	0.105	0.145	0.168	0.188	0.211	0.212	0.225	0.238	0.251	0.266
	(-2,0.4,-0.135)	0.110	0.138	0.164	0.182	0.201	0.213	0.224	0.227	0.248	0.266
	(-2,0.4,0.15)	0.104	0.130	0.147	0.165	0.183	0.208	0.213	0.222	0.237	0.266
	Regression Coefficients										
	(2.2-2x)	0.106	0.137	0.163	0.190	0.201	0.219	0.230	0.231	0.249	0.266
	(1.8-2x)	0.112	0.136	0.176	0.187	0.202	0.208	0.221	0.231	0.250	0.266
	(-2-2x)	0.110	0.141	0.164	0.189	0.202	0.219	0.228	0.238	0.252	0.266
	(2-2.2x)	0.103	0.144	0.171	0.180	0.212	0.218	0.229	0.234	0.248	0.266
	(2-1.8x)	0.111	0.144	0.171	0.180	0.212	0.218	0.229	0.234	0.248	0.266
	(2+2x)	0.110	0.132	0.149	0.170	0.187	0.209	0.232	0.235	0.251	0.266
500	true optimal	0.174	0.269	0.339	0.408	0.429	0.458	0.495	0.506	0.526	0.537
	random	0.146	0.219	0.286	0.331	0.378	0.417	0.451	0.483	0.505	0.537
	Missing mechanism										
	(-2.2,0.4,-0.15)	0.170	0.268	0.339	0.393	0.420	0.463	0.477	0.495	0.510	0.537
	(-1.8,0.4,-0.15)	0.171	0.266	0.339	0.380	0.428	0.458	0.482	0.504	0.525	0.537
	(2,0.4,-0.15)	0.165	0.260	0.338	0.407	0.445	0.458	0.479	0.495	0.510	0.537
	(-2,0.44,-0.15)	0.166	0.267	0.316	0.379	0.425	0.446	0.482	0.492	0.512	0.537
	(-2,0.36,-0.15)	0.168	0.264	0.322	0.383	0.417	0.450	0.464	0.486	0.512	0.537
	(-2,-0.4,-0.15)	0.153	0.215	0.275	0.342	0.381	0.414	0.454	0.482	0.517	0.537
	(-2,0.4,-0.165)	0.161	0.267	0.329	0.379	0.426	0.437	0.472	0.484	0.516	0.537
	(-2,0.4,-0.135)	0.167	0.266	0.330	0.389	0.423	0.457	0.488	0.504	0.515	0.537
	(-2,0.4,0.15)	0.148	0.234	0.279	0.337	0.386	0.421	0.455	0.486	0.510	0.537
	Regression Coefficients										
	(2.2-2x)	0.172	0.250	0.327	0.382	0.418	0.450	0.475	0.500	0.518	0.537
	(1.8-2x)	0.171	0.260	0.339	0.383	0.422	0.454	0.474	0.485	0.507	0.537
	(-2-2x)	0.170	0.266	0.338	0.391	0.429	0.462	0.493	0.504	0.509	0.537
	(2-2.2x)	0.169	0.266	0.338	0.387	0.424	0.453	0.476	0.491	0.521	0.537
	(2-1.8x)	0.168	0.267	0.328	0.379	0.425	0.441	0.481	0.496	0.520	0.537
	(2+2x)	0.156	0.225	0.283	0.347	0.380	0.424	0.468	0.505	0.524	0.537
1000	true optimal	0.296	0.469	0.597	0.677	0.729	0.766	0.788	0.797	0.818	0.828
	random	0.243	0.395	0.487	0.574	0.636	0.690	0.741	0.777	0.802	0.828
	Missing mechanism										
	(-2.2,0.4,-0.15)	0.275	0.474	0.580	0.668	0.732	0.764	0.774	0.798	0.811	0.828
	(-1.8,0.4,-0.15)	0.278	0.460	0.587	0.677	0.727	0.770	0.785	0.794	0.810	0.828
	(2,0.4,-0.15)	0.290	0.379	0.592	0.653	0.717	0.751	0.774	0.789	0.806	0.828
	(-2,0.44,-0.15)	0.276	0.467	0.587	0.671	0.719	0.759	0.791	0.794	0.812	0.828
	(-2,0.36,-0.15)	0.288	0.464	0.606	0.674	0.735	0.765	0.783	0.791	0.809	0.828
	(-2,-0.4,-0.15)	0.241	0.376	0.488	0.574	0.646	0.697	0.731	0.774	0.800	0.828
	(-2,0.4,-0.165)	0.283	0.460	0.592	0.672	0.727	0.763	0.781	0.795	0.816	0.828
	(-2,0.4,-0.135)	0.281	0.469	0.584	0.672	0.725	0.763	0.786	0.794	0.806	0.828
	(-2,0.4,0.15)	0.246	0.395	0.499	0.595	0.659	0.725	0.758	0.782	0.802	0.828
	Regression Coefficients										
	(2.2-2x)	0.276	0.485	0.597	0.672	0.725	0.762	0.780	0.793	0.813	0.828
	(1.8-2x)	0.287	0.469	0.596	0.669	0.727	0.760	0.780	0.795	0.806	0.828
	(-2-2x)	0.291	0.469	0.580	0.587	0.708	0.741	0.778	0.786	0.810	0.828
	(2-2.2x)	0.293	0.460	0.587	0.636	0.714	0.749	0.777	0.797	0.810	0.828
	(2-1.8x)	0.287	0.465	0.596	0.669	0.713	0.725	0.765	0.792	0.813	0.828
	(2+2x)	0.253	0.392	0.501	0.593	0.656	0.706	0.758	0.781	0.814	0.828

