**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

# Optimal designs in the presence of missing responses

by

**Kim May Lee**

Thesis submitted for the degree of Doctor of Philosophy

May 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Mathematical Sciences

Doctor of Philosophy

# OPTIMAL DESIGNS IN THE PRESENCE OF MISSING RESPONSES

by Kim May Lee

Design of experiments is an approach that could minimise the costs of conducting experiments by maximising the information that could be obtained from the study prior to the implementation. It is crucial that the experimenters consider the impact of the presence of missing observations on the statistical power of the study and the precision of the inferences. This research incorporates the features of some missing data analysis approaches into the experimental design framework for finding a design that is robust to the presence of missing responses.

We propose optimal design framework for the linear regression models and the linear mixed models respectively. Assuming that missing responses are generated by a monotone missing at random (MAR) mechanism, we consider the features of complete case analysis and a multiple imputation respectively in the design framework for the linear regression models, and of available case analysis in the cohort design framework for the linear mixed models.

The optimal design framework for the linear regression models with complete case analysis is a generalisation to the work that is proposed by Imhof et al. (2002). Besides that, we believe we are the first who consider a multiple imputation approach at the design stage of an experiment. Moreover, having accounted for the presence of dropouts, we introduce two types of design regime in the cohort design framework for the linear mixed models.

Throughout this project we show that using the optimal designs that assume completely observed responses or the naive designs may not be the best option. There are statistical gains in accounting for the features of missing data analysis approaches at the design stage of an experiment, especially for the study that involves a small sample size and high costs. This novel research provides a new tool to tackle the presence of missing data in future experimental studies.

# Contents

# List of Figures

# List of Tables

## Declaration of Authorship

I, Kim May Lee, declare that the thesis entitled *Optimal designs in the presence of missing responses* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- 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;
- where I have consulted the published work of others, this is always clearly attributed;
- 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;
- I have acknowledged all main sources of help;
- 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;
- none of this work has been published before submission.

Signed:.....

Date:.....

# Acknowledgements

First of all, I would like to thank my supervisors, Dr. Stefanie Biedermann and Dr. Robin Mitra for their patience, guidance and continuous support throughout. The knowledge and research skills that I have gained from them are invaluable. This work was supported by the studentship funding from the Institute for Life Sciences at the University of Southampton.

I would also like to thank Professor Clive Holmes for his comments on designing clinical trials in practice, Robert Howard and Patrick Philips for providing the data from the Domino study RCTN49545035 which was funded by the MRC and Alzheimer's Society UK.

Finally, I would like to thank my family and the church members for their unconditional love and support.

# Chapter 1

# Introduction

The motivation for this project comes from clinical studies of Alzheirmer's disease. To slow down the deterioration of the disease, clinicians are often interested in the efficacy of new interventions. For example, Howard et al. (2012) study the benefits of administering the treatments donepezil, memantine, and the combination of the two, to those patients who suffer from moderate to severe Alzheimer's disease. In this study, 291 patients were being followed up for the duration of 52 weeks, whereby measurements on each patient were taken at week 6, 18, 30 and 52 respectively. This type of data is often called longitudinal data or repeated measurements in the literature. In practice, the presence of missing values is often unavoidable in longitudinal data especially when the duration of the study is long. Moreover, since the clinical studies involve human beings who are relatively old and unhealthy, it is particularly difficult to ensure that patients attend follow-up sessions. Hence, we aim to tackle the presence of missing data at the design stage of an experiment, by merging the notions of two major areas of statistical research, namely design of experiments and missing data analysis.

In statistics, design of experiments is an approach that maximises the information which could be obtained from a study. For a given goal, this approach provides experimental design settings for collecting observations. Some examples of design settings are the levels of factors, the number of groups of experimental units, and the time points of measuring observations. By designing an experiment from a statistical point of view, sufficient information for making inferences might be obtained with low experimental costs. The investigation in this research area has been focusing on the design framework for certain statistical models, algorithms for finding an optimal design, and robustness measures of designs when some underlying assumptions, such as the assumption of completely observed responses, do not hold. In general, the experimenters always assume that the same statistical method will be implemented in the data analysis.

In the research area of missing data analysis, many methods have been developed based on some assumptions about the process that generates missing values. By treating

the missing values as some realisations of a random process, Rubin (1976) has defined three types of missing data mechanisms, namely missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The differences between these mechanisms are in the conditions of the data governing parameter of the random process. In the literature, most of the proposed techniques tackle the presence of missing values by making assumptions for the missing data mechanism, and can be classified into one of the following three categories:

- analysis based on available information (e.g. complete case analysis, available case analysis);
- analyse the "complete-data set" after the missing values have been replaced by some artificial values (e.g. mean imputation, multiple imputation); and
- model the missing data mechanism at the same time of analysing the data (e.g. full information maximum likelihood).

At the moment, the traditional way to account for the impact of missing values at the design stage of an experiment is to scale up the total sample size of the study, regardless of the missing data mechanism. The features of the methodologies that tackle the presence of missing values are often ignored by the experimenters. The aim of our work is to fill this gap by considering the features of missing data analysis in the optimal design framework. We investigate the optimal design framework for the linear regression model and the linear mixed model respectively in the presence of missing responses. The following are some questions that we aim to address in our investigations, in the context of a clinical study:

- Provided that the missing responses are generated by a MAR mechanism that depends on the dose level, what is a suitable dose level for the treatment group?
- Given a MAR mechanism, how many subjects are required in each treatment group such that sufficient information is obtained from the experiment for making inferences?
- If time is one of the factors that causes a patient to be missing at random, when should the interim sessions be for collecting measurements on the different treatment groups?
- For the same statistical model, are optimal designs different if different missing data approaches are considered at the design stage of an experiment?

In statistical analysis, the linear regression model is one of the most commonly used models that is employed to study the linear relationship between an explanatory variable and a response variable. Some examples of the explanatory variable are the conditions

of an experiment such as the temperature in a biological study or the dose level in a dose-response study. Many research in design of experiments has been devoted to studying the theory of the optimal design framework for this model, but with the assumption of completely observed responses. Following the optimal design framework that is proposed by Imhof et al. (2002), in which complete case analysis is assumed implicitly, we generalise this framework to provide a more robust optimal design to the presence of responses that are missing at random. Complete case analysis is the simplest missing data analysis and is the default method in most of the statistical software, such as STATA and SPSS.

Following this we also consider multiple imputation in the optimal design framework for the linear regression model. In recent years, this analysis is becoming popular in practice as it is also available in most statistical software. Compared with the simple imputation techniques such as mean imputation and last observation carried forward, multiple imputation provides a more reliable confidence interval for the parameter estimates, which accounts for the variability that is caused by the missing values. The idea of multiple imputation is to replace the missing values by some realisations that are drawn from a plausible model repeatedly before analysing the complete-data sets. These complete-data sets contain the same observed values and the different imputed values that are obtained from each repeated imputation. Standard complete-data analysis is then applied to each of these complete-data sets respectively yielding a set of outputs. To make inferences, the variations that arose in the repeated imputations and in the standard analysis respectively are combined using rules proposed Rubin (1987). In our investigations, we aim to construct an optimal design which minimises the total variations of these estimates in multiple imputation. We compare these optimal designs with those that are obtained from the design framework which assumes complete case analysis for the linear regression model.

In view of longitudinal data analysis, there often exists correlation between the observations of the same experimental unit that are measured at some time points over the duration of the study. In the literature, this correlation is often called serial correlation and can be captured by the observational error of a statistical model. In many research areas, a linear mixed model is one of the popular models that is used to study the changes of an outcome variable over time based on the longitudinal data. This model provides flexibility to study both the fixed effects and the random effects of the explanatory variable, as well as capturing the serial correlation.

Since the cohort design framework for the linear mixed model in the literature is not comprehensive, we first revised the model formulation of Ouwens et al. (2002) and Schmelter (2007a) respectively with the assumption of completely observed repeated measurements. In our investigation, we aim to find the time points of measuring different groups of experimental units, taking into account the impact of the presence of dropouts. By dropout, we mean that once a patient does not attend a follow-up session, no further

information will be contributed by this patient for the remaining duration of the study. We therefore consider the approach that is employed in Ortega-Azurduy et al. (2008) to find the expected total information in the available case analysis for the longitudinal data. Our research problem is different from that of Ortega-Azurduy et al. (2008) in the context that their model formulation is restricted to studying the time effect on the response variable of a group of experimental units, whereas our model formulations allow the experimenters to study the time effect on more than one group of subjects.

The structure of this thesis is as follows. In Chapter 2, we present some backgrounds for the experimental design framework and the missing data analysis strategies. We introduce the linear regression model and the linear mixed model. We also depict complete case analysis, available case analysis, and multiple imputation for missing data analysis. In Chapter 3, we present the literature on robustness measures of design against missing responses, the framework that is proposed by Imhof et al. (2002) for the linear regression model with potentially missing responses, the construction of optimal designs for the experiments with correlated responses, and some remarks for missing data analysis strategies. In Chapter 4, we illustrate the optimal design framework for the linear regression model with complete case analysis. In Chapter 5, we study the cohort design framework for the linear mixed effect models with available case analysis. In Chapter 6, we describe the optimal design framework for the linear regression model with multiple imputation. We then conclude the main findings of this project and provide some possible future works to the readers in Chapter 7.

# Chapter 2

# Background

Suppose that we want to study the relationship between some explanatory variables and some response variables by considering statistical models, with the aim to obtain consistent and efficient estimators for the fixed and unknown parameters of the models. By consistent, we mean that an estimator is close to the true value of the model parameter as the sample size of the study increases; by efficient, we expect the estimator to be unbiased and to have minimal mean squared error. Here, we first illustrate the linear regression model for the experiments where each experimental subject is measured once. We discuss the conventional design framework and some theory for constructing an optimal design for this model. Following that, we introduce the linear mixed model for the analysis of longitudinal data. The optimal design framework for the linear mixed model can provide optimal designs for the experiments that have more than one group of experimental units and with repeated measurements. We then illustrate the missing data mechanisms that are defined by Rubin (1976) with some examples. Lastly, complete case analysis, available case analysis and multiple imputation are depicted respectively in the remaining part of this chapter.

## 2.1 Linear regression model

Let $y_i$ be the response of experimental unit $i$, $i = 1, ..., N$, denoting the index of $N$ experimental units, and $x_i$ be the corresponding explanatory variable. For $(p+1)$ known functions, $f_0(x), ..., f_p(x)$, the response of experimental unit $i$ can be represented by a linear regression model,

$$y_i = \beta_0 f_0(x_i) + ... + \beta_p f_p(x_i) + \epsilon_i, \tag{2.1}$$

where the observational error, $\epsilon_i$, follows a normal distribution with mean zero and variance $\sigma^2$. Assuming that subjects are independent and identically distributed, we have $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. For example, a $p^{th}$ polynomial normal linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \epsilon_i,$$

which can be written in the matrix form,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_N)^T$ is a vector of $N$ responses, $X$ is a $N \times (p+1)$ matrix with rows $f^T(x_i) = (f_0(x_i), ..., f_p(x_i))$, $\boldsymbol{\beta}$ is a column vector of $p+1$ unknown parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_N)^T$, $i = 1, ..., N$, is a vector of observational errors. The matrix $X$ is called a design matrix in the context of design of experiments because a design problem is to deduce the value of this matrix with respect to an objective of the experiment, i.e. set the values of the explanatory variable of a study. The fixed but unknown $\boldsymbol{\beta}$ reflect the effect of a unit change in the value of $x_i$ on the response $y_i$. Using the method of least squares, $\boldsymbol{\beta}$ can be estimated by

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y},$$

provided that $X$ has full column rank $p+1$, with variance-covariance matrix

$$\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}.$$

It can be shown that the least squares estimator $\hat{\boldsymbol{\beta}}$ is unbiased, and is an efficient estimator when $X$ yields minimal values for the elements of $\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}})$. A future response, $y_i = f^T(x_i)\hat{\boldsymbol{\beta}}$, at $x_i$, $i > N$ can be predicted using the above described linear model with variance

$$var(y_i) = f^T(x_i)\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}) f(x_i),$$

which represents the variation of making the prediction at $x_i$.

### 2.1.1 Experimental Design

Prior to conducting an experiment, we can maximise the information that can be obtained from it by determining the optimal setting of the design matrix $X$ with respect to the goal of the study. The experimental units are then being measured at this design setting in the experiment to provide information for making inferences. This approach is called experimental design, where the framework consists of three components: a statistical model, an objective function, i.e. a design criterion, and an algorithm that is used to search for an optimal design. A design problem searches an optimal solution over the range of $X$, i.e. a design region, and is subject to some constraints. We note that although a statistical model is assumed at the design stage of an experiment for finding an optimal design, a different model is often employed for data analysis in practice, especially, when the data are not fully observed.

We now introduce the notion of the conventional optimal design framework for linear regression model (2.1) in detail. Let $\mathfrak{X}$ be a design region such that all $x_i$ are bounded in a space, $N = \sum_{i=1}^{m} n_i$ be the total sample size where $n_i$ is the total number of observations that are measured at $x_i$, and $i = 1, ..., m$, denotes the index of the unique values of the explanatory variable. These unique values $x_i$ are the values that the explanatory variable can take. (Note that in the standard notations of some statistical models, we often have $x_i$, $i = 1, ..., N$, where some of these $x_i$ are the same. This reflects the replications of observations that are measured at the unique values of the explanatory variable.) By definition, a design is

$$\xi = \begin{Bmatrix} x_1 & \cdots & x_m \\ w_1 & \cdots & w_m \end{Bmatrix}$$

where $w_i = n_i/N$ is called the weight at support point $x_i$, with $w_i > 0 \; \forall i$ and $\sum_{i=1}^{m} w_i = 1$.

The approximate design theory for model (2.1) that is introduced by Kiefer (1959) defines a design, $\xi$, as an exact design if $w_i N$, $i = 1, ..., m$, are all integers; otherwise $\xi$ is called an approximate/ continuous design where $w_i N$ needs not be an integer. By considering the continuous design approach, we can avoid the problem of discrete optimisation. This approach has been widely employed to find optimal designs for the experiments. To implement a continuous design in practice, the weights are rounded such that the $w_i N$, $i = 1, ..., m$, are integers. See, for example, Pukelsheim and Rieder (1992), for an efficient rounding procedure. In our investigation, we consider finding a continuous design using the approximate design framework. This is because standard optimisation techniques can be applied easily to find the number of support points, $m$, the values of $x_i$ and $w_i$ such that the maximum information is obtained from an experiment. If necessary, some integer approximations can be applied to find the closest exact design for implementing the experiment.

In general, an optimal design is constructed by optimising a function of the Fisher information matrix of the chosen statistical model. This information matrix was introduced by R. A. Fisher in his pioneering work. It is independent of the responses of the experiment, and is determined by the setting of the controlled variables, i.e. the design matrix, $X$. According to Rao (1945) and Cramer (1946), who define the Cramér-Rao lower bound, the variance of an unbiased estimator is bounded below by the inverse of the Fisher information matrix. For example, for linear regression model (2.1), this lower bound is equal to $(X^TX)^{-1}\sigma^2$. With completely observed information, $(X^TX)^{-1}\sigma^2$ is proportional to the inverse of

$$M(\xi) = \sum_{i=1}^{m} f(x_i)f^T(x_i)\ w_i,$$

where $\xi$ is a continuous design and $f^T(x_i)$ is the $i^{th}$ unique row of $X$.

We now present an example to illustrate the above terminologies. Consider a dose-response study with total sample size, $N = 101$, where the dose, $x_i$ is a continuous variable that can take any values between 0 and 1, i.e. design region $\mathfrak{X} = [0, 1]$. A continuous design with two support points, i.e. $m = 2$, for the experiment of this study is

$$\xi = \begin{Bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{Bmatrix}.$$

The Fisher information matrix of model (2.1) with $p = 1$ in the continuous design form is

$$M(\xi) = f(0)f^T(0)\ 1/2 + f(1)f^T(1)\ 1/2,$$

where $f^T(0)$ and $f^T(1)$ are the two unique rows of the design matrix $X$, and $X$ has dimension $101 \times (p+1)$. This information matrix is equivalent to $\frac{X^TX}{101}$, where the exact design in this example would have either $n_1 = 50$, $n_2 = 51$ or $n_1 = 51$, $n_2 = 50$, subjects that are being measured at $x_1 = 0$ and $x_2 = 1$ respectively. We note that even though both exact designs approximate the continuous design for implementing the experiment, the objective functions of some design criteria computed using the respective exact designs might not be the same.

We now define a design criterion more formally, which can be employed to formulate an objective of the experiment. Let $\Xi$ be the class of all possible designs on $\mathfrak{X}$ and $\mathfrak{M}$ be the set of all information matrices with respect to $\Xi$, i.e. $\mathfrak{M} = \{M(\xi); \xi \in \Xi\}$. A design criterion is a continuous, differentiable real-valued function $\psi(M(\xi))$ on a square symmetry matrix, which is a decreasing and convex function over $\mathfrak{X}$ such that there is a critical point in the region. The technical explanation of these properties can be found, for example, in Silvey (1980) and Atkinson et al. (2007). A general design problem is to find the values of $x_i$ and $w_i$ of $\xi$ with respect to an objective of the study, such that the maximum information is attained if this design setting is used in the experiment. By

considering a function, $\psi(M(\xi))$, in planning an experiment, we can optimise an aspect of the information matrix of a statistical model. We can find an optimal design that contains values of $x_i$ and $w_i$ such that $\psi(M(\xi^*)) = \min_{\xi \in \Xi} \psi(M(\xi))$. If this continuous design exists, it is called an optimal design, which is often denoted by $\xi^*$. Note that this optimisation problem is often subject to some constraints, such as $w_i \geq 0 \ \forall i$ and $\sum_{i=1}^{m} w_i = 1$. The following are some examples of design criteria where $M^{-1}(\xi)$ is the inverse of the information matrix for the experiment with design setting $\xi$:

- $D$-optimality: $\psi(M(\xi)) = |M^{-1}(\xi)|$ minimises the generalised variance of the estimates.
- $G$-optimality: $\psi(M(\xi)) = \max_{x} \ f^T(x) M^{-1}(\xi) f(x)$ minimises the maximum variance of the predicted responses.
- $L$-optimality: $\psi(M(\xi)) = tr \ C \ M^{-1}(\xi)$ minimises some linear combinations of the variances of the estimates, where $tr$ stands for trace, and $C$ is a given non-negative definite symmetric matrix which is selected according to the relative importance of each estimate.

Before the introduction of the Cramér-Rao lower bound, Smith (1918) is one of the first to find an optimal design for a quadratic regression model, i.e. model (2.1) with $p = 2$. She finds the optimal setting of $X$ that minimises the maximum value of $var(\hat{y})$ across $\mathfrak{X}$. Concerning the role of the determinant of $(X^T X)^{-1}$ of model (2.1) in hypothesis testing, Wald (1943) assesses the efficiency of using different designs in the test statistic. These two criteria are later called $G$- and $D$-optimality respectively by Kiefer (1958). Among the many optimal criteria that are not being mentioned here, $D$-optimality is the most commonly used design criterion to these days. One of the reasons is that a $D$-optimal design is also $G$-optimal. This can be proven by the following equivalence theorem that is introduced by Kiefer and Wolfowitz (1960):

*General equivalence theorem.* Let $\phi(x, \xi)$ be the Fréchet derivative of the design criterion $\psi(M(\xi))$. The following statements are equivalent.

- $\xi^*$ minimises $\psi(M(\xi))$.
- $\xi^*$ maximizes the minimum over $\mathfrak{X}$ of $\phi(x, \xi)$.
- The minimum over $\mathfrak{X}$ of $\phi(x, \xi^*) = 0$, where this minimum occurs at the support points of $\xi^*$. For any non-optimum design $\xi$, the minimum over $\mathfrak{X}$ of $\phi(x, \xi^*) < 0$.

By definition, the Fréchet derivative of $\psi(M(\xi))$ at matrix $M(\xi) = M_1$ in the direction of $M(\xi) = M_2$ is

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left[ \psi((1-\varepsilon) M_1 + \varepsilon M_2) - \psi(M_1) \right].$$

This derivative verifies if the corresponding information matrix $M_1$ provides maximum information when the search is done in the direction of $M_2$.

To be more specific, the general equivalence theorem checks that a continuous design $\xi^*$ is $D$-optimal on a design region $\mathfrak{X} = [l, u]$ if and only if

$$f^T(x)\ M^{-1}(\xi^*)\ f(x) \leq p + 1 \quad \forall x \in [l, u], \tag{2.2}$$

where $p+1$ is the number of unknown parameters in model (2.1) with degree $p$ (see Kiefer (1974)). Notice that the left hand side of (2.2) is equivalent to the $G$-optimality for model (2.1), where the interest lies in minimising the maximum $var(\hat{y})$ at any $x \in [l, u]$. This shows that a $D$-optimal design for model (2.1) is minimising the volume of the confidence ellipsoid of $\hat{\boldsymbol{\beta}}$, as well as making predictions with high precision, i.e. small variance. Moreover, the largest standardised variance over $\mathfrak{X} = [l, u]$ is achieved at the support points of $\xi^*$ when equality is achieved in the expression. Hence, employing such an optimal design protects against the worse outcome over $\mathfrak{X}$.

The development of this theorem is based on the contributions of De la Garza (1954) and Hoel (1958). For a given spacing of total information at the support points of a design, De la Garza (1954) shows that $p + 1$ support points are sufficient for estimating $\boldsymbol{\beta}$ of model (2.1) with degree $p$. This result is found by comparing the values of $\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}})$ of different designs. Employing this result, Hoel (1958) attains the optimal setting of $X$ for $D$-optimality, whereby $G$-optimality was used as a measure of efficiency for the designs. Assuming fully observed responses for model (2.1), Kiefer and Wolfowitz (1960) generalise these results in their investigation, yielding the above general equivalence theorem. Whittle (1973) also proves the general equivalence theorem by considering the Fréchet derivative of the design criterion $\psi(M(\xi))$.

On the other hand, by considering differential calculus in the approximate design theory, Silvey (1980) has verified that the $D$- and $G$-optimal design have equal weights, $w_i = \frac{1}{p+1}$, $i = 1, ..., m$, at each of the support points (see Silvey (1980) Lemma 5.1.3 on pg 42). Using these theoretical results, a $D$-optimal design for the linear regression model can be obtained easily.

Apart from the fact that a $D$-optimal design is also $G$-optimal for the linear regression model, $D$-optimal designs also have an invariance property with respect to re-parameterisation and re-scaling. More specifically, the design problem is invariant under non-singular linear transformations of the design space. This property is useful in the context that for model (2.1) with a given number of unknown parameters, the support points of a $D$-optimal design for a given $\mathfrak{X}$ can be transformed accordingly to the support points of a $D$-optimal design for another design region. Hence, we do not need to consider the optimisation problem again if a different design region is of interest. A detailed review of the $D$-optimal design can be found in John and Draper (1975).

We now give the background for $L$-optimality. By specifying the diagonal elements of the non-negative definite symmetric matrix $C$, $L$-optimality can be used to find the optimal setting of an experiment that minimises some combinations of the variance of $\hat{\boldsymbol{\beta}}$. Two special cases of this design criterion are

- $A$-optimality which minimises the sum of the variances of all elements of $\hat{\boldsymbol{\beta}}$, where $C$ is the identity matrix;
- $c$-optimality which minimises a linear combination of the variances of the elements of $\hat{\boldsymbol{\beta}}$, where $C = cc^T$, $c$ is a vector of some constants.

In the literature, $A$- and $c$-optimality are introduced by Chernoff (1953) and Elfving (1952) respectively. These criteria are generalised to $L$-optimality by Fedorov (1972).

We note that one of the drawbacks of employing a $c$-optimal design is that the information matrix of model (2.1) might become singular. This is because the design problem focuses only on the variance of one of the estimated parameters out of $p+1$ parameters. Consequently, the design might not provide sufficient information for estimating all other parameters of the model especially when $p$ is large. Moreover, unlike $D$-optimality, this class of design criteria do not possess the invariance property, and the design problem varies according to the combination of the variance. Hence, an optimal design needs to be constructed accordingly to find the design setting that minimises a combination of the variances of $\hat{\boldsymbol{\beta}}$ over a given design region.

In summary, we have presented the background for the optimal design framework for linear regression model (2.1). One of the underlying assumptions of these results is that the observations are assumed to be fully observed in the experiment. In our investigation, we study the impact of the presence of missing responses on the optimal designs for model (2.1). We consider two different techniques which analyse the incomplete data, i.e. complete case analysis and multiple imputation, at the design stage of an experiment. The optimal design framework that accounts for these missing data analysis techniques are presented in Chapter 4 and Chapter 6 of this thesis respectively.

## 2.2 Linear Mixed Model

We now give the background for the linear mixed model. This model is one of the commonly used models that analyse the experimental data which has repeated measurements. Let $\boldsymbol{y}_i^T = (y_{i1}, y_{i2}, ..., y_{iq})$ be the $q$ repeated measurements of subject $i$, $i = 1, .., N$, denoting the index of $N$ experimental units, and $\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, ..., x_{iq})$ be the corresponding set of explanatory variable. The responses of subject $i$ can be

represented by the linear mixed model,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed effect parameters, $\boldsymbol{b}_i$ is a $s \times 1$ vector of unknown individual effect parameters (also called random coefficients), $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices with dimension $q \times p$ and $q \times s$ respectively, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{iq})^T$ is a vector of observational errors.

By modelling longitudinal data with the linear mixed model, we can study the fixed effects and the random effects of the explanatory variables on the responses of experimental units. These effects are reflected by $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ respectively in the model. Since experimental subjects are chosen from a population, the variations of choosing these subjects can be captured by the variability of the $s$ random coefficients $\boldsymbol{b}_i$. In other words, $\boldsymbol{b}_i$ reflect how different subjects response differently to the explanatory variable in the study.

In the literature, it is often assumed that the experimental units are independent and identically distributed, whereby each $\boldsymbol{b}_i$ is normally distributed with mean zero and covariance matrix $\boldsymbol{D}$, i.e. $\boldsymbol{b}_i \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{D})$ where $\boldsymbol{0}$ is a vector of zeros, and $\boldsymbol{D}$ is a $s \times s$ matrix, $i = 1, .., N$; and each $\boldsymbol{\epsilon}_i$ is assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\sigma^2\boldsymbol{\psi}$, i.e. $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N(\boldsymbol{0}, \sigma^2\boldsymbol{\psi})$, $\boldsymbol{\psi}$ is a positive definite matrix. For the experiments that measure the outcome variable on the same experimental units repeatedly over time, there often exists correlation between the observations of the same experimental unit. This correlation is often known as serial correlation, which can be captured by the structure of $\boldsymbol{\psi}$ with elements equal to $cov(\epsilon_{ij}, \epsilon_{ij'}) \neq 0$ where $\epsilon_{ij}$ and $\epsilon_{ij'}$ correspond to the observational errors of experimental unit $i$ at time $t_j$ and $t_{j'}$ respectively. The dimension of $\boldsymbol{\psi}$ is dependent on the number of repeated measurements. In general, since $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ reflect different types of randomness in the model, they are often assumed to be independent. We note that linear regression model (2.1) is a special case of the linear mixed model with $q = 1$, $\boldsymbol{b}_i \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{D})$, and $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N(\boldsymbol{0}, \sigma^2)$, $i = 1, .., N$, when $\boldsymbol{D}$ is known to be equal to a zero matrix.

To make inference about the fixed effect parameters $\boldsymbol{\beta}$ of the linear mixed model, we consider the marginal distribution of the repeated measurements, i.e. it can be shown that $\boldsymbol{y}_i$ is normally distributed with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{\psi}$. Using the maximum likelihood method, an estimator of $\boldsymbol{\beta}$ is given by the generalised least squares estimates,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right) \sum_{i=1}^{N} \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{y}_i,$$

with

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \left( \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1}.$$

This estimator is the best linear unbiased estimator; best in the context that $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ is bounded below by the inverse of the Fisher information matrix,

$$\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i.$$

On the other hand, to make inference about the random coefficients $\boldsymbol{b}_i$, such as predicting subject specific evolution, another approach is required. For example, a natural frequentist approach estimates $\boldsymbol{b}_i$ by minimising the mean squared prediction error conditional on the observed responses; a Bayesian framework could assign a prior distribution to the unknown parameters in the covariance matrix $\boldsymbol{D}$, whereby an estimator for $\boldsymbol{b}_i$ can be obtained by sampling from the posterior distribution of $\boldsymbol{b}_i$ using Markov Chain Monte Carlo methods such as Gibbs sampling.

### 2.2.1 Experimental design with repeated measurements

For the experiment which objective is to study the fixed effect parameters $\boldsymbol{\beta}$ of the linear mixed model, the structure of $\boldsymbol{V}_i$ is often assumed to be known at the design stage of an experiment. Even though in practice the elements of $\boldsymbol{D}$ and $\boldsymbol{\psi}$ are estimated using the experimental data, for example by the restricted maximum likelihood approach, assuming some values for these elements at the design stage of an experiment results in a locally optimal design. In our context, a locally optimal design that is found by a design framework means that the design is only optimal for a given structure of $\boldsymbol{V}_i$. We note that these optimal designs are approximately optimal if the estimated structure of $\boldsymbol{V}_i$ (for example using the historical data or data from pilot study) is close to the true structure of $\boldsymbol{V}_i$. Another way to deal with the unknown $\boldsymbol{V}_i$ is to consider a Bayesian design framework (see Chaloner and Verdinelli (1995)).

Considering that $\boldsymbol{\beta}$ of the linear mixed model are estimated by the generalised least squares estimates, we can find an optimal design by optimising a function of the Fisher information matrix,

$$\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i,$$

for a given structure of $\boldsymbol{V}_i$. Since this matrix sums the total information which is being contributed by each experimental unit $i$, the matrix can be re-expressed in terms of weights for the construction of a continuous design. A design problem is to find the setting of the design matrix $\boldsymbol{X}_i$ and the corresponding proportion of experimental units who have this design setting, such that an aspect of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ is minimised over the design

region. In terms of application, the elements of $\boldsymbol{X}_i$ often correspond to the time points of measuring the outcome variable on experimental unit $i$.

In general, the notion of design criteria, i.e. the goals of experiments, for the linear mixed model are similar to those for linear regression model (2.1). However, due to complicated structure of the information matrix, numerical methods are often employed to find the optimal designs with respective to the design criteria. Note that for $D$-optimality, the optimisation problem can be simplified to minimising the negative determinant of the Fisher information matrix, instead of minimising the determinant of the inverse matrix in the optimisation problem. Moreover, since only locally optimal designs are available for the linear mixed model, the verification of an optimal design is also done numerically. For example, different initial designs can be employed in the optimisation algorithm to check if a design setting converges to some values, and the objective values of a design criterion that are achieved by the design candidates can be compared for evaluating the efficiency of the designs.

To the best of our knowledge, the literature on the design theory for the linear mixed model is relatively limited, see for example Fedorov and Nachtsheim (1995), Fedorov and Hackl (1997) and Fedorov et al. (2002). We present the literature that is relevant to our design problem for the linear mixed model in Chapter 3 of this thesis. In our investigation, we aim to find the design framework for the linear mixed models in the presence of dropouts. By dropout, we mean that once an experimental subject is not being observed at a support point, no further observations will be contributed by this subject for the remaining duration of the study. In Chapter 5 of this thesis, we consider various settings for the experiments that have more than one cohort, and assume that available case analysis will be implemented by the data analyst.

## 2.3 Missing Data Mechanisms

We now present the missing data mechanisms that have been defined in the literature. To account for the presence of missing data within a statistical framework, a missing value could be considered as a realization of a random variable. Let $\mathcal{M}_i$ be a Bernoulli random variable where

$$\mathcal{M}_i = \begin{cases} 1, & \text{for } y_i \text{ is missing;} \\ 0, & \text{otherwise,} \end{cases}$$

$i = 1, ..., N$, denoting the index of $N$ experimental units. These $\mathcal{M}_i$ are independent provided that the experimental units are independent. From a frequentist point of view, there is a data governing parameter in the random process that generates the value of $\mathcal{M}_i$. By imposing some conditions on the data governing parameter, Rubin (1976) has classified three types of missing data mechanism in the literature. He names the random process of $\mathcal{M}_i$ as missing at random (MAR) when the data governing parameter is

dependent on the available/observed information, given the data. In terms of notations, for given information $x_i$, a MAR mechanism generates a missing observation $y_i$ using the following probability model,

$$Pr(\mathcal{M}_i = 1|x_i) = P(x_i)\ \forall x_i,$$

where $P(x_i)$ is dependent on $x_i$ and nothing else. A special case of MAR, namely missing completely at random (MCAR), is defined for the process when the data governing parameter places no restrictions on the realisation of $\mathcal{M}_i$. Therefore all observations have the same probability of being missing, i.e. $Pr(\mathcal{M}_i = 1|x_i)$ is equal to a constant which lies between 0 and 1 $\forall i$, and the probability is independent of the data governing parameter. For missing values that are not being generated by these two types of mechanism, the random process is classified as missing not at random (MNAR). The data governing parameter of MNAR might now depend on missing values or other unmeasured variables, given the data.

We now illustrate the different types of missing data mechanism in the context of an Alzheimer's disease clinical trial where there are missing responses in the primary outcome measurement. The primary outcome measurement is assumed to be missing at random if the probability that an observation is missing depends only on other fully observed information, such as the administered treatment doses and the availability of a carer for each patient. It is assumed to be missing completely at random if the missing value is just a simple random sample from the clinical trial. Suppose that the availability of a carer for each patient is not measured in the study, the primary outcome measurement could be missing not at random because a patient may not be able to attend the follow-up session when the carer is not available.

In general, the presence of missing values is not restricted to the response variable only. For example in survey research, there could be missing information on the explanatory variables such as age and income of the participants due to privacy. In practice, the data governing parameter of a missing data process could be considered as a nuisance parameter in the analysis of the study. One can test the missing data mechanism via sensitivity analysis. Nevertheless, there is no consensus on handling missing data in practice. Moreover in some cases the pattern of missing values plays a role in modelling the incomplete-data set. See Figure (2.1) for some examples of missing data patterns in repeated measurements. In certain industries, such as confirmatory clinical trials, a guideline on missing data analysis has been established. Broadly speaking, for a given incomplete-data set, different approaches may yield different inferential findings due to the underlying assumptions for the missing values.

In our investigation, we study the impact of the presence of missing values on the optimal designs. Assuming that the missing values present only in the response variable, we

**Monotone dropout pattern**

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $y_{i1}$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ? |
| $y_{i2}$ | ✓ | ✓ | ✓ | ✓ | ✓ | ? | ? |
| $y_{i3}$ | ✓ | ✓ | ✓ | ? | ? | ? | ? |

**Intermittent missing data pattern**

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $y_{i1}$ | ? | ✓ | ✓ | ✓ | ✓ | ? | ✓ |
| $y_{i2}$ | ✓ | ✓ | ? | ✓ | ? | ✓ | ✓ |
| $y_{i3}$ | ✓ | ? | ? | ? | ✓ | ? | ✓ |

Figure 2.1: Example of a monotone dropout pattern (left) and an intermittent missing data pattern (right).

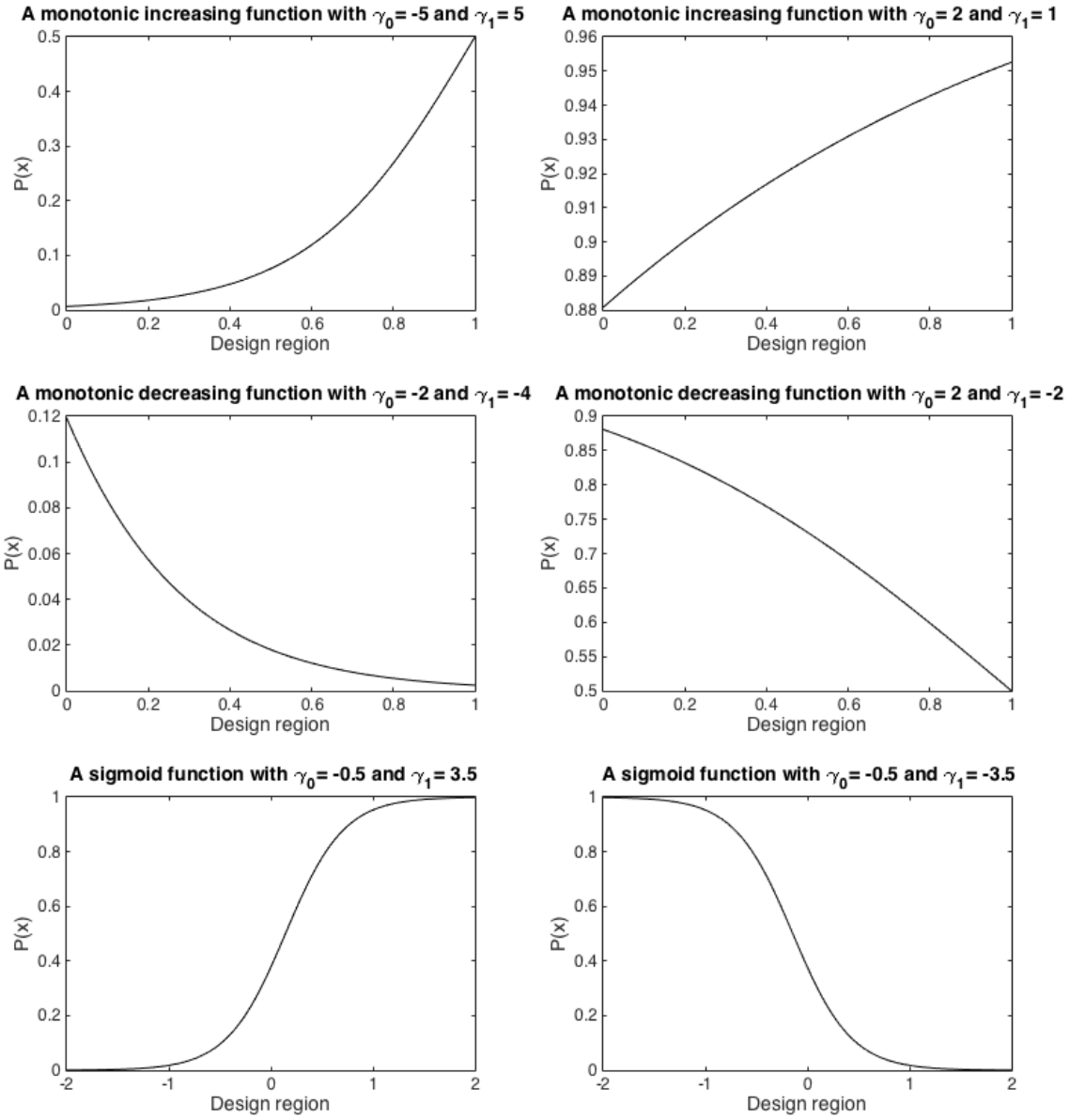



Figure 2.2: Examples of monotone functions for design region $\mathfrak{X} = [0, 1]$ that are plotted with the inverse logit link functions.

incorporate a monotone MAR mechanism into the design framework for the linear regression model and the design framework for the linear mixed model respectively. In the design framework for the linear mixed model, a monotone dropout pattern is assumed. More specifically, we assume that once an experimental unit is not being measured at a particular time point, no further observations from this subject are possible for the remaining duration of the experiment. For simplicity, we employ a response probability function instead of a MAR mechanism for a dropout process using the complement rule for probability in our investigation. Figure (2.2) illustrates some examples of monotone functions that are plotted using the inverse logit link function,

$$P(x_i) = \frac{exp(\gamma_0 + \gamma_1 x_i)}{1 + exp(\gamma_0 + \gamma_1 x_i)},$$

and the corresponding values of $\gamma_0$ and $\gamma_1$. This function is attractive as we can obtain a monotone probability function over a design region by setting the value of $\gamma_1$. For example, it is monotone increasing with $x_i$ for $\gamma_1 > 0$, and monotone decreasing with $x_i$ for $\gamma_1 < 0$. The bottom plots in Figure (2.2) show the sigmoid shapes of the logistic functions over a wider range of $x_i$.