TABLE A.4: MSE for different designs in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
200	true optimal	0.0643	0.0356	0.0267	0.0221	0.0202	0.0186	0.0171	0.0160	0.0146	0.0142
	random	0.0837	0.0457	0.0329	0.0265	0.0226	0.0198	0.0179	0.0163	0.0151	0.0142
	Missing mechanism										
	(-2.2,0.4,-0.15)	0.0657	0.0367	0.0269	0.0225	0.0197	0.0180	0.0172	0.0160	0.0150	0.0142
	(-1.8,0.4,-0.15)	0.0685	0.0356	0.0279	0.0233	0.0194	0.0178	0.0171	0.0161	0.0150	0.0142
	(2,0.4,-0.15)	0.0835	0.0357	0.0266	0.0224	0.0199	0.0182	0.0169	0.0161	0.0150	0.0142
	(-2,0.44,-0.15)	0.0660	0.0361	0.0268	0.0221	0.0196	0.0191	0.0166	0.0160	0.0151	0.0142
	(-2,0.36,-0.15)	0.0653	0.0362	0.0267	0.0222	0.0196	0.0186	0.0172	0.0160	0.0151	0.0142
	(-2,-0.4,-0.15)	0.0831	0.0455	0.0328	0.0263	0.0214	0.0191	0.0171	0.0162	0.0151	0.0142
	(-2,0.4,-0.165)	0.0648	0.0366	0.0267	0.0221	0.0195	0.0179	0.0171	0.0160	0.0151	0.0142
	(-2,0.4,-0.135)	0.0665	0.0365	0.0267	0.0235	0.0196	0.0179	0.0171	0.0160	0.0150	0.0142
	(-2,0.4,0.15)	0.0815	0.0456	0.0323	0.0265	0.0223	0.0189	0.0167	0.0162	0.0150	0.0142
	Regression Coefficients										
	(2.2-2x)	0.0651	0.0377	0.0267	0.0223	0.0196	0.0179	0.0166	0.0160	0.0150	0.0142
	(1.8-2x)	0.0677	0.0386	0.0267	0.0222	0.0195	0.0186	0.0171	0.0162	0.0150	0.0142
	(-2-2x)	0.0654	0.0377	0.0277	0.0229	0.0201	0.0183	0.0170	0.0159	0.0150	0.0142
	(2-2.2x)	0.0647	0.0361	0.0267	0.0234	0.0196	0.0179	0.0166	0.0160	0.0150	0.0142
	(2-1.8x)	0.0653	0.0361	0.0267	0.0234	0.0196	0.0179	0.0166	0.0160	0.0150	0.0142
	(2+2x)	0.0818	0.0436	0.0316	0.0253	0.0225	0.0193	0.0172	0.0160	0.0150	0.0142
500	true optimal	0.0241	0.0134	0.0102	0.0084	0.0076	0.0069	0.0067	0.0061	0.0058	0.0055
	random	0.0306	0.0172	0.0126	0.0102	0.0087	0.0077	0.0070	0.0064	0.0059	0.0055
	Missing mechanism										
	(-2.2,0.4,-0.15)	0.0248	0.0134	0.0102	0.0082	0.0074	0.0067	0.0065	0.0061	0.0058	0.0055
	(-1.8,0.4,-0.15)	0.0246	0.0134	0.0099	0.0085	0.0075	0.0069	0.0064	0.0061	0.0059	0.0055
	(2,0.4,-0.15)	0.0243	0.0136	0.0100	0.0082	0.0073	0.0068	0.0064	0.0062	0.0059	0.0055
	(-2,0.44,-0.15)	0.0242	0.0135	0.0108	0.0086	0.0076	0.0069	0.0065	0.0063	0.0059	0.0055
	(-2,0.36,-0.15)	0.0241	0.0134	0.0104	0.0086	0.0076	0.0070	0.0067	0.0063	0.0059	0.0055
	(-2,-0.4,-0.15)	0.0293	0.0192	0.0125	0.0098	0.0086	0.0077	0.0069	0.0064	0.0059	0.0055
	(-2,0.4,-0.165)	0.0241	0.0134	0.0102	0.0086	0.0076	0.0072	0.0067	0.0062	0.0059	0.0055
	(-2,0.4,-0.135)	0.0245	0.0134	0.0102	0.0086	0.0076	0.0070	0.0065	0.0062	0.0058	0.0055
	(-2,0.4,0.15)	0.0297	0.0158	0.0124	0.0099	0.0083	0.0075	0.0068	0.0063	0.0059	0.0055
	Regression Coefficients										
	(2.2-2x)	0.0242	0.0134	0.0102	0.0086	0.0076	0.0070	0.0064	0.0061	0.0059	0.0055
	(1.8-2x)	0.0243	0.0134	0.0100	0.0087	0.0077	0.0070	0.0067	0.0065	0.0062	0.0055
	(-2-2x)	0.0242	0.0135	0.0101	0.0084	0.0074	0.0068	0.0064	0.0061	0.0058	0.0055
	(2-2.2x)	0.0243	0.0134	0.0102	0.0086	0.0076	0.0069	0.0065	0.0062	0.0059	0.0055
	(2-1.8x)	0.0245	0.0134	0.0102	0.0086	0.0076	0.0070	0.0065	0.0061	0.0059	0.0055
	(2+2x)	0.0300	0.0166	0.0124	0.0097	0.0084	0.0074	0.0067	0.0063	0.0059	0.0055
1000	true optimal	0.0118	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0029	0.0028	0.0027
	random	0.0149	0.0085	0.0063	0.0051	0.0044	0.0039	0.0035	0.0032	0.0030	0.0027
	Missing mechanism										
	(-2.2,0.4,-0.15)	0.0124	0.0067	0.0049	0.0041	0.0036	0.0036	0.0031	0.0030	0.0029	0.0027
	(-1.8,0.4,-0.15)	0.0121	0.0066	0.0049	0.0040	0.0035	0.0032	0.0031	0.0030	0.0029	0.0027
	(2,0.4,-0.15)	0.0118	0.0084	0.0049	0.0042	0.0036	0.0033	0.0033	0.0030	0.0029	0.0027
	(-2,0.44,-0.15)	0.0119	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0030	0.0029	0.0027
	(-2,0.36,-0.15)	0.0119	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0030	0.0029	0.0027
	(-2,-0.4,-0.15)	0.0154	0.0086	0.0063	0.0051	0.0043	0.0038	0.0034	0.0031	0.0029	0.0027
	(-2,0.4,-0.165)	0.0119	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0030	0.0029	0.0027
	(-2,0.4,-0.135)	0.0118	0.0066	0.0048	0.0042	0.0036	0.0033	0.0031	0.0030	0.0029	0.0027
	(-2,0.4,0.15)	0.0141	0.0083	0.0061	0.0048	0.0041	0.0036	0.0033	0.0031	0.0030	0.0027
	Regression Coefficients										
	(2.2-2x)	0.0119	0.0066	0.0049	0.0040	0.0036	0.0033	0.0031	0.0029	0.0029	0.0027
	(1.8-2x)	0.0118	0.0065	0.0048	0.0040	0.0036	0.0033	0.0031	0.0030	0.0029	0.0027
	(-2-2x)	0.0118	0.0066	0.0050	0.0049	0.0037	0.0035	0.0031	0.0030	0.0029	0.0027
	(2-2.2x)	0.0120	0.0066	0.0049	0.0042	0.0037	0.0034	0.0032	0.0030	0.0029	0.0027
	(2-1.8x)	0.0118	0.0066	0.0048	0.0040	0.0036	0.0034	0.0032	0.0030	0.0029	0.0027
	(2+2x)	0.0143	0.0081	0.0059	0.0048	0.0041	0.0036	0.0033	0.0031	0.0029	0.0027

TABLE A.5: Power for extreme designs for $n = 1000$ in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Missing Mechanism	true optimal	0.296	0.469	0.597	0.677	0.729	0.766	0.788	0.797	0.818	0.828
	random	0.243	0.395	0.487	0.574	0.636	0.690	0.741	0.777	0.802	0.828
	(-2,0.4,0.15)	0.246	0.395	0.499	0.595	0.659	0.725	0.758	0.782	0.802	0.828
	(-2,0.4,0.3)	0.245	0.393	0.506	0.591	0.645	0.694	0.742	0.779	0.802	0.828
	(-2,0.4,0.6)	0.236	0.383	0.503	0.580	0.638	0.691	0.741	0.778	0.802	0.828
	(-2,0.4,0.9)	0.239	0.383	0.505	0.589	0.645	0.692	0.721	0.770	0.802	0.828
Regression Coefficient	(2+2x)	0.253	0.392	0.501	0.593	0.656	0.706	0.758	0.781	0.814	0.828
	(2+4x)	0.260	0.390	0.501	0.594	0.657	0.705	0.757	0.781	0.809	0.828
	(2+6x)	0.247	0.389	0.500	0.584	0.653	0.696	0.740	0.776	0.806	0.828
	(2+8x)	0.249	0.389	0.492	0.576	0.639	0.692	0.738	0.772	0.800	0.828