## 2.4 Missing data analysis

In this section, we illustrate the missing data analysis strategies that we consider at the design stage of an experiment. We assume complete case analysis and multiple imputation respectively in the design framework for linear regression model, and available case analysis in the design framework for linear mixed model.

### 2.4.1 Complete case analysis/ Listwise deletion

Complete case analysis (aka listwise deletion) is the simplest and the easiest method that can be employed to analyse the incomplete-data sets. This method is the default setting in many statistical packages for regression analysis, such as in SPSS, SAS and STATA. The idea of this method is to exclude the partially observed information of the experimental units in the analysis. After the 'data cleaning' procedure, the information is treated as a complete set where standard statistical procedures for making inferences are applied. Figure (2.3) shows the data that is used in the analysis for the missing data examples in Figure (2.1).

As can be noticed from the examples, listwise deletion discards most of the information in the data analysis of the intermittent missing data pattern example. Nevertheless, we consider this method in the experimental design framework for the linear regression model in Chapter 4 for cross-sectional data, where only one observation is measured on each experimental unit in the study.

**Monotone dropout**

| $i$ | 1 | 2 | 3 |
|---|---|---|---|
| $y_{i1}$ | ✓ | ✓ | ✓ |
| $y_{i2}$ | ✓ | ✓ | ✓ |
| $y_{i3}$ | ✓ | ✓ | ✓ |

**Intermittent missing data pattern**

| $i$ | 7 |
|---|---|
| $y_{i1}$ | ✓ |
| $y_{i2}$ | ✓ |
| $y_{i3}$ | ✓ |

Figure 2.3: The information that is used in complete case analysis for the examples in Figure (2.1).

### 2.4.2 Available case analysis/ Pairwise deletion

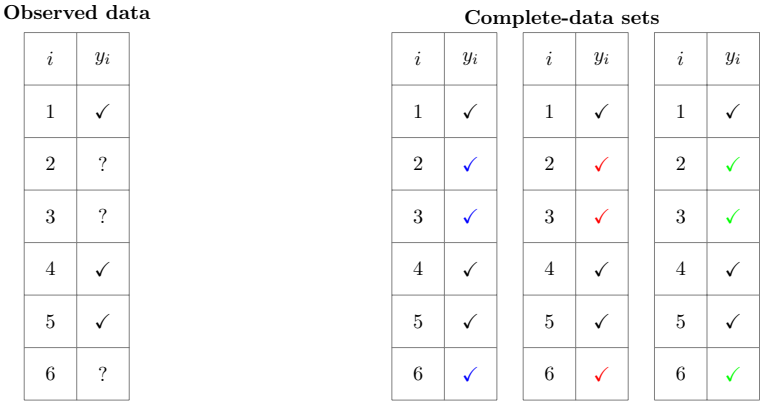
Instead of discarding the partially observed data in the data analysis, pairwise deletion makes inferences based on all available data (hence, it is also called available case analysis). This is a simple method that is also available in most of the statistical packages. To illustrate this method, consider the example of intermittent missing data pattern in Figure (2.1) and find the expected observation at each time point, i.e. the mean of $y_{i1}, y_{i2}$ and $y_{i3}$. Using complete case analysis, the mean of the three variables are $y_{71}, y_{72}$ and $y_{73}$ respectively because the information that is contributed by other subjects contains at least one missing value and hence is discarded in the analysis. On the other hand, pairwise deletion uses all the available data, giving $(y_{21} + y_{31} + y_{41} + y_{51} + y_{71})/5$, $(y_{12} + y_{22} + y_{42} + y_{62} + y_{72})/5$, and $(y_{13} + y_{53} + y_{73})/3$ respectively for the mean of the three response variables.

In our investigation, we consider this missing data analysis for the experiment that has repeated measurements and a monotone dropout pattern in Chapter 5.

### 2.4.3 Multiple imputation

We now illustrate a multiple imputation approach that we consider in the design framework for linear regression model (2.1). Rather than ignoring the presence of missing values in the analysis, multiple imputation replaces missing values repeatedly with some simulated values that are drawn from a plausible distribution before standard complete-data analysis is carried out for each complete-data set. The complete-data sets contain the same observed values and the different imputed values that are obtained from the repeated imputations. Figure (2.4) shows an example of three complete-data sets. To make inferences, the outputs of all complete-data analyses are combined using some combining rules.

In our investigation, we impute missing values by the realisations of a posterior predictive distribution which is constructed using the likelihood function for the unknown




| $i$ | $y_i$ |
|---|---|
| 1 | ✓ |
| 2 | ? |
| 3 | ? |
| 4 | ✓ |
| 5 | ✓ |
| 6 | ? |

| $i$ | $y_i$ |
|---|---|
| 1 | ✓ |
| 2 | ✓ |
| 3 | ✓ |
| 4 | ✓ |
| 5 | ✓ |
| 6 | ✓ |

| $i$ | $y_i$ |
|---|---|
| 1 | ✓ |
| 2 | ✓ |
| 3 | ✓ |
| 4 | ✓ |
| 5 | ✓ |
| 6 | ✓ |

| $i$ | $y_i$ |
|---|---|
| 1 | ✓ |
| 2 | ✓ |
| 3 | ✓ |
| 4 | ✓ |
| 5 | ✓ |
| 6 | ✓ |

Figure 2.4: Illustration of multiple imputation where the coloured checkmarks represent the imputed values which are drawn from a plausible distribution.

parameters $\boldsymbol{\beta}$ of model (2.1) and a noninformative prior distribution. This posterior distribution is a normal distribution. We note that other imputation models might be adopted in the multiple imputation but we do not consider them here. After imputing the missing values, each complete-data set is analysed for finding the least squares estimates of $\boldsymbol{\beta}$ of the regression model (2.1). The variations that arose in multiply imputing the missing values and in the analyses are then combined using the combining rules that are proposed by Rubin (1987) for making inferences.

#### 2.4.3.1 Rubin's combining rules

Suppose that we want to make inference on the unknown parameters $\boldsymbol{\beta}$ of the regression model (2.1). To combine the analyses of $t$ complete-data sets in the multiple imputation, where $t$ is the number of repeated imputation, Rubin (1987) defined the estimates for the unknown parameters of the model as

$$\bar{\boldsymbol{\beta}} = \sum_{l=1}^{t} \frac{1}{t} \hat{\boldsymbol{\beta}}^{(l)},$$

where $\hat{\boldsymbol{\beta}}^{(l)}$ is the least squares estimates that are computed for the $l^{th}$ complete-data set with variance-covariance $\boldsymbol{cov}(\hat{\boldsymbol{\beta}}^{(l)})$;

$$\bar{\boldsymbol{U}}_{\boldsymbol{t}} = \sum_{l=1}^{t} \frac{\boldsymbol{cov}(\hat{\boldsymbol{\beta}}^{(l)})}{t}$$

as the within imputation variance-covariance that accounts for the uncertainty which arose in analysing each complete-data set; and

$$\boldsymbol{B}_t = \sum_{l=1}^{t} \frac{(\hat{\boldsymbol{\beta}}^{(l)} - \bar{\boldsymbol{\beta}})^T (\hat{\boldsymbol{\beta}}^{(l)} - \bar{\boldsymbol{\beta}})}{t-1}$$

as the between imputation variance-covariance due to the presence of the imputed values. The total variance-covariance of $(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})$ is then defined as

$$\mathcal{T}_t = \bar{\boldsymbol{U}}_{\boldsymbol{t}} + \left(1 + \frac{1}{t}\right) \boldsymbol{B}_t,$$

which measures how far the estimates are from the true but unknown parameters.

To construct the confidence intervals for a scalar $\beta$, we can employ an approximated student $t$-distribution with degrees of freedom

$$v_t = (t-1)\left(1 + \frac{t}{t+1}\frac{\bar{U}_t}{B_t}\right)^2.$$

For example, a $100(1-\alpha)\%$ interval estimate for $\beta$ is

$$\left[\bar{\beta} \pm t_{v_t}(\alpha/2)\sqrt{\mathcal{T}_t}\right]$$

where $\bar{\beta}$ is a point estimate for $\beta$ and $t_{v_t}(\alpha/2)$ is the upper $100\alpha/2$ percentage point of the student $t$-distribution on $v_t$ degrees of freedom. We note that as $t$ approaches to $\infty$, the student $t$-distribution with degrees of freedom $v_t = \infty$ converges to a normal distribution, e.g. for $1 - \alpha = 0.95$ with $t = \infty$, $t_{v_t}(\alpha/2) = 1.96$ can be used in the construction of confidence interval. Given a null value, denoted by $\beta_o$, the associated significant level is

$$P\left(F_{1,v_t} > (\beta_o - \bar{\beta}_t)^2/\mathcal{T}_t\right)$$

where $F_{1,v_t}$ is a $F$ random variable with one and $v_t$ degrees of freedom. More details of the derivation of these results are available in Rubin (1987).

In Chapter 6 of this thesis, we propose an optimal design framework for the linear regression model which assumes the implementation of multiple imputation. We are interested in the optimal designs that minimise some aspects of $\mathcal{T}_t$ over a design region. The key challenge in this investigation is to find the expected value of $\mathcal{T}_t$ at the design stage of an experiment when no observations are available. In the next chapter, we present some literature that is relevant to our research problems.

# Chapter 3

# Literature review

Apart from the missing data analysis strategies that are illustrated in the previous chapter, there is extensive literature on the methodology that tackles the presence of missing data. On the contrary, there is limited research that considers the presence of missing values at the design stage of an experiment. Here, we present some literature that is relevant to our research topic.

We first review some measures that are employed by the experimenters to study the robustness of a design. Robustness measures assess the performance of a design against the worse outcomes when the underlying assumptions of the design fail to hold. We then look at the optimal design framework for the two statistical models that we study in our investigations. For the linear regression model, we only illustrate the framework that is proposed by Imhof et al. (2002), who consider the potentially missing responses at the design stage of an experiment. The background for the optimal design framework for the linear regression model with the assumption of completely observed responses can be found in the previous chapter. Following that, we assess the literature that finds an optimal design for the experiment that has correlated observations. The presence of correlation between the observations can arise when repeated measurements are taken on the same experimental unit over time. We then present some remarks for the missing data analysis methods that we employ in our research. Since the focus of our investigation is on the optimal design framework, we do not review the strengths and the weaknesses of the missing data analysis. Lastly, we end this chapter by depicting the idea of accounting for the presence of missing responses at the design stage of an experiment. The work of Imhof et al. (2002) and Ortega-Azurduy et al. (2008) are the key references to our investigations.

## 3.1 Robustness against missing observations

Provided that responses are fully observed from the experiment, most of the available design framework provide a good basis to finding an optimal design for some real world applications. Often before the implementation of a design, some measures of robustness are used to estimate how poor the design performs when the underlying assumptions of the design are not valid. Some examples of the underlying assumptions are completely observed responses, number of parameters of the statistical model, and the structure of the observational errors. By comparing the robustness of some design candidates with respect to some assumptions, the design that has the highest efficiency can be chosen for implementing an experiment.

Among the many concerns, the presence of missing observations often causes some problems in some types of design. For instance in a balanced block design, the blocks and the treatment groups may become unbalanced when some values are missing in the design. Focusing on the analysis of variance, several authors such as Hedayat et al. (1974), Most (1975), John (1976) and Kageyama (1980) investigate the robustness of a block design. Some of the authors introduce some upper and lower bounds for the efficiency of a design in order to obtain a design that is robust to the unknown missing observations.

On the other hand, when the rate of missing responses is high, some matrices which are involved in the data analysis may become singular, i.e. non-invertible. As a consequence, the analysis fails to provide estimates for the unknown parameters of the statistical model. For example, complete case analysis may fail to provide the least squares estimates for the linear regression model when the loss of information is too severe that the Fisher information matrix becomes non-invertible. In the literature, Ghosh (1979), Ghosh (1980), and Ghosh (1982) consider the robustness of a factorial design from this perspective. To overcome the issue of having a singular matrix, Ghosh incorporates extra treatments into the factorial design such that the design matrix remains invertible in the analysis.

For a response surface study, some authors such as Andrews and Herzberg (1979), Akhtar and Prescott (1986), Herzberg et al. (1987), and Ahmad and Gilmour (2010) have studied the loss of information of a design due to the presence of missing responses in the experiment. Herzberg and Andrews (1976) propose three measures of robustness for an exact design that has only one replication at each discrete point. The three respective criteria are

- the probability of $|X^T(I - \mathcal{M})X|$ being zero, which can be employed to compare designs;
- maximise the expected value of $|X^T(I - \mathcal{M})X|^{1/p}$ where $p$ is the number of unknown parameter in the model; and

- minimise the expected value of the maximum variance of the predicted observation, given that $|X^T(I - \mathcal{M})X| \neq 0$,

where $X$ is the design matrix of the model in the response surface study, $I$ is an identity matrix and $\mathcal{M}$ is a diagonal matrix containing the individual missing data indicators $\mathcal{M}_i$, which is equal to one if response $i$ is missing, zero otherwise. The first measure checks the chances of getting little information if an exact design with the corresponding information matrix is employed in the experiment, whereas the other two robustness measures have similar ideas to the $D$- and the $G$-optimality, which are named by Kiefer (1958). Herzberg and Andrews (1976) do not consider the inverse of the matrix in the investigation as the information matrix may become singular due to the presence of missing observations.

In view of an exact design for the linear regression model, Hackl (1995) introduces $D$-optimality that accounts for the issue of having a singular information matrix. Since only one replication is allowed at each support point of an exact design, the presence of singular information matrix in the analysis is unavoidable when there are missing responses. Nevertheless, the author considers two aspects in the determinant of the design criterion matrix: a weighted sum of covariance matrix of the design candidates that have the same number of missing observations but being missing at different locations over the design space; and a penalty term that is multiplied by a probability that a design results in a singular information matrix.

In spite of the nice features that have been considered by Herzberg and Andrews (1976) and Hackl (1995) respectively, using the above described criteria to find a robust design against the presence of missing values is not pragmatic, especially when the sample size of an experiment is greater than the number of discrete points in the design region. Compare this approach to the continuous design framework where the replications are represented by the weights of an approximate design, one replication at each support point falls short of the purpose of implementing an experiment. Moreover, the number of designs to be compared increases with the sample size of the experiment. This is because the robustness measures consider all the possible designs that have potentially missing observations at different support points. To obtain the most robust design given a missing data mechanism, the experimenter needs to consider all the possible designs and compute the objective value of the design criterion for each of these designs. This process is repeated if a different sample size is used for the experiment, or if a different missing data mechanism is employed. Low et al. (1999) is an example that finds an $A$-optimal design for crossover trials by considering a robustness measure.

## 3.2 Optimal design for potentially missing responses

We now present the optimal design framework that is proposed by Imhof et al. (2002) for the linear regression model,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{y}$ is a vector of $N$ responses, $X$ is a $N \times (p+1)$ design matrix, $\boldsymbol{\beta}$ is a column vector of $p+1$ unknown parameters, and $\boldsymbol{\epsilon}$ is a vector consisting of independent and identically distributed observational errors, which follows a normal distribution with mean zero and variance $\sigma^2$. Recall that with fully observed responses, the Fisher information matrix of this model is

$$M(\xi) = \sum_{i=1}^{m} f(x_i) f^T(x_i)\ w_i,$$

where $\xi$ is a continuous design, $f^T(x_i)$ is the $i^{th}$ unique row of design matrix $X$, and $x_i$ is a support point of $\xi$, $i = 1, ..., m$. Provided that some information about the missing responses is available, Imhof et al. (2002) propose an optimal design framework for the linear regression model which is based on a function of

$$\sum_{i=1}^{m} f(x_i) f^T(x_i)\ w_i(1 - P(x_i)),$$

where $P(x_i)$ is the probability that a response is missing at $x_i$, which can depend on $x_i$ but not $\boldsymbol{y}$, i.e. the missing value is generated by a MAR mechanism. The authors use $w_i(1 - P(x_i))$ in the augmented information matrix to reflect the expected number of experimental units who have observations in the experiment. More details are given in Chapter 4.

Similar to the investigation that is done by Herzberg and Andrews (1976) in deriving the robustness measures, Imhof et al. (2002) do not consider the inverse of information matrix in the approximation to the covariance matrix of the least squares estimates. It is believed that if the presence of missing observations is too severe that it causes the matrix to become singular, the implementation of the design provides no benefit to the goal of the study. However, employing the notion of the continuous design theory, this framework allows for having replications of observations at the support points of a design. Having accounted for the impact of missing responses, the authors show that the sample size of the experiment has no impact on the optimal design, and the same optimal design can be chosen for the experiment which has a different sample size and the same MAR mechanism. For some classes of MAR mechanisms, Imhof et al. (2002) and Imhof et al. (2004) propose some closed form solutions for the support points of some optimal designs. These solutions are some functions that involve the nuisance parameters of the considered missing data mechanisms. However, there is no guidance

for finding the nuisance parameters of the missing data mechanisms in terms of practical usages of the optimal designs.

Moreover, Imhof et al. (2004) argue that the proposed design framework is sufficient for finding an optimal design for the generalized Michaelis-Menten model, where the variance of the responses is dependent on the explanatory variable. This is because the proposed information matrix is analogous to the information matrix of the design framework for a weighted polynomial regression model. In the weighted polynomial regression model, each observational error follows a normal distribution with zero mean and a variance function that is dependent on the explanatory variable, i.e. heteroscedastic variance. By treating this variance function as a MAR mechanism, the information matrix of the weighted polynomial regression model is equivalent to the expected information matrix that is considered in the framework that is proposed by Imhof et al. (2002). As a result, we can infer that the proposed framework can be applied to two contexts, i.e. one for the linear regression model with a pre-specified monotone MAR mechanism, the other for the experiment where the responses are normally distributed with heteroscedastic variance. See for example, Antille et al. (2003), Chang and Jiang (2007) and Dette and Trampisch (2010) for the literature on the optimal designs for the weighted polynomial regression model.

In Chapter 4 of this thesis, we revise the optimal design framework that is proposed by Imhof et al. (2002). We consider the inverse of the information matrix in the presence of missing responses, as opposed to the expectation of the matrix itself. This is because in the presence of missing values, the information matrix might not provide a good approximation to the variance-covariance matrix of $\boldsymbol{\beta}$ of the linear regression model. Assuming that complete case analysis will be implemented at the data analysis stage of a study, we provide a better design framework for the linear regression model, whereby the framework that is proposed by Imhof et al. (2002) can be considered as a special case of ours.

## 3.3 Optimal design for correlated observations

We now consider the optimal designs for the experiments that have repeated measurements. This type of experiment studies the changes of an outcome variable on the same experimental units over time, and is often known as a longitudinal study. For example, to study the seed germination and growth rate of a type of plant under different conditions, observations are taken on a daily basis on the same seeds/plants that are placed under different experimental conditions; to study the efficacy of new interventions on a disease, such as Alzheimer's disease, an outcome variable is measured on the patients at several follow-up sessions. In longitudinal study, the observations that are measured on

the same subject are often correlated with one another, but the observations of different subjects are usually independent. Here we first present the literature on the design framework for the linear regression models that analyse correlated observations of one experimental unit. We then review the optimal design framework for the linear mixed models which analyse the repeated measurements of more than one experimental unit. In these works, we have found only one paper that investigates the efficiency loss of some optimal designs due to the presence of missing responses.

To study the behaviour of a process, such as the chemical reaction of a kind of mixture over time, linear regression model (3.1) can be employed to analyse the repeated measurements of the one experimental subject, i.e. the mixture. The total number of observations of model (3.1) in this context correspond to the number of repeated measurements of an experimental subject, and the observational error $\boldsymbol{\epsilon}$ follows a stochastic process that captures the dependency of an observation on the previous responses, i.e. $cov(\epsilon_j, \epsilon_{j'}) \neq 0$ where $\epsilon_j$ and $\epsilon_{j'}$ correspond to the observation errors of the experimental unit at time $t_j$ and $t_{j'}$ respectively. In the literature, some researchers study the design problem for model (3.1) with some functions for the stochastic process that are dependent on time and a correlation parameter $\rho$, $0 \leq \rho \leq 1$. The support points of an optimal design in this context correspond to the time points of measuring the outcome variable on the experimental unit.

For example, Sacks and Ylvisaker (1966), Sacks and Ylvisaker (1968), Bickel and Herzberg (1979), and Zhigljavsky et al. (2010) find an optimal design density based on some asymptotic arguments for the variance-covariance matrix of the least squares estimates. Having accounted for the corresponding function of $cov(\epsilon_j, \epsilon_{j'})$, an exact optimal design for a given number of observations can be obtained by employing the quantiles of the optimal design density after some transformations. In recent years, Dette et al. (2013) have provided some necessary conditions for $cov(\epsilon_j, \epsilon_{j'})$ and identified universally optimal design densities for some multi-parameter regression models. Following these works, Dette et al. (2014) have proposed a design criterion which measures the distance between a given design and an universally optimal design.

On the other hand, with some known functions for $cov(\epsilon_j, \epsilon_{j'})$, other authors such as Näther (1985), Dette et al. (2008), and Harman and Štulajter (2010) consider weighted least squares estimation for $\boldsymbol{\beta}$ of model (3.1), and find exact optimal designs for the experiments that have one experimental unit and repeated observations. For example, Dette et al. (2008) consider a first order autoregressive process for the stochastic process of the observational error. This process is often known as an AR(1) process in the literature, and has covariance function $cov(\epsilon_j, \epsilon_{j'}) = \rho^{|t_j - t_{j'}|}\sigma^2$. In the investigation of the design problem for the weighted least squares analysis, Dette et al. (2008) present some properties for the exact $D$-optimal designs. They make comparisons with the aforementioned exact designs that are found from the asymptotic arguments for the ordinary least squares analysis, and show that both designs are similar when $\rho$ is small,

but differ substantially as $\rho$ increases. Moreover, the optimal designs for the weighted least square analysis approach to equally spaced designs as $\rho$ approaches to one.

Instead of assuming that all experimental units respond in the same way to the experimental conditions, Tan and Berger (1999) consider the linear model,

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{y}_i$ is a vector of repeated measurements of subject $i$, $\boldsymbol{\beta}_i$ is a vector of $p \times 1$ random regression coefficients with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\boldsymbol{D}$, and the observational error, $\boldsymbol{\epsilon}_i$, follows a normal distribution with mean zero and variance-covariance matrix $\sigma^2\boldsymbol{\psi}$, $\boldsymbol{\psi}$ is a positive definite matrix with elements $\boldsymbol{\psi}(t_j, t_{j'}) = \rho^{|t_j - t_{j'}|}$. This covariance structure is the same as the AR(1) process which is considered in Dette et al. (2008). Recall that a general formulation of a linear mixed model is

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{y}_i$ is a vector of repeated measurements of subject $i$, $\boldsymbol{\beta}$ is the fixed effect parameters, $\boldsymbol{b}_i$ is the random coefficients with variance-covariance matrix $\boldsymbol{D}$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices, and $\boldsymbol{\epsilon}_i$ is the observational error of subject $i$. The model that is considered by Tan and Berger (1999) can be considered as a special case of this linear mixed model, where the mean and the covariance matrix of $\boldsymbol{\beta}_i$ in the former model correspond to the fixed effect parameters $\boldsymbol{\beta}$, and the covariance matrix of $\boldsymbol{b}_i$ in the later model respectively.

However, the model which is considered by Tan and Berger (1999) requires two stage analysis for making inferences. More specifically, considering that $\boldsymbol{\beta}_i$ are fixed and $\boldsymbol{\epsilon}_i$ are independent, each $\boldsymbol{\beta}_i$ is fitted with the corresponding repeated measurements of experimental unit $i$ in the first stage of analysis. The covariance matrix of $\boldsymbol{\beta}_i$ can then be obtained in the second stage of analysis by regressing on the individual estimate of $\boldsymbol{\beta}_i$ (see for example, Laird and Ware (1982) for more details). Considering the general Gauss-Markov estimator for the unknown mean parameter $\boldsymbol{\beta}$ of the considered model, Tan and Berger (1999) find the time points of measuring the outcome variable on the experimental subjects by minimising the determinant of the variance-covariance matrix of the estimates over the time space for some $\rho$, $0 \leq \rho \leq 1$. Given the number of repeated measurements, and fixing the upper and the lower bound of a design region as the first and the last time point of a design, the optimisation problem finds the middle time points of an optimal design with respective to the design criterion.

In view of having move than one group of experimental units, Ouwens et al. (2002) investigate the $D$-optimal design and maximin design respectively for the above described general linear mixed model. Like others, the authors assume an AR(1) process for $\boldsymbol{\epsilon}_i$ to capture the correlation between the repeated measurements of the same subject. Recall

that the information matrix of the general linear mixed model,

$$\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i,$$

sums the total information that is being contributed by each subject, the authors consider a continuous design framework that groups the subjects who have the same design setting using the notion of weights. Since the formulation of the design problem does not account for the experimental conditions of different cohorts, most of the illustration of $D$-optimal design in Ouwens et al. (2002) show that the optimal time points of collecting measurements on different groups of subjects are equivalent.

On the other hand, assuming constant variance for the observational error $\boldsymbol{\epsilon}_i$, Schmelter (2007a) and Schmelter (2007b) find the optimal time points of measuring the outcome variable on the experimental units for a special case of the above mentioned general linear mixed model. In the investigations, the author incorporates an indicator matrix into the model, such that all groups of experimental subjects have the same baseline measurements at the onset of the study. Moreover, the subjects in every group are allowed to be measured at different time points in the continuous design problem. Nevertheless, the continuous optimal design that is found by this framework cannot easily be implemented because the optimal design requires a number or a fraction of repeated measurements at the same time point on the same experimental unit.

Considering the general formulation of the linear mixed model but a different design problem for the longitudinal study, Tekle et al. (2008) assume non-overlapping measurement time for different cohorts, i.e. the experiment on a cohort can only be started after the last observations of one group of experimental units are measured. Having fixed the number of repeated measurements and the number of cohorts in the experiment, the authors employ an algorithm that is similar to an exchange algorithm to find the optimal time points of measuring the outcome variable on the cohorts. Their algorithm compares a set of discrete time points and the weights of the cohorts in order to obtain a design that has the best objective value for the design criterion. They infer that the number of repeated measurements in the first cohort should not exceed the number of fixed effect parameters in the linear mixed model. Moreover, comparing the efficiency of the optimal designs that have different numbers of cohorts, Tekle et al. (2008) and Tekle et al. (2011) show that having too many cohorts can be a waste of resources. Nonetheless, these findings are not surprising since the time points of measuring the outcome variable on the different cohorts are assumed to be non-overlapping, and the information which is being contributed by the first cohort is sufficient for estimating the fixed effect parameters of the linear mixed model.

Concerning the impact of the presence of missing values, Ortega-Azurduy et al. (2008) study the efficiency loss of $D$-optimal designs for the linear mixed model. They assume

the presence of dropouts in the study, and each subject has at least one observation at the first time point of the experiment. The definition of dropout is that once a subject is not being observed in the study, no further information will be contributed by this subject. Employing the AR(1) process for the observational error $\boldsymbol{\epsilon}_i$, the authors find the expected number of repeated measurements that are being contributed by each experimental unit for computing the expected total information matrix of the linear mixed model. The inverse of the expected total information matrix is then assumed to be providing a good approximation to the variance-covariance matrix of the estimated fixed effect parameters when there are dropouts. Given the value of $\rho$, the number of repeated measurements and the response probability function, a $D$-optimal design, i.e. the optimal time points of measuring the outcome variable on a group of subjects, can be obtained by minimising the determinant of this matrix. The loss of $D$-efficiency due to the presence of dropouts is then studied by comparing the optimal design that assumes the presence of dropouts with the optimal design that assumes completely observed responses.

In Chapter 5 of this thesis, we first augment the linear mixed model that is studied by Ouwens et al. (2002), such that there are clear distinctions between the groups of experimental units in the formulation of the design problem. Besides that, assuming an AR(1) process for the observational error $\boldsymbol{\epsilon}_i$, we find the optimal cohort design for the special case of the linear mixed model that is studied by Schmelter (2007a) and Schmelter (2007b). With the assumption of fully observed repeated measurements, the design problem is to find the middle time points of measuring the outcome variable on the groups of experimental units, having chosen the bounds of the design region as the first and the last time point of an optimal design. Following that, we consider the presence of dropouts at the design stage of an experiment. We employ the approach that is used in Ortega-Azurduy et al. (2008) to find the expected total information matrices of our models. In Ortega-Azurduy et al. (2008), the design problem is to find the optimal time points of measuring the outcome variable on a group of experimental units for the general linear mixed model. Considering for the two models which are similar to those that are studied by Ouwens et al. (2002) and Schmelter (2007a) respectively, we find the optimal time points of measuring the outcome variable on different groups of experimental units. Moreover, different cohorts are allowed to have different response probability functions in our optimal design framework.

## 3.4 Missing data analysis strategies

So far we have presented the literature that is related to design of experiments and our research problems. We now look at the missing data analysis in the medical research, and present some remarks for the missing data analysis techniques that we employ in our

investigations. Since the focus of our investigation is on the experimental design framework, we do not discuss the criticisms of the methodologies which tackle the presence of missing values in the data set.

In a randomised control trial, intention-to-treat (ITT) analysis is often employed in the study. The idea of ITT is to analyse the information based on the allocated groups of patients without accounting for the impact of having any possible changes to the treatment regimes of the trial. However, in practice, due to non-adherence, other treatment may be administered to the patients in correspondent to health related and ethical issues. As a result, ITT analysis may not reflect the true findings for the clinical study. For more details of ITT analysis, see for example Gupta (2011).

In a systematic review on the principle of ITT analysis in relation to missing data, Alshurafa et al. (2012) have found that only 36 out of 66 methodological articles have discussed the relationship of missing data with ITT analysis. Out of 78% of these articles suggested that ITT requires a specific strategy to account for the impact of the presence of missing values. White et al. (2011) have proposed some strategies to tackle the presence of missing values for ITT analysis in randomised trials. Other authors such as Thomas et al. (2000), Gardette et al. (2006), Salim et al. (2008) and Coley et al. (2011) have made some comparisons of missing data strategies in the context of clinical trials. The conclusion of these studies is that simple ad hoc methods do not yield reliable inference especially when the missing values are not missing completely at random. On the other hand, Mackinnon (2008) and Altman (2009) suggest that some kind of imputation for missing values using the available data might be required for ITT analysis, and Sterne et al. (2009) have described the potential and pitfalls of employing multiple imputation in the clinical research analysis.

In our investigation, we assume a reasonable rate of missing responses across the design region and the missing responses are generated by a monotone MAR mechanism. We consider complete case analysis, multiple imputation and available case analysis respectively at the design stage of some experiments. In the literature on data analysis, Yates (1933) is one of the earliest works that studies the analysis of variance for a randomised block and Latin square design respectively in the presence of missing data. Rather than ignoring the blocks that have missing responses, the author considers substituting a value into the incomplete data such that the corresponding sum of squares is minimised. Concerning the parameter estimation problem, Afifi and Elashoff (1966) show that approach that is considered in Yates (1933) gives the same estimated parameters as using the complete case analysis. Provided that the missing values are missing completely at random, i.e. the unobserved values are generated by a MCAR mechanism, most authors agree that complete case analysis does not cause bias in estimation. Little (1992) argues that if missing values are generated by a MAR mechanism and present only in the response variable of a linear regression model, complete case analysis yields reliable inferences for the regression analysis.

In Chapter 4 of this thesis, we aim to find an optimal design that minimises a function of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ are the least squares estimates for the unknown parameters of the linear regression model. These estimates are computed using complete observations that are being contributed by the experimental subjects. In Chapter 6, we want to find an optimal design that minimises a function of the total variance-covariance matrix of $\bar{\boldsymbol{\beta}}$, where $\bar{\boldsymbol{\beta}}$ are computed in the multiple imputation (Rubin (1987)) for the unknown parameters of the linear regression model. The general idea of multiple imputation and the combining rules for making inferences are available in the previous chapter. In general, a multiple imputation approach is an attractive approach especially in the context of multivariate data. Most of the literature on multiple imputation propose strategies to impute the missing responses when the incomplete data set has various types of variables, such as the categorical and the continuous variables. In our investigation, we assume the same model for both the imputation model and the analysis model. Concerning the number of repeated imputations, several authors such as Bodner (2008), Graham et al. (2007), and Kenward and Carpenter (2007) have discussed the issues on selecting this number. They show that the rate of missing responses and the tolerance of the power for hypothesis testing may play a role in choosing this value. In our investigation, we study the role of the number of repeated imputations in the design framework for the linear regression model. The optimal designs that are found by the two respective design framework for the linear regression model are compared in Chapter 6 of this thesis.

In Chapter 5 of this thesis, we investigate the optimal design framework for linear mixed models in the presence of dropouts. We employ linear and quadratic probability functions for the dropout processes of different cohorts. Assuming that available case analysis is employed to model the longitudinal data, we want to find the optimal designs that minimise some aspects of the covariance matrix of fixed effect parameters of the linear mixed models. To explore the design framework for different classes of the linear mixed models, we vary the structure of the covariance matrix of random coefficients. The probability functions and the covariance structures of random coefficients that are employed in our investigation are the same as those that are used in Ortega-Azurduy et al. (2008).

We now give some remarks about the missing probability functions. In our investigation of the design framework for the linear regression model, we employ the following inverse logit link function,

$$P(x_i) = \frac{exp(\gamma_0 + \gamma_1 x_i)}{1 + exp(\gamma_0 + \gamma_1 x_i)},$$

for some chosen values of $\gamma_0$ and $\gamma_1$. The inverse logit link function is a commonly used choice for modelling the missing data mechanism ( Ibrahim et al. (1999), Bang and Robins (2005), Mitra and Reiter (2011), Mitra and Reiter (2012)) as in practical situations, a logistic regression can be employed to estimate the parameters in the missing data model. Note that $\gamma_0 + \gamma_1 x_i$ in the exponential terms of $P(x_i)$ can be extended to higher order necessarily. For example, we could have $\gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2$ in the exponential

terms if the quadratic explanatory variable $x_i^2$ plays a role in causing a response to be missing at random, given the data.

Other commonly used MAR mechanisms that are being considered in design of experiments are the exponential probability function and the symmetric probability function (Baek et al. (2006), Hu et al. (2010)). We note that there are many other choices for modelling the missing data mechanism (Little (1995)), and our approach would be compatible with any choice of missing data model that is monotone. In view of designing a future experiment, one can use the data that are obtained from some historical/pilot studies to estimate the parameters of the missing data model. Alternatively, the missing proportion at some points in the design region can be elicited by the practitioners to find the corresponding MAR mechanism. For example, consider the missing proportions at two support points within the design region, the value of $\gamma_0$ and $\gamma_1$ in the above inverse logit link function can be found by solving the simultaneous equations.

## 3.5 Missing data problem at the design stage of a study

In some developmental research where the collected data plays a major role in the study, some research has been done to tackle the presence of missing values at the design stage of the study. For example, multiple matrix sampling, sometimes referred to as a split questionnaire, is a widely used method to construct a pre-planned missing data design, see for example Gonzalez and Eltinge (2007) for the review of this technique. The general idea of multiple matrix sampling is to spread the lengthy surveys or assessments across the participants of the study in order to reduce the rate of missing responses. To make inferences based on the information that are contributed by each group of participants, the strategies of missing data analysis can be employed to combined the information in the data analysis stage (Raghunathan and Grizzle (1995) and Little and Rhemtulla (2013)). Some examples of pre-planned missing data designs that employ multiple matrix sampling are multiform designs, wave missing, and two-method design.

Analogous to the idea of a pre-planned missing data design, our investigation on the optimal design framework accounts for the presence of missing responses and some features of the missing data analysis methods. The following shows the key contributions of our research.

- In Chapter 4, we propose a design framework for the linear regression model that accounts for complete case analysis. This framework includes the framework that is proposed by Imhof et al. (2002) as a special case.
- In Chapter 5, we propose cohort design framework for two special types of linear mixed models in the presence of dropouts. The design framework is different

from Ortega-Azurduy et al. (2008) in the context of model formulations and the applications.

- In Chapter 6, we assess the role of multiple imputation on the optimal designs for the linear regression model. To the best of our knowledge, no literature on design of experiments has considered multiple imputation when finding an optimal design.

The conclusions of this research and some suggestions for future research are presented in Chapter 7 of this thesis.

# Chapter 4

# Optimal designs for linear regression model with missing responses

To study the relationship between an explanatory variable and a response variable, a linear regression model is often employed to analyse the data of the experiment. In the situation where all outcomes will be observed, the inverse of the total information matrix of the model is commonly used as an approximation to the variance-covariance matrix of the estimated unknown parameters. However, when some of the responses may be missing, it is not clear if the inverse information matrix will be a good approximation to the observed variance-covariance matrix, i.e. the variance-covariance matrix after the experiment has been carried out. Hence it is not known if a design which is optimal with respect to a function of the expected information matrix will actually make the (observed) covariance matrix (or a function thereof) small.

In this chapter, we aim to tackle the issues of the presence of missing observations at the design stage of an experiment. We assume that missing responses are generated by a missing at random (MAR) mechanism. Considering complete case analysis and continuous design theory, we find optimal designs for the linear regression models based on an approximated information matrix (for the variance-covariance matrix of the least squares estimates). Here we first discuss the motivation and the set-up of our research problem. Following that, we present some analytical explanations about the expected information matrix in the presence of missing responses. Assuming that complete case analysis will be implemented for the incomplete data, we propose an optimal design framework for the linear regression models, which includes the method by Imhof et al. (2002) as a special case. We then give some examples of optimal designs for a simple linear regression model, which are found by the framework that assumes complete observations and our framework which accounts for the impact of missing responses

respectively. The performances of the designs are verified through simulation studies. Using the information that is obtained from an Alzheimer's disease study, we illustrate the application of our optimal design framework. Lastly we end this chapter with some discussion.

## 4.1 Set-up of problem

Let $\boldsymbol{y}$ be a vector of responses, $\boldsymbol{\beta}$ be a vector of the true but unknown parameters, $X$ be the design matrix that contains the values of the explanatory variable, and $\boldsymbol{\epsilon}$ be a vector of observational errors that are independent and identically distributed with a normal distribution that has mean zero and variance $\sigma^2$. The linear regression model is

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $X$ has rows $f^T(x_i) = (f_0(x_i), ..., f_p(x_i))$. An example of the model is a polynomial model of order $p$, which has the form

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_2} \\ \vdots \\ \vdots \\ y_{n_{m-1}+1} \\ \vdots \\ y_{n_m} \end{pmatrix} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_1 & \cdots & x_1^p \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_m & \cdots & x_m^p \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_m & \cdots & x_m^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_1} \\ \epsilon_{n_1+1} \\ \vdots \\ \epsilon_{n_2} \\ \vdots \\ \vdots \\ \epsilon_{n_{m-1}+1} \\ \vdots \\ \epsilon_{n_m} \end{pmatrix}$$

where

$$f^T(x_i) = \begin{pmatrix} 1 & x_i & \cdots & x_i^p \end{pmatrix},$$

$i = 1, ..., m$, is the index of the unique values of the explanatory variable, i.e. the unique support point $\{x_1, x_2, ..., x_m\}$ of an experimental design, $\xi$, and $\sum_{i=1}^{m} n_i = N$ is the total number of experimental units in the experiment.

Recall that the Fisher information matrix of a continuous design $\xi$ for the linear regression model that assumes fully observed responses is

$$M(\xi) = \sum_{i=1}^{m} f(x_i) f^T(x_i) \ \frac{n_i}{N} = \sum_{i=1}^{m} f(x_i) f^T(x_i) \ w_i,$$

where $w_i = n_i/N$ with $\sum_{i=1}^{m} w_i = 1$. For the fully efficient least squares estimator,

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y},$$

of the linear regression model, the conventional optimal design framework considers a function of $M(\xi)$ to minimise certain aspect of the precision of the estimates. This is because the inverse of $M(\xi)$ yields a good approximation to the variance of $\hat{\boldsymbol{\beta}}$, i.e.

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1}\sigma^2 \propto M^{-1}(\xi),$$

when information is fully observed. For example, a simple linear model with $f^T(x_i) = \begin{pmatrix} 1 & x_i \end{pmatrix}$ has

$$\begin{pmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) \\ cov(\hat{\beta}_0, \hat{\beta}_1) & var(\hat{\beta}_1) \end{pmatrix} \propto M^{-1}(\xi) = \frac{1}{|M(\xi)|} \begin{pmatrix} \sum_{i=1}^{m} w_i x_i^2 & -\sum_{i=1}^{m} w_i x_i \\ -\sum_{i=1}^{m} w_i x_i & \sum_{i=1}^{m} w_i \end{pmatrix} \quad (4.1)$$

where

$$|M(\xi)| = \sum_{i=1}^{m} w_i \sum_{i=1}^{m} w_i x_i^2 - (\sum_{i=1}^{m} w_i x_i)^2.$$

For nonlinear models, this equality is not exact but $(M(\xi))^{-1}/N$ is a good approximation to $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$. In particular for large sample sizes, we have

$$\boldsymbol{cov}(\sqrt{N}\hat{\boldsymbol{\beta}}) \to (M(\xi))^{-1}$$

as $N \to \infty$. Hence, to account for the uncertainty of the missing responses in the design framework, we consider the law of total variance that gives

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{E}(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})) + \boldsymbol{cov}(\boldsymbol{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})),$$

where $\boldsymbol{\mathcal{M}} = (\mathcal{M}_1, ..., \mathcal{M}_N)$ is a vector of missing data indicator,

$$\mathcal{M}_i = \begin{cases} 1, & \text{for } y_i \text{ is missing;} \\ 0, & \text{otherwise,} \end{cases}$$

in the least squares estimates. The inner moments on the right hand side of the above expression consider the conditional distribution of the estimates given the observed experimental data, whereas the outer moments average the uncertainty across the possible outcome of the random missing data indicators when data is unobserved.

Notice that if the values of $\mathcal{M}_i$ are such that $X^T X$ is non-singular, we have $\boldsymbol{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})) = \boldsymbol{\beta}$ by the property of the unbiased estimators. Hence we have $\boldsymbol{cov}(\boldsymbol{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})) = \boldsymbol{cov}(\boldsymbol{\beta}) = 0$ as there is no uncertainty in the fixed but unknown parameter $\boldsymbol{\beta}$, resulting in

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{E}(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})).$$

Otherwise, an experiment provides no benefits when $\boldsymbol{\beta}$ is not estimable by the singular Fisher information matrix that arose from the presence of missing observations. In our investigation, we assume a moderate rate of missing responses such that $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ exists, i.e. $M(\xi)$ is invertible. The aim of this chapter is to find $\boldsymbol{E}(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}}))$ before the consideration of a design criterion in the design framework. We assume that complete case analysis will be implemented at the data analysis stage of an experiment. This approach discards partially observed information of the experimental units in the analysis. It is an appealing method due to its simplicity, and is available in most of the statistical software. Moreover, Little (1992) suggests that a complete case analysis will still result in unbiased estimates for the regression coefficients even when the outcome values are missing at random.

In the next section, we present some analytical explanations about the expected information matrix, and justify an optimal design framework for general MAR mechanisms. This optimal design framework includes the method by Imhof et al. (2002) as a special case.

## 4.2 Optimal design with complete case analysis

Recognising that the total information that is being collected at a support point $x_i$ is

$$f(x_i)f^T(x_i)\ w_i = f(x_i)f^T(x_i)\ \frac{n_i}{N}$$

where $n_i$ is the number of observed responses at $x_i$ and $N$ is the total number of experimental units, the presence of missing observations causes $n_i$ to be a summation of the random missing data indicators when the responses are yet to be observed. Hence, we can express the information at $x_i$ as

$$f(x_i)f^T(x_i)\ \frac{\sum_{r=1}^{n_i}(1-\mathcal{M}_r)}{N} = f(x_i)f^T(x_i)\ \frac{Z_i}{N},$$

where $\mathcal{M}_r$ is a Bernoulli random variable with a probability, say $P(x_i)$, and $Z_i = \sum_{r=1}^{n_i}(1-\mathcal{M}_r)$ follows a Binomial distribution with mean $n_i = Nn_i/N = Nw_i$ and probability $(1-P(x_i))$.

Since the inverse of the information matrix conditional on all the missing data indicators is approximately proportional to $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ provided that the information matrix is invertible, we consider

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{E}(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})) = E\{[M(\xi, Z_i)]^{-1}\}\sigma^2$$

where

$$M(\xi, Z_i) = \sum_{i=1}^{m} f(x_i)f^T(x_i)\ Z_i \tag{4.2}$$

is a revised version of the Fisher information matrix at the design stage of an experiment. This matrix is equivalent to the matrix $X^TX$ when the missing data indicators are incorporated into the design matrix $X$.

As the inversion of a matrix involves the division of its determinant, $[M(\xi, Z)]^{-1}$ contains ratios of random variables. For example, we have

$$[M(\xi, Z)]^{-1} = \frac{1}{(x_1 - x_2)^2 Z_1 Z_2} \begin{pmatrix} x_1^2 Z_1 + x_2^2 Z_2 & -x_1 Z_1 - x_2 Z_2 \\ -x_1 Z_1 - x_2 Z_2 & Z_1 + Z_2 \end{pmatrix} \tag{4.3}$$

for the simple linear regression analysis with two support points $x_1$ and $x_2$. To find $\boldsymbol{E(cov(\hat{\beta}|\mathcal{M}))}$, an approximation of the inverse of (4.2) is required before the expectation. This is because the distribution of $Z_i/(Z_iZ_j)$ for $i \neq j$ is not analytically available due to the small probability that $Z_i$, $i = 1, 2$, becomes zero (when all observations at the support point are missing). To tackle this issue, we employ a multivariate second-order Taylor series expansion in the approximation, in the hope that the approximated $[M(\xi, Z)]^{-1}\sigma^2$ is close to the observed value $\boldsymbol{cov(\hat{\beta}|\mathcal{M})}$.

As an example, consider $Z_i/(Z_iZ_j)$ as a function that depends on two variables $Z_i$ and $Z_iZ_j$, the Taylor series to a second order about their mean, $E\{Z_i\}$ and $E\{Z_iZ_j\} = E\{Z_i\}E\{Z_j\}$, is

$$\begin{aligned} &\frac{E\{Z_i\}}{E\{Z_i\}E\{Z_j\}} + (Z_i - E\{Z_i\}) \left( \frac{1}{E\{Z_i\}E\{Z_j\}} \right) \\ &\quad - (Z_iZ_j - E\{Z_i\}E\{Z_j\}) \left( \frac{E\{Z_i\}}{(E\{Z_i\}E\{Z_j\})^2} \right) \\ &\quad + (Z_iZ_j - E\{Z_i\}E\{Z_j\})^2 \left( \frac{E\{Z_i\}}{(E\{Z_i\}E\{Z_j\})^3} \right) \\ &\quad - (Z_i - E\{Z_i\})(Z_iZ_j - E\{Z_i\}E\{Z_j\}) \left( \frac{1}{(E\{Z_i\}E\{Z_j\})^2} \right). \end{aligned}$$

Taking expectation with respect to all the random missing data indicators in the above expression and with some algebraic simplification, we obtain

$$\begin{aligned} E\left(\frac{Z_i}{Z_iZ_j}\right) &\approx \frac{1}{E\{Z_j\}} + \frac{E\{Z_i^2\}Var(Z_j)}{(E\{Z_i\})^2(E\{Z_j\})^3} \\ &= \frac{1}{Nw_j(1 - P(x_j))} + \frac{P(x_i)P(x_j)}{Nw_i(1 - P(x_i))(Nw_j(1 - P(x_j)))^2} + \frac{P(x_j)}{(Nw_j(1 - P(x_j)))^2} \end{aligned} \tag{4.4}$$

for $i \neq j$ and $P(x_i)$ is the MAR mechanism. A full derivation of this result is given in Appendix A.1. If responses are missing completely at random (MCAR), this expression

simplifies to

$$E\left(\frac{Z_i}{Z_iZ_j}\right) \approx \frac{1}{Nw_j(1-P)} + \frac{P^2}{Nw_i(1-P)(Nw_j(1-P))^2} + \frac{P}{(Nw_j(1-P))^2}$$
$$= \frac{1}{N'w_j} + \frac{P^2}{N'w_i(N'w_j)^2} + \frac{P}{(N'w_j)^2} \tag{4.5}$$

where $N' = N(1-P)$ corresponds to a reduced total sample size of the experiment and $P$ is the probability that a response is missing completely at random.

Note that when the fraction of random variables is simplified algebraically as if in the deterministic situation, and apply the second order Taylor series approximation, we have

$$E\left(\frac{Z_i}{Z_iZ_j}\right) = E\left(\frac{1}{Z_j}\right) \approx \frac{1}{E\{Z_j\}} + \frac{Var(Z_j)}{(E\{Z_j\})^3} = \frac{1}{Nw_j(1-P(x_j))} + \frac{P(x_j)}{(Nw_j(1-P(x_j)))^2}$$

for the experiment with a MAR mechanism;

$$E\left(\frac{Z_i}{Z_iZ_j}\right) = E\left(\frac{1}{Z_j}\right) \approx \frac{1}{E\{Z_j\}} + \frac{Var(Z_j)}{(E\{Z_j\})^3} = \frac{1}{Nw_j(1-P)} + \frac{P}{(Nw_j(1-P))^2}$$

for the experiment with a MCAR mechanism. Comparing these expressions with (4.4) and (4.5) respectively, cancelling the common random variable in the fraction will give relatively smaller values in the approximation. We also note that the expression with a MCAR mechanism is independent of the support points in both situations. This suggests that the conventional way of scaling up the total sample size for accounting the presence of missing responses is valid when the responses are MCAR.

In the literature, Imhof et al. (2002) consider an expectation of (4.2) with respect to the random missing data indicators. They argue that in the presence of missing responses, the expected information that is to be collected at a support point $x_i$ is

$$E\left\{f(x_i)f^T(x_i)\ \frac{Z_i}{N}\right\} = \frac{f(x_i)f^T(x_i)}{N}E\{Z_i\}$$
$$= \frac{f(x_i)f^T(x_i)}{N}n_i(1-P(x_i))$$
$$= f(x_i)f^T(x_i)w_i(1-P(x_i)),$$

which yields

$$E\{M(\xi,Z)\} = \sum_{i=1}^{m} f(x_i)f^T(x_i)\ w_i(1-P(x_i)) = M(\xi, E\{Z\})$$

as the total expected Fisher information. Hence they consider that

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{E}(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})) \propto [E\{M(\xi,Z)\}]^{-1}$$

by the notion of Cramér-Rao inequality. Maximising a function of $[E\{M(\xi, Z)\}]^{-1}$, such as the determinant of $E\{M(\xi, Z)\}$ for $D$-optimality, the authors employ the continuous design theory to find an optimal design. This approach implicitly employs a first-order Taylor series expansion in the approximation, where

$$E\left\{[M(\xi, Z)]^{-1}\right\} \approx [E\{M(\xi, Z)\}]^{-1},$$

in which $E(Z_i/(Z_iZ_j)) = 1/E(Z_j)$ and only the first element on the right hand side of (4.4) and (4.5) are involved in the framework.

To vindicate the work that is proposed by Imhof et al. (2002) on $D$-optimality, we establish the following theorem and lemma so that the general equivalence theorem which is introduced in Chapter 2 for the conventional optimal design can be used to verify the design when a monotone MAR mechanism is considered at the design stage of an experiment. Theorem (1) shows that for a large class of MAR mechanisms and polynomial models, the $D$-optimal design that is found by using a first order approximation in the design framework has the same number of support points as the number of parameters of the polynomial model. This result corresponds to the contribution of De la Garza (1954) and Silvey (1980) in the conventional optimal design framework for finding the number and the weight of support points of an optimal design. The proof of Theorem (1) can be found in Appendix A.2.

*Theorem* 1. Let $h(x) = \frac{1}{1-P(x)}$ and $h^{(2p)}(x)$ be the derivative of order $(2p)$, assume that the equality $h^{(2p)}(x) = c$ has at most one solution for every constant $c \in \Re$. For the polynomial normal linear regression model of degree $p$ with a monotone MAR mechanism $P(x)$, a $D$-optimal design has exactly $p + 1$ support points with equal weights at each of them.

Hence design search can be restricted to $(p + 1)$-point designs, with known weights $w_i = 1/(p + 1)$, $i = 1, ..., p + 1$. A further simplification is given in Lemma (2), which shows that under the assumptions of Theorem (1), if the MAR mechanism is monotone, one of the bounds of the design region is a support point of the $D$-optimal design. The idea of the lemma is that since we know the monotone MAR mechanism, we know the location over the design region where the maximum information can be collected. Other support points of the optimal design can be found by considering the design problem as an optimisation problem, with constraints $\sum_{i=1}^{m} w_i = 1$ and the range of $\mathfrak{X}$.

*Lemma* 2. Let $P(x)$ be a probability function that satisfies the conditions in Theorem 1 and $[l, u]$ be the design interval $\mathfrak{X}$. If $P(x)$ is strictly increasing, then the lower bound, $l$, is a support point. If $P(x)$ is strictly decreasing, then the upper bound, $u$, is a support point.

*Proof.* For a continuous design $\xi$ with $p+1$ support points, we have

$$|E\{M(\xi, Z)\}| = |M(\xi, E\{Z\})| \quad = \quad \prod_{i=1}^{p+1} w_i(1 - P(x_i)) \prod_{1 \le i < j \le p+1} (x_i - x_j)^2 \quad (4.6)$$

where we order the $p+1$ values for $x$ by size:

$$l \le x_1 < x_2 < ... < x_{p+1} \le u.$$

If $P(x)$ is a monotonic increasing function, $(1 - P(x))$ will be the largest at $x_1 = l$ and $(x_1 - x_j)^2$ will also be the largest for $x_1 = l$, for all values of $x_j$ where $j \le p+1,\ j \ne 1$. If $P(x)$ is a monotonic decreasing function, we get the largest $(1 - P(x))$ at $x_{p+1} = u$, and $(x_i - x_{p+1})^2$ will be the largest for $x_{p+1} = u$, for all values of $x_i$ where $i < p+1$. Hence, we are maximising $|E\{M(\xi, Z)\}|$ in both of these situations. □

As mentioned in Chapter 3, the consideration of a function of $M(\xi, E\{Z\})$ in the design framework that is proposed by Imhof et al. (2002) coincides with those of the design framework for the weighted polynomial regression model. Consequently, the conventional optimal design theory that is illustrated in Chapter 2 can be applied accordingly to find the optimal design when a MAR mechanism is considered at the design stage of an experiment. For example, the general equivalence theorem can be used to verify if $\xi^*$ is $D$-optimal for the experiment that assumes a monotone MAR mechanism over a design region $[l, u]$, i.e. this is true if and only if

$$f^T(x)\ [M(\xi^*, E\{Z\})]^{-1}\ f(x)(1 - P(x)) \le p+1 \quad \forall x \in [l, u].$$

To illustrate this, consider a two-point $D$-optimal design that is constructed based on $|M(\xi, E\{Z\})|$ for a simple linear model with a monotonic increasing MAR mechanism. Figure 4.1 gives a geometric illustration of the revised version of the general equivalence theorem. The plot on the left shows three inverse logit link MAR mechanisms with a common $\gamma_0$; the plot on the right shows the corresponding standardised variances of the estimates that are computed using the corresponding $D$-optimal designs. In all of these cases, the lower bound of $\mathfrak{X}$ is chosen as one of the support points according to Lemma (2). The other support point is found for the respective experiments with the MAR mechanisms, and is denoted by the marker on the corresponding functions on the left plot in Figure 4.1. The interpretation of this revised version of the general equivalence theorem is analogous to the one for the conventional design framework, i.e. the largest standardised variance over the design region is obtained at the support points of the $D$-optimal design, and it is always less than or equal to the number of parameters of the model elsewhere within $\mathfrak{X}$.

In the next section, we first illustrate the $c$- and $D$-optimal design for the simple linear regression model with the assumption of fully observed responses. We then compare

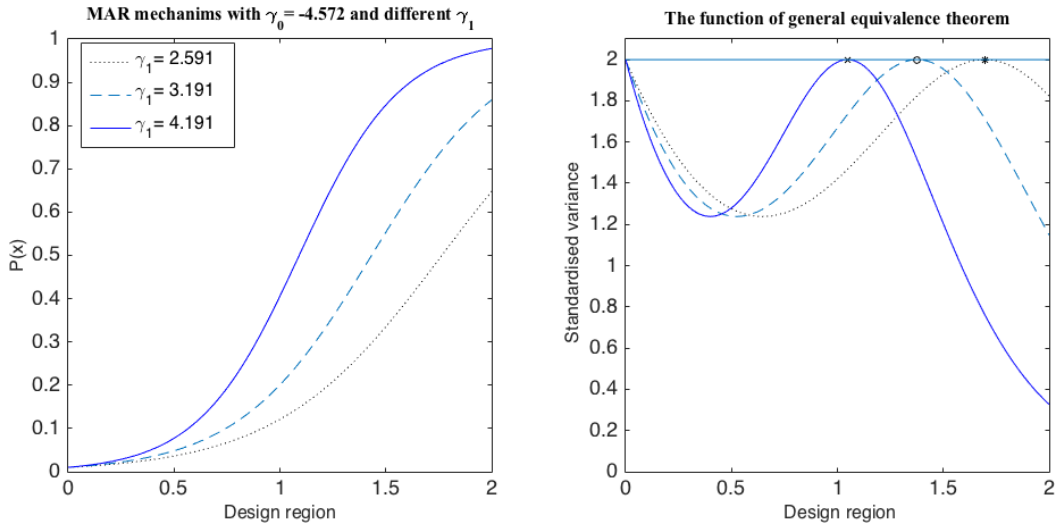



Figure 4.1: Illustration of the revised general equivalence theorem (right) for $D$-optimal designs that assume monotonic increasing MAR mechanisms (left).

and contrast the design framework that is proposed by Imhof et al. (2002) with our design framework, where we approximate $[M(\xi, Z)]^{-1}$ before considering a function of $E\{[M(\xi, Z)]^{-1}\}$ for finding an optimal design. In particular, we consider two support points are sufficient for the parameter estimation of the simple linear regression model, and from Theorem 1, the $D$-optimal designs based on $E\{M(\xi, Z)\}$ are two-point designs for a large variety of MAR mechanisms. Hence finding the best two-point design for the second order approximation facilitates the comparison of the two approaches.

## 4.3 Illustration and simulation: simple linear regression model

We now illustrate the optimal design framework for the simple linear regression model where

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.7}$$

is the single response of experimental unit $i$, $i = 1, ..., n_m$. With the assumption of fully observed responses, the design problem which minimises $var(\hat{\beta}_1)$ of this model employs a $c$-optimality in finding $x_i$ and $w_i$ such that the second diagonal element of $M^{-1}(\xi)$ in (4.1), i.e.

$$M^{-1}(\xi)_{[2,2]} = \frac{1}{|M(\xi)|} \sum_{i=1}^{m} w_i = \frac{1}{|M(\xi)|},$$

is minimum over the design region $\mathfrak{X}$ at the support points of the optimal design; and minimises the volume of the confidence ellipse due to $D$-optimality, which minimises

$$|M^{-1}(\xi)| = \frac{1}{|M(\xi)|^2}|M(\xi)| = \frac{1}{|M(\xi)|}$$

over $\mathfrak{X}$. These optimisation problems are subject to the constraint $\sum_{i=1}^{m} w_i = 1$.

Note that from the second equality of $D$-optimality, an optimal design can be found without computing the inverse of the Fisher information matrix due to the above property of the determinant of an invertible matrix that contains no randomness. Assuming responses are fully observed, the general equivalence theorem and the theorem that is derived by Silvey (1980) (see Lemma 5.1.3 on pg 42) show that a $D$-optimal design for the simple linear model with two support points has $w_1 = 1/2 = w_2$, and the bounds of $\mathfrak{X}$ are the support points of the optimal design. In particular, for $\mathfrak{X} = [-1, 1]$, having half replication of observations at $x_1 = -1$ and $x_2 = 1$ respectively will grant maximum information for the experiment that minimises $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ of the simple linear regression model.

On the other hand, since the objective function of $c$-optimality for minimising $var(\hat{\beta}_1)$ is equivalent to the objective function of $D$-optimality for the simple linear model, the aforementioned $D$-optimal design is also $c$-optimal. Note that this is not true for the linear regression model with $p > 1$ when $c$-optimality that minimises other linear combination of the variances of the estimated parameters is considered, such as minimising $var(\hat{\beta}_1)$ and $var(\hat{\beta}_3)$ for the cubic regression model.

Accounting for the presence of missing responses, we now illustrate the optimal design framework for the simple linear regression model. To fix ideas, for a design region $\mathfrak{X} = [l, u]$ where $l < u$, consider total sample size $N$ and two support points $x_1$ and $x_2$. Let $n_1 = Nw_1$ responses, $\{y_1, \ ..., \ y_{n_1}\}$, be taken at experimental condition $x_1$, and $n_2 = N - n_1 = N(1 - w_1) = Nw_2$ responses, $\{y_{n_1+1}, \ ..., \ y_N\}$, be measured at $x_2$. We seek an optimal design

$$\xi^* = \begin{Bmatrix} x_1 & x_2 \\ w_1 & w_2 \end{Bmatrix}$$

by minimising a function of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{E}(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}}))$, after selecting a specific missing data mechanism $P(x)$. The optimal design $\xi^*$ can be found by taking derivatives of the criterion with respect to the support points and the weights respectively, with constraints $w_1 + w_2 = 1$ and $x_2 > x_1 \in \mathfrak{X}$. For example, a $D$-optimal design minimises the determinant of the expected value of (4.3), which is

$$\frac{1}{(x_1 - x_2)^2} E\left(\frac{Z_1}{Z_1 Z_2}\right) E\left(\frac{Z_2}{Z_1 Z_2}\right), \tag{4.8}$$

over $\mathfrak{X}$; a $c$-optimal design for minimising $var(\hat{\beta}_1)$ minimises

$$\frac{1}{(x_1 - x_2)^2}\left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) \tag{4.9}$$

over $\mathfrak{X}$; an $A$-optimal design for minimising $var(\hat{\beta}_1) + var(\hat{\beta}_2)$ minimises

$$\frac{1}{(x_1 - x_2)^2}\left((x_1^2+1)E\left(\frac{Z_1}{Z_1 Z_2}\right) + (x_2^2+1)E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) \tag{4.10}$$

over $\mathfrak{X}$, where the expectations are approximated by either a first-order (i.e. the framework that is proposed by Imhof et al. (2002)) or a second-order Taylor series expansion, and are depending on the form of the missing data mechanism.

Theorem 3 shows that the $D$-, $c$- and $A$-optimal two-point designs that are found based on the second order expansion have a similar structure to the corresponding first order designs. The $c$-optimal design minimises the estimated slope parameter of the simple linear model.

*Theorem* 3. For the simple linear regression model (4.7), assume we approximate $E\{[M(\xi, Z)]^{-1}\}$ by a second order Taylor expansion, and let the design interval $\mathfrak{X} = [l, u]$. Restrict design search to two-point designs.

(a) If the missing data mechanism is MCAR, then the $D$- and the $c$-optimal design are equally weighted on $l$ and $u$. If, in addition, $l \geq 0$ or $u \leq 0$, the $A$-optimal design will also have support points $l$ and $u$.

(b) If the missing data mechanism is MAR and monotone increasing (decreasing), then $l$ ($u$) is a support point of the $D$- and the $c$-optimal design. If, in addition $l \geq 0$ ($u \leq 0$), this result also holds for $A$-optimality.

The proof of part (a) of Theorem 3 is given in Appendix A.3.

**Proof of Theorem 3 (b).** Let without loss of generality $x_1 < x_2$, and assume $P(x)$ is monotone increasing. From (4.4), it can be seen that the second order approximation for $E[Z_1/(Z_1Z_2)]$ and for $E[Z_2/(Z_1Z_2)]$ are both increasing in $x_1$ and are hence minimised when $x_1 = l$. Since $x_1 = l$ also minimises $1/(x_1 - x_2)^2$, and all expressions are non-negative, the objective functions in (4.8) and (4.9) are both minimised when $x_1 = l$. If $l \geq 0$, $(x_1^2 + 1)$ is also increasing in $x_1$, and the result for $A$-optimality follows.

An analogous argument shows that $x_2 = u$ minimises (4.8), (4.9) and, for $u \leq 0$, also (4.10) if $P(x)$ is monotone decreasing. □

From part (a), we find that the optimal designs for the simple linear regression model with MCAR assumption are the same as those that assume complete observations. In part (b), we find that the lower/upper limit of the design interval is a support point of

the optimal designs that assume responses are missing at random. We infer that these optimal designs have the same support structure as the first order design from Lemma 2. However, the weights and the other support point may have different values and may depend on the total sample size $N$.

We now demonstrate some designs for the two respective approximation strategies and illustrate their performance through simulations. Assuming that a simple linear regression model will be fitted to the complete case data of an experiment, obtaining estimates of the coefficients, $(\hat{\beta}_0, \hat{\beta}_1)$, and their variances, from the available cases, i.e. using only those units for which $y_i$ is observed, we simulate a response variable by

$$y_i = 1 + x_i + \epsilon_i,$$

i.e. $\beta_0 = \beta_1 = 1$, where $\epsilon_i \sim N(0, \sigma^2)$, $x_i$ are the support points of a given two-point design within the design region $\mathfrak{X} = [0, u]$, sample size $N = 30$, and $\sigma^2 = 1$. We then introduce missing values into the simulated $y_i$, $i = 1, \ldots, 30$, by specifying a MAR mechanism through the following inverse logit link model,

$$P(x_i) = \frac{exp(\gamma_0 + \gamma_1 x_i)}{1 + exp(\gamma_0 + \gamma_1 x_i)},$$

with $\gamma_0 = -4.572$ and $\gamma_1 = 3.191$. The positive value of $\gamma_1$ indicates the mechanism is monotone increasing with $x_i$. Note that there are many other choices for modelling the missing data mechanism, see for example Little (1995). Our approach would be compatible with any monotone missing data mechanisms.

We first consider several designs $\xi = \{0, x_2\}$, and under each design, compare the two proposed approaches for approximating elements of the matrix that are specified in $E\{[M(\xi, Z)^{-1}]\}$, as well as some other relevant functions of this matrix. The two proposed approaches approximate $E\left(\frac{Z_i}{Z_i Z_j}\right)$ in the elements of $E\{[M(\xi, Z)^{-1}]\}$ with first order and second order Taylor expansions respectively, where the first order expansion corresponds to the approach that is proposed by Imhof et al. (2002) and the latter corresponds to (4.4). Specifically for each design, we repeatedly simulate incomplete data using the above described models and empirically obtain the estimates for $E\{[M(\xi, Z)^{-1}]\}$ by averaging the elements in $M(\xi, Z)^{-1}$ (see (4.3)), across the simulated data. Treating these as the true values, we can then compare the two approximations.

Table 4.1 presents the simulation results over 200 000 sets of simulated data from two different designs where $x_2 = 1$ and $x_2 = 1.5$ respectively. Consider the design where $x_2 = 1.5$, we find that for the $[2, 2]$ element in $E\{[M(\xi, Z)^{-1}]\}$ which is the objective function of $c$-optimality that minimises $var(\hat{\beta}_1)$, the first order approximation has a bias of 7.2%, while for the second order approximation this bias has reduced to 1.9%. For this same design, the trace of $E\{[M(\xi, Z)^{-1}]\}$ (which corresponds to the objective function of $A$-optimality) has a bias of 4.4% and the determinant of the matrix (which

Table 4.1: Comparison of approximations for the two designs with weight $w_1 = 0.5 = w_2$, $P(x_1) = P(0) = 0.01$, and $N = 30$. First line in each row corresponds to the simulation output.

| $\xi$ | $\{0, 1\}$ | $\{0, 1.5\}$ |
|---|---|---|
| $[1, 1]$ element of expected value of (4.3) | 0.06740 | 0.06740 |
| First order Taylor series approximation | 0.06736 | 0.06736 |
| Second order Taylor series approximation | 0.06740 | 0.06741 |
| $[2, 2]$ element of expected value of (4.3) | 0.15242 | 0.10375 |
| First order Taylor series approximation | 0.15078 | 0.09628 |
| Second order Taylor series approximation | 0.15222 | 0.10179 |
| $[1, 2]$ element of expected value of (4.3) | -0.06740 | -0.04494 |
| First order Taylor series approximation | -0.06736 | -0.04490 |
| Second order Taylor series approximation | -0.06740 | -0.04494 |
| Determinant of expected value of (4.3) | 0.00573 | 0.00497 |
| First order Taylor series approximation | 0.00562 | 0.00447 |
| Second order Taylor series approximation | 0.00572 | 0.00484 |
| Frequency of $M(\xi, Z_i)$ becomes singular | 0 | 23 |
| $P(x_2)$ | 0.20085 | 0.55342 |

corresponds to the objective function of $D$-optimality) has a bias of 10.1% when using the first order approximation, while the biases are reduced to 1.1% and 2.6% respectively when using the second order approximation. In general, we can see that using the second approximation yields better approximations of the elements of $E\{[M(\xi, Z)^{-1}]\}$ and some relevant functions of the matrix.

We now consider the optimal design problem for the experiment with the aforementioned missing data mechanism over the design region $\mathfrak{X} = [0, u]$. Using Theorem 1 and Lemma 2, the lower bound of $\mathfrak{X}$, 0, is chosen as one of the support points of the two-point optimal design, denoted by $x_1$ here. Substituting the MAR mechanism, the value of $x_1$ and $w_1 = 1 - w_2$ into the corresponding elements of $E\{[M(\xi, Z)^{-1}]\}$, we find the optimal values for $x_2$ and $w_2$ by minimising a function of this matrix with respect to $x_2$ and $w_2$ in *Mathematica* with the *Minimize* function. Table 4.2 presents these optimal values for the $A$-, $c$- and $D$-optimal designs respectively for the experiment that has $N = 30$. We see that using the second order approximation results in an upper design point that is smaller than the upper design point when using the first order approximation. Figure 4.2 shows more examples of the optimal values for the experiments that have different sample sizes and different $\gamma_1$ in the logistic link function with $\gamma_0 = -4.572$. When $N$ is large, we find that the optimal design settings that are found based on the second order expansion are very similar to the those that are obtained using first order approximation.

To illustrate performance, for each design given in Table 4.2, we repeatedly simulate the incomplete data 200 000 times as described above. In each incomplete data set, we compute the sample estimate $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ from the complete cases, i.e. where the design matrix $\boldsymbol{X}$ and response vector $\boldsymbol{y}$ comprise only units with observed response

Table 4.2: Optimal designs that are found by the two respective design framework, for $N = 30$, $\gamma_0 = -4.572$, $\gamma_1 = 3.191$, $x_1 = 0$, $P(x_1) = 0.01$, and $w_1 = 1 - w_2$.

| | $\xi^*_{A\ 2nd}$ | $\xi^*_{A\ 1st}$ | $\xi^*_{c\ 2nd}$ | $\xi^*_{c\ 1st}$ | $\xi^*_{d\ 2nd}$ | $\xi^*_{d\ 1st}$ |
|---|---|---|---|---|---|---|
| $x_2$ | 1.46206 | 1.51466 | 1.54924 | 1.60059 | 1.33597 | 1.37660 |
| $w_2$ | 0.4665 | 0.4539 | 0.6257 | 0.6208 | 0.5110 | 0.5 |
| $P(x_2)$ | 0.5233 | 0.5650 | 0.5919 | 0.6308 | 0.4234 | 0.4553 |

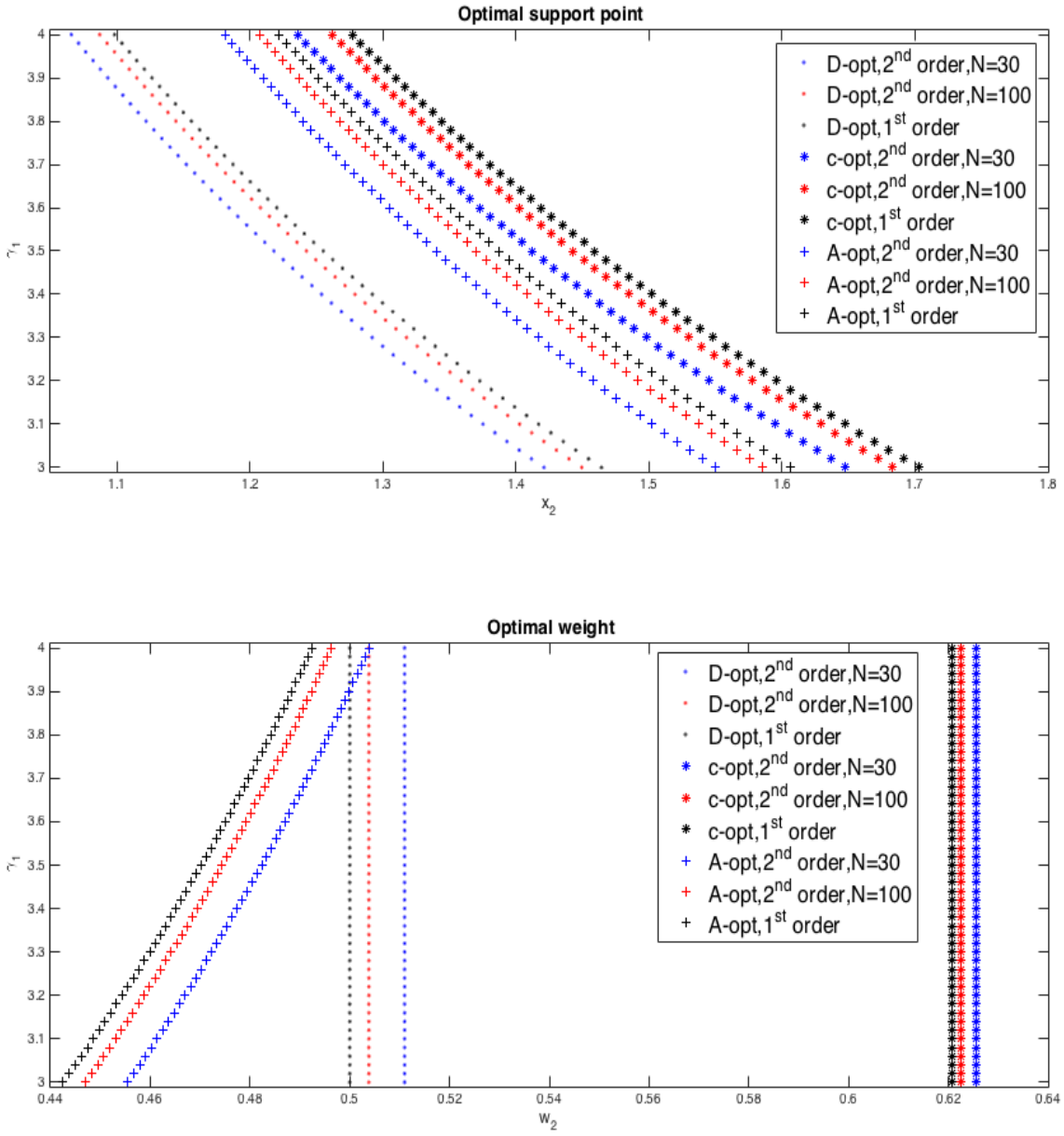



Figure 4.2: Optimal values for the experiments that have different $\gamma_1$ in the inverse logit link function with $\gamma_0 = -4.572$ (MAR mechanism). The other support point is $x_1 = 0$ with weight $w_1 = 1 - w_2$.

values. We can then empirically obtain the covariance matrix for $\hat{\boldsymbol{\beta}}$ across the replications. Table 4.3 summarises the performance of the designs derived under the different optimality criteria and approximations. We see that the designs that are obtained under $A$-optimality have the smallest trace of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$, as expected. Further, this trace is smaller when using the design that is found from the second order approximation rather than the first order approximation. This pattern is repeated for the other optimality criteria. The design that is obtained under $c$-optimality from the 2nd order approximation results in the smallest variance for $\hat{\beta}_1$, and the design that is found under $D$-optimality from the 2nd order approximation results in the smallest determinant of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$. In addition, in general the optimal designs that are constructed based on the second order approximation resulted in fewer cases where it was not possible to estimate the parameters, due to singular matrix that is caused by the presence of missing data.

Table 4.3: Simulation outputs across 200 000 simulated data for different designs. The last row indicates the frequency where $M(\xi, Z)$ becomes singular.

| | $\xi^*_{A\ 2nd}$ | $\xi^*_{A\ 1st}$ | $\xi^*_{c\ 2nd}$ | $\xi^*_{c\ 1st}$ | $\xi^*_{d\ 2nd}$ | $\xi^*_{d\ 1st}$ |
|---|---|---|---|---|---|---|
| sample $var(\hat{\beta}_1)$ (e-01) | 1.0687 | 1.0823 | **0.97349** | 0.98102 | 1.0401 | 1.0486 |
| $tr$(sample $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$) (e-01) | **1.6988** | 1.7123 | 1.8893 | 1.8968 | 1.7590 | 1.7197 |
| \|sample $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$\| (e-03) | 4.8764 | 5.0880 | 5.4165 | 5.7121 | **4.5809** | 4.6526 |
| No. of cases failed | 19 | 67 | 16 | 35 | 0 | 2 |

We have empirically evaluated the framework that is proposed to construct optimal designs in the presence of missing values under the assumption of a complete case analysis. We have seen that this framework worked well in the simulations, with evidence suggesting that the second order approximation, at least in these simulations, had the potential to provide better approximations and hence result in better designs. We note that in this example, the optimal designs found by employing the second order approximation to $\frac{1}{Z_j}$ are very similar to those that consider second order approximation to $\frac{Z_i}{Z_i Z_j}$ before taking the expectation, i.e. $x_2$ for the design criteria illustrated here found by both approaches are equivalent up to three significant figures. For this reason, we do not present the simulation output here.

## 4.4 Application: Redesigning a study on Alzheimer's disease

In this section we consider a scenario that is motivated from an application which concerned with designing a clinical trial to treat Alzheimer's disease. To illustrate an application of our approach, we use data from an Alzheimer's disease study which investigated the benefits of administering the treatments donepezil, memantine, and the combination of the two, to patients over a period of 52 weeks, on various quality of life measures. See Howard et al. (2012) for full details of the study. The total number

of patients included in the primary intention-to-treat sample was 291, with 72 in the placebo group (Group 1), 74 in the memantine treatment group (Group 2), 73 in the donepezil treatment group (Group 3), and 72 in the donepezil-memantine group (Group 4).

In the per-protocol analysis, 43 patients were excluded in Group 1, 32 in Group 2, 23 in Group 3 and 21 in Group 4. Considering these patients as data missing at random, a logistic regression model is fitted to the data, specifically

$$P(\mathcal{M}_i = 1|x_i, v_i) = \frac{exp(\gamma_0 + \gamma_1 x_i + \gamma_2 v_i)}{1 + exp(\gamma_0 + \gamma_1 x_i + \gamma_2 v_i)}$$

where $x_i, v_i \in \{0, 1\}$ represent the level of donepezil and memantine respectively (with 1 indicating the treatment is applied) for patient $i$. From the data the regression coefficients were estimated to be $\hat{\gamma}_0 = 0.2636472$, $\hat{\gamma}_1 = -0.8988845$ and $\hat{\gamma}_2 = -0.4108504$.

For illustration purposes, we assume a multiple regression model without an interaction term will fit the data, i.e.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{4.11}$$

where $Y_i$ corresponds to the outcome value for patient $i$, and $\sigma^2$ is assumed to be known and fixed to be 1 without loss of generality. We define the four groups ($G_1$ - $G_4$) the units are allocated to in terms of the design variables $x$ and $v$:

- $G_1$: $x_i = 0, v_i = 0$ with $n_1$ experimental units;
- $G_2$: $x_i = 0, v_i = 1$ with $n_2$ experimental units;
- $G_3$: $x_i = 1, v_i = 0$ with $n_3$ experimental units;
- $G_4$: $x_i = 1, v_i = 1$ with $n_4$ experimental units.