TABLE A.6: MSE for extreme designs for $n = 1000$ in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Missing Mechanism	true optimal	0.0118	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0029	0.0028	0.0027
	random	0.0149	0.0085	0.0063	0.0051	0.0044	0.0039	0.0035	0.0032	0.0030	0.0027
	(-2,0.4,0.15)	0.0141	0.0083	0.0061	0.0048	0.0041	0.0036	0.0033	0.0031	0.0030	0.0027
	(-2,0.4,0.3)	0.0146	0.0084	0.0060	0.0042	0.0040	0.0038	0.0034	0.0032	0.0029	0.0027
	(-2,0.4,0.6)	0.0152	0.0085	0.0060	0.0049	0.0043	0.0039	0.0034	0.0032	0.0029	0.0027
	(-2,0.4,0.9)	0.0151	0.0085	0.0061	0.0049	0.0042	0.0039	0.0036	0.0032	0.0029	0.0027
Regression Coefficient	(2+2x)	0.0143	0.0081	0.0059	0.0048	0.0041	0.0036	0.0033	0.0031	0.0029	0.0027
	(2+4x)	0.0141	0.0083	0.0061	0.0048	0.0042	0.0037	0.0034	0.0031	0.0029	0.0027
	(2+6x)	0.0144	0.0082	0.0061	0.0050	0.0043	0.0036	0.0034	0.0031	0.0029	0.0027
	(2+8x)	0.0147	0.0081	0.0062	0.0050	0.0043	0.0038	0.0035	0.0032	0.0029	0.0027

TABLE A.7: Power and MSE for different designs for $n = 1000$ in 10000 replicates

Sample size	Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Power	true optimal	0.296	0.469	0.597	0.677	0.729	0.766	0.788	0.797	0.818	0.828
	random	0.243	0.395	0.487	0.574	0.636	0.690	0.741	0.777	0.802	0.828
	Highest values	0.088	0.114	0.120	0.165	0.244	0.333	0.463	0.594	0.725	0.826
MSE	true optimal	0.0118	0.0066	0.0048	0.0040	0.0036	0.0033	0.0031	0.0029	0.0028	0.0027
	random	0.0149	0.0085	0.0063	0.0051	0.0044	0.0039	0.0035	0.0032	0.0030	0.0027
	Highest values	0.1021	0.0414	0.0243	0.0160	0.0114	0.0084	0.0063	0.0048	0.0037	0.0028

A.3 Tables relating to Chapter 6

This sections shows further results on the robustness of Algorithm 1 in comparison to the random design and misspecified designs as discussed in Subsection 6.3.1

TABLE A.8: Power for different designs in 10000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.369	0.648	0.763	0.814	0.844	0.852	0.857	0.859	0.867
Random	0.274	0.460	0.565	0.641	0.697	0.751	0.791	0.816	0.838
Missing mechanism									
(-2,0.4,-0.15)	0.362	0.620	0.755	0.812	0.838	0.840	0.849	0.854	0.858
(2,-0.4,-0.15)	0.223	0.329	0.463	0.528	0.651	0.653	0.719	0.785	0.801
(2,0.4,0.15)	0.318	0.537	0.651	0.739	0.767	0.821	0.837	0.844	0.858
(1,0.4,-0.15)	0.359	0.629	0.746	0.812	0.839	0.850	0.855	0.857	0.861
(3,0.4,-0.15)	0.354	0.630	0.755	0.801	0.843	0.849	0.854	0.858	0.859
(2,0.6,-0.15)	0.368	0.620	0.751	0.807	0.837	0.849	0.853	0.860	0.862
(2,0.2,-0.15)	0.357	0.588	0.729	0.808	0.829	0.847	0.854	0.859	0.863
(2,0.4,-0.3)	0.362	0.641	0.756	0.809	0.837	0.850	0.854	0.856	0.859
(2,0.4,-0.45)	0.369	0.647	0.755	0.809	0.839	0.852	0.853	0.854	0.863
Regression Coefficients									
(2+2x)	0.305	0.536	0.652	0.750	0.760	0.806	0.827	0.855	0.861
(-2-2x)	0.349	0.618	0.755	0.812	0.839	0.851	0.852	0.859	0.867
(4-2x)	0.356	0.618	0.754	0.812	0.836	0.851	0.854	0.858	0.862
(2-4x)	0.363	0.641	0.755	0.810	0.836	0.851	0.857	0.859	0.866

TABLE A.9: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.6986	0.7555	0.9725	0.8830	0.9498	0.9380	0.8325	0.8862	0.6052
(2,0.3,-0.15)	0.5461	0.7824	0.6292	0.7192	0.8431	0.8440	0.9061	0.8461	0.9045
(2,0.2,-0.15)	0.6624	0.4870	0.5826	0.7951	0.7477	0.7517	0.9167	0.9574	0.7712
(2,0.1,-0.15)	0.5876	0.5707	0.5596	0.6238	0.7024	0.8087	0.8266	0.9141	0.8503
(2,-0.1,-0.15)	0.3874	0.2361	0.2057	0.4101	0.3647	0.5485	0.5886	0.5037	0.5223
(2,-0.2,-0.15)	0.3097	0.1090	0.3190	0.1719	0.3229	0.1663	0.3585	0.5541	0.5261
(2,-0.3,-0.15)	0.2415	0.0040	0.0034	0.0349	0.0106	-0.0075	0.0309	0.1404	0.2361
(2,-0.4,-0.15)	-0.4787	-0.4249	-0.5279	-0.1021	-0.3001	-0.2096	-0.5542	-0.7082	-0.2814
(2,0.4,-0.25)	0.8520	0.9190	0.9284	0.9818	0.9424	0.9326	0.9603	0.7872	0.8020
(2,0.4,-0.35)	0.8623	0.9566	0.9574	0.9650	0.9554	0.9705	0.9113	0.7338	0.7551
(2,0.4,-0.45)	0.9274	0.9721	0.9431	0.9811	0.9541	0.9762	0.8346	0.7914	0.8037
(2,0.4,0.15)	0.2672	0.2834	0.2369	0.2361	0.2376	0.3637	0.4325	0.3780	0.7249
Regression Coefficients									
(1.777-1.971x)	0.5255	0.5665	0.7467	0.8499	0.7941	0.8584	0.8421	0.9253	0.9671

TABLE A.10: Power for different designs in 10000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.369	0.648	0.763	0.814	0.844	0.852	0.857	0.859	0.867
Random	0.274	0.460	0.565	0.641	0.697	0.751	0.791	0.816	0.838
Missing mechanism									
(2,0.3,-0.15)	0.348	0.614	0.733	0.801	0.830	0.850	0.856	0.860	0.865
(2,0.2,-0.15)	0.357	0.588	0.729	0.808	0.829	0.847	0.854	0.859	0.863
(2,0.1,-0.15)	0.351	0.588	0.718	0.793	0.824	0.848	0.853	0.860	0.862
(2,-0.1,-0.15)	0.331	0.529	0.650	0.763	0.795	0.841	0.849	0.851	0.861
(2,-0.2,-0.15)	0.334	0.483	0.674	0.710	0.783	0.788	0.841	0.854	0.861
(2,-0.3,-0.15)	0.317	0.461	0.565	0.656	0.695	0.748	0.790	0.833	0.848
(2,-0.4,-0.15)	0.223	0.329	0.463	0.528	0.651	0.653	0.719	0.785	0.801
(2,0.4,-0.25)	0.368	0.639	0.756	0.809	0.837	0.851	0.855	0.857	0.863
(2,0.4,-0.35)	0.369	0.641	0.754	0.807	0.840	0.849	0.853	0.856	0.860
(2,0.4,-0.45)	0.369	0.647	0.755	0.809	0.839	0.852	0.853	0.854	0.863
(2,0.4,0.15)	0.318	0.537	0.651	0.739	0.767	0.821	0.837	0.844	0.858
Regression Coefficients									
(1.777-1.971x)	0.345	0.588	0.751	0.813	0.831	0.849	0.854	0.858	0.867