In this situation we have thus fixed the design points, which are defined by the values of $(x, v)$ and are equal to $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$ respectively. The design problem is then to find the optimal number of patients to allocate to Groups $G_1$ - $G_4$, denoted by $n_1$, $n_2$, $n_3$, and $n_4$ respectively, under the assumption hat the analyst fits a linear regression model of the form described in (4.11) using the complete cases. The diagonal elements of the (conditional) covariance matrix of the least squares estimators for this model are

$$var(\hat{\beta}_0|\boldsymbol{\mathcal{M}}) = \frac{Z_2Z_3 + Z_2Z_4 + Z_4Z_3}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$var(\hat{\beta}_1|\boldsymbol{\mathcal{M}}) = \frac{(Z_2 + Z_4)(Z_1 + Z_3)}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$var(\hat{\beta}_2|\boldsymbol{\mathcal{M}}) = \frac{(Z_3 + Z_4)(Z_1 + Z_2)}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

where $Z_k = \sum_{r \in G_k}(1 - \mathcal{M}_r)$ is the sum of the response indicators for Group $G_k$, $k = 1, \ldots, 4$. The $A$-optimal design for this model minimises an appropriate approximation to

$$E\{var(\hat{\beta}_0|\boldsymbol{\mathcal{M}})\} + E\{var(\hat{\beta}_1|\boldsymbol{\mathcal{M}})\} + E\{var(\hat{\beta}_2|\boldsymbol{\mathcal{M}})\}$$

subject to the constraints $\sum_{k=1}^{4} w_k = 1$ (equivalent to the constraint $n_1 + n_2 + n_3 + n_4 = N$) and $w_k \geq 0,\ k = 1, \ldots, 4$. See Appendix A.4 for the analytical expression of the objective function for $A$-optimality. The corresponding expression for $D$-optimality is not given here, but it can be easily obtained through the use of analytical software such as *Maple 17* or *Mathematica*.

Setting $N = 291$ and using the above estimated MAR mechanism, the optimal design is found by using the *Minimize* function in *Mathematica*, subject to the integer constraint. Table 4.4 shows the allocation scheme of an $A$- and a $D$-optimal design, which are denoted by $\xi_A^*$ and $\xi_D^*$ respectively. In this example, due to the large sample size, we did not find any significant differences between the designs that are obtained through the first and the second order approximations and so we have not distinguished between both designs here.

Table 4.4: $A$- and $D$-optimal designs for the Alzheimer's disease clinical study.

| | $n_1$ | $n_2$ | $n_3$ | $n_4$ | $N$ |
|---|---|---|---|---|---|
| | (expected number of missing values) | | | | |
| $\xi_A^*$ | 108 | 64 | 64 | 55 | 291 |
| | (61.1) | (29.6) | (22.2) | (14.3) | |
| $\xi_D^*$ | 60 | 72 | 78 | 81 | 291 |
| | (33.9) | (33.4) | (27.0) | (21.1) | |

Using the same procedure as in the previous, we assess the performance of the optimal designs by simulating incomplete data from the different designs using (4.11) above, choosing values of $\beta_0, \beta_1, \beta_2$ to be $1, 1, 1$ respectively. Note that the specific values of $\beta_0, \beta_1, \beta_2$ will not affect the performance of the different designs in the simulation. The missing values are introduced into the response variable using the above specified MAR mechanism with parameters $\hat{\gamma}_0 = 0.2636472$, $\hat{\gamma}_1 = -0.8988845$ and $\hat{\gamma}_2 = -0.4108504$ which are estimated from the data. From each incomplete data set, regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are estimated from the complete cases. We repeat this process 350 000 times to generate 350 000 incomplete data sets, which allows us to empirically obtain the covariance matrix for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for each design across the simulated sets. The original design, i.e. $\xi_{ori} = (n_1, n_2, n_3, n_4) = (72, 74, 73, 72)$ with expected missing observations $(40.7, 34.3, 25.3, 18.7)$ is also considered as a candidate design here.

Table 4.5 presents the simulated values for the $A$- and the $D$-optimality objective functions of the different designs. As expected, $\xi_A^*$ has the smallest value for $tr\left(\text{sample } \boldsymbol{cov}(\hat{\boldsymbol{\beta}})\right)$ as this optimal design minimises the trace of our approximation to this matrix. Similarly, $\xi_D^*$ has the smallest determinant of the simulated covariance matrix. Both designs result in an improved criterion value over the original design and so could potentially have improved performance if they had been applied. For example, the $A$-optimal design would be expected to achieve a similar trace of the sample covariance matrix as the original design, while requiring only 95.55% of the overall sample size, or 13 fewer patients.

Table 4.5: Empirical values for the $A$- and the $D$-optimality objective functions for different designs.

| | $tr\left(\text{sample } \boldsymbol{cov}(\hat{\boldsymbol{\beta}})\right)$ | $\lvert\text{sample } \boldsymbol{cov}(\hat{\boldsymbol{\beta}})\rvert$ |
|---|---|---|
| $\xi_A^*$ | **0.066327** | 3.722e-06 |
| $\xi_D^*$ | 0.072111 | **3.3028e-06** |
| $\xi_{ori}$ | 0.069416 | 3.3439e-06 |

## 4.5 Conclusion and discussion

In this chapter we have proposed a theoretical framework for designing experiments that takes into account the possibility of missing values. By incorporating a model for the missing data mechanism, we are able to create designs that optimise an objective function of interest. For various commonly used objective functions, we have shown that our designs have the potential to improve performance over designs that do not necessarily incorporate the missing data model into the optimisation.

Our framework has broadened the approach that is proposed by Imhof et al. (2002), which is in fact a special case of the suggested framework here that only takes a Taylor expansion of order one. Besides, we have illustrated the potential benefits of extending the first order approach through a simulation study. However, in some situations that have large sample sizes, the first and the second order approaches may lead to very similar designs in which case the first order approach will be preferred for practical reasons. We have noted some further theoretical properties of using an approach based on the first order expansion and derived the necessary results in this chapter.

Moreover, since a MCAR mechanism has less influence on the quality of the experiment, we infer that the effect of a MCAR mechanism on the optimal designs is negligible. To obtain an optimal design for this situation, the optimal design that assumes complete observations can be employed because the conventional design framework is independent of the total sample size. To overcome the reduction of the power of a study due to the presence of responses that are missing completely at random, we suggest to scale up the total sample size of the experiment.

Concerning the invariance property of $D$-optimality for the linear regression models, Imhof et al. (2004) show that the $D$-optimal design that is found by the framework with a first order approach inherits the scale invariance property. By investigating several design regions using the corresponding MAR mechanisms, we ascertain that the scale invariance property of $D$-optimality is preserved in the suggested framework that employs second order approximation. Table 4.6 presents some $D$-optimal designs for different design regions to reflect this finding. This suggests that it is possible to provide a theoretical proof for future work using the transformation matrix associated with the location-scale transformation. Besides, researchers could also investigate the asymptotic covariance matrix, i.e. the covariance matrix of the limiting distribution in the presence of the missing at random mechanism for finding an optimal design, as an alternative approach to our design framework.

Table 4.6: $D$-optimal designs that are found by the second order approximation framework for different $\mathfrak{X}$, with $N = 60$.

| $\mathfrak{X}$ | $[-1, 1]$ | $[0, 1]$ | $[0, 2]$ | $[0, 3]$ | $[1, 2]$ | $[1, 3]$ | $[1, 4]$ |
|---|---|---|---|---|---|---|---|
| $x_1$ | -1 | 0 | 0 | 0 | 1 | 1 | 1 |
| $x_2$ | 0.92915 | 0.96457 | 1.92915 | 2.89372 | 1.96457 | 2.92915 | 3.89372 |
| $w_2$ | 0.50438 | 0.50438 | 0.50438 | 0.50438 | 0.50438 | 0.50438 | 0.50438 |
| $P(x_1)$ | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 | 0.0025 |
| $P(x_2)$ | 0.35360 | 0.35360 | 0.35360 | 0.35360 | 0.35360 | 0.35360 | 0.35360 |
| $\gamma_0$ | -3.19721 | -5.98896 | -5.98896 | -5.98896 | -11.5725 | -8.7807 | -7.85013 |
| $\gamma_1$ | 2.79175 | 5.58350 | 2.79175 | 1.86117 | 5.58350 | 2.79175 | 1.86117 |

# Chapter 5

# Optimal cohort designs for repeated measurements

In this chapter, we consider the optimal design framework for those experiments that take repeated measurements on different groups of experimental units. This type of experiment is often known as a longitudinal study, which assesses the changes of an outcome variable on experimental units over time. For example, some outcome variables are measured on the placebo group and the treatment group at several follow-up sessions to study the efficacy of new intervention on patients who suffer from moderate to severe Alzheimer's disease; temperature are measured on the people who are at risk of getting an infection on a daily basis to investigate the infection rate of a disease; and the weight of two groups of participants are observed on a weekly basis to evaluate the efficacy of some diet products. To analyse the longitudinal data, a linear mixed model is one of the common choices that has been considered in many research areas. This model has a flexible feature that allows for studying the different type of variations in a longitudinal study, i.e. within-person variation and between-person variation.

In practice, the presence of missing values in the longitudinal study has become a prominent issue especially when the duration of the experiment is long. Using the naive complete case analysis at the cost of discarding the units where data have been partially observed would cause reduction in the power of the study, and bias in the estimation goal of statistical analysis if care is not taken. In the literature on design of experiments, we have found only one paper, i.e. Ortega-Azurduy et al. (2008), who study the efficiency loss of $D$-optimal designs due to dropouts in a longitudinal study for a group of experimental units, provided that the covariance structure of the responses is known. By definition, a dropout refers to an experimental unit whose information is not being observed further once the outcome variable on the subject is not being measured at a time point. To capture the correlation between the repeated measurements of an experimental subject (which is sometimes known as serial correlation in some research areas),

the authors assume a first order autoregressive process, i.e. an AR(1) process, for the observational errors in a linear mixed model. The AR(1) process is often used to model time series for the experiments where the recent observations are less correlated with the responses that are observed in the past. To make inferences based on the incomplete longitudinal data, Ortega-Azurduy et al. (2008) assume that available case analysis will be conducted at the data analysis stage of a study.

Following the approach that is considered in Ortega-Azurduy et al. (2008), we investigate the optimal design framework for the experiments that have more than one group of experimental units. Accounting for the impact of dropouts and the features of available case analysis, we aim to find the best time points of measuring an outcome variable on the groups of subjects for a longitudinal study. In our investigation, we assume that the covariance structure of the responses is known or can be estimated by historical data. We consider two different model formulations of linear mixed models, which are augmented models of those studied in Ouwens et al. (2002) and Schmelter (2007a) respectively. In both of these works, the design framework assumes that completely observed repeated measurements are available from the longitudinal study. In Ouwens et al. (2002), an AR(1) process is assumed for the observational errors of the general linear mixed model. However, the formulation of their design problem does not account explicitly for the experimental conditions of different groups. On the other hand, assuming constant variance for the observational errors, Schmelter (2007a) considers a continuous optimal design framework for a special case of the general linear mixed model. The author shows that an optimal design requires a non-integer replication of observations that is collected from the same subject at the same time point. Nevertheless, this optimal design is not pragmatic in the context of real life applications. We note that the optimal design for the linear mixed model is dependent on the covariance structure of the responses.

Here we first present the standard formulation of a linear mixed model and the two extended models that we employ in our investigation. For each type of linear mixed model formulations, we find optimal cohort designs for four classes of model, namely fixed effect model, random intercept model, random intercept and slope model, and random intercept and slope model with correlated random coefficients. Assuming that repeated measurements are fully observed, we illustrate some examples of optimal cohort designs across the range of the correlation parameter of an AR(1) process. In the subsequent section, we assess the optimal design framework for the two extended models in the presence of dropouts. Considering that different cohorts have different dropout probability functions, we assess two prospects of the design problem: (1) different cohorts are allowed to have different sets of time points of measuring the outcome variable, and (2) all experimental units are restricted to have the same set of time points of measuring the outcome variable. To illustrate our design framework, we revisit the design of an Alzheimer's disease clinical trial (Howard et al. (2012)) using the real experimental data

to elicit the dropout probability functions. Lastly, we conclude this chapter with some discussion and research directions for future investigations.

## 5.1 Set-up of linear mixed models

We now present the general formulation of a linear mixed model, and introduce four classes of this model which we study in our investigation. Let $\boldsymbol{y}_i^T = (y_{i1}, y_{i2}, ..., y_{iq})$ be the $q$ repeated measurements of subject $i$, $i = 1, .., N$, denoting the index of $N$ experimental units. The responses of subject $i$ can be represented by the linear mixed model,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed parameters, $\boldsymbol{b}_i$ is a $s \times 1$ vector of unknown individual effects (also called random coefficients), $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices with dimension $q \times p$ and $q \times s$ respectively, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{iq})^T$ is a vector of observational errors. To reflect the presence of the correlation that arose from measuring the outcome variable on the same experimental unit, i.e. within-person variation, an AR(1) process can be considered for $\boldsymbol{\epsilon}_i$. In our investigation, we assume that experimental subjects are identical and independently distributed, and that the observational errors of subject $i$ is normally distributed with mean zero and covariance matrix $\sigma^2\boldsymbol{\psi}$, i.e. $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N(\boldsymbol{0}, \sigma^2\boldsymbol{\psi})$ where $\boldsymbol{\psi}$ is the covariance structure of an AR(1) process that has elements $\boldsymbol{\psi}(t_j, t_{j'}) = \rho^{|t_j - t_{j'}|}$, $0 \leq \rho \leq 1$, $t_j$ and $t_{j'}$ are some time points, and $j = 1, .., q$. To reflect the fact that different subjects respond differently to the explanatory variable in the study, i.e. between-person variation, the random coefficients $\boldsymbol{b}_i$ are assumed to be normally distributed with mean zero and variance-covariance matrix $\boldsymbol{D}$, i.e. $\boldsymbol{b}_i \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{D})$. For reading purpose, we refer to this model as $\mathsf{M}_o$ in the remaining part of this chapter.

Assuming that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are independent, we can make inferences for the fixed effect parameters $\boldsymbol{\beta}$ of the linear mixed model base on the marginal distribution of the repeated measurements. It can be shown that the marginal distribution of $\boldsymbol{y}_i$ is a multivariate normal distribution with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{V}_i$, where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{\psi}$. By fixing the values of $\boldsymbol{D}$, a few classes of the linear mixed model can be studied. For example, consider a linear model with fixed effect parameters $\boldsymbol{\beta} = (\beta_0,\ \beta_1)^T$, let $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$ be the random intercept and the random slope parameter of the model, a fixed effect model has

$$\boldsymbol{D} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix};$$

a random intercept model has

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix};$$

a random intercept and random slope model with independent random coefficients has

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix};$$

and a random intercept and random slope model with correlated random coefficients has

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix},$$

for some non zero values $d_{11}$, $d_{22}$ and $d_{12}$. For simplicity, we define the last model as a correlated random intercept and slope model.

Considering the maximum likelihood method, we can obtain an unbiased estimator for the fixed effect parameters $\boldsymbol{\beta}$ of the linear mixed model, which is

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i,$$

with covariance matrix

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \left( \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1}.$$

Recall that in the absence of missing observations, the Cramér-Rao lower bound of $\hat{\boldsymbol{\beta}}$, i.e. the lower bound of the variance of the estimator is the inverse of the following total information matrix,

$$\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i, \tag{5.1}$$

which sums the individual total information that is being contributed by each experimental unit. To consider a continuous design framework for finding an optimal design, matrix (5.1) can be rewritten as follows,

$$\begin{aligned} \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i &= \sum_{k=1}^{c} n_k \boldsymbol{X}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{X}_k \\ &= N \sum_{k=1}^{c} \frac{n_k}{N} \boldsymbol{X}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{X}_k \\ &= N \sum_{k=1}^{c} w_k \boldsymbol{X}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{X}_k \end{aligned} \tag{5.2}$$

where $n_k$ is the number of replications of the unique design setting $k$ in the experiment (i.e. the number of experimental subjects who have the same design matrix $\boldsymbol{X}_k$), and $w_k = n_k/N$ is the weight of the corresponding design setting in the study, with $\sum_{k=1}^{c} w_k = 1$. Since a function of matrix (5.2) is often not analytically available, a numerical optimisation is employed to find the optimal setting of $\boldsymbol{X}_k$ and the values of $w_k$ such that a function of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ is minimised.

In the literature, assuming an AR(1) process for the observational errors, Ouwens et al. (2002) study the cohort design problem for $\mathsf{M}_o$ with the assumption of completely observed repeated measurements. Considering time as the only explanatory variable, the authors find an optimal design,

$$\xi^* = \begin{Bmatrix} \boldsymbol{t_1'} & \boldsymbol{t_2'} & \cdots & \boldsymbol{t_c'} \\ w_1 & w_2 & \cdots & w_c \end{Bmatrix}, \tag{5.3}$$

where $\boldsymbol{t_k'} = \{t_{k1}, t_{k2}, ..., t_{kq}\}$ is the optimal allocation of unique time points for measuring an outcome variable on cohort $k$, $k = 1, ..., c$, such that a function of matrix (5.2) is minimised over a design region of time, $\mathfrak{X}$. In their investigation, $\boldsymbol{t_k'}$ are the elements of $\boldsymbol{X}_k$ in model $\mathsf{M}_o$, and no replication of observations that is being measured on the same subject at the same time point is allowed. In particular, replications of observations at the same time point are obtained by having more experimental units in the study. Moreover, the optimisation problem is subject to

$$t_{k1} < t_{k2} < ... < t_{kq} \tag{5.4}$$

and

$$\sum_{k=1}^{c} w_k = 1. \tag{5.5}$$

Since matrix (5.2) does not account explicitly for the experimental conditions of different groups, the optimal cohort designs which are found by Ouwens et al. (2002) with the assumption of fully observed responses for model $\mathsf{M}_o$ consist of the same allocation of time points $\boldsymbol{t_k'}$ for all groups, given the number of cohorts, $c$.

In our investigation, we study the impact of dropouts on the time points of measuring the outcome variable on each group. Before that, we extend model $\mathsf{M}_o$ here such that the effect of experimental conditions on the outcome variable of different cohorts is accounted in the design problem. We introduce an extra variable, $\boldsymbol{\delta}$, which reflects the experimental conditions of different cohorts, into the model formulation such that the repeated measurements of subject $i$ is

$$\boldsymbol{y}_i = \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i$ is an extended matrix of $\boldsymbol{X}_i$ that consists of an extra column for the values of $\boldsymbol{\delta}$. To be more specific, we have

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iq} \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} & \cdots & t_{i1}^p & \delta_{i1} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & t_{iq} & \cdots & t_{iq}^p & \delta_{iq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ \beta_{p+1} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{iq} \end{pmatrix}$$

for the repeated measurements on subject $i$. We note that $\boldsymbol{\delta}$ could either be a categorical variable or a continuous variable that is at most time dependent (i.e. $\delta_{ij}$ are different for $j = 1, ..., q$). For illustration purposes, we treat $\boldsymbol{\delta}$ as a constant over time and that $\delta_{ij}$ are the same for $j = 1, .., q$, and $\delta_{ij} \neq \delta_{i'j}$ when subject $i$ and subject $i'$ are in different cohorts. In the remaining part of this chapter, we refer to this model as $\mathsf{M}_d$ for simplicity. To study the effect of time and the effect of experimental conditions on the outcome variable of the experiment, the maximum likelihood estimation gives

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^{N} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i$$

with variance-covariance matrix

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \left( \sum_{i=1}^{N} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i \right)^{-1}.$$

Since the inverse of the information matrix of $\mathsf{M}_d$,

$$N \sum_{k=1}^{c} w_k \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_k^T \boldsymbol{V}_k^{-1} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_k \tag{5.6}$$

that is presented in terms of weights, is the Cramér-Rao lower bound of the estimated fixed effect parameters, our design problem is to find

$$\xi^* = \begin{Bmatrix} (\boldsymbol{t_1'}, \boldsymbol{\delta_1}) & (\boldsymbol{t_2'}, \boldsymbol{\delta_2}) & \cdots & (\boldsymbol{t_c'}, \boldsymbol{\delta_c}) \\ w_1 & w_2 & \cdots & w_c \end{Bmatrix} \tag{5.7}$$

where the extra elements, $\boldsymbol{\delta_k} = \{\delta_{k1}, \delta_{k2}, ..., \delta_{kq}\}$, reflect the experimental conditions of cohort $k$ at each time point, such that a function of (5.6) is optimised over the design regions, i.e. the design region of time points, $\mathfrak{X}$, and the design region of $\boldsymbol{\delta_k}$ (if $\boldsymbol{\delta_k}$ is to be optimised in the design problem). Like Ouwens et al. (2002), our optimisation problem is subject to constraints (5.4) and (5.5). In the next section, we assess the impact of dropouts on the optimal designs for our model. We do not consider this for the model in Ouwens et al. (2002), i.e. $\mathsf{M}_o$, because the experimental conditions of different cohorts are not distinguishable in the design problem.

On the other hand, for the longitudinal study where different cohorts have the same baseline measurements at the onset of the experiment, Schmelter (2007a) and Schmelter (2007b) study the continuous design theory for a special case of $\mathsf{M}_o$. The author assumes a constant variance for $\boldsymbol{\epsilon}_i$, and the correlation between the repeated measurements of an experimental unit is captured by the variance of $\boldsymbol{b}_i$. It is also assumed that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{b}_i$ are independent. The work aims to propose an optimal design framework for the following model,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{K}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

where the extra $\boldsymbol{K}_i$ is a matrix that indicates the group of experimental units. In the remaining part of this chapter, we refer to this model as $\mathsf{M}_g$ for simplicity. In the investigation of the design framework for this model, the author considers the information matrix,

$$\sum_{i=1}^{N} \boldsymbol{K}_i^T\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\boldsymbol{K}_i = N\sum_{k=1}^{c} w_k\boldsymbol{K}_k^T\boldsymbol{X}_k^T\boldsymbol{V}_k^{-1}\boldsymbol{X}_k\boldsymbol{K}_k, \tag{5.8}$$

with the assumption of fully observed repeated measurements. Moreover, every individual in each cohort is allowed to have different design settings, i.e. $\boldsymbol{t}_{\boldsymbol{k}}^{'}$ of a design $\xi$ may now consist of different individual allocations of time points for each subject who is in cohort $k$. He shows that an optimal design is achieved when every subject in each group have the same allocation of time points. However, a non-integer replication of observations that is being measured on the same experimental unit at the same time point is required in order to achieve a goal of an experiment, such as minimising the determinant of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ for model $\mathsf{M}_g$.

To investigate the impact of dropouts on the optimal cohort designs for this model, we restrict the design problem to the more practically relevant situation where experimental subjects are measured once at each time point. We replicate the observation at a certain time point by having more experimental subjects in the longitudinal study instead of repeatedly taking measurements on the same subject at the same time point. Moreover in our investigation, as opposed to the constant variance of $\boldsymbol{\epsilon}_i$ that is considered in Schmelter (2007a) and Schmelter (2007b), we capture the correlation between the repeated measurements of an experimental unit by an AR(1) process in the observational error of model $\mathsf{M}_g$. In the next section, we study the impact of dropouts on the optimal design framework for model $\mathsf{M}_g$.

In general, both model $\mathsf{M}_d$ and model $\mathsf{M}_g$ are special cases of the linear mixed model $\mathsf{M}_o$. We note that model $\mathsf{M}_d$ with $\boldsymbol{Z}_i = \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i$ is equivalent to model $\mathsf{M}_g$ when $\boldsymbol{K}_i$ is an identity matrix and $\boldsymbol{X}_i = \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_i$ in our notations. To illustrate the applications of the models, we refer to model $\mathsf{M}_d$ for analysing the experiments that have different baseline

measurements for different groups, such as a clinical trial that compares new intervention on a group of patients who have not received preoperative care with the standard treatment on another group of patients who have received preoperative care; and model $\mathsf{M}_g$ with $\boldsymbol{K}_i$ not equal to an identity matrix for the longitudinal study where all subjects have the same baseline measurements at the onset of the study, such as those studies that investigate the efficacy of a new product on different groups of subjects who have the same health status. In the illustrations of this chapter, we assume random effects are caused by the time factor only in both models. We now illustrate the information matrices of the above described three types of linear mixed model, where model $\mathsf{M}_o$ has two fixed effect parameters whereas the other models have three fixed effect parameters respectively.

**Example 5.1.** *Consider an experiment that has $c = 2$ cohorts and three repeated measurements, i.e. $q = 3$ in each of the two cohorts with fully observed measurements.*

*For model $\mathsf{M}_o$ with fixed effect parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, the information matrix is*

$$N\sum_{k=1}^{2} w_k \boldsymbol{X}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{X}_k$$
$$= Nw_1 \begin{pmatrix} 1 & 1 & 1 \\ t_{11} & t_{12} & t_{13} \end{pmatrix} \boldsymbol{V}_i^{-1} \begin{pmatrix} 1 & t_{11} \\ 1 & t_{12} \\ 1 & t_{13} \end{pmatrix} + N(1-w_1) \begin{pmatrix} 1 & 1 & 1 \\ t_{21} & t_{22} & t_{23} \end{pmatrix} \boldsymbol{V}_i^{-1} \begin{pmatrix} 1 & t_{21} \\ 1 & t_{22} \\ 1 & t_{23} \end{pmatrix}.$$

*For model $\mathsf{M}_d$ which allows for the estimation of group effect, we have fixed effect coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, where $\beta_1$ is the fixed effect of time and $\beta_2$ is the group effect. The total information matrix is*

$$N\sum_{k=1}^{2} w_k \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_k^T \boldsymbol{V}_k^{-1} \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_k$$
$$= Nw_1 \begin{pmatrix} 1 & 1 & 1 \\ t_{11} & t_{12} & t_{13} \\ \delta_{11} & \delta_{12} & \delta_{13} \end{pmatrix} \boldsymbol{V}_i^{-1} \begin{pmatrix} 1 & t_{11} & \delta_{11} \\ 1 & t_{12} & \delta_{12} \\ 1 & t_{13} & \delta_{13} \end{pmatrix} + N(1-w_1) \begin{pmatrix} 1 & 1 & 1 \\ t_{21} & t_{22} & t_{23} \\ \delta_{21} & \delta_{22} & \delta_{23} \end{pmatrix} \boldsymbol{V}_i^{-1} \begin{pmatrix} 1 & t_{21} & \delta_{21} \\ 1 & t_{22} & \delta_{22} \\ 1 & t_{23} & \delta_{23} \end{pmatrix}.$$

*For illustration purposes, we assume $\delta_{11} = \delta_{12} = \delta_{13} = \delta_1$, $\delta_{21} = \delta_{22} = \delta_{23} = \delta_2$, and $\delta_1 \neq \delta_2$, are some given values that are not being involved in the optimisation problem.*

*For model $\mathsf{M}_g$ which has the same intercept parameter for all groups, we have $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ for cohort 1 and $\boldsymbol{\beta} = (\beta_0, \beta_2)^T$ for cohort 2 as the fixed effect parameters, with*

$$\boldsymbol{K}_i = \begin{cases} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} & \text{if experimental unit } i \text{ is in group 1;} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \text{if experimental unit } i \text{ is in group 2.} \end{cases}$$

*The total information that is being contributed by these groups are*

$$N\sum_{k=1}^{2} w_k \boldsymbol{K}_k^T \boldsymbol{X}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{X}_k \boldsymbol{K}_k$$
$$= Nw_1 \begin{pmatrix} 1 & 1 & 1 \\ t_{11} & t_{12} & t_{13} \\ 0 & 0 & 0 \end{pmatrix} \boldsymbol{V}_i^{-1} \begin{pmatrix} 1 & t_{11} & 0 \\ 1 & t_{12} & 0 \\ 1 & t_{13} & 0 \end{pmatrix} + N(1-w_1) \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ t_{21} & t_{22} & t_{23} \end{pmatrix} \boldsymbol{V}_i^{-1} \begin{pmatrix} 1 & 0 & t_{21} \\ 1 & 0 & t_{22} \\ 1 & 0 & t_{23} \end{pmatrix}.$$

*In all of these three cases, a design problem is to find* $\boldsymbol{t}_1' = (t_{11}, t_{12}, t_{13})$, $\boldsymbol{t}_2' = (t_{21}, t_{22}, t_{23})$, *and* $w_1$, *such that an aspect of the corresponding* $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ *is minimised over the design region. In this example, Cohort 1 has size* $Nw_1$ *with three measurements per experimental unit at* $\boldsymbol{t}_1'$, *and Cohort 2 has size* $N(1-w_1)$ *with three measurements per subject at* $\boldsymbol{t}_2'$.

Assuming that responses are fully observed, we now depict the optimal design framework for the above described three types of models, i.e. $\mathsf{M}_o$, $\mathsf{M}_d$, and $\mathsf{M}_g$ respectively, which correspond to the model in Ouwens et al. (2002), our extended model and the model in Schmelter (2007a). We want to find the design settings for the longitudinal study that has more than one cohort. In all of these models, the within-subject correlation is assumed to be captured by an AR(1) process, i.e. the elements of the variance-covariance matrix of $\boldsymbol{\epsilon}_i$ are proportional to $\boldsymbol{\psi}(t_j, t_{j'}) = \rho^{|t_j - t_{j'}|}$, $0 \leq \rho \leq 0.9$. In general, an optimal design is dependent on the values of $\rho$. Given the number of cohorts, $c$, the number of repeated measurements, $q$, and the structure of $\boldsymbol{V}_i$, a design problem is to find the time points of measuring an outcome variable on the cohorts and the proportion of cohorts who have such design setting, such that a function of the corresponding $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ is minimised over the design region $\mathfrak{X}$. This design problem is subject to some constraints. In our investigation, we employ the notion of locally optimal designs because $\boldsymbol{V}_i$ is not known at the design stage of an experiment. To construct a future experiment, the structure of $\boldsymbol{V}_i$ can be estimated using some historical data or the information that is obtained from some pilot studies. By fixing $\rho$ and the values of $\boldsymbol{D}$ in each type of the linear mixed models, some optimal cohort designs for the aforementioned classes of linear mixed models, namely fixed effect model, random intercept model, random intercept and slope model, and correlated random intercept and slope model can be found.

To illustrate this, consider an example where an experiment has $c = 2$ and $q = 4$ in each of the two cohorts. Assume that the design region of time points $\mathfrak{X} = [-1, 1]$, $\sigma^2 = 1$, the fixed effect parameters of each type of models are those that are presented in Example 1, and the random effects of each model are

$$\boldsymbol{Z}_i \boldsymbol{b}_i = b_{0i} + t_{kj} b_{1i}$$

Table 5.1: The value of $D$, i.e. covariance of $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$, for different classes of linear mixed models.

| Class of linear mixed model | $d_{11}$ | $d_{22}$ | $d_{12}$ |
|---|---|---|---|
| Fixed effect model (FE) | 0 | 0 | 0 |
| Random intercept model (RI) | 1 | 0 | 0 |
| Random intercept and slope model (RIRS) | 1 | 3 | 0 |
| Correlated random intercept and slope model (RIRSc) | 1 | 3 | $0.8\sqrt{3}$ |

for all subjects, where $t_{kj}$ is the $j^{th}$ time point of measuring the outcome variable on cohort $k$, $k = 1, 2$, and the random coefficients $\boldsymbol{b}_i = (b_{0i}, b_{1i})$ have variance-covariance,

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix},$$

with some known values for $d_{11}$, $d_{22}$ and $d_{12}$; for model $\mathsf{M}_d$, we set $\boldsymbol{\delta_1} = \{\delta_1,\ \delta_1,\ \delta_1,\ \delta_1\}$ and $\boldsymbol{\delta_2} = \{\delta_2,\ \delta_2,\ \delta_2,\ \delta_2\}$ with $\delta_1 = 1$ and $\delta_2 = 0$ for illustration purposes. An example of a design problem is to find $\boldsymbol{t_1'} = (t_{11}, t_{12}, t_{13}, t_{14})$, $\boldsymbol{t_2'} = (t_{21}, t_{22}, t_{23}, t_{24})$ and $w_1$, such that the negative determinant of the corresponding total information matrix is minimised over $\mathfrak{X}$ (note that this is equivalent to minimising the determinant of the inverse matrix). This design problem is subject to constraints (5.4) and (5.5). Following the convention (see Tan and Berger (1999)) and in accordance with common practice, we fix the bounds of $\mathfrak{X}$ as the first and the last time point of the design, i.e. $t_{11} = t_{21} = -1$ and $t_{14} = t_{24} = 1$ in this example, and find the middle time points of measuring the outcome variable on the groups and the corresponding weights by using some optimisation algorithms. In our investigation, we use the function *fmincon* in *Matlab* to find the optimal cohort designs, with equidistant time points and $w_1 = 0.5$ as the initial points for the optimisation problem.

Following the graphical presentations in Ouwens et al. (2002), Ortega-Azurduy et al. (2008), Berger and Wong (2009), we now present some optimal designs for the experiments that have different values of $\rho$. We note that this presentation is different from those used in some research fields, i.e. the $x$, $y$ axes of our plots are interchangeable with the presentation in other areas. Figure 5.1 shows the two middle time points of locally $D$-optimal cohort designs for the experiment that has two cohorts and four repeated measurements. The four plots in the figure correspond to the four classes of the models, i.e. deduced by the values of $\boldsymbol{D}$ (see Table 5.1). In each of these plots, the $y$-axis shows the considered value of $\rho$ (with 0.1 between each case) in the design problems; the two black-dotted lines, red-dashed lines and blue-solid lines correspond to the second and the third optimal time point of measuring the outcome variable for model $\mathsf{M}_o$, $\mathsf{M}_g$ and $\mathsf{M}_d$ respectively across $\rho$. In these design problems, both cohorts have the same optimal time points of measuring the outcome variable as completely observed repeated measurements are assumed. For example, the second and the third time point of the $D$-optimal cohort design for model $\mathsf{M}_d$ with fixed effect and $\rho = 0.1$ are approximately

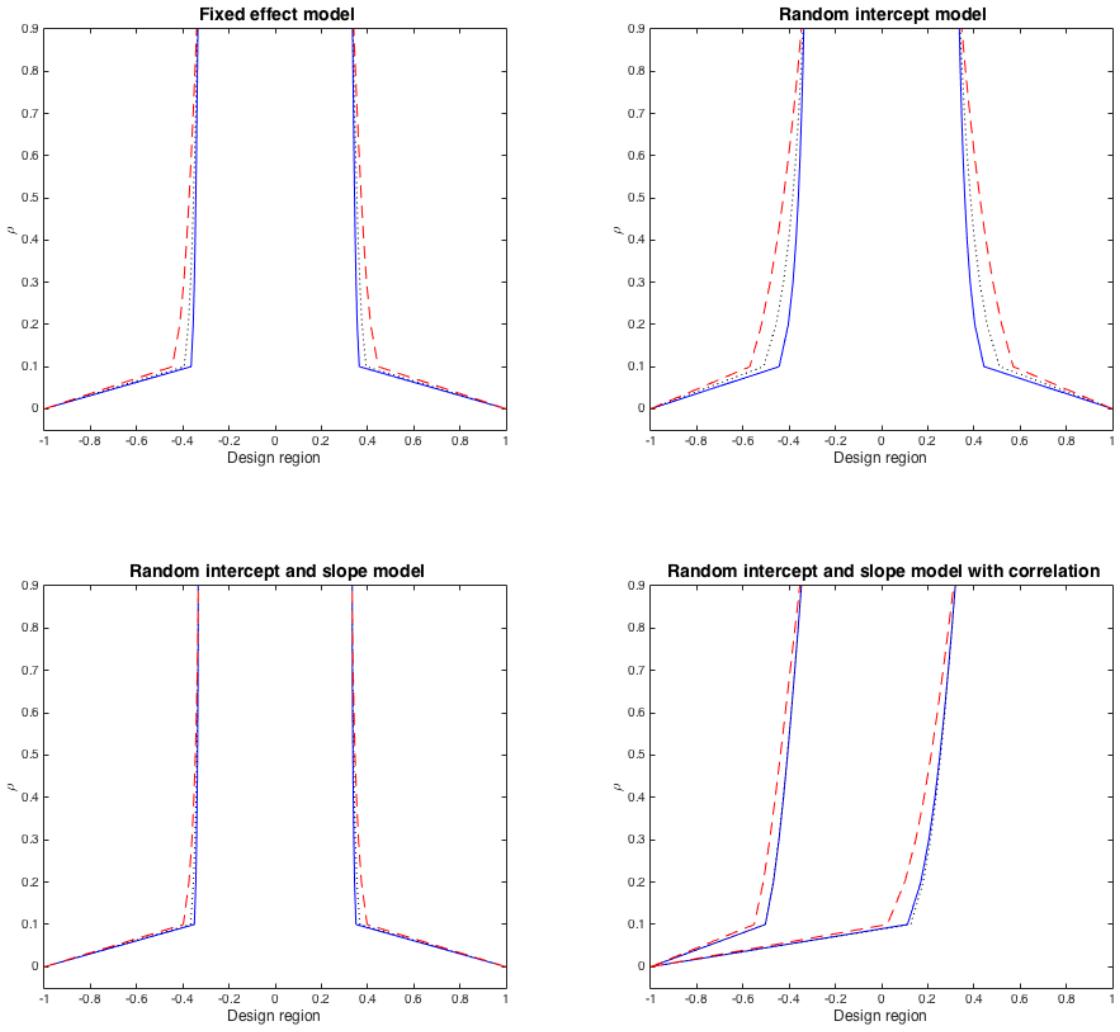



Figure 5.1: The pair of black-dotted line, red-dashed line and blue-solid line in each plot correspond to the second and the third optimal time points of $D$-optimal cohort designs for $\mathsf{M}_o$, $\mathsf{M}_g$ and $\mathsf{M}_d$ respectively.

equal to -0.38 and 0.38 respectively for both groups. In general, these plots show that small values of $\rho$ have greater effect on the time points of the $D$-optimal cohort designs for the four classes of linear mixed models. As $\rho$ approaches to one, the equally spaced design, i.e. $\{t_1, t_2, t_3, t_4\} = \{-1, -1/3, 1/3, 1\}$ for both cohorts, is a robust design for the experiments that have fully observed repeated measurements. Apart from the correlated random intercept and slope model, the shape of the optimal designs across $0 \leq \rho \leq 0.9$ for the other three classes of linear mixed models are very similar. In our investigation, we have tried using different initial points in the optimisation algorithm to verify the convergence of the solutions. Moreover, using different sets of values for $d_{11}$, $d_{22}$ and $d_{12}$ for the four classes of linear mixed model, we find that the trends of the optimal time points across $\rho$ are similar. The chosen values (see Table 5.1) for the illustrations here correspond to the values that are considered in Ortega-Azurduy et al. (2008) for the investigation of efficiency loss of $D$-optimal designs due to the presence of dropouts in a group of subjects.

In the next section, we find optimal cohort designs for model $\mathsf{M}_d$, i.e. the extended model, and for model $\mathsf{M}_g$ which has the same intercept parameter for different groups, respectively with some pre-specified dropout processes for different cohorts. We do not consider the model that is studied by Ouwens et al. (2002), i.e. model $\mathsf{M}_o$, because the model formulation does not account for the experimental conditions of different cohorts. Following the approach that is considered in Ortega-Azurduy et al. (2008), and employing an AR(1) process for the serial correlation of the repeated measurements of the same subject, we examine the impact of dropouts on the design settings for the longitudinal study that has more than one group of subjects.

## 5.2 Optimal cohort design for repeated measurements with dropouts

We now describe the approach that is considered in Ortega-Azurduy et al. (2008). Assuming the presence of dropouts at the design stage of an experiment and an AR(1) process to capture the within-subject correlation, Ortega-Azurduy et al. (2008) study the efficiency loss of $D$-optimal designs for the four classes of linear mixed model with one cohort, i.e. $c = 1$ in model $\mathsf{M}_o$. By definition, a dropout refers to a subject who is lost to be followed up once the measurement on the subject is not being observed at a time point. As a consequence, the number of experimental units who remain in the study decreases with time, i.e. if $n_j$ is the number of subjects who remain in the experiment at time point $j$, the dropout process causes $n_1 \geqslant n_2 \geqslant ... \geqslant n_q$, where $q$ is the number of repeated measurements. Before implementing an experiment, these $n_j$ are unknown and can be estimated by

$$c_j = \begin{cases} N p_{obs}(t_j) & \text{if } j = q, \\ N\ p_{obs}(t_j) - N\ p_{obs}(t_{j+1}) & \text{if } j < q, \end{cases} \tag{5.9}$$

which is the expected number of experimental units who have $j$ repeated measurements in the study, $N$ is the total sample size and $p_{obs}(t_j)$ is the probability of having a response at time point $j$, i.e. the complement of missing response probability function, which could be missing at random.

Considering available case analysis for the longitudinal data that has a monotone dropout pattern, Ortega-Azurduy et al. (2008) find the $D$-optimal time points, $\{t_1, t_2, ..., t_q\}$, for model $\mathsf{M}_o$ with $c = 1$ cohort. These optimal designs minimise the determinant of the inverse of

$$\sum_{j=1}^{q} c_j \boldsymbol{X}_{[j]}^T \boldsymbol{V}_{[j]}^{-1} \boldsymbol{X}_{[j]}, \tag{5.10}$$

where subscript $[j]$ indicates the dimension of the corresponding submatrix. For example, the design matrix of experimental unit $i$ with three fully observed repeated measurements is

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \end{pmatrix};$$

otherwise

$$\boldsymbol{X}_{[1]} = \begin{pmatrix} 1 & t_1 \end{pmatrix}, \quad \text{or} \quad \boldsymbol{X}_{[2]} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \end{pmatrix}$$

is employed in the computation of the total information matrix if subject $i$ is lost to be followed up after the first or the second time point. The outcome measurement is missing at random if $1 - p_{obs}(t_j)$ is dependent on the observed information given the data, e.g. the presence of dropouts in a group may be due to the experimental conditions that cause adverse effect on the patients. Note that when $p_{obs}(t_j) = 1,\ \forall t_j$, matrix (5.10) is a special case of matrix (5.2) with one cohort, i.e. $c = 1$.

Recall that the information that is being contributed by each experimental subject is additive in the linear mixed model. Therefore we can likewise incorporate the dropout processes of different cohorts into the total information matrix (5.6) of model $\mathsf{M}_d$, which yields

$$\sum_{k=1}^{c} w_k \left( \sum_{j=1}^{q} c_j \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_{[j]}^T \boldsymbol{V}_{[j]}^{-1} \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_{[j]} \right) = \sum_{k=1}^{c} \sum_{j=1}^{q} c_{kj} \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_{k[j]}^T \boldsymbol{V}_{k[j]}^{-1} \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_{k[j]}, \tag{5.11}$$

and into the total information matrix (5.8) of the special case of the linear mixed model, $\mathsf{M}_g$, which gives

$$\sum_{k=1}^{c} w_k \left( \sum_{j=1}^{q} c_j \boldsymbol{K}_k^T \boldsymbol{X}_{[j]}^T \boldsymbol{V}_{[j]}^{-1} \boldsymbol{X}_{[j]} \boldsymbol{K}_k \right) = \sum_{k=1}^{c} \sum_{j=1}^{q} c_{kj} \boldsymbol{K}_k^T \boldsymbol{X}_{k[j]}^T \boldsymbol{V}_{k[j]}^{-1} \boldsymbol{X}_{k[j]} \boldsymbol{K}_k. \tag{5.12}$$

The $c_j$ in the summation on the left hand side of both expressions is defined by (5.9), whereas

$$c_{kj} = w_k c_j = \begin{cases} N w_k p_{k,obs}(t_{kj}) & \text{if } j = q \\ N w_k \left[ p_{k,obs}(t_{kj}) - p_{k,obs}(t_{kj+1}) \right] & \text{if } j < q \end{cases} \tag{5.13}$$

on the right hand side of both expressions represents the expected total number of subjects who are in cohort $k$ and have $j$ repeated measurements in the study, and $p_{k,obs}(t_{kj})$ is the probability that a subject in cohort $k$ has an observation at time point $j$. To allow for having different allocations of time points for different groups, we use subscript $k[j]$ to indicate $j$ dimension of the corresponding submatrix of the subjects

who are in cohort $k$. For instance,

$$\boldsymbol{X}_{k[1]} = \begin{pmatrix} 1 & t_{k1} \end{pmatrix}, \quad \boldsymbol{X}_{k[2]} = \begin{pmatrix} 1 & t_{k1} \\ 1 & t_{k2} \end{pmatrix}, \quad \boldsymbol{X}_{k[3]} = \begin{pmatrix} 1 & t_{k1} \\ 1 & t_{k2} \\ 1 & t_{k3} \end{pmatrix}$$

are some submatrices of design matrix $\boldsymbol{X}_k$ of model $\mathsf{M}_g$.

We now consider two prospects of the design problem for model $\mathsf{M}_d$ and model $\mathsf{M}_g$ respectively in the presence of some dropout processes. Model $\mathsf{M}_d$ correspond to our extended model that is introduced in the previous section, and $\mathsf{M}_g$ correspond to the model that has the same intercept parameter for different groups of experimental units. For a flexible set-up of an experiment where different cohorts are allowed to have different sets of time points of measuring the outcome variable, we are interested in either

- an optimal design (5.7) which optimises a function of the right hand side of expression (5.11) for model $\mathsf{M}_d$; or
- an optimal design (5.3) which optimises a function of the right hand side of expression (5.12) for model $\mathsf{M}_g$.

We refer to these conditions as $\mathbb{F}_d$ and $\mathbb{F}_g$ respectively in the remaining part of this chapter. For a more restricted situation where the outcome variable of all cohorts must be measured at the same set of time points, i.e. $\boldsymbol{t}_{\boldsymbol{k}}'$, $k = 1, ..., c$, are the same for all cohorts, we want to find either

- an optimal design (5.7) which optimises a function of the left hand side of expression (5.11) for model $\mathsf{M}_d$; or
- an optimal design (5.3) which optimises a function of the left hand side of expression (5.12) for model $\mathsf{M}_g$.

These conditions are denoted by $\mathfrak{R}_d$ and $\mathfrak{R}_g$ respectively for simplicity. In any case, the optimisation problem is subject to constraints (5.4), (5.5) and (5.9) or (5.13), depending on the condition of the design problem.

Concerning the presence of dropouts in different groups of experimental units, we consider monotone dropout patterns in the optimal design framework for the two types of linear mixed models. We incorporate response probability functions of different cohorts, i.e. the complement of missing response probability functions, into the constraint of the above described optimisation problems. For example assuming that the probability that an observation is missing is dependent on the time points given the data, i.e. responses

are missing at random, we can incorporate a linear response probability function for one cohort, i.e.

$$p_{1,obs}(t_{1j}) = 0.65 - 0.35t_{1j}, \tag{5.14}$$

and a quadratic response probability function for another cohort, i.e.

$$p_{2,obs}(t_{2j}) = 0.5 - 0.35t_{2j} + 0.15t_{2j}^2, \tag{5.15}$$

into constraint (5.13) for the optimisation problem that considers condition $\mathbb{F}_d$ or $\mathbb{F}_g$ (depending on the chosen model). These response probability functions are decreasing with $t$ and lie between zero and one for design region $\mathfrak{X} = [-1, 1]$, and have been employed in Ortega-Azurduy et al. (2008) for the investigation of efficiency loss of $D$-optimal designs due to the presence of dropouts in one cohort. Figure 5.2 shows these probability response functions over $\mathfrak{X}$. We note that our design framework is compatible with a wide class of missing at random mechanisms that have monotone response/missing response probability functions.

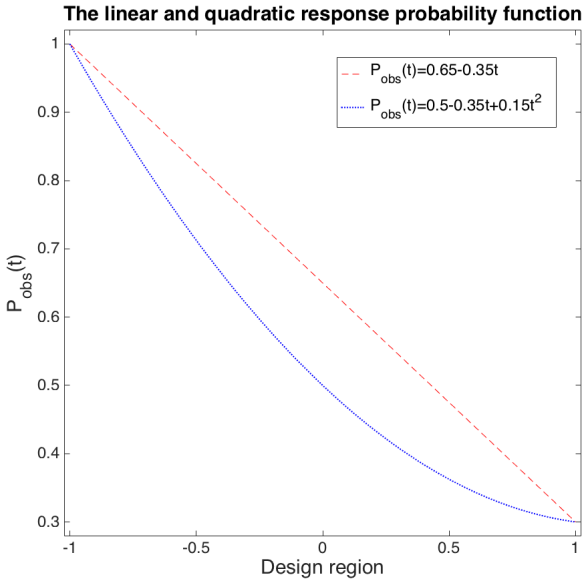



Figure 5.2: Linear response probability function (5.14) and quadratic response probability function (5.15) over the design region $\mathfrak{X} = [-1, 1]$.

Assuming that the within-subject correlation is captured by an AR(1) process, i.e. the elements of the covariance matrix of $\boldsymbol{\epsilon}_i$ are proportional to $\psi(t_j, t_{j'}) = \rho^{|t_j - t_{j'}|}$, we now depict the optimal cohort design framework for model $\mathsf{M}_d$ and model $\mathsf{M}_g$ respectively in the presence of dropouts. For each type and each class of the linear mixed models, given the number of cohorts, $c$, the number of repeated measurements, $q$, the structure of $\boldsymbol{V}_i$, and the response probability functions, a design problem is to find the time points of measuring the outcome variable on the cohorts and the corresponding weights, such that a function of the corresponding $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ is minimised over a design region $\mathfrak{X}$. Having

fixed the bounds of $\mathfrak{X}$ as the first and the last time point of the design, the design problem can be solved locally by using some optimisation algorithms with some initial designs. Like others, this design problem is subject to some constraints of the experiment.

We now present some examples of $D$-optimal cohort designs for an experiment that has $c = 2$ and $q = 4$ in each of the two groups. By varying the structure of $\boldsymbol{V}_i$, a locally optimal cohort design can be found for the different classes of model $\mathsf{M}_d$ and model $\mathsf{M}_g$ respectively. Assume the design region of time points $\mathfrak{X} = [-1, 1]$, $\sigma^2 = 1$, and the two cohorts have the linear and the quadratic response probability function respectively; and for model $\mathsf{M}_d$, we have $\boldsymbol{\delta_1} = \{\delta_1,\ \delta_1,\ \delta_1,\ \delta_1\}$ and $\boldsymbol{\delta_2} = \{\delta_2,\ \delta_2,\ \delta_2,\ \delta_2\}$ with $\delta_1 = 1$ and $\delta_2 = 0$. For illustration purposes, we assume the fixed effects of model $\mathsf{M}_d$ is

$$\beta_0 + t_{kj}\beta_1 + \delta_{kj}\beta_2 = \begin{cases} \beta_0 + t_{1j}\beta_1 + \beta_2 & \text{if experimental unit } i \text{ is in Cohort 1,} \\ \beta_0 + t_{2j}\beta_1 & \text{if experimental unit } i \text{ is in Cohort 2,} \end{cases}$$

and of model $\mathsf{M}_g$ is

$$\boldsymbol{X}_i\boldsymbol{K}_i\boldsymbol{\beta} = \begin{cases} \beta_0 + t_{1j}\beta_1 & \text{if experimental unit } i \text{ is in Group 1,} \\ \beta_0 + t_{2j}\beta_2 & \text{if experimental unit } i \text{ is in Group 2,} \end{cases}$$

with random effects

$$\boldsymbol{Z}_i\boldsymbol{b}_i = b_{0i} + t_{kj}b_{1i}$$

for both types of model, where $t_{kj}$ is the $j^{th}$ time point of measuring observation on cohort $k$, $k = 1, 2$, and the random coefficients $\boldsymbol{b}_i = (b_{0i}, b_{1i})$ have variance-covariance matrix

$$\boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix},$$

with some known values for $d_{11}$, $d_{22}$ and $d_{12}$. The difference between the two models is that model $\mathsf{M}_d$ has different intercept parameters but the same slope for both cohorts whereas model $\mathsf{M}_g$ has the same intercept parameter but different slopes for the two cohorts.

Having fixed the bounds of $\mathfrak{X}$ as the first and the last time point of the design, i.e. $t_{11} = t_{21} = -1$ and $t_{14} = t_{24} = 1$ in this example, we find the middle time points of measuring the outcome variable on the groups and the corresponding weights using the function *fmincon* in *Matlab*, with equidistant time points and $w_1 = 0.5$ as the initial points for the optimisation problem. For each class of the models, namely fixed effect model, random intercept model, random intercept and slope model, and correlated random intercept and slope model, we consider the four different design conditions, $\mathbb{F}_d$, $\mathbb{F}_g$, $\mathfrak{R}_d$, and $\mathfrak{R}_g$ respectively in the design problem, which is subject to constraints (5.4), (5.5) and (5.9) or (5.13).

Table 5.2: Details of the locally $D$-optimal cohort designs that are obtained under condition $\mathbb{F}_d$ and $\mathbb{F}_g$ respectively. In all cases, cohort 1 has a lower response probability than cohort 2 within $\mathfrak{X}$.

| | Figure | Cohort/ Group | $p_{k,obs}(t_{1j})$ | Plots |
|---|---|---|---|---|
| Model $\mathsf{M}_d$ | 5.3 | 1, with $\delta_1 = 1$ | (5.14) | blue-dotted line |
| | | 2, with $\delta_2 = 0$ | 1 | red-dashed line |
| | 5.5 | 1, with $\delta_1 = 1$ | (5.15) | blue-dotted line |
| | | 2, with $\delta_2 = 0$ | (5.14) | red-dashed line |
| Model $\mathsf{M}_g$ | 5.4 | 1 | (5.14) | blue-dotted line |
| | | 2 | 1 | red-dashed line |
| | 5.6 | 1 | (5.15) | blue-dotted line |
| | | 2 | (5.14) | red-dashed line |

For the experiment that has two cohorts and four repeated measurements in each of the two cohorts, Figure 5.3 and Figure 5.4 show the second and the third time points of locally $D$-optimal cohort designs for model $\mathsf{M}_d$ and $\mathsf{M}_g$ respectively, with linear response probability function (5.14) in one group, i.e. Cohort 1 with $\boldsymbol{\delta_1}$ in model $\mathsf{M}_d$ and Group 1 in model $\mathsf{M}_g$, and fully observed repeated measurements in the other group, i.e. Cohort 2 with $\boldsymbol{\delta_2}$ in model $\mathsf{M}_d$ and Group 2 in model $\mathsf{M}_g$; Figure 5.5 and Figure 5.6 correspond to those design points for model $\mathsf{M}_d$ and $\mathsf{M}_g$ respectively, with quadratic response probability function (5.15) in one group, i.e. Cohort 1 in model $\mathsf{M}_d$ and Group 1 in model $\mathsf{M}_g$, and linear response probability function (5.14) in the other group, i.e. Cohort 2 in model $\mathsf{M}_d$ and Group 2 in model $\mathsf{M}_g$. In these figures, each plot corresponds to the middle time points of the optimal cohort designs for each class of the models; the pair of dotted-lines in the first row of plots correspond to the second and the third time points of measuring the outcome variable on both groups, which are found under the restricted condition $\mathfrak{R}_d$ or $\mathfrak{R}_g$; the two pairs of lines in the second row of plots correspond to the two sets of time points of measuring the outcome variable on the different groups, which are found under the flexible condition $\mathbb{F}_d$ or $\mathbb{F}_g$. In each of these plots, the $y$-axis shows the considered value of $\rho$ (with 0.1 between each case) in the design problems. Table 5.2 summaries the details of the second rows of plots in these figures.

Comparing the shapes of the red-dashed lines in the second row of plots in Figure 5.3 and Figure 5.4 with the corresponding plots in Figure 5.1, we see that the trend of the second and the third time points of measuring the outcome variable on the group that has fully observed repeated measurement across $\rho$ are similar to those optimal time points that are found by the design framework which assumes completely observed repeated measurements. However, this is not true when there are dropouts in both groups, see Figure 5.5 and Figure 5.6. Moreover, the optimal time points for the experiments that have dropouts in both groups do not converge to the equidistant design as $\rho$ approaches to one, in particular for the random intercept and slope model, and the correlated random intercept and slope model. To be more specific, consider the second row of

plots in Figure 5.5 and Figure 5.6, we find that the optimal time points of measuring the outcome variable on the experimental units who remain longer in the study, i.e. the group with the linear response probability function, are larger than those of the subjects who are expected to be dropping out earlier from the study, i.e. the group with the quadratic response probability function. Comparing the first and the second row of plots in all of these figures, we learn that there exist some differences between the optimal time points that are found under different conditions, i.e. comparing the optimal cohort designs which subject to condition $\mathfrak{R}_d$ and $\mathbb{F}_d$ respectively for $\mathsf{M}_d$, or condition $\mathfrak{R}_g$ and $\mathbb{F}_g$ respectively for $\mathsf{M}_g$.

We now consider the weights of the locally $D$-optimal cohort designs for the above described experiments, i.e. the proportions of experimental units who have the corresponding sets of optimal time points that are presented in the four figures. Table 5.3 shows the maximum and the minimum weights across the ten considered cases with $\rho$, $0 \leq \rho \leq 0.9$ (and difference of 0.1 between each case). Notice that for the experiment that has the linear response probability function in one group and fully observed responses in the other group, the optimal design framework allocates more experimental units to the group that has completely observed repeated measurements, and measures the outcome variable of the other group at some time points which are further away from the upper bound of $\mathfrak{X}$ (see Figure 5.3 and Figure 5.4 for the corresponding optimal time points). Similarly when the quadratic and the linear response probability functions are assumed respectively for the two cohorts in the experiment with $\rho > 0$, the design framework maximises the expected total information by having more experimental units in the cohort that has a higher response rate within $\mathfrak{X}$, i.e. the group that has the linear response probability function (red-dashed line) in the plots in Figure 5.5 and Figure 5.6.

However for the experiment with the two response probability functions, $\rho = 0$ and the restricted condition $\mathfrak{R}_d$ or $\mathfrak{R}_g$, the locally $D$-optimal cohort designs for the corresponding fixed effect models and the random intercept models have $w_1 = 0.5 = w_2$, as all the experimental units have the same response rate at the third/ fourth optimal time points, i.e. at the upper bound of $\mathfrak{X}$ (see the first two plots from the left in the first row of plots in Figure 5.5 and in Figure 5.6 respectively). The same reason also applies to the locally $D$-optimal cohort design for model $\mathsf{M}_g$ with fixed effect parameters and $\rho = 0$, which is subject to condition $\mathbb{F}_g$ (see the first plot from the left in the second row of plots in Figure 5.6). Concerning model $\mathsf{M}_d$ with fixed effect parameters and the same model with a random intercept parameter for the experiment that has the two response probability functions, $\rho = 0$, and the flexible condition $\mathbb{F}_d$, the locally $D$-optimal cohort designs of these two classes of model $\mathsf{M}_d$ have $w_1 = 0.5064$ and $w_1 = 0.4984$ respectively for the group that has the quadratic response probability function (blue-dotted line in the first two plots from the left in the second row of plots in Figure 5.5). This could be because there is less variability in estimating the fixed effect parameters of the former model than of the model with a random intercept, and hence the optimal design framework

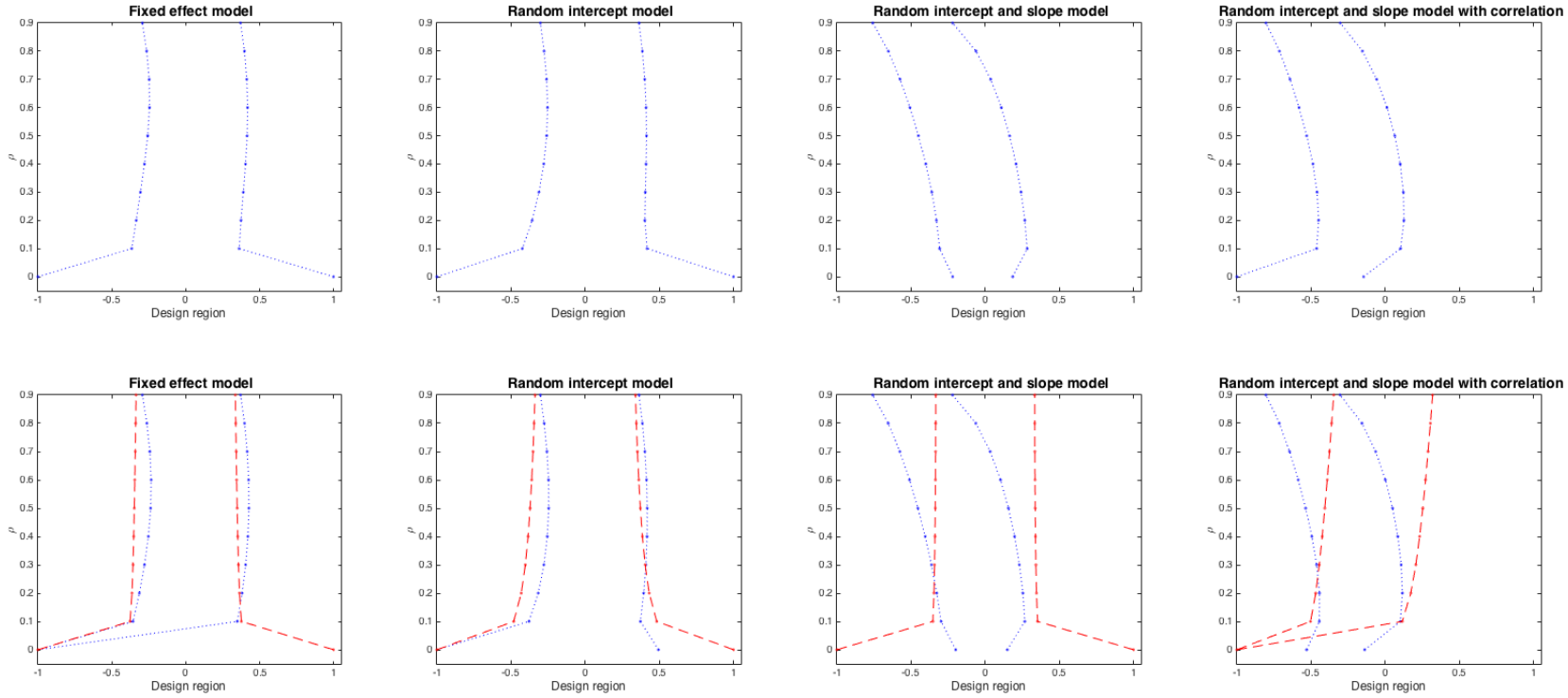



Figure 5.3: The middle two $D$-optimal time points for model $\mathsf{M}_d$ with $c = 2$, $q = 4$, designs condition $\mathfrak{R}_d$ (top row) and $\mathbb{F}_d$ (bottom row) respectively. In the bottom plots, Cohort 1 (blue-dotted lines) has $\delta_1 = 1$ and linear response probability function (5.14); Cohort 2 (red-dashed lines) has $\delta_2 = 0$ and completely observed responses.

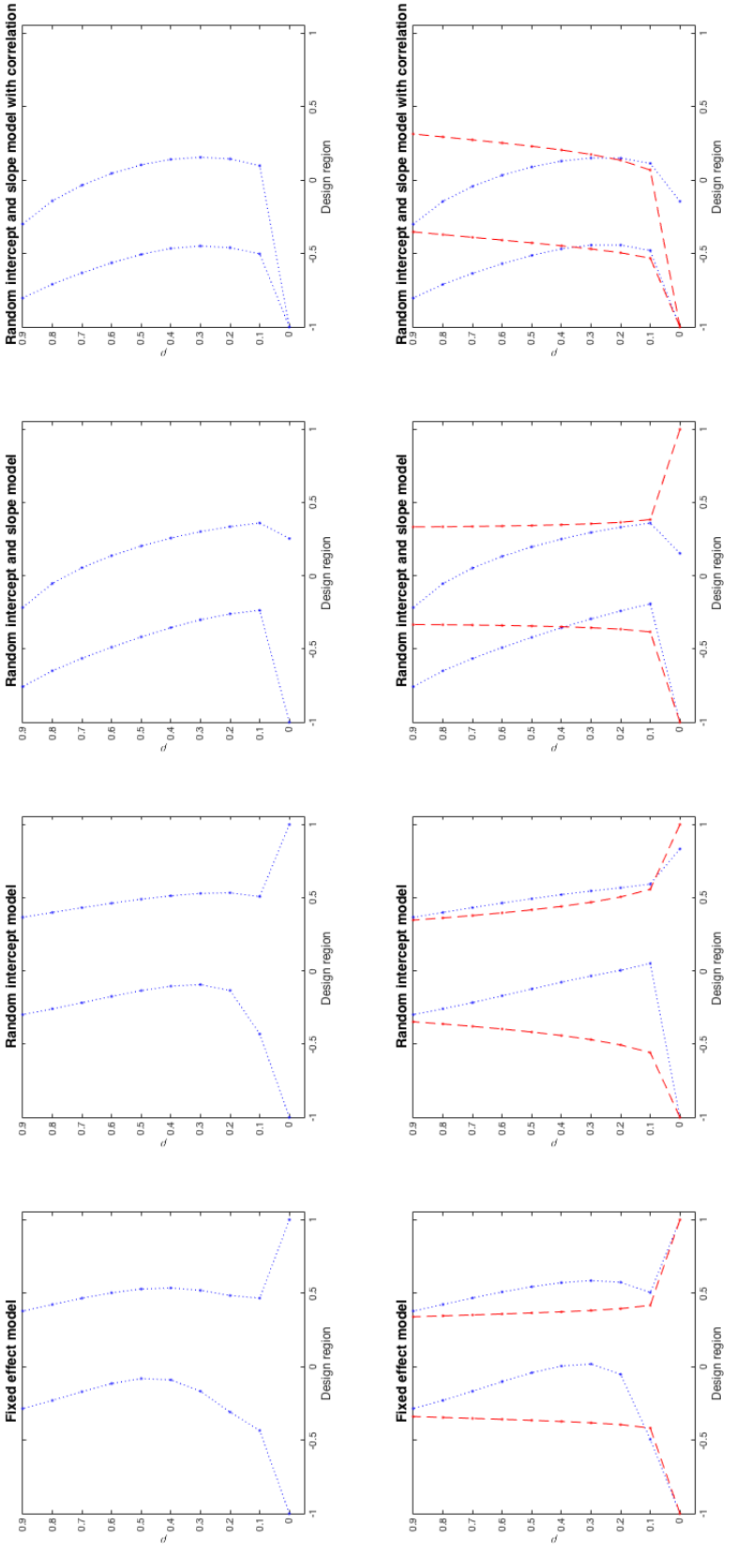



Figure 5.4: The middle two $D$-optimal time points for model $\mathsf{M}_g$ with with $c = 2$, $q = 4$, designs condition $\mathfrak{R}_g$ (top row) and $\mathbb{F}_g$ (bottom row) respectively. In the bottom plots, Group 1 (blue-dotted lines) has linear response probability function (5.14); Group 2 (red-dashed lines) has completely observed responses.

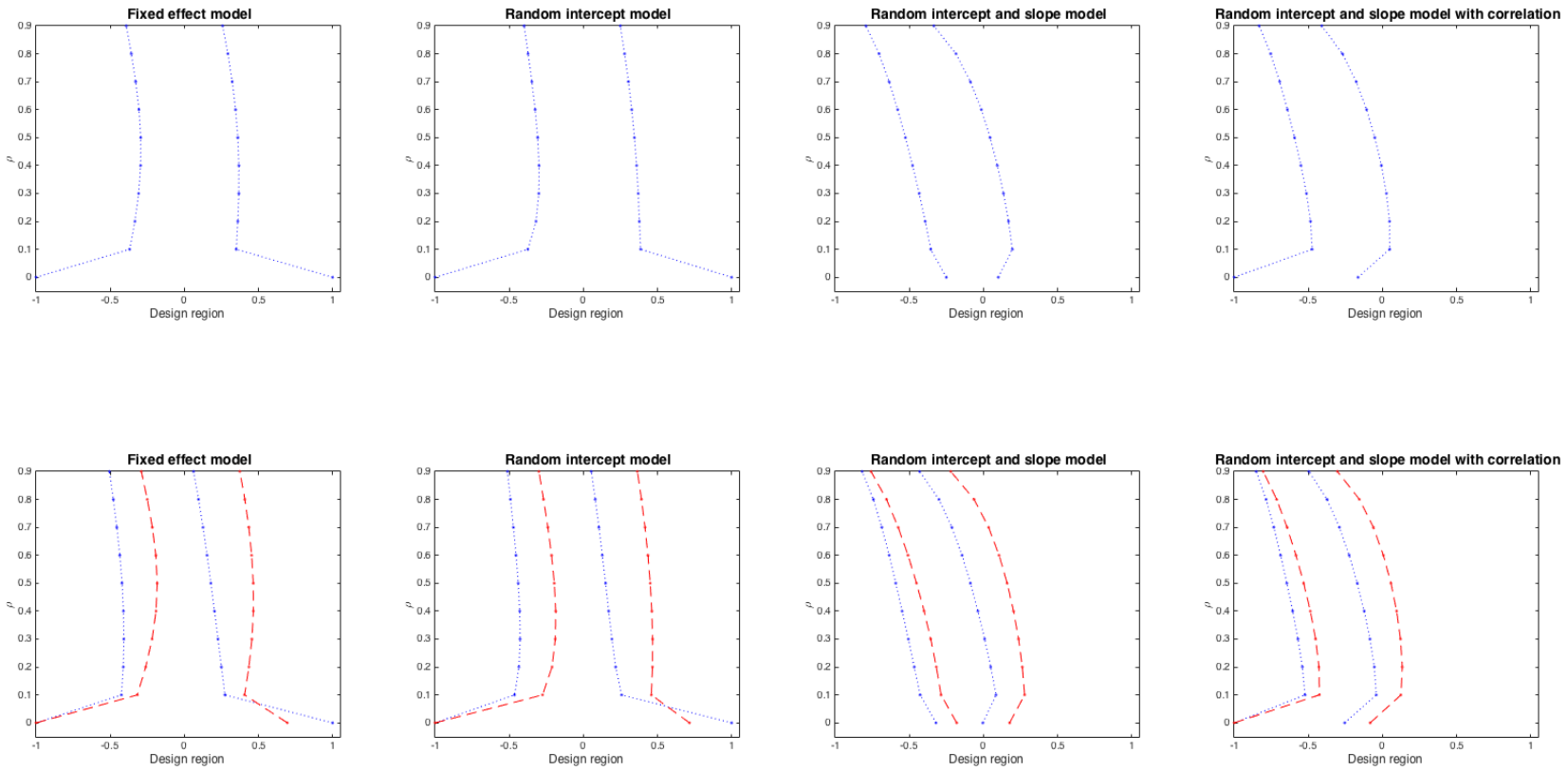



Figure 5.5: The middle two $D$-optimal time points for model $\mathsf{M}_d$ with $c = 2$, $q = 4$, designs condition $\mathfrak{R}_d$ (top row) and $\mathbb{F}_d$ (bottom row) respectively. In the bottom plots, Cohort 1 (blue-dotted lines) has $\delta_1 = 1$ and quadratic response probability function (5.15); Cohort 2 (red-dashed lines) has $\delta_2 = 0$ and linear response probability function (5.14).

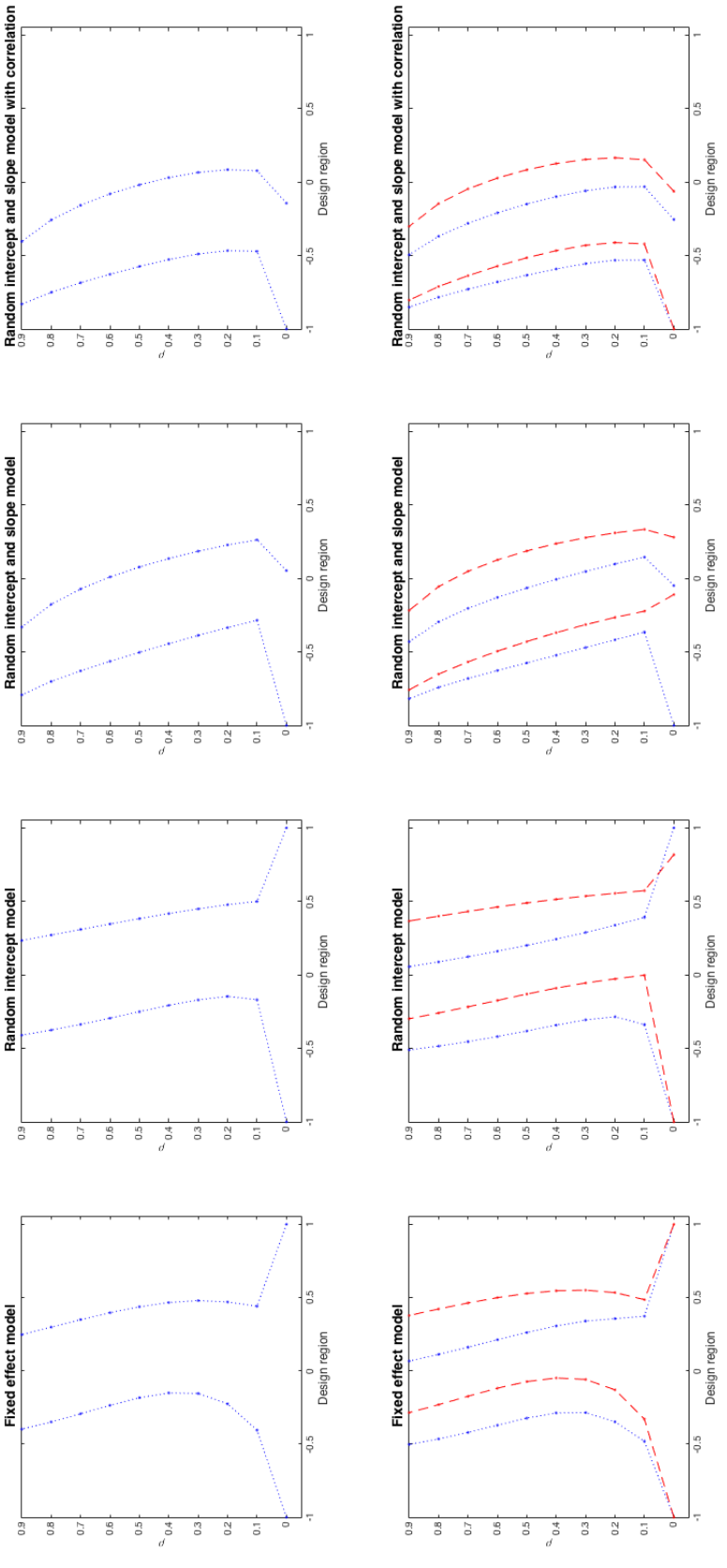



Figure 5.6: The middle two $D$-optimal time points for model $\mathsf{M}_g$ with with $c = 2$, $q = 4$, designs condition $\mathfrak{R}_g$ (top row) and $\mathbb{F}_g$ (bottom row) respectively. In the bottom plots, Group 1 (blue-dotted lines) has quadratic response probability function (5.15); Group 2 (red-dashed lines) has linear response probability function (5.14).

Table 5.3: The weight under two design schemes corresponds to the second row of plots whereas the weight under one design schemes corresponds to the first row of plots in Figure 5.3, 5.4, 5.5 and 5.6.

| | Two design schemes | | One design schemes | |
|---|---|---|---|---|
| | Max $w_1$ | Min $w_1$ | Max $w_1$ | Min $w_1$ |
| $\mathsf{M}_d$ in Figure 5.3, $w_1$ for the cohort with linear response | | | | |
| FE | 0.4331 | 0.3813* | 0.4329 | 0.4157* |
| RI | 0.4280 | 0.3921* | 0.4280 | 0.3935* |
| RIRS | 0.4879 | 0.4594* | 0.4879 | 0.4676* |
| RIRSc | 0.4856 | 0.4477* | 0.4856 | 0.4571* |
| $\mathsf{M}_g$ in Figure 5.4, $w_1$ for the group with linear response | | | | |
| FE | 0.4896 | 0.4157* | 0.4896 | 0.4157* |
| RI | 0.4948 | 0.4624* | 0.4948 | 0.4601* |
| RIRS | 0.4814 | 0.4328* | 0.4814 | 0.4324* |
| RIRSc | 0.4779 | 0.4302* | 0.4780 | 0.3872* |
| $\mathsf{M}_d$ in Figure 5.5, $w_1$ for the cohort with quadratic response | | | | |
| FE | 0.5064* | 0.4808 | 0.5000* | 0.4804 |
| RI | 0.4984* | 0.4803 | 0.5000* | 0.4799 |
| RIRS | 0.4948 | 0.4885* | 0.4948 | 0.4906 |
| RIRSc | 0.4938 | 0.4869* | 0.4939 | 0.4890 |
| $\mathsf{M}_g$ in Figure 5.6, $w_1$ for the group with quadratic response | | | | |
| FE | 0.5000* | 0.4821 | 0.5000* | 0.4828 |
| RI | 0.4981 | 0.4901 | 0.5000* | 0.4878 |
| RIRS | 0.4921 | 0.4624 | 0.4921 | 0.4781 |
| RIRSc | 0.4907 | 0.4761 | 0.4907 | 0.4773 |

* *these weights are obtained under the case with* $\rho = 0$.

maximises the information for the former model by having more experimental units for the group that has a lower response rate within $\mathfrak{X}$ and measuring the outcome variable on the other group at some optimal time points that have higher response rates.

In general, the optimal design framework that accounts for the impact of dropouts overcome the extra variability that is due to the presence of random parameters in the models by allocating more subjects to the group that has a higher response rate within $\mathfrak{X}$. Here we have presented some examples of locally $D$-optimal cohort designs in the presence of some dropout processes for some classes of model $\mathsf{M}_d$ and model $\mathsf{M}_g$, that are subject to condition $\mathfrak{R}_d$, $\mathfrak{R}_g$, $\mathbb{F}_d$ and $\mathbb{F}_g$ respectively for some values of $\rho$, $0 \leq \rho \leq 0.9$. In the above described examples, we have employed the values that are presented in Table 5.1 for $\boldsymbol{D}$ to study the locally optimal cohort designs for each class of model $\mathsf{M}_d$ and model $\mathsf{M}_g$ respectively. Considering the same set-up of the design problems but different sets of values for $\boldsymbol{D}$, we find that the trends of the optimal time points across $0 \leq \rho \leq 0.9$ for each class of the models are similar to those that are illustrated here. These locally optimal designs have been verified by using different initial points in the optimisation algorithm. In the next section, we illustrate the performance of some designs through simulation studies.

Table 5.4: Notations of optimal designs and the details of the design problems.

| | Restricted design right hand side | Flexible design left hand side | Completely observed repeated measurements |
|---|---|---|---|
| notations | $\xi^*_{D,1}$, $\xi^*_{A,1}$ | $\xi^*_{D,2}$ , $\xi^*_{A,2}$ | $\xi^*_{D,f}$ , $\xi^*_{A,f}$ |
| model $\mathsf{M}_d$ | information matrix (5.11) | | information matrix (5.6) |
| model $\mathsf{M}_g$ | information matrix (5.12) | | information matrix (5.8) |
| constraints | (5.4), (5.5), (5.9) | (5.4), (5.5), (5.13) | (5.4), (5.5) |

## 5.3 Simulation study

Consider the above described experiment that has two cohorts and four repeated measurement with $\mathfrak{X} = [-1, 1]$, $\rho = 0.5$, and the presented values in Table 5.1 for $\boldsymbol{D}$. Assume that the two groups of experimental units have linear and quadratic response probability function, (5.14) and (5.15) respectively. We now evaluate the performance of the equally spaced time point design, i.e. $\xi_{eq} = \{t_1, t_2, t_3, t_4\} = \{-1, -1/3, 1/3, 1\}$, the locally $A$- and $D$-optimal cohort designs that assume completely observed responses, and the locally $A$- and $D$-optimal cohort designs which account for the presence of dropouts, for each class of model $\mathsf{M}_d$ and $\mathsf{M}_g$ respectively. Table 5.4 shows the notations of the optimal designs where the first subscript in $\xi^*$ denotes the design criterion; the corresponding total information matrices; and the conditions that we employ for finding the optimal designs for the experiments that assume the presence of dropouts and those that assume completely observed repeated measurements. We consider two prospects for the optimal designs that account for the impact of dropouts, i.e. the flexible design setting where the outcome variable on different cohorts may be measured at different sets of time points and the restricted design setting where the measurements of all cohorts must be collected at the same set of time points. These design conditions are denoted by $\mathbb{F}_d$ and $\mathfrak{R}_d$ respectively for model $\mathsf{M}_d$; $\mathbb{F}_g$ and $\mathfrak{R}_g$ respectively for model $\mathsf{M}_g$ in the previous section.