TABLE A.11: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
True optimal	0.742	0.708	0.751	0.907	0.691	0.919	0.931	0.711	0.702
(-2,1.1,0.15)	0.745	0.888	0.960	0.962	0.993	0.949	0.956	0.888	0.978
(-2,0.9,0.15)	0.534	0.481	0.835	0.528	0.794	0.553	0.822	0.931	0.610
(-2,0.8,0.15)	0.708	0.527	0.708	0.963	0.485	0.895	0.648	0.528	0.959
(-2,0.7,0.15)	0.528	0.891	0.528	0.783	0.934	0.669	0.708	0.692	0.787
(-2,0.5,0.15)	0.746	0.707	0.777	0.502	0.704	0.708	0.800	0.639	0.820
(-2,0.3,0.15)	0.376	0.358	0.751	0.236	0.389	0.236	0.528	0.708	0.387
(-2,0.1,0.15)	0.528	0.638	0.524	0.874	0.528	0.515	0.485	0.403	0.684
(-2,-0.1,0.15)	-0.528	-0.491	-0.228	0.004	-0.528	-0.658	0.224	-0.480	-0.234
(-2,-0.3,0.15)	-0.158	-0.235	-0.267	-0.323	-0.223	-0.592	-0.555	-0.528	-0.229
(-2,-0.5,0.15)	-0.514	-0.538	-0.528	-0.617	-0.708	-0.528	-0.506	-0.560	-0.485
(-2,-0.7,0.15)	-0.607	-0.692	-0.708	-0.808	-0.962	-0.582	-0.820	-0.820	-0.836
(-2,-0.8,0.15)	-0.597	-0.658	-0.793	-0.669	-0.825	-0.688	-0.817	-0.905	-0.391
(-2,-0.9,0.15)	-0.602	-0.885	-0.582	-0.906	-0.948	-0.820	-0.894	-0.935	-0.848
(-2,-1.1,0.15)	-0.754	-0.466	-0.931	-0.639	-0.639	-0.708	-0.682	-0.889	-0.673
(-2,-1.3,0.15)	-0.885	-0.867	-0.778	-0.708	-0.523	-0.720	-0.226	-0.708	-0.847
(-2,1.3,-0.15)	0.371	0.942	0.617	0.686	0.584	0.925	0.519	0.775	0.931
(-2,1.3,0.1)	0.822	0.527	0.393	0.855	0.951	0.669	0.748	0.708	0.955
(-2,1.3,-0.1)	0.820	0.889	0.708	0.527	0.818	0.647	0.943	0.513	0.708
(-2,1.3,0.05)	0.846	0.543	0.820	0.979	0.708	0.775	0.906	0.708	0.886
(-2,1.3,-0.05)	0.498	0.564	0.587	0.531	0.927	0.770	0.973	0.905	0.618
(2,-1.3,-0.15)	-0.666	-0.708	-0.814	-0.889	-0.880	-0.683	-0.885	-0.917	-0.446
Regression Coefficients									
(2-x)	0.267	0.525	0.685	0.888	0.656	0.883	0.820	0.699	0.762
(-2-x)	0.586	0.744	0.866	0.860	0.849	0.528	0.840	0.708	0.808
(-2+x)	0.820	0.712	0.820	0.704	0.891	0.878	0.820	0.888	0.924

TABLE A.12: Power for different designs in 10000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
True optimal	0.358	0.582	0.718	0.818	0.869	0.915	0.930	0.948	0.955	0.963
Random	0.305	0.501	0.669	0.776	0.832	0.879	0.909	0.926	0.949	0.959
Missing mechanism										
(-2,1.1,0.15)	0.350	0.571	0.736	0.813	0.876	0.911	0.926	0.942	0.952	0.959
(-2,0.9,0.15)	0.337	0.568	0.728	0.793	0.869	0.890	0.924	0.941	0.953	0.955
(-2,0.8,0.15)	0.354	0.566	0.715	0.824	0.856	0.910	0.923	0.939	0.951	0.957
(-2,0.7,0.15)	0.348	0.590	0.705	0.821	0.870	0.901	0.923	0.945	0.955	0.954
(-2,0.5,0.15)	0.349	0.581	0.717	0.793	0.870	0.897	0.928	0.942	0.947	0.963
(-2,0.3,0.15)	0.352	0.554	0.721	0.786	0.860	0.882	0.917	0.939	0.948	0.961
(-2,0.1,0.15)	0.355	0.579	0.711	0.814	0.865	0.896	0.919	0.940	0.943	0.956
(-2,-0.1,0.15)	0.306	0.493	0.654	0.763	0.787	0.846	0.915	0.921	0.949	0.955
(-2,-0.3,0.15)	0.324	0.514	0.652	0.744	0.816	0.847	0.887	0.917	0.945	0.960
(-2,-0.5,0.15)	0.301	0.495	0.609	0.709	0.777	0.845	0.897	0.922	0.949	0.959
(-2,-0.7,0.15)	0.296	0.468	0.600	0.702	0.772	0.853	0.876	0.908	0.940	0.956
(-2,-0.8,0.15)	0.295	0.474	0.601	0.727	0.774	0.836	0.885	0.915	0.949	0.958
(-2,-0.9,0.15)	0.299	0.452	0.617	0.687	0.764	0.825	0.872	0.915	0.940	0.955
(-2,-1.1,0.15)	0.288	0.484	0.584	0.717	0.788	0.850	0.889	0.910	0.943	0.962
(-2,-1.3,0.15)	0.274	0.456	0.605	0.711	0.796	0.848	0.901	0.921	0.936	0.957
(-2,1.3,-0.15)	0.350	0.575	0.723	0.809	0.871	0.903	0.922	0.941	0.951	0.960
(-2,1.3,0.1)	0.346	0.568	0.700	0.815	0.864	0.893	0.927	0.939	0.952	0.958
(-2,1.3,-0.1)	0.359	0.558	0.719	0.797	0.874	0.900	0.926	0.943	0.953	0.956
(-2,1.3,0.05)	0.354	0.579	0.717	0.822	0.868	0.903	0.932	0.936	0.947	0.962
(-2,1.3,-0.05)	0.348	0.573	0.711	0.797	0.872	0.901	0.926	0.945	0.951	0.961
(2,-1.3,-0.15)	0.287	0.465	0.594	0.692	0.777	0.847	0.876	0.911	0.946	0.960
Regression Coefficients										
(2-x)	0.339	0.551	0.726	0.821	0.868	0.906	0.928	0.945	0.956	0.962
(-2-x)	0.351	0.573	0.731	0.807	0.875	0.898	0.927	0.940	0.950	0.962
(-2+x)	0.358	0.580	0.724	0.807	0.876	0.910	0.921	0.945	0.952	0.960

A.4 Additional examples for Chapter 7

In this section, further examples under the conjectured designs are provided.

In example (a) below, we introduced a quadratic term in the missing mechanism only to see if this affects the performance of the test. Tables A.13 and A.14 show the Type I error and the corresponding γ_1 values respectively. All designs have Type I error values that approximate 0.05.