We first consider the simulation set-up for model $\mathsf{M}_d$ where Cohort 1 has $\boldsymbol{\delta_1} = \{\delta_1,\ \delta_1,\ \delta_1,\ \delta_1\}$, $\delta_1 = 1$, with sample size $n_1$ and quadratic response probability function (5.15); Cohort 2 has $\boldsymbol{\delta_2} = \{\delta_2,\ \delta_2,\ \delta_2,\ \delta_2\}$, $\delta_2 = 0$, with sample size $n_2$ and linear response probability function (5.14). The locally optimal cohort designs for each class of model $\mathsf{M}_d$ are presented in Table 5.5, where $\xi^*_{D,2}$ and $\xi^*_{A,2}$ minimise the negative determinant and the trace respectively, of the right hand side of matrix (5.11) under condition $\mathbb{F}_d$; $\xi^*_{D,1}$ and $\xi^*_{A,1}$ minimise the corresponding functions of the left hand side of (5.11) under condition $\mathfrak{R}_d$; $\xi^*_{D,f}$ and $\xi^*_{A,f}$ minimise the corresponding functions of the total information matrix (5.6) in the framework that assumes completely observed responses. The constraints of these optimisation problems are presented in Table 5.4, and the design settings are found by fixing the lower and the upper bound of $\mathfrak{X}$ as the first and the last time point of the design in the optimisation problem.

For each considered design, we simulate

$$y_{ij} = \begin{cases} 1 + t_{1j} + 1 + b_{0i} + t_{1j}b_{1i} + \epsilon_{1j} & \text{for subject } i \text{ in Cohort 1,} \\ 1 + t_{2j} + b_{0i} + t_{2j}b_{1i} + \epsilon_{2j} & \text{for subject } i \text{ in Cohort 2,} \end{cases}$$

where $t_{1j}$ and $t_{2j}$ correspond to the $j^{th}$ time point of measuring the outcome variable on Cohort 1 that has size $n_1$ and Cohort 2 that has size $n_2$ respectively, $\epsilon_{ij} \sim N(0, \sigma^2\boldsymbol{\psi})$, $\boldsymbol{\psi} = 0.5^{|t_j - t_{j'}|}$, $\sigma^2 = 1$, and $\boldsymbol{b_i} \sim N(0, \boldsymbol{D})$ with the corresponding values of $\boldsymbol{D}$ for each class of model $\mathsf{M}_d$. The optimal time points and the size of each cohort, i.e. $n_1$ and $n_2$, for each design are presented in Table 5.5. Missing values in the vector of observed responses of subject $i$, i.e. $\boldsymbol{y}_i = \{y_{i1}, ..., y_{i4}\}$, are then introduced by specifying quadratic response probability function (5.15) in the responses of Cohort 1 and linear response probability function (5.14) in the responses of Cohort 2. Using these response probability functions, we can see in Figure 5.2 that Cohort 1 has a lower response rate than Cohort 2 within $\mathfrak{X} = [-1, 1]$, but both groups have the same probability of getting a response at the bounds of $\mathfrak{X}$. This could be an example of a clinical study which compares the efficacy of a standard treatment (i.e. $\delta_2 = 0$) to a new intervention (i.e. $\delta_1 = 1$), which might cause more adverse effects than the standard treatment and hence resulting in more dropouts in Cohort 1.

To compare the performance of each design in Table 5.5 and the equally spaced time point design $\xi_{eq}$, we repeatedly simulate the incomplete data 300 000 times as described above. In each incomplete data set, we compute the sample estimates for the fixed effect coefficients,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \left( \sum_{i=1}^{N} \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_{i[j]}^T \boldsymbol{V}_{i[j]}^{-1} \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_{i[j]} \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta})_{i[j]}^T \boldsymbol{V}_{i[j]}^{-1} \boldsymbol{y}_{i[j]},$$

from the available cases where subscript $i[j]$ denotes the number of observed responses that are being contributed by subject $i$. To be more specific, the dimensions of the matrices vary according to the number of observed repeated measurements of each subject in the simulation. For example, we have

$$\boldsymbol{Y}_{i[2]} = \begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix}, \quad \boldsymbol{X}(\boldsymbol{t}, \boldsymbol{\delta} = \boldsymbol{1})_{i[2]} = \begin{pmatrix} 1 & t_{i1} & 1 \\ 1 & t_{i2} & 1 \end{pmatrix},$$

$$\text{with} \quad \boldsymbol{V}_i = \begin{pmatrix} 1 & 1 \\ t_{i1} & t_{i2} \end{pmatrix} \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix} \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \end{pmatrix} + \begin{pmatrix} 1 & \rho^{|t_{i1} - y_{i2}|} \\ \rho^{|t_{i2} - y_{i1}|} & 1 \end{pmatrix}$$

for subject $i$ who is in Cohort 1 and drops out from the study after the second time point. We then compute the elements of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ using the sample estimates that are

Table 5.5: Locally optimal designs for the four classes of model $\mathsf{M}_d$ with $\rho = 0.5$.

| | $t_{11}$ | $t_{12}$ | $t_{13}$ | $t_{14}$ | $t_{21}$ | $t_{22}$ | $t_{23}$ | $t_{24}$ | $n_1$ | $n_2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | Quadratic response probability | | | | Linear response probability | | | | $\delta_1 = 1$ | $\delta_2 = 0$ |
| Fixed effect model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.4201 | 0.1789 | 1 | -1 | -0.1829 | 0.4650 | 1 | 29 | 31 |
| $\xi^*_{D,1}$ | -1 | -0.2942 | 0.3609 | 1 | -1 | -0.2942 | 0.3609 | 1 | 29 | 31 |
| $\xi^*_{D,f}$ | -1 | -0.3438 | 0.3438 | 1 | -1 | -0.3438 | 0.3438 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.4659 | 0.1294 | 1 | -1 | -0.2505 | 0.4127 | 1 | 25 | 35 |
| $\xi^*_{A,1}$ | -1 | -0.3325 | 0.3309 | 1 | -1 | -0.3325 | 0.3309 | 1 | 25 | 35 |
| $\xi^*_{A,f}$ | -1 | -0.3356 | 0.3356 | 1 | -1 | -0.3356 | 0.3356 | 1 | 35 | 25 |
| Random intercept model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.4396 | 0.1496 | 1 | -1 | -0.1962 | 0.4510 | 1 | 29 | 31 |
| $\xi^*_{D,1}$ | -1 | -0.3084 | 0.3448 | 1 | -1 | -0.3084 | 0.3448 | 1 | 29 | 31 |
| $\xi^*_{D,f}$ | -1 | -0.3605 | 0.3605 | 1 | -1 | -0.3605 | 0.3605 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.5133 | 0.0671 | 1 | -1 | -0.3102 | 0.3620 | 1 | 25 | 35 |
| $\xi^*_{A,1}$ | -1 | -0.3872 | 0.2769 | 1 | -1 | -0.3872 | 0.2769 | 1 | 25 | 35 |
| $\xi^*_{A,f}$ | -1 | -0.3356 | 0.3356 | 1 | -1 | -0.3356 | 0.3356 | 1 | 35 | 25 |
| Random intercept and slope model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.5910 | -0.0872 | 1 | -1 | -0.4511 | 0.1604 | 1 | 29 | 31 |
| $\xi^*_{D,1}$ | -1 | -0.5261 | 0.0450 | 1 | -1 | -0.5261 | 0.0450 | 1 | 29 | 31 |
| $\xi^*_{D,f}$ | -1 | -0.3365 | 0.3365 | 1 | -1 | -0.3365 | 0.3365 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.5934 | -0.0908 | 1 | -1 | -0.4550 | 0.1552 | 1 | 25 | 35 |
| $\xi^*_{A,1}$ | -1 | -0.5189 | 0.0589 | 1 | -1 | -0.5189 | 0.0589 | 1 | 25 | 35 |
| $\xi^*_{A,f}$ | -1 | -0.3356 | 0.3356 | 1 | -1 | -0.3356 | 0.3356 | 1 | 35 | 25 |
| Correlated random intercept and slope model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.6450 | -0.1695 | 1 | -1 | -0.5312 | 0.0576 | 1 | 29 | 31 |
| $\xi^*_{D,1}$ | -1 | -0.5927 | -0.0513 | 1 | -1 | -0.5927 | -0.0513 | 1 | 29 | 31 |
| $\xi^*_{D,f}$ | -1 | -0.4074 | 0.2527 | 1 | -1 | -0.4074 | 0.2527 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.6399 | -0.1615 | 1 | -1 | -0.5214 | 0.0715 | 1 | 25 | 35 |
| $\xi^*_{A,1}$ | -1 | -0.5705 | -0.0105 | 1 | -1 | -0.5705 | -0.0105 | 1 | 25 | 35 |
| $\xi^*_{A,f}$ | -1 | -0.3877 | 0.2785 | 1 | -1 | -0.4077 | 0.2537 | 1 | 35 | 25 |

obtained from each simulated data set, and the expected values of

$$\left(\sum_{i=1}^{N} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_{i[j]}^{T} \boldsymbol{V}_{i[j]}^{-1} \boldsymbol{X}(\boldsymbol{t},\boldsymbol{\delta})_{i[j]}\right)^{-1}$$

across the simulated data for each design. The later matrix is the empirical values for the inverse total information matrix (5.11), which is the Cramér-Rao lower bound of the estimates, i.e. the lower bound of the variance of an unbiased estimator.

Table 5.6 summarises the performance of $\xi_{eq}$ and the designs that are presented in Table 5.5 for the four classes of model $\mathsf{M}_d$ with serial correlation parameter $\rho = 0.5$ and the corresponding values of $\boldsymbol{D}$ that are presented in Table 5.1. The first and the

Table 5.6: Simulation outputs which are averaged across 300 000 number of simulated sets using each considered design for model $\mathsf{M}_d$.

| | $\lvert(5.11)^{-1}\rvert$ | $\lvert\boldsymbol{cov}(\hat{\boldsymbol{\beta}})\rvert$ | Trace of (5.11) | $Tr(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}))$ |
|---|---|---|---|---|
| Fixed effect model | | | | |
| $\xi^*_{D,2}$ | **7.60046e-06** | **7.56294e-06** | 8.81412e-02 | 8.79916e-02 |
| $\xi^*_{D,1}$ | 7.65384e-06 | 7.62967e-06 | 8.82360e-02 | 8.81433e-02 |
| $\xi^*_{D,f}$ | 7.66426e-06 | 7.70704e-06 | 8.89497e-02 | 8.91742e-02 |
| $\xi^*_{A,2}$ | 7.75019e-06 | 7.74907e-06 | **8.67705e-02** | **8.67855e-02** |
| $\xi^*_{A,1}$ | 7.80595e-06 | 7.80580e-06 | 8.69374e-02 | 8.69462e-02 |
| $\xi^*_{A,f}$ | 7.96147e-06 | 7.99317e-06 | 9.52478e-02 | 9.55065e-02 |
| $\xi_{eq}$ | 8.01876e-06 | 8.02518e-06 | 9.02794e-02 | 9.02158e-02 |
| Random intercept model | | | | |
| $\xi^*_{D,2}$ | **4.58579e-05** | **4.57402e-05** | 1.90617e-01 | 1.90693e-01 |
| $\xi^*_{D,1}$ | 4.61590e-05 | 4.60394e-05 | 1.90707e-01 | 1.90711e-01 |
| $\xi^*_{D,f}$ | 4.62757e-05 | 4.65289e-05 | 1.92449e-01 | 1.92777e-01 |
| $\xi^*_{A,2}$ | 4.67747e-05 | 4.68448e-05 | **1.87238e-01** | **1.87291e-01** |
| $\xi^*_{A,1}$ | 4.70530e-05 | 4.71735e-05 | 1.87430e-01 | 1.87510e-01 |
| $\xi^*_{A,f}$ | 4.81471e-05 | 4.82363e-05 | 2.07460e-01 | 2.07768e-01 |
| $\xi_{eq}$ | 4.80829e-05 | 4.83048e-05 | 1.93907e-01 | 1.94117e-01 |
| Random intercept and slope model | | | | |
| $\xi^*_{D,2}$ | **3.95065e-04** | **3.95828e-04** | 2.91511e-01 | 2.91627e-01 |
| $\xi^*_{D,1}$ | 3.98026e-04 | 3.98885e-04 | 2.92193e-01 | 2.92374e-01 |
| $\xi^*_{D,f}$ | 4.13703e-04 | 4.17913e-04 | 2.97731e-01 | 2.98298e-01 |
| $\xi^*_{A,2}$ | 4.03958e-04 | 4.07186e-04 | **2.88330e-01** | **2.89158e-01** |
| $\xi^*_{A,1}$ | 4.07153e-04 | 4.11549e-04 | 2.89081e-01 | 2.90423e-01 |
| $\xi^*_{A,f}$ | 4.28860e-04 | 4.34749e-04 | 3.15527e-01 | 3.16958e-01 |
| $\xi_{eq}$ | 4.35252e-04 | 4.38606e-04 | 3.03564e-01 | 3.04191e-01 |
| Correlated random intercept and slope model | | | | |
| $\xi^*_{D,2}$ | **1.33389e-04** | **1.33750e-04** | 2.19538e-01 | 2.19726e-01 |
| $\xi^*_{D,1}$ | 1.34172e-04 | 1.34255e-04 | 2.20017e-01 | 2.20110e-01 |
| $\xi^*_{D,f}$ | 1.39569e-04 | 1.40619e-04 | 2.23957e-01 | 2.24217e-01 |
| $\xi^*_{A,2}$ | 1.36245e-04 | 1.37407e-04 | **2.17131e-01** | 2.18010e-01 |
| $\xi^*_{A,1}$ | 1.37185e-04 | 1.37514e-04 | 2.17629e-01 | **2.17769e-01** |
| $\xi^*_{A,f}$ | 1.45620e-04 | 1.47751e-04 | 2.35286e-01 | 2.36178e-01 |
| $\xi_{eq}$ | 1.46402e-04 | 1.47648e-04 | 2.28921e-01 | 2.29479e-01 |

second column of values correspond to the empirical values of the determinant of the inverse total information matrix (5.11) and the empirical values of $|(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}))|$; the third and the fourth column of values correspond to the empirical values of the trace of the inverse matrix (5.11) and the empirical values of $var(\hat{\beta}_0) + var(\hat{\beta}_1) + var(\hat{\beta}_2)$. These values reflect the objective functions of the design criteria, where $D$-optimality minimises the volume of the confidence ellipsoid of the estimated fixed effect parameters, and $A$-optimality minimises the total variance of the estimated fixed effect parameters of the model. Inspecting the empirical values of the functions of the inverse total information matrix (5.11) for each class of model $\mathsf{M}_d$, we see that the optimal designs, $\xi^*_{D,2}$ and $\xi^*_{A,2}$, that are obtained under the flexible condition $\mathbb{F}_d$, i.e. different sets of time points of measuring the outcome variable on different cohorts are allowed, achieve the smallest objective value among the considered designs for the corresponding design criterion (see the first and the third column of values in Table 5.6). Following that, the optimal designs that are subjected to the restrictive design condition $\mathfrak{R}_d$, i.e. $\xi^*_{D,1}$ and $\xi^*_{A,1}$ respectively, have the second smallest optimality objective values among the considered designs. This finding is analogous when the empirical values of $|(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}))|$ and trace of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ are compared respectively among the considered designs for each class of the model, except for the $A$-optimal cohort design for model $\mathsf{M}_d$ that has correlated random intercept and slope parameters. Comparing the empirical values of trace of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ that are obtained from using the considered designs for the correlated random intercept and slope model, the simulation studies show that $\xi^*_{A,1}$ has the smallest value (2.17769e-01) across the considered designs, which is 0.000241 smaller than $\xi^*_{A,2}$'s (2.18010e-01). This small difference that arose in the comparison of the empirical values may be due to the Monte Carlo simulation error.

We now assess the performance of some cohort designs for each class of model $\mathsf{M}_g$, for the above described experiment that has two cohorts and four repeated measurement, with $\mathfrak{X} = [-1, 1]$, $\rho = 0.5$, the presented values in Table 5.1 for $\boldsymbol{D}$, and the linear and the quadratic response probability function for the two groups of experimental units respectively. Recall that model $\mathsf{M}_g$ analyses the longitudinal data where all experimental units have the same baseline measurements. Table 5.7 shows the locally $A$- and $D$-optimal cohort designs for the corresponding classes of model $\mathsf{M}_g$, where the approach for finding the optimal cohort designs for the previously considered model is employed. In the table, $\xi^*_{D,2}$ and $\xi^*_{A,2}$ minimise the negative determinant and the trace respectively, of the right hand side of matrix (5.12) under condition $\mathbb{F}_g$; $\xi^*_{D,1}$ and $\xi^*_{A,1}$ minimise the corresponding functions of the left hand side of matrix (5.12) under condition $\mathfrak{R}_g$; and $\xi^*_{D,f}$ and $\xi^*_{A,f}$ minimise the corresponding functions of the total information matrix (5.8) of the framework that assumes completely observed responses. The corresponding total information matrices and the constraints are summarized in Table 5.4.

Table 5.7: Locally optimal designs for the four classes of model $\mathsf{M}_g$ with $\rho = 0.5$.

| | $t_{11}$ | $t_{12}$ | $t_{13}$ | $t_{14}$ | $t_{21}$ | $t_{22}$ | $t_{23}$ | $t_{24}$ | $n_1$ | $n_2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | Quadratic response probability | | | | Linear response probability | | | | | |
| Fixed effect model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.3229 | 0.2615 | 1 | -1 | -0.0736 | 0.5277 | 1 | 29 | 31 |
| $\xi^*_{D,1}$ | -1 | -0.1838 | 0.4369 | 1 | -1 | -0.1838 | 0.4369 | 1 | 29 | 31 |
| $\xi^*_{D,f}$ | -1 | -0.3729 | 0.3729 | 1 | -1 | -0.3729 | 0.3729 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.3047 | 0.2710 | 1 | -1 | -0.0542 | 0.5348 | 1 | 30 | 30 |
| $\xi^*_{A,1}$ | -1 | -0.1704 | 0.4390 | 1 | -1 | -0.1704 | 0.4390 | 1 | 30 | 30 |
| $\xi^*_{A,f}$ | -1 | -0.3821 | 0.3821 | 1 | -1 | -0.3821 | 0.3821 | 1 | 30 | 30 |
| Random intercept model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.3815 | 0.2007 | 1 | -1 | -0.1306 | 0.4895 | 1 | 30 | 30 |
| $\xi^*_{D,1}$ | -1 | -0.2503 | 0.3827 | 1 | -1 | -0.2503 | 0.3827 | 1 | 30 | 30 |
| $\xi^*_{D,f}$ | -1 | -0.4255 | 0.4255 | 1 | -1 | -0.4255 | 0.4255 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.3972 | 0.1862 | 1 | -1 | -0.1523 | 0.4759 | 1 | 30 | 30 |
| $\xi^*_{A,1}$ | -1 | -0.2771 | 0.3591 | 1 | -1 | -0.2771 | 0.3591 | 1 | 30 | 30 |
| $\xi^*_{A,f}$ | -1 | -0.3821 | 0.3821 | 1 | -1 | -0.3821 | 0.3821 | 1 | 30 | 30 |
| Random intercept and slope model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.5744 | -0.0643 | 1 | -1 | -0.4290 | 0.1874 | 1 | 29 | 31 |
| $\xi^*_{D,1}$ | -1 | -0.5026 | 0.0778 | 1 | -1 | -0.5026 | 0.0778 | 1 | 29 | 31 |
| $\xi^*_{D,f}$ | -1 | -0.3458 | 0.3458 | 1 | -1 | -0.3458 | 0.3458 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.5602 | -0.0465 | 1 | -1 | -0.4059 | 0.2129 | 1 | 30 | 30 |
| $\xi^*_{A,1}$ | -1 | -0.4881 | 0.0937 | 1 | -1 | -0.4881 | 0.0937 | 1 | 30 | 30 |
| $\xi^*_{A,f}$ | -1 | -0.3821 | 0.3821 | 1 | -1 | -0.3821 | 0.3821 | 1 | 30 | 30 |
| Correlated random intercept and slope model | | | | | | | | | | |
| $\xi^*_{D,2}$ | -1 | -0.6336 | -0.1497 | 1 | -1 | -0.5152 | 0.0835 | 1 | 29 | 31 |
| $\xi^*_{D,1}$ | -1 | -0.5744 | -0.0187 | 1 | -1 | -0.5744 | -0.0187 | 1 | 29 | 31 |
| $\xi^*_{D,f}$ | -1 | -0.4383 | 0.2144 | 1 | -1 | -0.4383 | 0.2144 | 1 | 30 | 30 |
| $\xi^*_{A,2}$ | -1 | -0.6171 | -0.1270 | 1 | -1 | -0.4904 | 0.1158 | 1 | 29 | 31 |
| $\xi^*_{A,1}$ | -1 | -0.5539 | 0.0096 | 1 | -1 | -0.5539 | 0.0096 | 1 | 29 | 31 |
| $\xi^*_{A,f}$ | -1 | -0.4524 | 0.2132 | 1 | -1 | -0.4524 | 0.2132 | 1 | 30 | 30 |

For each considered design, we simulate

$$y_{ij} = \begin{cases} 1 + t_{1j} + b_{0i} + t_{1j}b_{1i} + \epsilon_{ij} & \text{for subject } i \text{ in Group 1,} \\ 1 + t_{2j} + b_{0i} + t_{2j}b_{1i} + \epsilon_{ij} & \text{for subject } i \text{ in Group 2,} \end{cases}$$

where $t_{1j}$ and $t_{2j}$ correspond to the $j^{th}$ time point of measuring the outcome variable on Group 1 that has size $n_1$ and Group 2 that has size $n_2$ respectively, $\epsilon_{ij} \sim N(0, \sigma^2\boldsymbol{\psi})$, $\boldsymbol{\psi} = 0.5^{|t_j - t_{j'}|}$, $\sigma^2 = 1$, and $\boldsymbol{b_i} \sim N(0, \boldsymbol{D})$ with the corresponding values of $\boldsymbol{D}$ for each class of model $\mathsf{M}_g$. The optimal time points, the size of each group of experimental units, i.e. $n_1$ and $n_2$, for each design are presented in Table 5.7. Missing values in the vector of observed responses of subject $i$, i.e. $\boldsymbol{y}_i = \{y_{i1}, ..., y_{i4}\}$, are then introduced by specifying the corresponding response probability functions for each group of experimental units.

In this example, Group 1 has quadratic response probability function (5.15) and Group 2 has linear response probability function (5.14).

Similar to the previously considered model, for each design in Table 5.7 and $\xi_{eq}$, we repeatedly simulate the incomplete data 300 000 times as described above. In each incomplete data set, we compute the sample estimates for the fixed effect coefficients,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \left( \sum_{i=1}^{N} \boldsymbol{K}_i^T \boldsymbol{X}_{i[j]}^T \boldsymbol{V}_{i[j]}^{-1} \boldsymbol{X}_{i[j]} \boldsymbol{K}_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{K}_i^T \boldsymbol{X}_{i[j]}^T \boldsymbol{V}_{i[j]}^{-1} \boldsymbol{y}_{i[j]},$$

from the available cases where subscript $i[j]$ denotes the number of observed responses that are being contributed by subject $i$, and

$$\boldsymbol{K}_i = \begin{cases} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} & \text{if experimental unit } i \text{ is in Group 1;} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \text{if experimental unit } i \text{ is in Group 2.} \end{cases}$$

We then compute the elements of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ using the sample estimates that are obtained from each simulated data set, and the expected values of

$$\left( \sum_{i=1}^{N} \boldsymbol{K}_i^T \boldsymbol{X}_{i[j]}^T \boldsymbol{V}_{i[j]}^{-1} \boldsymbol{X}_{i[j]} \boldsymbol{K}_i \right)^{-1}$$

across the simulated data for each design. The later matrix is the empirical values for the inverse total information matrix (5.12), which is the Cramér-Rao lower bound of the estimates, i.e. the lower bound of the variance of an unbiased estimator.

Table 5.8 shows the performance of the equally spaced time point design, $\xi_{eq}$, and of the designs that are presented in Table 5.7 for each class of model $\mathsf{M}_g$. As expected, the locally $D$-optimal cohort designs, $\xi^*_{D,2}$, which are found under the flexible design condition $\mathbb{F}_g$, yield the smallest empirical values of $D$-optimality objective values and the empirical values of $|\boldsymbol{cov}(\hat{\boldsymbol{\beta}})|$ for each class of model $\mathsf{M}_g$. For $A$-optimality, $\xi^*_{A,2}$ of all classes of model $\mathsf{M}_g$ yield the smallest empirical values of $A$-optimality objective values and the empirical values of trace of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$, except for model $\mathsf{M}_g$ that has a random intercept parameter, and model $\mathsf{M}_g$ that has correlated random intercept and slope parameters. The designs $\xi^*_{A,2}$ for these two classes of model $\mathsf{M}_g$ have empirical values, 8.36500e-02 and 2.66690e-01 respectively for the trace of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$, which are 0.0000237 and 0.000321 larger than the empirical values of the trace of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ that are obtained from using $\xi^*_{D,2}$ for the corresponding classes of model $\mathsf{M}_g$. On the other hand, comparing the simulation outputs and the optimality objective values (which are not presented here) for each design, we find that the performance of the locally $A$-

Table 5.8: Simulation outputs which are averaged across 300 000 number of simulated sets using each considered design for model $\mathsf{M}_g$.

| | $\lvert(5.12)^{-1}\rvert$ | $\lvert\boldsymbol{cov}(\hat{\boldsymbol{\beta}})\rvert$ | Trace of (5.12) | $Tr(\boldsymbol{cov}(\hat{\boldsymbol{\beta}}))$ |
|---|---|---|---|---|
| Fixed effect model | | | | |
| $\xi^*_{D,2}$ | **6.21178e-06** | **6.19010e-06** | 5.96127e-02 | 5.95632e-02 |
| $\xi^*_{D,1}$ | 6.25394e-06 | 6.26074e-06 | 5.97801e-02 | 5.98266e-02 |
| $\xi^*_{D,f}$ | 6.30747e-06 | 6.31969e-06 | 5.99563e-02 | 5.99723e-02 |
| $\xi^*_{A,2}$ | 6.21608e-06 | 6.21549e-06 | **5.95501e-02** | **5.95090e-02** |
| $\xi^*_{A,1}$ | 6.25903e-06 | 6.27858e-06 | 5.97240e-02 | 5.97829e-02 |
| $\xi^*_{A,f}$ | 6.31543e-06 | 6.32112e-06 | 5.99908e-02 | 5.99896e-02 |
| $\xi_{eq}$ | 6.60018e-06 | 6.59978e-06 | 6.11550e-02 | 6.11309e-02 |
| Random intercept model | | | | |
| $\xi^*_{D,2}$ | **1.89489e-05** | **1.90291e-05** | 8.35567e-02 | **8.36263e-02** |
| $\xi^*_{D,1}$ | 1.91258e-05 | 1.92770e-05 | 8.38519e-02 | 8.41215e-02 |
| $\xi^*_{D,f}$ | 1.94567e-05 | 1.95737e-05 | 8.43589e-02 | 8.45077e-02 |
| $\xi^*_{A,2}$ | 1.89446e-05 | 1.90270e-05 | **8.35406e-02** | 8.36500e-02 |
| $\xi^*_{A,1}$ | 1.91249e-05 | 1.91755e-05 | 8.38315e-02 | 8.38686e-02 |
| $\xi^*_{A,f}$ | 1.92587e-05 | 1.93293e-05 | 8.40284e-02 | 8.41222e-02 |
| $\xi_{eq}$ | 2.03302e-05 | 2.04521e-05 | 8.57588e-02 | 8.59024e-02 |
| Random intercept and slope model | | | | |
| $\xi^*_{D,2}$ | **7.05637e-04** | **7.06181e-04** | 3.16752e-01 | 3.16618e-01 |
| $\xi^*_{D,1}$ | 7.09584e-04 | 7.10538e-04 | 3.17335e-01 | 3.17420e-01 |
| $\xi^*_{D,f}$ | 7.26840e-04 | 7.29370e-04 | 3.19022e-01 | 3.18926e-01 |
| $\xi^*_{A,2}$ | 7.06675e-04 | 7.08598e-04 | **3.16404e-01** | **3.16483e-01** |
| $\xi^*_{A,1}$ | 7.10600e-04 | 7.15250e-04 | 3.17002e-01 | 3.17368e-01 |
| $\xi^*_{A,f}$ | 7.28626e-04 | 7.31955e-04 | 3.19337e-01 | 3.19443e-01 |
| $\xi_{eq}$ | 7.59178e-04 | 7.63680e-04 | 3.23854e-01 | 3.24456e-01 |
| Correlated random intercept and slope model | | | | |
| $\xi^*_{D,2}$ | **2.31907e-04** | **2.32342e-04** | 2.66417e-01 | **2.66369e-01** |
| $\xi^*_{D,1}$ | 2.32867e-04 | 2.33925e-04 | 2.66968e-01 | 2.66967e-01 |
| $\xi^*_{D,f}$ | 2.36923e-04 | 2.36508e-04 | 2.68610e-01 | 2.68588e-01 |
| $\xi^*_{A,2}$ | 2.32008e-04 | 2.32881e-04 | **2.66397e-01** | 2.66690e-01 |
| $\xi^*_{A,1}$ | 2.32969e-04 | 2.33116e-04 | 2.66949e-01 | 2.67059e-01 |
| $\xi^*_{A,f}$ | 2.36666e-04 | 2.39207e-04 | 2.68512e-01 | 2.69179e-01 |
| $\xi_{eq}$ | 2.47324e-04 | 2.49400e-04 | 2.74618e-01 | 2.75535e-01 |

and $D$-optimal designs that are obtained under condition $\mathfrak{R}_g$ are not better than the performance of some designs. For example in Table 5.8, the optimal design $\xi^*_{A,2}$ achieves the second smallest $|(5.12)^{-1}|$ and $\xi^*_{D,2}$ achieves the second smallest trace of $(5.12)^{-1}$ among the considered designs for each class of model $\mathsf{M}_g$. Nevertheless, comparing the performance of $\xi^*_{A,1}$ and $\xi^*_{D,1}$ for each class of model $\mathsf{M}_g$, we find that the optimal cohort designs achieve the corresponding optimality objective values in the simulation, except $\xi^*_{D,1}$ for model $\mathsf{M}_g$ that has a random intercept parameter. For this class of model, $\xi^*_{D,1}$ has empirical value 1.91258e-05 for $|(5.12)^{-1}|$, which is 9e-10 larger than the empirical value of $|(5.12)^{-1}|$ when $\xi^*_{A,1}$ (1.91249e-05) is considered. All of these small differences that arose in the comparisons of the empirical values may be due to the simulation noise.

To summarise the investigation on the performance of the locally $A$- and $D$-optimal cohort designs for each class of model $\mathsf{M}_d$ and $\mathsf{M}_g$, we consider the loss of efficiency of the worse design when instead an optimal design $\xi^*_{D,2}$ or $\xi^*_{A,2}$ should have been chosen. These efficiency loss are

$$RE_D = \left(\frac{|I_{\xi^*_{D,2}}|}{|I_\xi|}\right)^{1/p} \quad \text{and} \quad RE_A = \frac{\text{Trace of } I_{\xi^*_{A,2}}}{\text{Trace of } I_\xi}$$

for $D$-efficiency and $A$-efficiency respectively, where $I$ denotes the corresponding inverse total information matrices and $p$ is the number of fixed effect coefficients in the model. In the examples that are considered here, we have $1/p = 1/3$ in $RE_D$ for both model $\mathsf{M}_d$ and $\mathsf{M}_g$.

Table 5.9 shows the efficiency loss of the worst design for each scenario, where the efficiency loss are computed using the largest and the smallest optimality objective values, i.e. the empirical values of the determinant or the trace of the corresponding inverse total information matrix that are obtained from the simulation studies. For example we compute $RE_A$ of $\xi^*_{A,f}$ for each class of model $\mathsf{M}_d$ with the empirical values of trace of (5.11) in Table 5.6, that are obtained from using $\xi^*_{A,2}$ and $\xi^*_{A,f}$ respectively, as the later design has the largest total variance of the estimated fixed effect parameters among the considered designs. We observe that for the examples that we have considered here, the $A$-efficiency loss due to the presence of dropouts is the greatest when the design that assumes completely observed repeated measurements is employed for model $\mathsf{M}_d$ with the fixed effect parameters, i.e. the performance of $\xi^*_{A,2}$ for the fixed effect model is expected to be about 9% better than the performance of $\xi^*_{A,f}$ when there are dropouts in the longitudinal study according to the considered dropout processes. Moreover for each class of model $\mathsf{M}_d$, we notice that the locally $A$-optimal cohort design that assumes completely observed repeated measurements performs worse than the equally spaced time point designs when there are dropouts in the longitudinal study that has different baseline measurements for different groups. On the other hand for $D$-optimality, we find that there are not more than 5% efficiency loss when the performances of the worst

Table 5.9: $A$- and $D$-efficiency loss for each class of model $\mathsf{M}_d$ and $\mathsf{M}_g$.

| | FE | RI | RIRS | RIRSc |
|---|---|---|---|---|
| $RE_D$ of $\mathsf{M}_d$ | 0.982 | 0.984 | 0.968 | 0.969 |
| $RE_D$ of $\mathsf{M}_g$ | 0.980 | 0.977 | 0.976 | 0.979 |
| $RE_A$ of $\mathsf{M}_d$ | 0.911 | 0.903 | 0.914 | 0.923 |
| $RE_A$ of $\mathsf{M}_g$ | 0.974 | 0.974 | 0.977 | 0.970 |

designs are compared with the performances of the corresponding $\xi^*_{D,2}$ for each class of model $\mathsf{M}_d$ and model $\mathsf{M}_g$.

## 5.4 Application: Redesigning a study on Alzheimer's disease

We now illustrate an application of our design framework using the data from an Alzheimer's disease study. Full details of the study are depicted in Howard et al. (2012). In general, this study has four groups of patients with five repeated measurements in each group at week 0, 6, 18, 30, and 52 respectively. The aim of the study is to explore the changes from baseline measurement in each group over the period of 52 weeks. For illustration purposes, we only consider the experimental units in the placebo group and the donepezil-memantine group, who were included in the primary intention-to-treat sample. Here we treat the rate of change of the primary outcome measures, SMMSE score (higher score indicates better cognitive function), as the response variable of our model. Since every subject has rate zero at week 0, we employ model $\mathsf{M}_g$ (which has the same intercept parameter but different slopes parameters for different groups) for this study.

Fitting the rate of change of SMMSE score of the two groups to the possible classes of model $\mathsf{M}_g$ with an AR(1) process to capture the within-subject correlation in $R$ program, we find that the fixed effect model

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{K}_i \hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon_i},$$

with

$$\boldsymbol{X}_i \boldsymbol{K}_i \hat{\boldsymbol{\beta}} = \begin{cases} 0.01504 - 0.003730\ t_{1j} & \text{if subject } i \text{ is in placebo group,} \\ 0.01504 - 0.001137\ t_{2j} & \text{if subject } i \text{ is in donepezil-memantine group,} \end{cases}$$

and $\boldsymbol{\epsilon_i}$ is normally distributed with $\hat{\sigma}^2 = 0.2465^2$ and $\hat{\rho} = 0.1562$, has the smallest AIC and BIC values among the possible classes of models (see Table 5.10). The total sample size of the data that is used in this illustration is $N = 72 + 72 = 144$. From the fitted values of the fixed effect coefficients, we see that one unit change in time results in more

Table 5.10: AIC and BIC of the possible classes of model $\mathsf{M}_g$, which are computed using the considered data.

| | FE | RI | RIRS |
|---|---|---|---|
| AIC | 45.16 | 45.65 | 47.65 |
| BIC | 65.01 | 69.46 | 75.43 |

Table 5.11: Number of subjects who remain in the study at each time point (extracted from Howard et al. (2012)).

| No. of Patients still Receiving Study Drug | | | | | |
|---|---|---|---|---|---|
| Week | 0 | 6 | 18 | 30 | 52 |
| Placebo group | 72 | 67 | 41 | 32 | 20 |
| Donepezil-memantine group | 72 | 68 | 63 | 56 | 38 |

Table 5.12: Middle time points of some designs for model $\mathsf{M}_g$ with fixed effect parameters, $N$=144, $q = 5$. The expected number of observations at $t_{ij}$ is shown under the support point. The last column shows the empirical values of $|\boldsymbol{cov}(\hat{\boldsymbol{\beta}})|$ that are averaged over 300 000 simulated sets.

| | $t_{12}$ | $t_{13}$ | $t_{14}$ | $t_{22}$ | $t_{23}$ | $t_{24}$ | $n_1$ | $n_2$ | $\lvert\boldsymbol{cov}(\hat{\boldsymbol{\beta}})\rvert$ |
|---|---|---|---|---|---|---|---|---|---|
| | placebo group | | | donepezil-memantine group | | | | | (e-17) |
| $\xi_{ori}$ | 6 | 18 | 32 | 6 | 18 | 32 | 72 | 72 | 4.67040 |
| | 62.6 | 52.7 | 38.1 | 68.5 | 64.8 | 58.0 | | | |
| $\xi^*_{D,2}$ | 2.3793 | 38.7863 | 42.4455 | 2.4408 | 46.4454 | 49.3676 | 71 | 73 | **2.30383** |
| | 64.6 | 26.6 | 22.2 | 69.2 | 42.5 | 39.2 | | | |
| $\xi^*_{D,1}$ | 2.4345 | 44.6476 | 48.2657 | 2.4345 | 44.6476 | 48.2657 | 71 | 73 | 2.42169 |
| | 64.6 | 19.8 | 16.2 | 69.2 | 44.5 | 40.5 | | | |
| $\xi^*_{D,f}$ | 2.4621 | 47.4557 | 49.9046 | 2.4621 | 47.4557 | 49.9046 | 72 | 72 | 2.53927 |
| | 64.5 | 16.9 | 14.7 | 69.2 | 41.4 | 38.6 | | | |

changes in the SMMSE score of the placebo group than the changes of the score in the donepezil-memantine group.

To obtain some response probability functions for the two groups, we use the information in Table 5.11, i.e. the numbers of subjects who remain in the study over the period of 52 weeks (extracted from Howard et al. (2012)), for fitting a logistic regression model for the two groups respectively. We obtain the following response probability

$$P_{obs}(t_{ij}) = \frac{1}{1 + exp(\alpha_0 + \alpha_1 t_{ij})}$$

with $\{\alpha_0, \alpha_1\} = \{-2.3403, 0.07418\}$ for the placebo group, and $\{\alpha_0, \alpha_1\} = \{-3.3514, 0.06427\}$ for the donepezil-memantine group. Employing these response probability functions in the design framework, we find the locally $D$-optimal cohort design for the above described fixed effect model over the design region $\mathfrak{X} = [0, 52]$, by fixing week 0 and week 52 as the first and the last time point of measuring the SMMSE scores on both groups.

Table 5.12 shows the original design that is used in the clinical study, $\xi_{ori}$, the $D$-optimal cohort design $\xi^*_{D,2}$, $\xi^*_{D,1}$ and $\xi^*_{D,f}$ which correspond to the design that allows for having different sets of time points for different cohorts (i.e. under condition $\mathbb{F}_g$), the restricted design that has the same set of time points for both groups (i.e. under condition $\mathfrak{R}_g$), and the optimal design which assumes completely observed repeated measurements respectively. The number under each time point reflects the expected number of subjects who remain in the study at the time point, which is computed using the corresponding fitted inverse logit link response probability function. If $\xi^*_{D,2}$ was employed in the Alzheimer's disease clinical trial, the experimental units in the placebo group would be measured on week 2.4, 38.8, 42.4 and 52 respectively, and the donepezil-memantine group would be measured on week 2.4, 46.4, 49.4 and 52 respectively after the baseline measurement in week 0. In this case, the responses of the placebo group would have been observed earlier (before week 52) to avoid having large proportion of dropouts in the study. For a restricted experiment that employs $\xi^*_{D,1}$, the experimenters would collect measurements on both groups at week 2.4, 44.6, 48.3 and 52 respectively after week 0. In this case, the donepezil-memantine group would be measured at some earlier weeks than those of $\xi^*_{D,2}$ before week 52, to offset the dropout effect in the placebo group.

In the last column of the table, we show the empirical value of the determinant of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$. These values are computed using the similar set-up of the simulation study that is described in the previous section, but with the above reported $\hat{\boldsymbol{\beta}}$ in generating the observations and the corresponding inverse logit link probability functions in generating the missing values. Together with the fitted values of the AR(1) process, i.e. $\hat{\sigma}^2$ and $\hat{\rho}$, we obtain very small magnitudes for $|\boldsymbol{cov}(\hat{\boldsymbol{\beta}})|$ in the simulation studies of the considered designs. Nonetheless, comparing these values, we see that $\xi_{ori}$ is not $D$-optimal even though the expected number of subjects who remain in the study at each time point is larger than those of other design candidates. Moreover, we find that the loss of efficiency of the original design, $\xi_{ori}$, when instead $\xi^*_{D,2}$ or $\xi^*_{D,1}$ should have been chosen are $(2.30383/4.6704)^{1/3} = 0.79$ and $(2.42169/4.6704)^{1/3} = 0.80$ respectively, implying that five replicates of $\xi_{ori}$ would be as efficient as four replicates of the $D$-optimal cohort designs, $\xi^*_{D,2}$ or $\xi^*_{D,1}$.

## 5.5 Conclusion and discussion

In summary, we have shown that the presence of dropouts has impact on the locally optimal cohort designs for longitudinal studies. Only locally optimal designs are available for the linear mixed models as the covariance structure of the repeated measurements is often unknown at the design stage of an experiment. Using the data from some historical/ pilot studies, we can estimate this covariance structure for constructing an optimal design for a future experiment. In our investigations, we assume an AR(1) process for

the observational errors of the experimental subjects to capture the within-subject correlation. By trying different sets of values for the variance of random coefficients, we find that the structure of $\boldsymbol{D}$ rather than the values of its elements have more impact on the trends of the optimal time points across $\rho$. Moreover, we learn that the invariance property of the $D$-optimal cohort designs vanishes for the linear mixed models with an AR(1) process. This finding applies to both the design framework that accounts for the impact of dropouts and the design framework that assumes completely observed repeated measurements, reflecting the fact that the structure of $\sigma^2\boldsymbol{\psi}$ plays a role in finding a locally optimal cohort design.

In general, to find an optimal design, the upper and the lower bound of design region $\mathfrak{X}$ are always chosen as two of the support points of the design in the optimisation problem. To account for the monotone dropout mechanism at the design stage of an experiment, it is beneficial to fix the lower bound as the first time point because that is the point where maximum number of observations is available over the design region. As an extension to the work that is proposed by Ortega-Azurduy et al. (2008) and the above described framework that assumes the presence of dropouts, we suggest to treat the last support point as a variable to be found in the optimisation problem. Intuitively for a monotone decreasing response probability function, it is expected to have relatively more observations at a support point within the design region than at the upper bound of $\mathfrak{X}$. Here we illustrate the extension to the design framework of Ortega-Azurduy et al. (2008), i.e. model $\mathsf{M}_o$ with one cohort, for the four classes of linear mixed model with four repeated measurements.

Consider linear response probability function (5.14) for two design regions, $\mathfrak{X}_1 = [-1, 1]$ and $\mathfrak{X}_2 = [-1, 1.5]$ respectively, and the presented values of $\boldsymbol{D}$ in Table 5.1 for model $\mathsf{M}_o$ with one cohort and four repeated measurements. Figure 5.7 and Figure 5.8 show the locally $D$-optimal designs that are found within the two respective design regions for each class of model $\mathsf{M}_o$ across $\rho$, $0 \leq \rho \leq 0.9$. In all of these optimal designs, we fix the lower bound as the first time point (not being shown in the plots) because that is the point which has the highest response rate within the design regions. In the first row of plots in both figures, the two lines correspond to the second time point, $t_2$, and the third time point, $t_3$, of the locally $D$-optimal designs for each $\rho$, with the upper bound of the corresponding design regions being fixed as the last support point, $t_4$, of the optimal designs. In the second row of plots, the third line to the right in each plot corresponds to $t_4$ of the locally $D$-optimal designs for each $\rho$, where $t_4$ is being optimised in the design problem and constrained to be greater than both $t_2$ and $t_3$, and within the corresponding design regions. Comparing the top and the bottom plots in Figure 5.7, i.e. the $D$-optimal designs within $\mathfrak{X}_1 = [-1, 1]$, we see that the presence of dropouts affects the values of $t_4$ of the optimal designs for the random intercept and slope model, and the correlated random intercept and slope model. In the second row of plots in Figure 5.8, i.e. the $D$-optimal designs within $\mathfrak{X}_2 = [-1, 1.5]$, we see that the linear

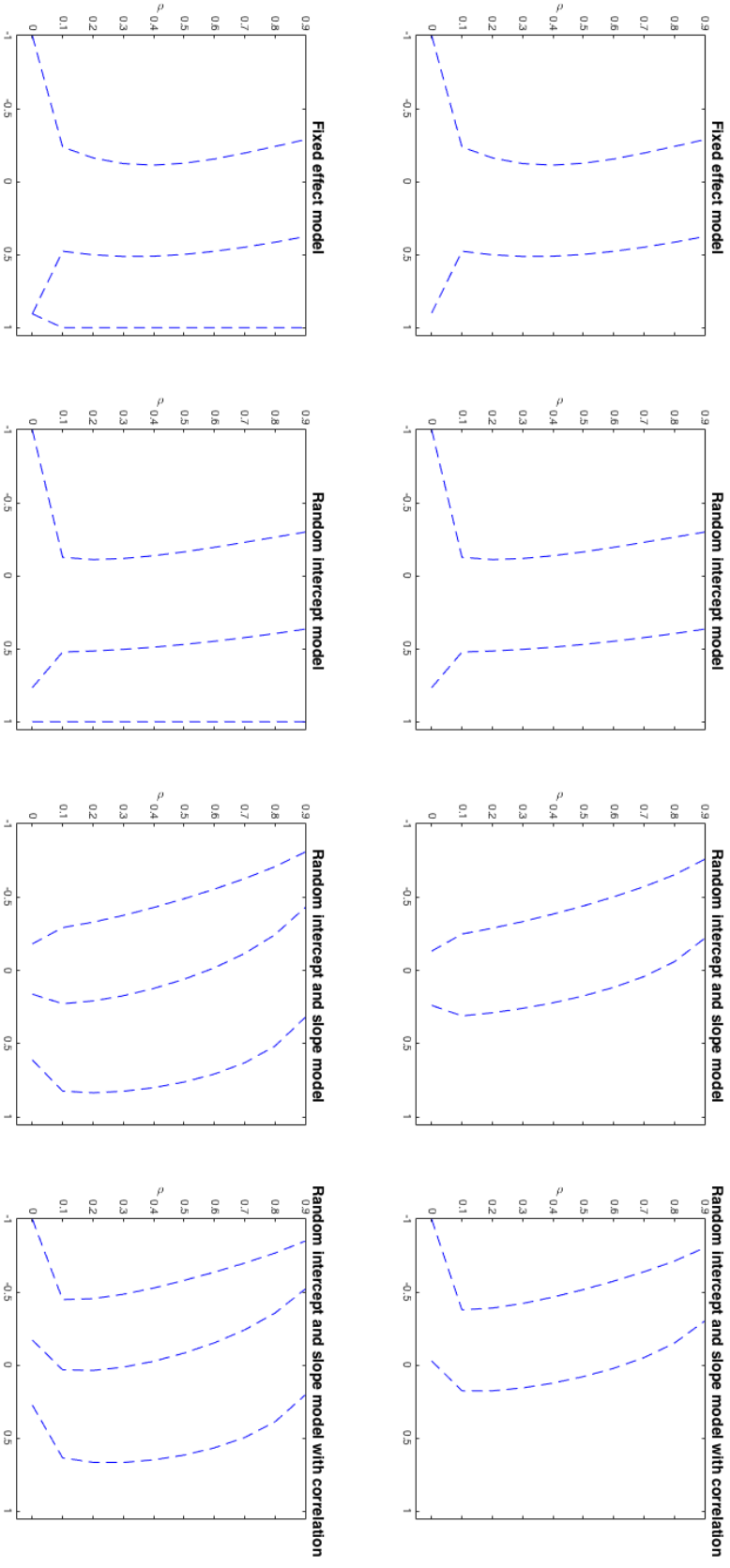



Figure 5.7: $D$-optimal time points for model $\mathsf{M}_o$ with $c = 1$ and linear response probability function (5.14), within $\mathfrak{X}_1 = [-1, 1]$. First row of plots corresponds to $t_2$ and $t_3$, with $t_4 = 1$; second row of plots corresponds to $t_2$, $t_3$ and $t_4$ across $\rho$.

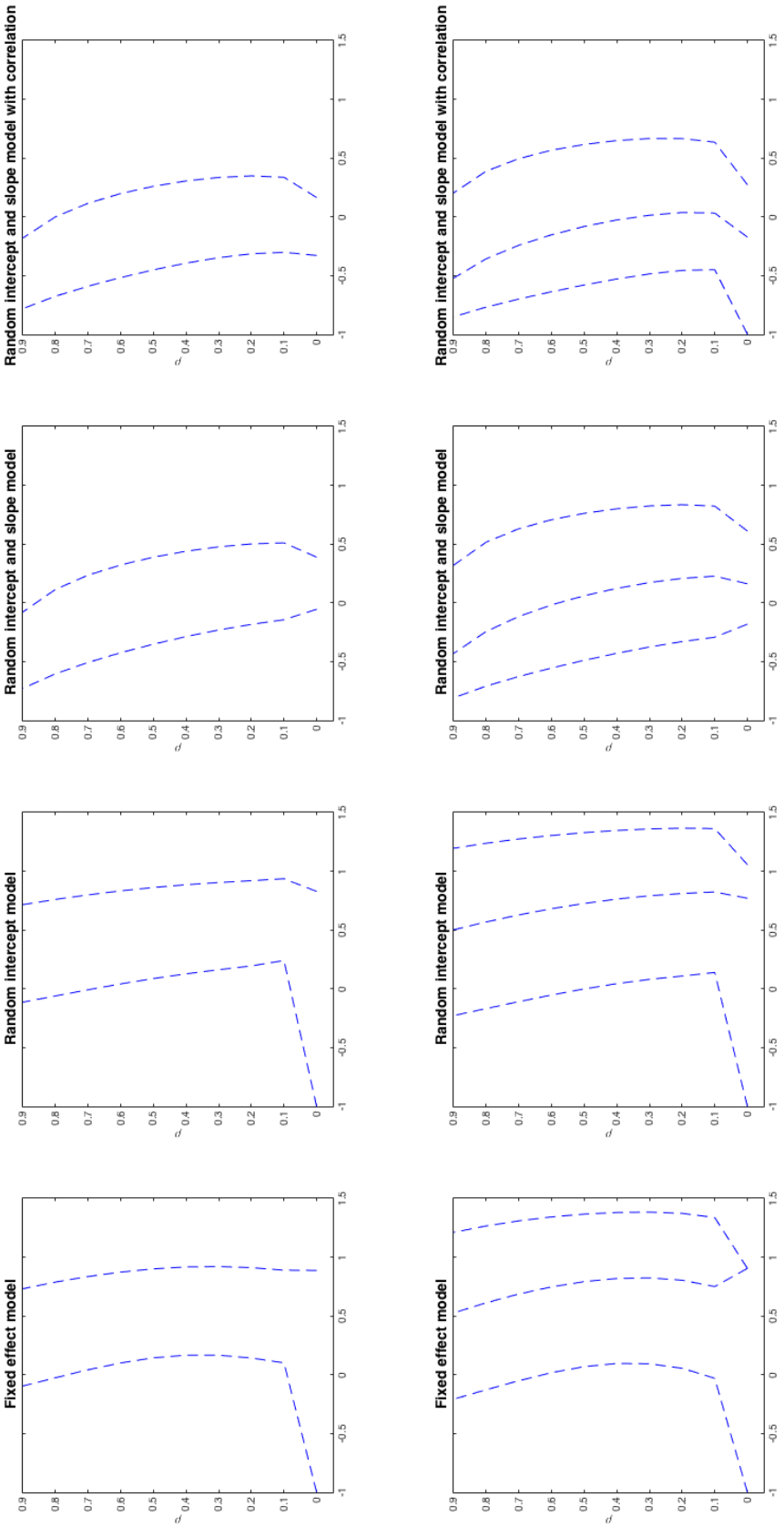



Figure 5.8: $D$-optimal time points for model $\mathsf{M}_o$ with $c = 1$ and linear response probability function (5.14), within $\mathfrak{X}_1 = [-1, 1.5]$. First row of plots corresponds to $t_2$ and $t_3$, with $t_4 = 1$; second row of plots corresponds to $t_2$, $t_3$ and $t_4$ across $\rho$.

response probability function has caused the last support points of the optimal designs to be shifted away from the upper bound of $\mathfrak{X}_2$ for all classes of model $\mathsf{M}_o$. Similar findings that are not being presented here are obtained when the optimal cohort designs for model $\mathsf{M}_d$ and $\mathsf{M}_g$ respectively are assessed with a few different design regions.

From these findings, we infer that for the optimal design framework that accounts for the impact of dropouts, choosing the upper bound of a design region as the last support point in the optimisation problem is not always the best option. Moreover, as the $D$-optimal designs for the linear mixed models with an AR(1) covariance structure for the observational errors do not possess the invariance property, it is necessary to construct an optimal design within the corresponding design region. Besides, when the design problem concerns with finding the time points of measuring the experimental units, employing a design region $[0, T]$ for time variable would facilitate the interpretation of the parameter estimates.

In this work, we have considered the optimal cohort design framework in the presence of dropouts for each class of the linear mixed models, namely fixed effect model, random intercept model, random intercept and slope model, and random intercept and slope model with correlated random coefficient. We have assessed the optimal cohort designs for two special formulations of the linear mixed model, i.e. model $\mathsf{M}_g$ that has the same intercept parameter but different slopes for different cohorts, and model $\mathsf{M}_d$ that has different intercepts but the same slope parameter for different groups. Moreover, we have studied two different experimental conditions, i.e. a restricted condition where all experimental subjects must have the same set of optimal time points of measuring the outcome variable (condition $\mathfrak{R}_g$ and $\mathfrak{R}_d$), and a flexible condition that allows for having different sets of optimal time points of measuring the outcome variable on different cohorts (condition $\mathbb{F}_g$ and $\mathbb{F}_d$). We note that in practice, clinical studies often require blinding of the experimental units. Hence the restricted design condition would be more appropriate for implementation purposes.

For future research, we suggest to consider different structures for the serial correlations and the impact of an intermittent missing data pattern on the optimal cohort designs for the linear mixed models. Some measures of robustness could also be studied against the covariance structure of responses, such as considering the efficiency of the optimal designs when a different structure is assumed or in the limiting cases of the covariance structure, e.g. when $\rho = 0$ or $\rho = \infty$ in the AR(1) process. Besides that, the extra variable that differentiates the conditions of different cohorts in $\mathsf{M}_d$ could be considered as a variable in the optimisation problem, with a missing data mechanism that depends on both the time variable and the extra variable. Apart from the available case analysis, the design framework could also account for other missing data analysis approaches such as multiple imputation as the extensions to the current work. Moreover, the design framework for more complicated models such as the generalized linear mixed models might be investigated in the presence of dropouts. The key challenge to considering these

suggestions is to find the expected total information for the corresponding models, having accounted for the features of the missing data analysis approach at the design stage of an experiment. More sophisticated optimisation algorithm might also be required to solve the complex objective function for a goal of a longitudinal study.

# Chapter 6

# Optimal designs when missing values are imputed repeatedly

In this chapter, we consider the optimal design framework for the linear regression model where the missing responses are assumed to be repeatedly replaced by some random samples of a plausible distribution before standard data analysis is implemented. This approach is called multiple imputation, which is a flexible and appealing approach for dealing with the presence of missing data. In the literature, extensive work has been done on the imputation models based on the assumptions for the incomplete data. However, to the best of our knowledge, none of the literature on design of experiments has considered the role of multiple imputation in the optimal design framework. Here we assess the consequences of employing multiple imputation on the optimal designs for the linear regression models.

In general, a multiple imputation approach consists of three steps:

1. Replace the missing values in the data.
2. Analyse the data using standard analysis approach by treating the observed and the imputed values as a complete-data set.
3. Repeat the above procedures for $t$ times. Combine the $t$ outputs.

In our investigation, we impute the missing responses with a linear regression model in Step 1, and analyse the complete-data sets with the same formulation of the linear regression model in Step 2. We then combine the outputs using Rubin's combining rules in Step 3.

To be more specific, before imputing the missing values we construct a posterior distribution, $P(\boldsymbol{\beta}^{(*)}|\boldsymbol{x_{obs}}, \boldsymbol{y_{obs}})$, for the missing data governing parameter, $\boldsymbol{\beta}^{(*)}$, of the imputation model. Employing a non-informative prior distribution (an improper prior),

i.e. $P(\boldsymbol{\beta^{(*)}}) \propto 1$, and the likelihood function of $\boldsymbol{\beta^{(*)}}$ which is constructed using the observed data $\boldsymbol{x_{obs}}$ and $\boldsymbol{y_{obs}}$, we have

$$P(\boldsymbol{\beta^{(*)}}|\boldsymbol{x_{obs}}, \boldsymbol{y_{obs}}) \sim \boldsymbol{N}(\hat{\boldsymbol{\beta}}_{obs}, V_{\hat{\boldsymbol{\beta}}_{obs}}\sigma^2)$$

where $\boldsymbol{N}$ stands for a multivariate normal distribution,

$$\hat{\boldsymbol{\beta}}_{obs} = (X_{obs}^T X_{obs})^{-1} X_{obs}^T \boldsymbol{y_{obs}},$$

$$V_{\hat{\boldsymbol{\beta}}_{obs}} = (X_{obs}^T X_{obs})^{-1},$$

$X_{obs}$ is the observed design matrix containing the rows of $f^T(x_{obs})$ with dimension $n_{obs} \times p$, $x_{obs}$ is the value of explanatory variable that has an observation, $n_{obs}$ is the number of subjects who have observed responses, $\boldsymbol{y_{obs}}$ is a vector containing the observed responses, and $\sigma^2$ is the variance of the error term $\boldsymbol{\epsilon}$ in the linear regression model that is assumed to be known in the study.

In practice, different prior distributions might be employed for the construction of the imputation model. Nonetheless, we only consider the non-informative prior distribution in our investigation such that the imputation model coincides with the analysis model. In the remaining part of this chapter, we shall denote $(X_{obs}^T X_{obs})^{-1} = (X^T X)_{obs}^{-1}$. Note that this matrix is proportional to $\boldsymbol{cov(\hat{\beta})}$ in a complete case analysis for the linear regression model, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_p)^T$ are the least squares estimates that are computed using completely observed information of the subjects.

Assuming missing values present only in the response variable of the model and that they are missing at random, we impute the missing response of subject $i$ by

$$y_{i(*)} = f^T(x_i)\boldsymbol{\beta}^* + \epsilon_i$$

where $\boldsymbol{\beta^*} = (\beta_0^*, ..., \beta_p^*)^T$ are the $(*)^{th}$ independent draw from $P(\boldsymbol{\beta^{(*)}}|\boldsymbol{x_{obs}}, \boldsymbol{y_{obs}})$, and $\epsilon_i$ is a random draw from $N(0, \sigma^2)$ that is independent of $\boldsymbol{\beta^*}$. In our investigation, we assume this imputation model has the same formulation as the linear regression model that analyses the complete-data sets in Step 2. In the remaining part of this chapter, for the experimental units who have missing observations, we denote the set of imputed responses by $\{y_{i(l)}\}$, $l = 1, .., t$ is the index of the repeated imputations; two different imputed sets at the corresponding value of the explanatory variable by $\{y_{i(*)}\}$ and $\{y_{i(**)}\}$ respectively, $i = 1, ..., N$.

In Step 2 of the multiple imputation, a complete-data set, $\boldsymbol{y}_{(*)}$ that contains $\boldsymbol{y_{obs}}$ and $\{y_{i(*)}\}$, is employed in the computation of the $(*)^{th}$ repeated imputation least squares estimates,

$$\hat{\boldsymbol{\beta}}^{(*)} = (X^T X)^{-1} X^T \boldsymbol{y_{(*)}},$$

for the linear regression model. These estimates have

$$\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}^{(*)}) = \boldsymbol{cov}\left((X^TX)^{-1}X^T\boldsymbol{y_{(*)}}\right).$$

Using Rubin's combining rules, which are proposed by Rubin (1987), we get the parameters of interest,

$$\bar{\boldsymbol{\beta}} = \frac{\sum_{l=1}^{t}\hat{\boldsymbol{\beta}}^{(l)}}{t}$$

in Step 3, with total variance-covariance

$$\mathcal{T}_t = \bar{\boldsymbol{U}}_t + \left(1+\frac{1}{t}\right)\boldsymbol{B}_t$$

where

$$\bar{\boldsymbol{U}}_t = \frac{\sum_{l=1}^{t}\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}^{(l)})}{t}$$

is the within imputation variance that accounts for the uncertainty which arose in analysing each complete-data set, and

$$\boldsymbol{B}_t = \sum_{l=1}^{t}\frac{(\hat{\boldsymbol{\beta}}^{(l)}-\bar{\boldsymbol{\beta}})^T(\hat{\boldsymbol{\beta}}^{(l)}-\bar{\boldsymbol{\beta}})}{t-1}$$

is the between imputation variance that measures the variability of the $t$ complete-data least squares estimates, which is due to the presence of the $t$ imputed sets in the analysis.

In this chapter, we propose an optimal design framework for the linear regression model that employs the above described multiple imputation. We are interested in the optimal designs which minimise some aspects of $\mathcal{T}_t$ over a design region. For example, an $A$-optimal design that minimises the trace of $\mathcal{T}_t$ can be employed for the experiment that aims to minimise the total variance of the individual $\bar{\boldsymbol{\beta}}$. Since there is no data at the design stage of the experiment, the main challenge of this investigation is to find an expectation of $\mathcal{T}_t$ with respect to the presence of missing responses and the multiple imputation respectively. Here we first address the present of potential correlations between the observed responses and the imputed values when we find the expectation of $\mathcal{T}_t$. We give some assumptions for $\bar{\boldsymbol{U}}_t$ which complement our investigation on the design framework. Following that, we scrutinise the expectation of $\boldsymbol{B}_t$ in two aspects. By focusing on the common elements that are present in $\boldsymbol{B}_t$, we find the expectation of these elements with respect to multiple imputation given the observed information, and an iterative expectation to account for the uncertainty of not seeing the data at the design stage of the experiment. In the later part of this chapter, we illustrate the application of the design framework for a simple linear regression model, and verify the performance of the designs through simulation. Moreover, we compare these optimal designs with the optimal designs that assume complete case analysis. We then end this chapter by concluding the finding of this investigation.

## 6.1 Within imputation variance-covariance

Considering the role of complete case analysis in the design framework for the linear regression model, the investigation in Chapter 4 focused on the approximation of

$$\boldsymbol{E\{cov\ (\hat{\beta})|\mathcal{M}\}}$$

where some rows of design matrix that correspond to missing responses are discarded in the analysis. More specifically, the design framework finds an optimal design based on a function of

$$E(V_{\hat{\boldsymbol{\beta}}_{obs}}) = E\{(X_{obs}^T X_{obs})^{-1}\}$$

of the posterior distribution, where the expectation is taken with respect to the presence of missing responses, and $X_{obs}$ is unobserved at the design stage of the experiment. On the contrary, there is no information being discarded in the multiple imputation. In particular having observed the experiment, we have

$$\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}^{(*)}) = \boldsymbol{cov}\left((X^TX)^{-1}X^T\boldsymbol{y_{(*)}}\right) = (X^TX)^{-1}\sigma^2,$$

where the design matrix $X$ is a full rank matrix consisting all rows of $f^T(x_i)$, $i = 1, ..., N$, and is independent of the response variable. Using the observed information in the construction of $P(\boldsymbol{\beta^{(*)}}|\boldsymbol{x_{obs}}, \boldsymbol{y_{obs}})$, different imputed sets, i.e. $\{y_{i(*)}\}$ and $\{y_{i(**)}\}$, are independent because the realisations of $P(\boldsymbol{\beta^{(*)}}|\boldsymbol{x_{obs}}, \boldsymbol{y_{obs}})$ for imputing the missing responses repeatedly are independent conditional on $\boldsymbol{y_{obs}}$. Hence given the data of the experiment, the least squares estimates $\hat{\boldsymbol{\beta}}^{(l)}$, $l = 1, ..., t$, are independent in the multiple imputation and $(X^TX)^{-1}\sigma^2$ provides a good approximation to the variance of $\hat{\boldsymbol{\beta}}^{(l)}$.

However at the design stage of an experiment, these $\hat{\boldsymbol{\beta}}^{(l)}$ are not independent when data are not available. Recall that

$$\hat{\boldsymbol{\beta}}^{(*)} = (X^TX)^{-1}X^T\boldsymbol{y}_{(*)}$$

where the complete-data $\boldsymbol{y}_{(*)}$ consists of the observed values $\boldsymbol{y_{obs}}$ and the imputed values $\{y_{i(*)}\}$. There exists correlation between $\boldsymbol{y_{obs}}$ (which are unseen at the design stage) and $\{y_{i(*)}\}$ as the latter quantities depend on the realisations of $P(\boldsymbol{\beta^{(*)}}|\boldsymbol{x_{obs}}, \boldsymbol{y_{obs}})$, which involves the former values in its construction. Consequently $(X^TX)^{-1}\sigma^2$ may not provide a good approximation to $\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}^{(*)})$ when we do not know which observation are missing and that require imputations. Nevertheless, to inspect the role of multiple imputation in the optimal design framework, we approximate

$$E\left(\bar{\boldsymbol{U}}_t\right) = \frac{E\left(\sum_{l=1}^t \boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}^{(l)})\right)}{t} = \frac{\sum_{l=1}^t E\left(\boldsymbol{cov}\ (\hat{\boldsymbol{\beta}}^{(l)})\right)}{t} = \frac{t(X^TX)^{-1}\sigma^2}{t} = (X^TX)^{-1}\sigma^2,$$

for $\sigma^2$ known. We verify this assumption through simulation in the later part of this chapter by comparing the empirical value of

$$E(\bar{\boldsymbol{U}}_t) = E(\mathcal{T}_t) - \left(1 + \frac{1}{t}\right) E(\boldsymbol{B}_t)$$

with the theoretical value of $(X^T X)^{-1}\sigma^2$. We obtain the empirical values of $E(\mathcal{T}_t)$ and $E(\boldsymbol{B}_t)$ respectively, by averaging the variances across the simulated data in the simulation studies of different designs. A small relative difference, i.e.

$$RD = 1 - \frac{(X^T X)^{-1}\sigma^2}{E(\mathcal{T}_t) - \left(1 + \frac{1}{t}\right) E(\boldsymbol{B}_t)}$$

would reflect the validity of this assumption.

## 6.2 Between imputation variance-covariance

We now assess the elements of the between imputation variance-covariance $\boldsymbol{B}_t$ in terms of the explanatory variable and the response variables of the linear regression model. Recall that Rubin's combining rules give

$$\boldsymbol{B}_t = \sum_{l=1}^{t} \frac{(\hat{\boldsymbol{\beta}}^{(l)} - \bar{\boldsymbol{\beta}})^T (\hat{\boldsymbol{\beta}}^{(l)} - \bar{\boldsymbol{\beta}})}{t-1},$$

where $\bar{\boldsymbol{\beta}} = \sum_{l=1}^{t} \hat{\boldsymbol{\beta}}^{(l)}/t$ and $\hat{\boldsymbol{\beta}}^{(l)}$ is the $l^{th}$ complete-data least squares estimates in the multiple imputation. For example, a simple linear regression model has

$$\begin{aligned}
\boldsymbol{B}_t &= \frac{1}{t-1} \sum_{l=1}^{t} \begin{pmatrix} \hat{\beta}_0^{(l)} - \bar{\beta}_0 \\ \hat{\beta}_1^{(l)} - \bar{\beta}_1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0^{(l)} - \bar{\beta}_0 & \hat{\beta}_1^{(l)} - \bar{\beta}_1 \end{pmatrix} \\
&= \frac{1}{t-1} \sum_{l=1}^{t} \begin{pmatrix} \left(\hat{\beta}_0^{(l)} - \bar{\beta}_0\right)^2 & \left(\hat{\beta}_0^{(l)} - \bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)} - \bar{\beta}_1\right) \\ \left(\hat{\beta}_0^{(l)} - \bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)} - \bar{\beta}_1\right) & \left(\hat{\beta}_1^{(l)} - \bar{\beta}_1\right)^2 \end{pmatrix},
\end{aligned}$$

where the diagonal elements of $\boldsymbol{B}_t$ correspond to $var(\bar{\beta}_0)$ and $var(\bar{\beta}_1)$ respectively, and the off-diagonal elements correspond to $cov(\bar{\beta}_0, \bar{\beta}_1)$. To express $\boldsymbol{B}_t$ in terms of the variables of the model, we expand the cross product of the elements in the summation of $\boldsymbol{B}_t$ to several summations of individual product of the estimates. For example, we

have

$$\begin{aligned}
&\sum_{l=1}^{t}\left(\hat{\beta}_0^{(l)}-\bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)}-\bar{\beta}_1\right)=\sum_{l=1}^{t}\left\{\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}-\hat{\beta}_0^{(l)}\bar{\beta}_1-\hat{\beta}_1^{(l)}\bar{\beta}_0+\bar{\beta}_1\bar{\beta}_0\right\}\\
&=\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}-\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\bar{\beta}_1-\sum_{l=1}^{t}\hat{\beta}_1^{(l)}\bar{\beta}_0+t\ \bar{\beta}_1\bar{\beta}_0\\
&=\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}-\frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\sum_{l=1}^{t}\hat{\beta}_1^{(l)}-\frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_1^{(l)}\sum_{l=1}^{t}\hat{\beta}_0^{(l)}+t\ \frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_1^{(l)}\frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\\
&=\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}-\frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\sum_{l=1}^{t}\hat{\beta}_1^{(l)}-\frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_1^{(l)}\sum_{l=1}^{t}\hat{\beta}_0^{(l)}+\frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_1^{(l)}\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\\
&=\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}-\frac{1}{t}\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\sum_{l=1}^{t}\hat{\beta}_1^{(l)}\\
&=\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}-\frac{1}{t}\left(\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}+\sum_{\substack{l=1,\ldots,t-1\\ q>l}}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(q)}+\sum_{\substack{l=1,\ldots,t-1\\ q>l}}\hat{\beta}_1^{(l)}\hat{\beta}_0^{(q)}\right)\\
&=\left(1-\frac{1}{t}\right)\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}-\frac{1}{t}\left(\sum_{\substack{l=1,\ldots,t-1\\ q>l}}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(q)}+\sum_{\substack{l=1,\ldots,t-1\\ q>l}}\hat{\beta}_1^{(l)}\hat{\beta}_0^{(q)}\right)
\end{aligned}$$

for the off diagonal elements of $(t-1)\boldsymbol{B}_t$ of the simple linear model. In the third line, we express the mean of the estimates in terms of its summation to complement the algebraic simplification. Dividing both sides by $t-1$, $cov(\bar{\beta}_0,\bar{\beta}_1)$ of the simple linear regression model is

$$\begin{aligned}
\frac{\sum_{l=1}^{t}\left(\hat{\beta}_0^{(l)}-\bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)}-\bar{\beta}_1\right)}{t-1}&=\left(\frac{1}{t}\right)\sum_{l=1}^{t}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}\\
&\quad-\frac{1}{t(t-1)}\underbrace{\left(\sum_{\substack{l=1,\ldots,t-1\\ q>l}}\hat{\beta}_0^{(l)}\hat{\beta}_1^{(q)}+\sum_{\substack{l=1,\ldots,t-1\\ q>l}}\hat{\beta}_1^{(l)}\hat{\beta}_0^{(q)}\right)}_{t(t-1)\ terms}.
\end{aligned}$$

A similar expression for the variance of each estimate can be obtained by changing the indices of the parameters in the above expression, yielding

$$\frac{\sum_{l=1}^{t}\left(\hat{\beta}^{(l)}-\bar{\beta}\right)^2}{t-1}=\left(\frac{1}{t}\right)\sum_{l=1}^{t}\left(\hat{\beta}^{(l)}\right)^2-\frac{2}{t(t-1)}\underbrace{\sum_{\substack{l=1,\ldots,t-1\\ q>l}}\hat{\beta}^{(l)}\hat{\beta}^{(q)}}_{t(t-1)/2\ terms}$$

with an appropriate index for the parameters (see Appendix B.1 for the analytical work).

We now consider expressing the repeated complete-data least squares estimates in terms of the variables of the linear regression model. Notice that in the summation of $\boldsymbol{B}_t$, i.e. the left hand side of the above expressions, for each

$$(\hat{\beta}^{(l)} - \bar{\beta}) = (\hat{\beta}^{(l)} - \frac{\sum_{l=1}^{t} \hat{\beta}^{(l)}}{t})$$

with an appropriate index for the parameters, we can simplify usefully if we partition the data into two parts that contain the observed data and the missing information respectively using a Bernoulli missing data indicator,

$$\mathcal{M}_i = \begin{cases} 1, & \text{for } y_i \text{ is missing with probability } P(x_i); \\ 0, & \text{otherwise with probability } 1 - P(x_i), \end{cases}$$

where $P(x_i)$ is a monotone MAR mechanism. For instance, consider the intercept parameter of the simple linear regression model, i.e.

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

in the standard form when all observations are observed. Incorporating the missing data indicators and the index of the repeated imputations accordingly, we have

$$\begin{aligned}
&\sum_{l=1}^{t} \hat{\beta}_0^{(l)} \\
&= \sum_{l=1}^{t} \left( \frac{\sum_{i=1}^{N} {x_i}^2 \left( \sum_{i=1}^{N} y_i(1 - \mathcal{M}_i) + \sum_{i=1}^{N} y_{i(l)} \mathcal{M}_i \right)}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \right) \\
&\quad - \sum_{l=1}^{t} \left( \frac{\sum_{i=1}^{N} x_i \left( \sum_{i=1}^{N} x_i y_i(1 - \mathcal{M}_i\ ) + \sum_{i=1}^{N} x_i y_{i(l)} \mathcal{M}_i\ \right)}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \right) \\
&= \sum_{l=1}^{t} \left( \frac{\sum_{i=1}^{N} {x_i}^2 \left( \sum_{i=1}^{N} y_{i\ obs} + \sum_{i=1}^{N} y_{i(l)\ imp} \right)}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \right) \\
&\quad - \sum_{l=1}^{t} \left( \frac{\sum_{i=1}^{N} x_i \left( \sum_{i=1}^{N} x_i y_{i\ obs} + \sum_{i=1}^{N} x_i y_{i(l)\ imp} \right)}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \right) \\
&= t \left( \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_{i\ obs} - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_{i\ obs}}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \right) \\
&\quad + \sum_{l=1}^{t} \left( \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_{i(l)\ imp} - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_{i(l)\ imp}}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \right),
\end{aligned}$$

and

$$
\begin{aligned}
&\hat{\beta}_0^{(l)} - \frac{\sum_{l=1}^{t} \hat{\beta}_0^{(l)}}{t} \\
&= \frac{\sum_{i=1}^{N} {x_i}^2 \left( \sum_{i=1}^{N} y_{i\ obs} + \sum_{i=1}^{N} y_{i(l)\ imp} \right) - \sum_{i=1}^{N} x_i \left( \sum_{i=1}^{N} x_i y_{i\ obs} + \sum_{i=1}^{N} x_i y_{i(l)\ imp} \right)}{N \sum_{i=1}^{N} {x_i}^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \\
&\quad - \frac{t}{t} \left( \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_{i\ obs} - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_{i\ obs}}{N \sum_{i=1}^{N} {x_i}^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \right) \\
&\quad - \sum_{l=1}^{t} \frac{1}{t} \left( \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_{i(l)\ imp} - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_{i(l)\ imp}}{N \sum_{i=1}^{N} {x_i}^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \right) \\
&= \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_{i(l)\ imp} - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_{i(l)\ imp}}{N \sum_{i=1}^{N} {x_i}^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \\
&\quad - \sum_{l=1}^{t} \frac{1}{t} \left( \frac{\sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_{i(l)\ imp} - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_{i(l)\ imp}}{N \sum_{i=1}^{N} {x_i}^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \right),
\end{aligned}
$$

where $y_{i\ obs} = y_i(1 - \mathcal{M}_i)$ and $y_{i(l)\ imp} = y_{i(l)}\mathcal{M}_i$. From this example, we see that $\boldsymbol{B}_t$ is independent of the observed responses in general.

In what follows, we drop the products that contain the observed responses accordingly when we express the elements of $\boldsymbol{B}_t$ in terms of the variables of the model. For example, for the parameter estimates of the simple linear regression model, the part which involves the imputed values are

$$
\begin{aligned}
\begin{pmatrix} \hat{\beta}_0^{(l)} \\ \hat{\beta}_1^{(l)} \end{pmatrix} &= \frac{1}{N \sum_{i=1}^{N} {x_i}^2 - \left( \sum_{i=1}^{N} x_i \right)^2} \begin{pmatrix} \sum_{i=1}^{N} {x_i}^2 \sum_{i=1}^{N} y_{i(l)} \mathcal{M}_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_{i(l)} \mathcal{M}_i \\ N \sum_{i=1}^{N} x_i y_{i(l)} \mathcal{M}_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_{i(l)} \mathcal{M}_i \end{pmatrix} \\
&= \frac{1}{const} \begin{pmatrix} \mathbb{C}_{(*)} \sum {x_i}^2 - \mathbb{D}_{(*)} \sum x_i \\ \mathbb{D}_{(*)} N - \mathbb{C}_{(*)} \sum x_i \end{pmatrix},
\end{aligned}
$$

where $\mathcal{M}_i = 1$ indicates response $i$ is missing and being replaced by $y_{i(*)}$, $\mathbb{C}_l = \sum_{i=1}^{N} y_{i(l)} \mathcal{M}_i$, $\mathbb{D}_l = \sum_{i=1}^{N} x_i y_{i(l)} \mathcal{M}_i$, and $const = N \sum_{i=1}^{N} {x_i}^2 - \left( \sum_{i=1}^{N} x_i \right)^2$, $i = 1, ..., N$, is the index of the $N$ experimental units in the study. We can then find the products of the corresponding estimates, i.e.

$$
\begin{aligned}
\hat{\beta}_0^{(*)} \hat{\beta}_0^{(**)} = \frac{1}{const^{\,2}} \Big( & \mathbb{C}_{(*)} \mathbb{C}_{(**)} \left( \sum {x_i}^2 \right)^2 + \mathbb{D}_{(*)} \mathbb{D}_{(**)} \left( \sum x_i \right)^2 \\
& - \left( \mathbb{C}_{(*)} \mathbb{D}_{(**)} + \mathbb{C}_{(**)} \mathbb{D}_{(*)} \right) \sum {x_i}^2 \sum x_i \Big),
\end{aligned}
$$

$$\hat{\beta}_1^{(*)}\hat{\beta}_1^{(**)} = \frac{1}{const^{\,2}}\left(N^2\mathbb{D}_{(*)}\mathbb{D}_{(**)} + \mathbb{C}_{(*)}\mathbb{C}_{(**)}\left(\sum x_i\right)^2 - (\mathbb{D}_{(**)}\mathbb{C}_{(*)} + \mathbb{D}_{(*)}\mathbb{C}_{(**)})N\sum x_i\right),$$

and

$$\hat{\beta}_0^{(*)}\hat{\beta}_1^{(**)} = \frac{1}{const^{\,2}}\left(\mathbb{C}_{(*)}\mathbb{D}_{(**)}N\sum {x_i}^2 + \mathbb{C}_{(**)}\mathbb{D}_{(*)}\left(\sum x_i\right)^2 - \mathbb{C}_{(*)}\mathbb{C}_{(**)}\sum {x_i}^2\sum x_i - \mathbb{D}_{(*)}\mathbb{D}_{(**)}N\sum x_i\right),$$

where the summation sums from $i = 1, ..., N$, for the elements of $\boldsymbol{B}_t$.