(a) $n = 1000$, $(\beta_0, \beta_1) = (2, -2)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (3, 2)$, $(\alpha_0, \alpha_1, \alpha_2) = (2.9, -0.13, 0)$ with regression $\beta_0 + \beta_1 x$ and missing mechanism $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

TABLE A.13: Type I error for different designs in 2000 replicates

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
(2.9,-0.13,0)	0.052	0.051	0.050	0.050	0.051	0.053	0.057	0.053	0.052	0.048
Random	0.055	0.054	0.035	0.043	0.060	0.052	0.048	0.057	0.046	0.048
Conjecture	0.055	0.052	0.051	0.051	0.048	0.053	0.048	0.052	0.053	0.054

TABLE A.14: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(2.9,-0.13,0)	-0.229	-0.387	-0.185	-0.166	-0.201	-0.216	-0.745	0.046	0.011
Conjecture	-0.500	-0.480	-0.450	-0.420	-0.400	-0.350	-0.320	-0.300	-0.280

The example below shows the parameters for the MNAR mechanism with a quadratic term in the regression model, the power for this example is shown in Table 7.16 with the corresponding γ_1 values in Table 7.17. Figure 7.14 shows that at a smaller recovery proportion $c = 0.1, ..., 0.4$ the optimal design has significant power compared to the conjectured design, as c increases, the conjectured design has power values close to the optimal design. At all values of c, the optimal and conjectured design outperforms the random design.

(b) $n = 1000$, $(\beta_0, \beta_1) = (2, -2)$, $\sigma_y^2 = 1$, $(\mu_x, \sigma_x^2) = (3, 2)$, $(\alpha_0, \alpha_1, \alpha_2) = (2.9, -0.13, 0.3)$ with regression $\beta_0 + \beta_1 x$ and missing mechanism $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

TABLE A.15: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(2.9,-0.13,0.3)	-0.955	-0.867	-0.884	-0.773	-0.802	-0.905	-0.715	-0.495	-0.502
Conjecture	-0.180	-0.230	-0.280	-0.300	-0.330	-0.350	-0.370	-0.400	-0.450

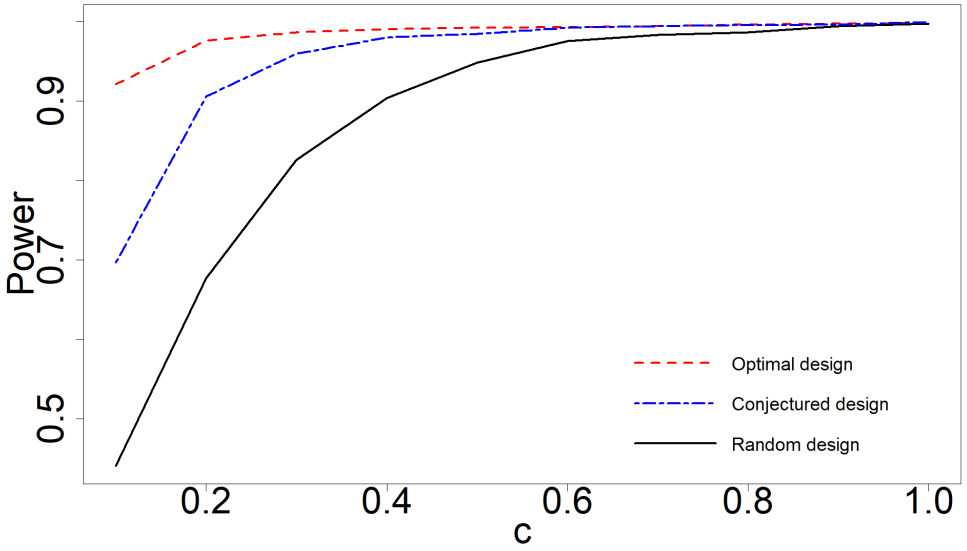

FIGURE A.1: Power plot using different designs.

Example (c) considers the power for the three different designs as shown in Table 7.18 and Figure 7.15. The optimal design outperforms the other designs and the conjectured design has better power than the random design. The conjectured design performs almost similar to the optimal design at $c = 0.7$ and above. The optimal design and conjectured design γ_1 values are shown in Table 7.19.

(c) $n = 1000$, $(\beta_0, \beta_1) = (0.5, -1.2)$, $\sigma_y^2 = 2$, $(\mu_x, \sigma_x^2) = (2, 3)$, $(\alpha_0, \alpha_1, \alpha_2) = (4, -0.45, -0.18)$ with regression $\beta_0 + \beta_1 x^2$ and missing mechanism $\alpha_0 + \alpha_1 x^2 + \alpha_2 y$.

TABLE A.16: γ_1 values

Design	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(4,-0.45,-0.18)	-0.204	-0.796	-0.889	-0.879	-0.949	-0.898	-0.889	-0.910	-0.936
Conjecture	-0.120	-0.180	-0.200	-0.230	-0.250	-0.270	-0.300	-0.330	-0.350

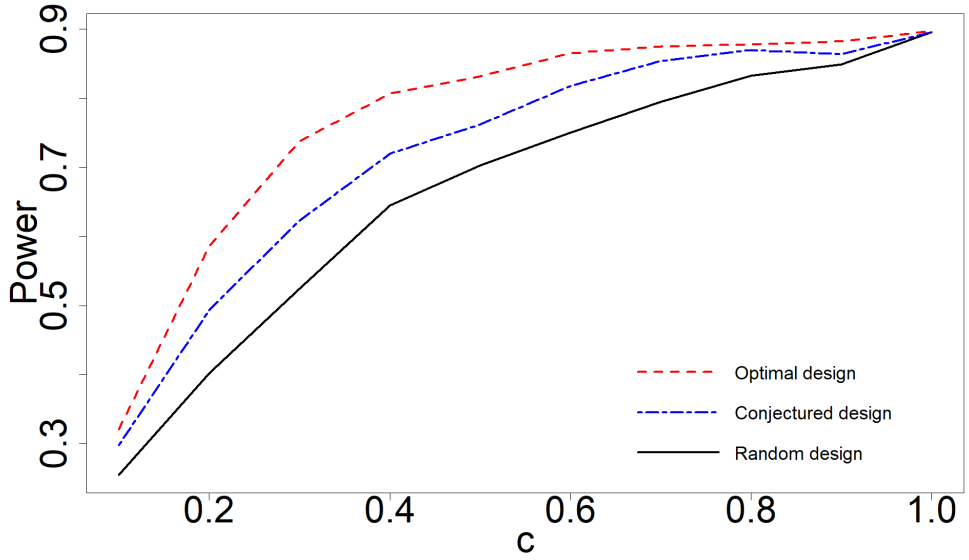

FIGURE A.2: Power plot using different designs.

In the example below with three covariates. The γ_i values are shown in Table 7.21 for the conjectured design. The power for both random and conjectured designs is shown in Table 7.20 with graphical representation in Figure 7.16. The conjectured design has better power than the random design. For both designs, the power increases as c increases.

(d) Generate 1000 points following a multiple linear regression model in 10000 replicates:

$$Y|(X_1 = x_1, X_2 = x_2, X_3 = x_3) \sim N(2x_1 - 0.8x_2 + 2x_3, 4)\,,$$

with $X_1 \sim N(2,1)$,$X_2 \sim N(1,4)$ and $X_3 \sim N(2,1)$. Introduce MNAR missingness into Y using:

$$P(M = 1|Y = y, X = x) = \frac{\exp(-2.8 - 0.25x_1 + 0.18x_2 + 1.4x_3 + 0.1y)}{1 + \exp(-2.8 - 0.25x_1 + 0.18x_2 + 1.4x_3 + 0.1y)}.$$

Using the random and conjectured designs, obtain the power of test.

TABLE A.17: γ values

γ_i	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
γ_1	0.05	0.07	0.10	0.15	0.18	0.20	0.23	0.26	0.28
γ_2	0.05	0.12	0.16	0.20	0.23	0.26	0.30	0.33	0.35
γ_3	0.23	0.26	0.30	0.34	0.37	0.39	0.60	0.65	0.78

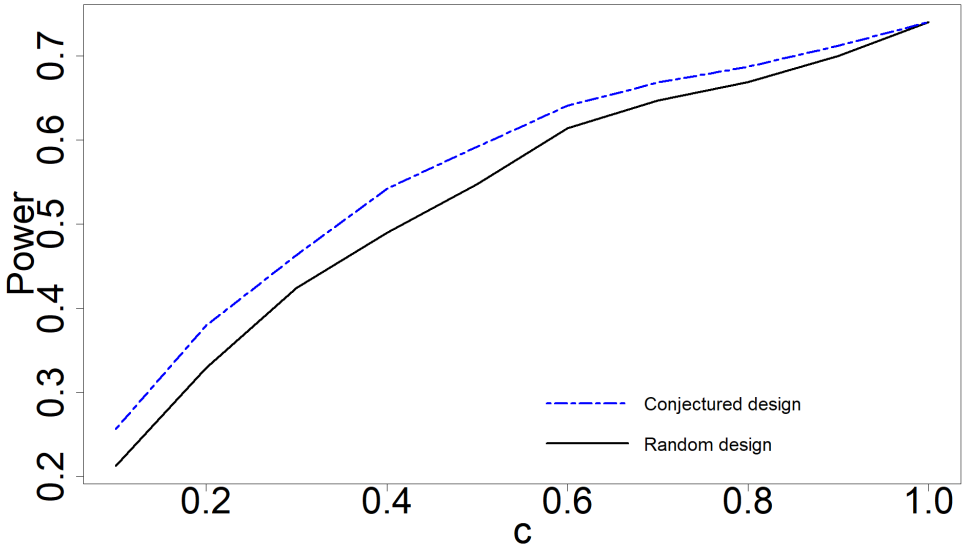

FIGURE A.3: Power plot using different designs.

A.5 Tables relating to Chapter 8

This section shows Tables relating to Chapter 8. Tables A.18 and A.19 show the monte carlo power, biases, variance and mean squared of using estimator $\hat{\theta}_{TE}$. Table A.20 shows the monte carlo power, biases, variance and mean squared of using estimators: $\hat{\theta}_{TE}$, $\hat{\theta}_{CE}$ and $\hat{\theta}_{SE}$. The monte Carlo power, biases, variances, and mean squared errors for different sample sizes using estimators: $\hat{\theta}_{TE}$, $\hat{\theta}_{CE}$ and $\hat{\theta}_{SE}$ are shown in Table A.21.

TABLE A.18: Monte Carlo power, biases, variances and mean squared errors for different combinations of MAR and MNAR using $\hat{\theta}_{TE}$.

c	Estimates	100% MNAR	10% MAR/90% MNAR	25% MAR/75% MNAR	50% MAR/50% MNAR
0.1	P	0.25030	0.22260	0.19450	0.14470
	Bias	0.05894	0.05756	0.05075	0.04037
	Var	0.01296	0.01222	0.01137	0.00983
	MSE	0.01643	0.01553	0.01395	0.01146
0.2	P	0.40770	0.36870	0.31060	0.22990
	Bias	0.03848	0.03734	0.03453	0.02892
	Var	0.00899	0.00861	0.00824	0.00725
	MSE	0.01047	0.01000	0.00943	0.00809
0.3	P	0.53740	0.48640	0.42110	0.29690
	Bias	0.02592	0.02454	0.02227	0.02004
	Var	0.00668	0.00673	0.00623	0.00560
	MSE	0.00735	0.00733	0.00672	0.00600
0.4	P	0.63360	0.56890	0.48550	0.34170
	Bias	0.01612	0.01693	0.01596	0.01447
	Var	0.00530	0.00518	0.00497	0.00471
	MSE	0.00556	0.00546	0.00523	0.00492
0.5	P	0.70310	0.64210	0.54710	0.37930
	Bias	0.01077	0.01039	0.01031	0.01079
	Var	0.00444	0.00440	0.00431	0.00405
	MSE	0.00456	0.00451	0.00442	0.00417
0.6	P	0.76770	0.68810	0.58340	0.40490
	Bias	0.00619	0.00699	0.00720	0.00782
	Var	0.00390	0.00384	0.00384	0.00378
	MSE	0.00394	0.00389	0.00390	0.00384
0.7	P	0.81250	0.73090	0.61110	0.42790
	Bias	0.00397	0.00440	0.00483	0.00543
	Var	0.00341	0.00348	0.00346	0.00336
	MSE	0.00343	0.00350	0.00348	0.00339
0.8	P	0.83790	0.76130	0.63910	0.44240
	Bias	0.00206	0.00238	0.00270	0.00337
	Var	0.00318	0.00327	0.00320	0.00326
	MSE	0.00318	0.00327	0.00321	0.00327
0.9	P	0.86130	0.77670	0.66130	0.45070
	Bias	0.00102	0.00115	0.00141	0.00166
	Var	0.00313	0.00308	0.00310	0.00311
	MSE	0.00313	0.00308	0.00311	0.00311
1.0	P	0.88120	0.80610	0.67380	0.46490
	Bias	-1.16e^{-7}	-7.22e^{-8}	-9.61e^{-8}	-1.76e^{-8}
	Var	0.00296	0.00299	0.00302	0.00299
	MSE	0.00296	0.00299	0.00302	0.00299

TABLE A.19: Monte Carlo power, biases, variances and mean squared errors for different combinations of MAR and MNAR using $\hat{\theta}_{TE}$.

c	Estimates	75% MAR/25% MNAR	90% MAR/10% MNAR	100% MAR
0.1	P	0.10220	0.06720	0.05220
	Bias	0.02926	0.02423	0.01981
	Var	0.00814	0.00672	0.00606
	MSE	0.00900	0.00730	0.00646
0.2	P	0.14130	0.09150	0.05400
	Bias	0.02266	0.01915	0.01774
	Var	0.00600	0.00544	0.00494
	MSE	0.00652	0.00581	0.00525
0.3	P	0.17290	0.10450	0.04900
	Bias	0.01858	0.01645	0.01559
	Var	0.00504	0.00449	0.00428
	MSE	0.00538	0.00476	0.00453
0.4	P	0.19360	0.11130	0.04800
	Bias	0.01412	0.01378	0.01355
	Var	0.00448	0.00417	0.00381
	MSE	0.00468	0.00436	0.00399
0.5	P	0.21780	0.12150	0.05250
	Bias	0.01115	0.01073	0.01065
	Var	0.00389	0.00377	0.00372
	MSE	0.00402	0.00388	0.00384
0.6	P	0.23030	0.12460	0.04680
	Bias	0.00845	0.00865	0.00924
	Var	0.00354	0.00354	0.00339
	MSE	0.00361	0.00362	0.00347
0.7	P	0.23980	0.13060	0.05090
	Bias	0.00602	0.00618	0.00657
	Var	0.00332	0.00336	0.00334
	MSE	0.00336	0.00340	0.00338
0.8	P	0.24390	0.13190	0.04700
	Bias	0.00390	0.00431	0.00460
	Var	0.00320	0.00314	0.00319
	MSE	0.00321	0.00316	0.00321
0.9	P	0.25940	0.13510	0.05390
	Bias	0.00200	0.00212	0.00211
	Var	0.00305	0.00303	0.00309
	MSE	0.00306	0.00303	0.00309
1.0	P	0.25990	0.13830	0.05150
	Bias	$-3.48e^{-8}$	$-1.21e^{-8}$	$-6.77e^{-10}$
	Var	0.00301	0.00292	0.00299
	MSE	0.00301	0.00292	0.00299

TABLE A.20: Monte Carlo power, biases, variances and mean squared errors for 90% MAR and 10% MNAR combination.

c	Estimates	$\hat{\theta}_{TE}$	$\hat{\theta}_{CE}$	$\hat{\theta}_{SE}$
0.1	Bias	0.0182	0.0251	0.0009
	Var	0.0037	0.0052	0.0053
	MSE	0.0041	0.0058	0.0053
0.2	Bias	0.0131	0.0217	-0.0011
	Var	0.0025	0.0025	0.0030
	MSE	0.0027	0.0030	0.0030
0.3	Bias	0.0182	0.0201	-0.0011
	Var	0.0021	0.0025	0.0023
	MSE	0.0022	0.0029	0.0023
0.4	Bias	0.0094	0.0162	0.0006
	Var	0.0019	0.0022	0.0020
	MSE	0.0020	0.0025	0.0020
0.5	Bias	0.0088	0.0141	0.0009
	Var	0.0017	0.0019	0.0018
	MSE	0.0018	0.0021	0.0018
0.6	Bias	0.0056	0.0118	-0.0007
	Var	0.0016	0.0019	0.0017
	MSE	0.0017	0.0020	0.0017
0.7	Bias	0.0037	0.0079	-0.0004
	Var	0.0016	0.0017	0.0016
	MSE	0.0016	0.0018	0.0016
0.8	Bias	0.0033	0.0057	0.0004
	Var	0.0016	0.0016	0.0016
	MSE	0.0016	0.0016	0.0016
0.9	Bias	0.0013	0.0027	-0.0002
	Var	0.0015	0.0015	0.0015
	MSE	0.0015	0.0015	0.0015
1.0	Bias	$-2.6364e^{-8}$	$-1.0614e^{-16}$	$-6.5010e^{e-8}$
	Var	0.0016	0.0016	0.0016
	MSE	0.0016	0.0016	0.0016

TABLE A.21: Monte Carlo power, biases, variances, and mean squared errors for different sample sizes using different estimators.

n	Estimates	$\hat{\theta}_{TE}$	$\hat{\theta}_{CE}$	$\hat{\theta}_{SE}$
500	Bias	0.02930	0.04340	0.00420
	Var	0.00650	0.00600	0.00900
	MSE	0.00730	0.00790	0.00900
1000	Bias	0.01550	0.02430	-0.00080
	Var	0.00320	0.00310	0.00410
	MSE	0.00350	0.00370	0.00410
2000	Bias	0.01110	0.02000	-0.00030
	Var	0.00150	0.00180	0.00190
	MSE	0.00160	0.00220	0.00190
3000	Bias	0.00980	0.01970	-0.00040
	Var	0.00110	0.00160	0.00140
	MSE	0.00120	0.00200	0.00140
4000	Bias	0.00630	0.01940	-0.00110
	Var	0.00070	0.00150	0.00090
	MSE	0.00070	0.00190	0.00090
5000	Bias	0.00650	0.01880	-0.00050
	Var	0.00050	0.00140	0.00070
	MSE	0.00050	0.00180	0.00070
6000	Bias	0.00490	0.01830	-0.00110
	Var	0.00047	0.00130	0.00060
	MSE	0.00049	0.00160	0.00060
7000	Bias	0.00480	0.01760	-0.00050
	Var	0.00046	0.00120	0.00054
	MSE	0.00048	0.00150	0.00054
8000	Bias	0.00380	0.01670	-0.00090
	Var	0.00039	0.00120	0.00050
	MSE	0.00041	0.00145	0.00050
9000	Bias	0.00480	0.01600	-0.00030
	Var	0.00034	0.00110	0.00048
	MSE	0.00036	0.00140	0.00048
10000	Bias	0.00450	0.01500	-0.00060
	Var	0.00030	0.00100	0.00040
	MSE	0.00030	0.00120	0.00040

References

F. B. Adebola and A. O. Adepetun. A new tripartite randomized response technique. *Journal of the Nigerian Association of Mathematical Physics*, 19, 2011.

F. B. Adebola, A. A. Adediran, and O. S. Ewemooje. Hybrid tripartite randomized response technique. *Communications in Statistics-Theory and Methods*, 46(23):11756–11763, 2017.

A. A. Adediran, F. B. Adebola, and O. S. Ewemooje. Unbiased estimator modeling in unrelated dichotomous randomized response. *Statistics in Transition New Series*, 21 (5), 2020.

V. O. Ajayi. Primary sources of data and secondary sources of data. *Benue State University*, 1(1):1–6, 2017.

J. M. Alho. Adjusting for nonresponse bias using logistic regression. *Biometrika*, 77(3): 617–624, 1990.

M. W. An, C. E. Frangakis, B. S. Musick, and C. T. Yiannoutsos. The need for double-sampling designs in survival studies: an application to monitor pepfar. *Biometrics*, 65(1):301–306, 2009.

R. R. Andridge and R. J. Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.

P. M. Aronow, A. S. Gerber, D. P. Green, H. Kern, and M. J. LaCour. Double sampling for nonignorable missing outcome data in randomized experiments, 2015.

A. Atkinson and V. Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975.

A. Atkinson, A. Donev, and R. Tobias. *Optimum experimental designs, with SAS*, volume 34. Oxford University Press, 2007.

I. Baek, W. Zhu, X. Wu, and W. K. Wong. Bayesian optimal designs for a quantal dose-response study with potentially missing observations. *Journal of Biopharmaceutical Statistics*, 16(5):679–693, 2006.

A. N. Baraldi and C. K. Enders. An introduction to modern missing data analyses. *Journal of school psychology*, 48(1):5–37, 2010.

K. Bhaskaran and L. Smeeth. What is the difference between missing completely at random and missing at random? *International journal of epidemiology*, 43(4):1336–1339, 2014.

A. Briggs, T. Clark, J. Wolstenholme, and P. Clarke. Missing.... presumed at random: cost-analysis of incomplete data. *Health economics*, 12(5):377–392, 2003.

B. W. Brown. Sample size requirements in full information maximum likelihood estimation. *International Economic Review*, pages 443–459, 1981.

J. Carpenter and M. Kenward. *Multiple Imputation and its Application*. John Wiley & Sons, 2012.

J. R. Carpenter and M. G. Kenward. Missing data in randomised controlled trials: a practical guide, 2007.

G. Casella and R. Berger. *Statistical inference*. CRC Press, 2024.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

P. E. Cheng. Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):81–87, 1994.

H. Chernoff. Locally optimal designs for estimating parameters. *The Annals of Statistics*, 24(4):586–602, 1953.

F. Cobben. *Nonresponse in sample surveys: methods for analysis and adjustment*. Statistics Netherlands The Hague, 2009.

K. Cooper and K. Stewart. Does household income affect children's outcomes? a systematic review of the evidence. *Child Indicators Research*, 14(3):981–1005, 2021.

A. Coppock, A. S Gerber, Donald P Green, and Holger L Kern. Combining double sampling and bounds to address nonignorable missing outcomes in randomized experiments. *Political Analysis*, 25(2):188–206, 2017.

G. Cordeiro. Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(3):404–413, 1983.

M. J. Daniels, D. Jackson, W. Feng, and I. R. White. Pattern mixture models for the analysis of repeated attempt designs. *Biometrics*, 71(4):1160–1167, 2015.

A. P. de Leon and A. C. Atkinson. The design of experiments to discriminate between two rival generalized linear models. In *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling, Munich, 13–17 July 1992*, pages 159–164. Springer, 1992.

A. Dean and D. Voss. *Design and analysis of experiments*. Springer, 1999.

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 (Methodological)*, 39(1):1–22, 1977.

Y. Dong and C. Y. Peng. Principled missing data methods for researchers. *SpringerPlus*, 2(1):1–17, 2013.

J. H. Drew and W. A. Fuller. Modeling nonresponse in surveys with callbacks. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pages 639–642, 1980.

M. R. Elliott, R. J. Little, and S. Lewitzky. Subsampling callbacks to improve survey efficiency. *Journal of the American Statistical Association*, 95(451):730–738, 2000.

C.K. Enders. *Applied missing data analysis*. Guilford Publications, 2022.

O. S. Ewemooje, F. B. Adebola, and A. A. Adediran. A stratified hybrid tripartite randomized response technique. *Gazi University Journal of Science*, 31(4):1246–1266, 2018.

O. A. Fasoranbaku and G. O. Daramola. Towards a better estimation of the parameters of linear regression models: The optimal designed experiment approach. *International Journal of New Technology and Research*, 4(5):263062, 2018.

S. Fielding, P. M. Fayers, and C. R. Ramsay. Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes*, 7(1):1–10, 2009.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

S. Ghosh. On robustness of designs against incomplete data. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 204–208, 1979.

P. Goos, B. Jones, and U. Syafitri. I-optimal design of mixture experiments. *Journal of the American Statistical Association*, 111(514):899–911, 2016.

B. G. Greenberg, A. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz. The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64(326):520–539, 1969.

Z. Guan, D. H. Leung, and J. Qin. Semiparametric maximum likelihood inference for nonignorable nonresponse with callbacks. *Scandinavian Journal of Statistics*, 45(4):962–984, 2018.

S. Hammon, A.and Zinn. Multiple imputation of binary multilevel missing not at random data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3): 547–564, 2020.

L.V. Hedges and H. Cooper. Research synthesis as a scientific process. *The handbook of research synthesis and meta-analysis*, 1, 2009.

D. F. Heitjan and S. Basu. Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3):207–213, 1996.

L. A. Imhof, D. Song, and W. K. Wong. Optimal design of experiments with possibly failing trials. *Statistica Sinica*, pages 1145–1155, 2002.

D. Jackson, I. R. White, and M. Leese. How much can we learn about missing data?: an exploration of a clinical trial in psychiatry. *Journal of the Royal Statistical Society, Series A*, 173(3):593–612, 2010.

M. Jamshidian and M. Mata. Advances in analysis of mean and covariance structure when data are incomplete. In *Handbook of latent variable and related models*, pages 21–44. Elsevier, 2007.

G. Kalton and L. Kish. Two efficient random imputation procedures. In *Proceedings of the survey research methods section*, pages 146–151. American Statistical Association, 1981.

H. Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402, 2013.

M. G. Kenward and G. Molenberghs. Likelihood based frequentist inference when data are missing at random. *Statistical Science*, pages 236–247, 1998.

J. K. Kim and C. L. Yu. A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106(493):157–165, 2011.

J. M. Kim and W. D. Warde. A stratified warner's randomized response model. *Journal of Statistical Planning and Inference*, 120(1-2):155–165, 2004.

K. M. Lang and T. D. Little. Principled missing data treatments. *Prevention Science*, 19 (3):284–294, 2018.

K. M. Lee, S. Biedermann, and R. Mitra. Optimal design for experiments with possibly incomplete observations. *Statistica Sinica*, 28(3):1611–1632, 2018a.

K. M. Lee, R. Mitra, and S. Biedermann. Optimal design when outcome values are not missing at random. *Statistica Sinica*, 28(4):1821–1838, 2018b.

K. M. Lee, S. Biedermann, and R. Mitra. D-optimal designs for multiarm trials with dropouts. *Statistics in Medicine*, 38(15):2749–2766, 2019.

B. Leurent, M. Gomes, R. Faria, S. Morris, R. Grieve, and J. R. Carpenter. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. *Pharmacoeconomics*, 36(8):889–901, 2018.

R. Little and D. Rubin. Statistical analysis with missing data. *Statistical Analysis with Missing Data*, pages 200–220, 2002.

R. J. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202, 1988.

R. J. Little and N. Schenker. Missing data. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 39–75. Springer, 1995.

R.J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

J. B. Lu, G.and Copas. Missing at random, likelihood ignorability and model completeness. *Annals of Statistics*, pages 754–765, 2004.

N. S. Mangat. An improved randomized response strategy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):93–95, 1994.

D. A. Marker, D. R. Judkins, and M. Winglee. Large-scale imputation for complex surveys. *Survey nonresponse*, 329341, 2002.

S. McPherson, C. Barbosa-Leiker, M. R. Mamey, M. McDonell, C. K. Enders, and J. Roll. A 'missing not at random'(mnar) and 'missing at random'(mar) growth model comparison with a buprenorphine/naloxone clinical trial. *Addiction*, 110(1):51–58, 2015.

W. Miao, X. Li, and B. Sun. A stableness of resistance model for nonresponse adjustment with callback data. *arXiv preprint arXiv:2112.02822*, 2021.

R. Mitra and J. Reiter. Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine*, 30(6):627–641, 2011.

R. Mitra and J. Reiter. A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical methods in medical research*, 25(1):188–204, 2016.

D. C. Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017.

G. W. Oehlert. *A first course in design and analysis of experiments*. 2010.

A. V. Oladugba and M. S. Madukaife. D-optimality and dl-optimality criteria for incomplete block designs. *Global Journal of Mathematical Sciences*, 8(2), 2009.

J. Qin and D. A. Follmann. Semiparametric maximum likelihood inference by using failed contact attempts to adjust for nonignorable nonresponse. *Biometrika*, 101(4): 985–991, 2014.

E. A. Rady, M. M. El-Monsef, and M. M. Seyam. Relationships among several optimality criteria. *Interstat*, 15(6):1–11, 2009.

D. B. Rubin. Missing at random-what does it mean? *ETS Research Bulletin Series*, 1973 (1):i–9, 1973.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association, 1978.

D. B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley Classics Library. John Wiley & Sons, 2004.

S. Self, R. Mauritsen, and J. Ohara. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, pages 31–39, 1992.

L. O. Silva and L. E. Zárate. A brief review of the main approaches for treatment of missing data. *Intelligent Data Analysis*, 18(6):1177–1198, 2014.

M. Soley-Bori. Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University*, 23:20, 2013.

J. A. Sterne, Carlin J. B. Spratt M. Royston P. Kenward M. G. Wood A. M. White, I. R., and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

E. J. Tchetgen Tchetgen and K. E. Wirth. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4):1123–1131, 2017.

C. Tommasi and J. López-Fidalgo. Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics & Data Analysis*, 54(1): 143–150, 2010.

B. A. Walther and J. L. Moore. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, 28(6):815–829, 2005.

S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

T. H. Waterhouse, D. C. Woods, J. A. Eccleston, and S. M. Lewis. Design selection criteria for discrimination/estimation for nested models and a binomial response. *Journal of statistical planning and inference*, 138(1):132–144, 2008.

T. Yan and R. Curtin. The relation between unit nonresponse and item nonresponse: A response continuum perspective. *International Journal of Public Opinion Research*, 22 (4):535–551, 2010.