In the next section, we assess the expectations of the elements of $\boldsymbol{B}_t$ before finding an optimal design that minimises a function of the expected value of $\mathcal{T}_t$. We study the expectations of the common elements of $\boldsymbol{B}_t$ in two contexts. Considering the imputation model, i.e. $\boldsymbol{N}(\hat{\boldsymbol{\beta}}_{obs}, V_{\hat{\boldsymbol{\beta}}_{obs}}\sigma^2)$ given the data, we find some results for the imputed values conditional on the observed information. These results are then averaged across the variability of not seeing the data at the design stage of an experiment. The aim of doing these is to find an analytical expression for $E(\boldsymbol{B}_t)$ in terms of the explanatory variable of the model. To achieve this, we use the law of total expectations and some iterative expectation in the investigation.

## 6.3 Expectation of between imputation variance-covariance

In general, it can be proof that

$$E(\boldsymbol{B}_t) = E\left\{(\hat{\boldsymbol{\beta}}^{(*)})^T\hat{\boldsymbol{\beta}}^{(*)}\right\} - E\left\{(\hat{\boldsymbol{\beta}}^{(*)})^T\hat{\boldsymbol{\beta}}^{(**)})\right\}$$

where superscript $(*)$ and $(**)$ correspond to different imputations. For example, following the previous derivation, the expected $cov(\bar{\beta}_0, \bar{\beta}_1)$ is

$$\begin{aligned}
&E\left(\frac{\sum_{l=1}^{t}\left(\hat{\beta}_0^{(l)} - \bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)} - \bar{\beta}_1\right)}{t-1}\right)\\
&= \frac{1}{t}\sum_{l=1}^{t}E\left(\hat{\beta}_0^{(l)}\hat{\beta}_1^{(l)}\right) - \frac{1}{t(t-1)}\left(\sum_{\substack{l=1,\ldots,t-1\\ q>l}}E\left(\hat{\beta}_0^{(l)}\hat{\beta}_1^{(q)}\right) + \sum_{\substack{l=1,\ldots,t-1\\ q>l}}E\left(\hat{\beta}_1^{(l)}\hat{\beta}_0^{(q)}\right)\right)\\
&= \frac{1}{t}t\ E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(*)}\right) - \frac{1}{t(t-1)}\left(\frac{t\ (t-1)}{2}E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(**)}\right) + \frac{t\ (t-1)}{2}E\left(\hat{\beta}_1^{(*)}\hat{\beta}_0^{(**)}\ \right)\right)\\
&= E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(*)}\right) - \left(\frac{1}{2}E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(**)}\right) + \frac{1}{2}E\left(\hat{\beta}_1^{(*)}\hat{\beta}_0^{(**)}\right)\right)\\
&= E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(*)}\right) - \ E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(**)}\right),
\end{aligned}$$

where the summation in the second line becomes a multiplier as the individual expected value obtained in the iterative expectation is the same for all repeated imputations. A similar result for the diagonal elements of $E(\boldsymbol{B}_t)$ is shown in Appendix B.2.

Nevertheless, at the design stage of an experiment, there exist two sources of randomness: (1) the uncertainty of not knowing which observations are missing at random, (2) the variability of imputing missing responses repeatedly. Therefore, we need to consider an iterative expectation, i.e. $E(\cdot) = E(E(\cdot|\boldsymbol{\mathcal{M}}))$, to tackle these uncertainties where the inner expectation is taken with respect to multiply imputing the missing responses conditional on the observations, and the outer expectation is taken with respect to the random missing data indicators. To illustrate this, consider $E(\boldsymbol{B}_t)$ for the estimates of the simple linear regression model:

$$
\begin{aligned}
&E\left(\frac{\sum_{l=1}^{t}\left(\hat{\beta}_0^{(l)}-\bar{\beta}_0\right)^2}{t-1}\right) = E\left(\hat{\beta}_0^{(*)}\hat{\beta}_0^{(*)}\right) - E\left(\hat{\beta}_0^{(*)}\hat{\beta}_0^{(**)}\right)\\
&= \frac{1}{const^{\,2}}\Big(E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}\left(\sum {x_i}^2\right)^2\\
&\quad + E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}\left(\sum x_i\right)^2\\
&\quad -2E\left\{E\left(\mathbb{C}_{(*)}\mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{C}_{(*)}\mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}\sum {x_i}^2\sum x_i\Big)
\end{aligned}
$$

for the between imputation variance of $\hat{\beta}_0$,

$$
\begin{aligned}
&E\left(\frac{\sum_{l=1}^{t}\left(\hat{\beta}_1^{(l)}-\bar{\beta}_1\right)^2}{t-1}\right) = E\left(\hat{\beta}_1^{(*)}\hat{\beta}_1^{(*)}\right) - E\left(\hat{\beta}_1^{(*)}\hat{\beta}_1^{(**)}\right)\\
&= \frac{1}{const^{\,2}}\Big(N^2E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}\\
&\quad + E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}\left(\sum x_i\right)^2\\
&\quad -2E\left\{E\left(\mathbb{D}_{(*)}\mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{D}_{(**)}\mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\}N\sum x_i\Big)
\end{aligned}
$$

for the between imputation variance of $\hat{\beta}_1$, and

$$
\begin{aligned}
&E\left(\frac{\sum_{l=1}^{t}\left(\hat{\beta}_0^{(l)}-\bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)}-\bar{\beta}_1\right)}{t-1}\right) = E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(*)}\right) - E\left(\hat{\beta}_0^{(*)}\hat{\beta}_1^{(**)}\right)\\
&= \frac{1}{const^{\,2}}\Big(E\left\{E\left(\mathbb{C}_{(*)}\mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{C}_{(*)}\mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}\left(N\sum {x_i}^2 + \left(\sum x_i\right)^2\right)\\
&\quad - E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}\sum {x_i}^2\sum x_i\\
&\quad -E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)\right\} - E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right)\right\}N\sum x_i\Big)
\end{aligned}
$$

for the between imputation covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$. Note that different subscripts of $\mathbb{C}$

and of $\mathbb{D}$ reflect the different repeated imputations. Due to this convention, given the information of the experiment, we have $E\left(\mathbb{C}_{(*)}\mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right) = E\left(\mathbb{C}_{(**)}\mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right)$ in the inner expectation of $cov(\hat{\beta}_0, \hat{\beta}_1)$.

On the contrary if having observed the experimental outcome, i.e. the missing data indicators $\mathcal{M}_i$ are observed, we can find an expression of $E(\boldsymbol{B}_t)$ without tackling the outer expectation. In this case, we know which responses are missing and replaced by the imputed values. In particular, for the common elements

$$\begin{aligned}
\mathbb{C}_{(*)}\ \mathbb{C}_{(**)} &= \left(\sum_{i=1}^{N} y_{i(*)}\mathcal{M}_i\right)\left(\sum_{i=1}^{N} y_{i(**)}\mathcal{M}_i\right) \\
&= \left(y_{1(*)}\mathcal{M}_1 + ... + y_{N(*)}\mathcal{M}_N\right)\left(y_{1(**)}\mathcal{M}_1 + ... + y_{N(**)}\mathcal{M}_N\right) \\
&= \sum_{i=1}^{N} y_{i(*)}\mathcal{M}_i\ y_{i(**)}\mathcal{M}_i + \sum_{\substack{i=1\\j>i}}^{N} y_{i(*)}\mathcal{M}_i\ y_{j(**)}\mathcal{M}_j + \sum_{\substack{i=1\\j>i}}^{N} y_{j(*)}\mathcal{M}_j\ y_{i(**)}\mathcal{M}_i \\
&= \sum_{i=1}^{N} y_{i(*)}y_{i(**)}\mathcal{M}_i + 2\sum_{\substack{i=1\\j>i}}^{N} y_{i(*)}y_{j(**)}\mathcal{M}_i\mathcal{M}_j,
\end{aligned}$$

$$\begin{aligned}
\mathbb{D}_{(*)}\ \mathbb{D}_{(**)} &= \left(\sum_{i=1}^{N} x_i y_{i(*)}\mathcal{M}_i\right)\left(\sum_{i=1}^{N} x_i y_{i(**)}\mathcal{M}_i\right) \\
&= \left(x_1 y_{1(*)}\mathcal{M}_1 + ... + x_N y_{N(*)}\mathcal{M}_N\right)\left(x_1 y_{1(**)}\mathcal{M}_1 + ... + x_N y_{N(**)}\mathcal{M}_N\right) \\
&= \sum_{i=1}^{N} x_i^2\ y_{i(*)}\mathcal{M}_i\ y_{i(**)}\mathcal{M}_i + \sum_{\substack{i=1\\j>i}}^{N} x_i y_{i(*)}\mathcal{M}_i\ x_j y_{j(**)}\mathcal{M}_j + \sum_{\substack{i=1\\j>i}}^{N} x_j y_{j(*)}\mathcal{M}_j\ x_i y_{i(**)}\mathcal{M}_i \\
&= \sum_{i=1}^{N} x_i^2\ y_{i(*)}y_{i(**)}\mathcal{M}_i + 2\sum_{\substack{i=1\\j>i}}^{N} x_i y_{i(*)}x_j y_{j(**)}\mathcal{M}_i\mathcal{M}_j,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{C}_{(*)}\ \mathbb{D}_{(**)} &= \left(\sum_{i=1}^{N} y_{i(*)}\mathcal{M}_i\right)\left(\sum_{i=1}^{N} x_i y_{i(**)}\mathcal{M}_i\right) \\
&= \left(y_{1(*)}\mathcal{M}_1 + ... + y_{N(*)}\mathcal{M}_N\right)\left(x_1 y_{1(**)}\mathcal{M}_1 + ... + x_N y_{N(**)}\mathcal{M}_N\right) \\
&= \sum_{i=1}^{N} x_i y_{i(*)}\mathcal{M}_i\ y_{i(**)}\mathcal{M}_i + \sum_{\substack{i=1\\j>i}}^{N} x_i y_{i(**)}\mathcal{M}_i\ y_{j(*)}\mathcal{M}_j + \sum_{\substack{i=1\\j>i}}^{N} x_j y_{j(**)}\mathcal{M}_j\ y_{i(*)}\mathcal{M}_i \\
&= \sum_{i=1}^{N} x_i y_{i(*)}y_{i(**)}\mathcal{M}_i + 2\sum_{\substack{i=1\\j>i}}^{N} x_i y_{i(**)}y_{j(*)}\mathcal{M}_i\mathcal{M}_j,
\end{aligned}$$

that are in the elements of $E(\boldsymbol{B}_t)$, we want to find

$$\begin{aligned}
&E\left(\mathbb{C}_{(*)}\ \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(\mathbb{C}_{(*)}\ \mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}}\right) \\
&= \sum_{i=1}^{N} \mathcal{M}_i \left(E\left(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\
&\quad +2\sum_{\substack{i=1 \\ j>i}}^{N} \mathcal{M}_i\mathcal{M}_j \left(E\left(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right),
\end{aligned}$$

$$\begin{aligned}
&E\left(\mathbb{D}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(\mathbb{D}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right) \\
&= \sum_{i=1}^{N} x_i^2 \mathcal{M}_i \left(E\left(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\
&\quad +2\sum_{\substack{i=1 \\ j>i}}^{N} x_i\ x_j \mathcal{M}_i\mathcal{M}_j \left(E\left(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
&E\left(\mathbb{C}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(\mathbb{C}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right) \\
&= \sum_{i=1}^{N} x_i \mathcal{M}_i \left(E\left(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\
&\quad + \sum_{\substack{i=1 \\ j>i}}^{N} x_i \mathcal{M}_i\mathcal{M}_j \left(E\left(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(**)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\
&\quad + \sum_{\substack{i=1 \\ j>i}}^{N} x_j \mathcal{M}_i\mathcal{M}_j \left(E\left(y_{j(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{j(**)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\
&= \sum_{i=1}^{N} x_i \mathcal{M}_i \left(E\left(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\
&\quad + \left(\sum_{\substack{i=1 \\ j>i}}^{N} x_i \mathcal{M}_i\mathcal{M}_j + \sum_{\substack{i=1 \\ j>i}}^{N} x_j \mathcal{M}_i\mathcal{M}_j\right) \left(E\left(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right),
\end{aligned}$$

to express $E(\boldsymbol{B}_t)$ in terms of the variables of the linear model given the values of $\mathcal{M}_i$. The expected product of the imputed values can be found by considering some results of the posterior distribution of $\boldsymbol{\beta}^{(*)}$, i.e. $\boldsymbol{N}(\hat{\boldsymbol{\beta}}_{obs}, V_{\hat{\boldsymbol{\beta}}_{obs}}\sigma^2)$ given the data. Note that since

$$E\left(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right) \tag{6.1}$$

and

$$E\left(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(**)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) \tag{6.2}$$

are the common elements in the above expressions, we only illustrate the results for them in the remaining part of this chapter. More specifically, using the law of total expectation, i.e. for two random variables $X$ and $Y$,

$$E(XY) = Cov(X,Y) + E(X)E(Y),$$

we study the relationship between

(i) $y_{i(*)}$ and $y_{j(*)}$, where the imputed responses are in the same imputed set;

(ii) $y_{i(*)}$ and $y_{j(**)}$, where the imputed responses are in two different imputed sets, i.e. the imputed response for experimental unit $i$ is in set $(*)$, whereas the imputed response for experimental unit $j$ is in set $(**)$.

### 6.3.1 Theoretical results given the observed data

Recall that using the non-informative prior distribution and the likelihood function of the parameters conditional on the observed data, we find that the data governing parameters $\boldsymbol{\beta^{(*)}}$ for imputing responses have a posterior distribution $\boldsymbol{N}(\hat{\boldsymbol{\beta}}_{obs}, V_{\hat{\boldsymbol{\beta}}_{obs}}\sigma^2)$ where $V_{\hat{\boldsymbol{\beta}}_{obs}} = (X^TX)^{-1}_{obs}$ is the inverse information matrix that is computed in the complete case analysis. Considering the property of this distribution, we can find the expected imputed value and its variance for the missing response of experimental unit $i$, given the information of the experiment. For example consider the simple linear regression model with observed data. We have

$$\begin{aligned} E\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) &= E\left(\beta_0^{(*)} + \beta_1^{(*)}\, x_i + \epsilon_i \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\ &= E\left(\beta_0^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) + x_i \cdot E\left(\beta_1^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right), \end{aligned}$$

for the mean of imputed response $y_{i(*)}$, and

$$\begin{aligned} var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) &= var\left(\beta_0^{(*)} + \beta_1^{(*)}\, x_i + \epsilon_i \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\ &= var\left(\beta_0^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) + x_i^2 \cdot var\left(\beta_1^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\ &\quad + 2 \cdot x_i \cdot Cov\left(\beta_0^{(*)},\ \beta_1^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) + \sigma^2 \\ &= \left(\begin{pmatrix} 1 & x_i \end{pmatrix} (X^TX)^{-1}_{obs} \begin{pmatrix} 1 \\ x_i \end{pmatrix} + 1\right)\sigma^2, \end{aligned}$$

for the variance of $y_{i(*)}$. The covariance between an imputed value $y_{i(*)}$ of subject $i$ and an imputed value $y_{j(*)}$ of subject $j$ is

$$\begin{aligned}
&Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\
&= Cov\left(\beta_0^{(*)} + \beta_1^{(*)} x_i + \epsilon_i,\ \beta_0^{(*)} + \beta_1^{(*)} x_j + \epsilon_j \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\
&= var\left(\beta_0^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) + x_i \cdot x_j \cdot var\left(\beta_1^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\
&\quad + (x_i + x_j) \cdot Cov\left(\beta_0^{(*)},\ \beta_1^{(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\
&= \begin{pmatrix} 1 & x_i \end{pmatrix} (X^T X)_{obs}^{-1} \begin{pmatrix} 1 \\ x_j \end{pmatrix} \sigma^2,
\end{aligned}$$

where $(X^T X)_{obs}$ is the information matrix containing only $f^T(x_i)$ of subjects who have observed responses after the implementation of the experiment. For the above mentioned relationship (ii) between the imputed values that are in different imputed sets, we have

$$Cov\left(y_{i(*)},\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) = 0$$

and

$$Cov\left(y_{i(*)},\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) = 0$$

respectively because given the observations of the experiment, the random draws of $\boldsymbol{\beta}_{obs}$ in the repeated imputations are independent.

Hence, given all the information after observing the experiment, we can use these results and the law of total expectation to obtain

$$\begin{aligned}
&E\left(y_{i(*)}\ y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\
&= var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) + E\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) E\left(y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)
\end{aligned}$$

and

$$\begin{aligned}
&E\left(y_{i(*)}\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) \\
&= Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) + E\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) E\left(y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)
\end{aligned}$$

for the imputed values of set $(*)$, and

$$E\left(y_{i(*)}\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) = E\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) E\left(y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)$$

and

$$E\left(y_{i(*)}\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) = E\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) E\left(y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)$$

for two imputed values that are in different imputed sets. Since there is no other randomness and the expected value of an imputed response is the same regardless of the

repeated imputation procedure, we have

$$E\left(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right) = var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)$$
$$= \left(\begin{pmatrix}1 & x_i\end{pmatrix}(X^TX)^{-1}_{obs}\begin{pmatrix}1\\ x_i\end{pmatrix} + 1\right)\sigma^2$$

for (6.1) and

$$E\left(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(y_{i(**)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right) = Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)$$
$$= \begin{pmatrix}1 & x_i\end{pmatrix}(X^TX)^{-1}_{obs}\begin{pmatrix}1\\ x_j\end{pmatrix}\sigma^2$$

for (6.2), to be employed in the corresponding elements of

$$E\left(\mathbb{C}_{(*)}\ \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(\mathbb{C}_{(*)}\ \mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}}\right),$$

$$E\left(\mathbb{D}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(\mathbb{D}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right),$$

and

$$E\left(\mathbb{C}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}}\right) - E\left(\mathbb{C}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}}\right).$$

These elements can then be substituted into the corresponding expressions of $E(\boldsymbol{B}_t)$ having observed the experiment.

### 6.3.2 Theoretical results prior to observing data

On the other hand, to account for the uncertainty of not seeing the data at the design stage of an experiment, we apply some laws of expectation to the above presented results. For random variable $W_1$, $W_2$, $W_3$, the law of total expectation gives

$$E(W_1) = E(E(W_1|W_3));$$

the law of total variance states

$$var(W_1) = var(E(W_1 \mid W_3)) + E(var(W_1 \mid W_3));$$

and the law of total covariances yields

$$cov(W_1, W_2) = cov(E(W_1W_2 \mid W_3)) + E(cov(W_1, W_2 \mid W_3)).$$

The idea of these laws is to average the variation of the unseen information ($W_3$ in this example), across the possible outcomes, i.e. the outer operation, after assessing the

variability of the parameters of interest given the data, i.e. the inner expectation/var/-cov. Here using the above presented results given the information, we can find the outer expectation for (6.1) and (6.2) respectively, where the expectation is taken with respect to the random missing data indicators, i.e.

$$E\left(E(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}})\right)$$

and

$$E\left(E(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(**)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right).$$

To illustrate this, we first apply the law of total expectation to the individual elements of the above expressions, where

$$\begin{aligned} &E\left(E(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) \\ &= E\left(var\left(y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left(E(y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) E\left(E(y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) \end{aligned}$$

and

$$\begin{aligned} &E\left(E(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) \\ &= E\left(Cov\left(y_{i(*)},\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left(E(y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) E\left(E(y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) \end{aligned}$$

correspond to the relationship between the imputed values that are in imputed set $(*)$, and

$$\begin{aligned} &E\left(E(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}})\right) \\ &= E\left(Cov\left(y_{i(*)},\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left(E(y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) E\left(E(y_{i(**)} \mid \boldsymbol{\mathcal{M}})\right) \end{aligned}$$

and

$$\begin{aligned} &E\left(E(y_{i(*)}\ y_{j(**)} \mid \boldsymbol{\mathcal{M}})\right) \\ &= E\left(Cov\left(y_{i(*)},\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left(E(y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) E\left(E(y_{j(**)} \mid \boldsymbol{\mathcal{M}})\right) \end{aligned}$$

correspond to the relationship between the imputed values that are in different imputed sets. Since the expected value of an imputed response given the data is the same for all $l = 1, ..., t$, and the law of total expectation grants the expected value of a random variable equals to the average of the conditional expected value, i.e. for example

$$E\left(E(y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) = E\left(E(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}})\right),$$

we only need to find an analytical expression for

$$\begin{aligned} & E\left(E(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}})\right) \\ & = E\left(var\left(y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) - E\left(Cov\left(y_{i(*)},\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \end{aligned}$$

and

$$\begin{aligned} & E\left(E(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(**)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) \\ & = E\left(Cov\left(y_{i(*)},\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) - E\left(Cov\left(y_{i(*)},\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \end{aligned}$$

respectively, to obtain the expressions of $E(\boldsymbol{B}_t)$ at the design stage of an experiment.

Using the law of variance and the law of covariance respectively, we have

$$\begin{aligned} & E\left(var\left(y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\ & = var\left(E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) + E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \end{aligned} \tag{6.3}$$

for the variance of an imputed response $y_{i(*)}$, and

$$\begin{aligned} & E\left(Cov\left(y_{i(*)},\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\ & = Cov\left(E\left[y_{i(*)}\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) + E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \end{aligned} \tag{6.4}$$

for the covariance between an imputed value $y_{i(*)}$ and another imputed response $y_{j(*)}$ that are in the same imputed set. On the other hand for the imputed responses which are in different imputed sets, even though there is no correlations between them given the observed data, we need to consider the expectation term in the law of total covariance to account for the uncertainty of not observing the data at the design stage. To be more specific, for the covariance between two imputed values of an experimental unit $i$ that are in different imputed sets, we have

$$\begin{aligned} & E\left(Cov\left(y_{i(*)},\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\ & = Cov\left(E\left[y_{i(*)}\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) + E\left[Cov\left(y_{i(*)},\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\ & = Cov\left(E\left[y_{i(*)}\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) \\ & = Cov\left(E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] E\left[\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) \\ & = Cov\left(E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] E\left[\ y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) \\ & = var\left(E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right); \end{aligned}$$

for the covariance between the imputed values of two experimental units $i$ and $j$ that are in different imputed sets, we have

$$\begin{aligned}
&E\left(Cov\left(y_{i(*)},\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\
&= Cov\left(E\left[y_{i(*)}\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) + E\left[Cov\left(y_{i(*)},\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= Cov\left(E\left[y_{i(*)}\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) \\
&= Cov\left(E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] E\left[\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) \\
&= Cov\left(E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] E\left[\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) \\
&= Cov\left(E\left[y_{i(*)}\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right).
\end{aligned}$$

In the second line of both expected covariance, we have

$$Cov\left(y_{i(*)},\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) = 0$$

and

$$Cov\left(y_{i(*)},\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right) = 0$$

respectively because given the data, the realisations of the posterior distribution in the repeated imputations are independent. Hence, the corresponding expectation term in the second line becomes zero. The same reason applies for rewriting

$$E\left[y_{i(*)}\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] = E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] E\left[\ y_{i(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]$$

and

$$E\left[y_{i(*)}\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] = E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right] E\left[\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]$$

that are in the fourth line of the respective expressions. Moreover, since the expectation of an imputed value given the data is the same for all repeated imputations, we have the results in the fifth line of the above expressions. By adding an appropriate expectation term to the first and the last line of the expected covariance, we have

$$\begin{aligned}
&E\left(Cov\left(y_{i(*)},\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= var\left(E\left[y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) + E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]
\end{aligned}$$

for the covariance between the imputed values of subject $i$ that are in different imputed sets, which is equivalent to $E\left(var\left(y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right)$ in (6.3); and

$$\begin{aligned}
&E\left(Cov\left(y_{i(*)},\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= Cov\left(E\left[y_{i(*)}\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right]\right) + E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]
\end{aligned}$$

for the covariance between the imputed values of subject $i$ and subject $j$ that are in

different imputed sets, which is equivalent to $E\left(Cov\left(y_{i(*)},\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right)\right)$ in (6.4). Rearranging

$$\begin{aligned} &E\left(Cov\left(y_{i(*)},\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\ &= E\left(var\left(y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) \end{aligned}$$

to

$$\begin{aligned} &E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\ &= E\left(var\left(y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) - E\left(Cov\left(y_{i(*)},\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right), \end{aligned} \tag{6.5}$$

and

$$\begin{aligned} &E\left(Cov\left(y_{i(*)},\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) + E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\ &= E\left(Cov\left(y_{i(*)},\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) \end{aligned}$$

to

$$\begin{aligned} &E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\ &= E\left(Cov\left(y_{i(*)},\ y_{j(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) - E\left(Cov\left(y_{i(*)},\ y_{j(**)} \mid \boldsymbol{\mathcal{M}}\right)\right), \end{aligned} \tag{6.6}$$

and applying these results to the outer expectation of (6.1) and (6.2) accordingly, we obtain

$$\begin{aligned} &E\left(E(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}})\right) \\ &= E\left(var\left(y_{i(*)} \mid \boldsymbol{\mathcal{M}}\right)\right) - E\left(Cov\left(y_{i(*)},\ y_{i(**)} \mid \boldsymbol{\mathcal{M}}\right)\right) \\ &= E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right], \end{aligned}$$

and

$$\begin{aligned} &E\left(E(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(**)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) \\ &= E\left(Cov(y_{i(*)},\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(Cov(y_{i(*)},\ y_{j(**)} \mid \boldsymbol{\mathcal{M}})\right) \\ &= E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]. \end{aligned}$$

These two results show that the expected value of the variance-covariance of the imputed responses in an imputed set given the observed data is equal to the difference between the expected value of the corresponding product of two imputed responses when the

experiment has not been observed. In particular, we can substitute

$$\begin{aligned}
&E\left(E(y_{i(*)}\ y_{i(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(*)}\ y_{i(**)} \mid \boldsymbol{\mathcal{M}})\right) \\
&= E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= \left(\begin{pmatrix} 1 & x_i \end{pmatrix} E\left((X^TX)^{-1}_{obs}\right) \begin{pmatrix} 1 \\ x_i \end{pmatrix} + 1\right)\sigma^2
\end{aligned}$$

and

$$\begin{aligned}
&E\left(E(y_{i(*)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(y_{i(**)}\ y_{j(*)} \mid \boldsymbol{\mathcal{M}})\right) \\
&= E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= \begin{pmatrix} 1 & x_i \end{pmatrix} E\left((X^TX)^{-1}_{obs}\right) \begin{pmatrix} 1 \\ x_j \end{pmatrix}\sigma^2
\end{aligned}$$

into the corresponding elements of $E(\boldsymbol{B}_t)$ for constructing an optimal design at the design stage of an experiment. To be more specific, we substitute these results into the three common elements that are in the expression of $E(\boldsymbol{B}_t)$ of the simple linear model, i.e.

$$\begin{aligned}
&E\left(E(\mathbb{C}_{(*)}\ \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(\mathbb{C}_{(*)}\ \mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}})\right) \\
&= E\left(\sum_{i=1}^{N} \mathcal{M}_i E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]\right. \\
&\qquad \left. +2\sum_{\substack{i=1 \\ j>i}}^{N} \mathcal{M}_i\mathcal{M}_j E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]\right) \\
&= \sum_{i=1}^{N} P(x_i) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&\quad + 2\sum_{\substack{i=1 \\ j>i}}^{N} P(x_i)P(x_j) E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= \sum_{i=1}^{N} P(x_i) \left(\begin{pmatrix} 1 & x_i \end{pmatrix} E\left((X^TX)^{-1}_{obs}\right) \begin{pmatrix} 1 \\ x_i \end{pmatrix} + 1\right)\sigma^2 \\
&\quad + 2\sum_{\substack{i=1 \\ j>i}}^{N} P(x_i)P(x_j) \begin{pmatrix} 1 & x_i \end{pmatrix} E\left((X^TX)^{-1}_{obs}\right) \begin{pmatrix} 1 \\ x_j \end{pmatrix}\sigma^2,
\end{aligned}$$

$$
\begin{aligned}
&E\left(E(\mathbb{D}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(\mathbb{D}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}})\right) \\
&= E\left(\sum_{i=1}^{N} x_i^2 \mathcal{M}_i E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]\right. \\
&\qquad \left. +2\sum_{\substack{i=1\\ j>i}}^{N} x_i\ x_j \mathcal{M}_i \mathcal{M}_j E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]\right) \\
&= \sum_{i=1}^{N} x_i^2 P(x_i) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&\quad + 2\sum_{\substack{i=1\\ j>i}}^{N} x_i\ x_j P(x_i) P(x_j) E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= \sum_{i=1}^{N} x_i^2 P(x_i) \left(\begin{pmatrix}1 & x_i\end{pmatrix} E\left((X^T X)_{obs}^{-1}\right) \begin{pmatrix}1\\ x_i\end{pmatrix} + 1\right)\sigma^2 \\
&\quad + 2\sum_{\substack{i=1\\ j>i}}^{N} x_i\ x_j P(x_i) P(x_j) \begin{pmatrix}1 & x_i\end{pmatrix} E\left((X^T X)_{obs}^{-1}\right) \begin{pmatrix}1\\ x_j\end{pmatrix}\sigma^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&E\left(E(\mathbb{C}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(\mathbb{C}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}})\right) \\
&= E\left(\sum_{i=1}^{N} x_i \mathcal{M}_i E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]\right. \\
&\qquad \left. + \left(\sum_{\substack{i=1\\ j>i}}^{N} x_i \mathcal{M}_i \mathcal{M}_j + \sum_{\substack{i=1\\ j>i}}^{N} x_j \mathcal{M}_i \mathcal{M}_j\right) E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]\right) \\
&= \sum_{i=1}^{N} x_i P(x_i) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&\quad + \left(\sum_{\substack{i=1\\ j>i}}^{N} x_i P(x_i) P(x_j) + \sum_{\substack{i=1\\ j>i}}^{N} x_j P(x_i) P(x_j)\right) E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right] \\
&= \sum_{i=1}^{N} x_i P(x_i) \left(\begin{pmatrix}1 & x_i\end{pmatrix} E\left((X^T X)_{obs}^{-1}\right) \begin{pmatrix}1\\ x_i\end{pmatrix} + 1\right)\sigma^2 \\
&\quad + \left(\sum_{\substack{i=1\\ j>i}}^{N} x_i P(x_i) P(x_j) + \sum_{\substack{i=1\\ j>i}}^{N} x_j P(x_i) P(x_j)\right) \begin{pmatrix}1 & x_i\end{pmatrix} E\left((X^T X)_{obs}^{-1}\right) \begin{pmatrix}1\\ x_i\end{pmatrix}\sigma^2.
\end{aligned}
$$

Note that $E\left((X^T X)_{obs}^{-1}\right)$ in the above expressions is the expected value of $V_{\hat{\boldsymbol{\beta}}_{obs}}$, i.e. the covariance matrix of the posterior distribution of $\boldsymbol{\beta}^{(*)}$. This matrix coincides with the one that occurs in a complete case analysis. Thus, the approximation result in the

optimal design framework for the linear regression model that assumes complete case analysis (Chapter 4) can be employed accordingly in these expressions.

## 6.4 Total imputation variance-covariance

We now combine the expected value of the between imputation variance-covariance and the expected value of the within imputation variance-covariance accordingly for the construction of an optimal design. More specifically, by considering a function of

$$E\left(\mathcal{T}_t\right) = E\left(\bar{\boldsymbol{U}}_t\right) + \left(1 + \frac{1}{t}\right) E(\boldsymbol{B}_t),$$

which is expressed in terms of the variables of the model, we can find an optimal design for a linear regression model that assumes multiple imputation. For example, the total imputation variance-covariance matrix of the least squares estimates (computed in the multiple imputation) for the simple linear regression model is

$$E\left(\mathcal{T}_t\right) = \begin{pmatrix} \mathcal{T}_{1,1} & \mathcal{T}_{1,2} \\ \mathcal{T}_{2,1} & \mathcal{T}_{2,2} \end{pmatrix},$$

where

$$\begin{aligned}
\mathcal{T}_{1,1} &= \frac{\sigma^2}{const^{\,2}} \left( N \left( \sum {x_i}^2 \right)^2 - \sum {x_i}^2 \left( \sum x_i \right)^2 \right) + \left(1 + \frac{1}{t}\right) E \left( \frac{\sum_{l=1}^{t} \left( \hat{\beta}_0^{(l)} - \bar{\beta}_0 \right)^2}{t-1} \right) \\
&= \frac{\sigma^2}{const^{\,2}} \left( N \left( \sum {x_i}^2 \right)^2 - \sum {x_i}^2 \left( \sum x_i \right)^2 \right) \\
&\quad + \left(1 + \frac{1}{t}\right) \frac{1}{const^{\,2}} \Big( E\left\{ E\left( \mathbb{C}_{(*)} \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}} \right) \right\} - E\left\{ E\left( \mathbb{C}_{(*)} \mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}} \right) \right\} \left( \sum {x_i}^2 \right)^2 \\
&\quad + E\left\{ E\left( \mathbb{D}_{(*)} \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}} \right) \right\} - E\left\{ E\left( \mathbb{D}_{(*)} \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}} \right) \right\} \left( \sum x_i \right)^2 \\
&\quad - 2E\left\{ E\left( \mathbb{C}_{(*)} \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}} \right) \right\} - E\left\{ E\left( \mathbb{C}_{(*)} \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}} \right) \right\} \sum {x_i}^2 \sum x_i \Big),
\end{aligned}$$

$$\begin{aligned}
\mathcal{T}_{2,2} &= \frac{\sigma^2}{const^{\,2}} \left( N^2 \sum {x_i}^2 - N \left( \sum x_i \right)^2 \right) + \left(1 + \frac{1}{t}\right) E \left( \frac{\sum_{l=1}^{t} \left( \hat{\beta}_1^{(l)} - \bar{\beta}_1 \right)^2}{t-1} \right) \\
&= \frac{\sigma^2}{const^{\,2}} \left( N^2 \sum {x_i}^2 - N \left( \sum x_i \right)^2 \right) \\
&\quad + \left(1 + \frac{1}{t}\right) \frac{1}{const^{\,2}} \Big( N^2 E\left\{ E\left( \mathbb{D}_{(*)} \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}} \right) \right\} - E\left\{ E\left( \mathbb{D}_{(*)} \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}} \right) \right\} \\
&\quad + E\left\{ E\left( \mathbb{C}_{(*)} \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}} \right) \right\} - E\left\{ E\left( \mathbb{C}_{(*)} \mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}} \right) \right\} \left( \sum x_i \right)^2 \\
&\quad - 2E\left\{ E\left( \mathbb{D}_{(*)} \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}} \right) \right\} - E\left\{ E\left( \mathbb{D}_{(**)} \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}} \right) \right\} N \sum x_i \Big),
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{T}_{1,2} &= -\frac{\sigma^2}{const^{\,2}}\left(N\sum x_i\sum {x_i}^2-\left(\sum x_i\right)^3\right)\\
&\quad+\left(1+\frac{1}{t}\right)E\left(\frac{\sum_{l=1}^{t}\left(\hat{\beta}_0^{(l)}-\bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)}-\bar{\beta}_1\right)}{t-1}\right)\\
&=-\frac{\sigma^2}{const^{\,2}}\left(N\sum x_i\sum {x_i}^2-\left(\sum x_i\right)^3\right)\\
&\quad+\left(1+\frac{1}{t}\right)\left(E\left\{E\left(\mathbb{C}_{(*)}\mathbb{D}_{(*)}\mid\boldsymbol{\mathcal{M}}\right)\right\}-E\left\{E\left(\mathbb{C}_{(*)}\mathbb{D}_{(**)}\mid\boldsymbol{\mathcal{M}}\right)\right\}\left(N\sum {x_i}^2+\left(\sum x_i\right)^2\right)\right.\\
&\quad-E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(*)}\mid\boldsymbol{\mathcal{M}}\right)\right\}-E\left\{E\left(\mathbb{C}_{(*)}\mathbb{C}_{(**)}\mid\boldsymbol{\mathcal{M}}\right)\right\}\sum {x_i}^2\sum x_i\\
&\quad\left.-E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(*)}\mid\boldsymbol{\mathcal{M}}\right)\right\}-E\left\{E\left(\mathbb{D}_{(*)}\mathbb{D}_{(**)}\mid\boldsymbol{\mathcal{M}}\right)\right\}N\sum x_i\right),
\end{aligned}$$

with $\sum=\sum_{i=1}^{N}$ in the corresponding elements and

$$const=N\sum_{i=1}^{N}{x_i}^2-\left(\sum_{i=1}^{N}x_i\right)^2.$$

Before we illustrate some two-point optimal designs for the simple linear regression model, we discuss the role of the number of imputations in $\mathcal{T}_t$. In the early days when computational tool is inadequate, most of the research investigate the validity of the inferences that is made based on Rubin's combining rules when $t$ is small. It was controversial that unnecessary variability might be incurred when $t$ is large, and initial research suggested that having five to twenty repeated imputations is sufficient for making valid inferences. However, recent investigations show that the rate of missing responses and the tolerance of the power for hypothesis testing may play a role in choosing $t$ in the multiple imputation (see Graham et al. (2007), Bodner (2008)). From the total variance-covariance of $\bar{\boldsymbol{\beta}}$, we see that the factor $1+\frac{1}{t}$ decreases to one as $t\to\infty$. Nonetheless, there is no consensus about the general behaviour of $\boldsymbol{B}_t$ as the empirical values of $\mathcal{T}_t$ also depend on the imputation model. In our investigation, we study the effect of $t=100,\ 200,\ 500$ and $\infty$ on the support points of some optimal designs. We compare the performance of these optimal designs with several values of $t$ in the simulation studies.

## 6.5 Two-point optimal design

In the class of a two point optimal design for the simple linear regression model, we can find the optimal setting of an experiment that employs the multiple imputation, given the total sample size $N$ and a monotone MAR mechanism. For a $c$-optimal design, the

objective function minimises

$$\mathcal{T}_{2,2}$$

over the design region $\mathfrak{X}$; for $A$-optimal design, the objective function minimises the trace of $E(\mathcal{T}_t)$, i.e.

$$tr(E(\mathcal{T}_t)) = \mathcal{T}_{1,1} + \mathcal{T}_{2,2}$$

over $\mathfrak{X}$ ; and for $D$-optimality, the optimal design minimises the determinant of the total variance-covariance matrix, i.e.

$$|E(\mathcal{T}_t)| = \mathcal{T}_{1,1}\mathcal{T}_{2,2} - \mathcal{T}_{1,2}\mathcal{T}_{2,1}$$

over $\mathfrak{X}$.

Considering that two support points $x_1$ and $x_2$ are sufficient to estimate $\beta_0$ and $\beta_1$ of the simple linear regression model, we index $n_1 = Nw_1$ responses, i.e. the group of responses $\{y_1,\ ...,\ y_{n_1}\} = G1$ that depends on $x_1$, and $n_2 = N(1-w_1)$ responses, i.e. the group of responses $\{y_{n_1+1},\ ...,\ y_N\} = G2$ that depends on $x_2$. To find $E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}\right)\right]$ before constructing an optimal design, we employ the second order Taylor series approximation to the elements of

$$\begin{aligned}
&E\left((X^TX)^{-1}_{obs}\right)\sigma^2 \\
&= \frac{\sigma^2}{(x_1-x_2)^2}\begin{pmatrix} x_1^2E\left(\frac{Z_1}{Z_1Z_2}\right) + x_2^2E\left(\frac{Z_2}{Z_1Z_2}\right) & -x_1E\left(\frac{Z_1}{Z_1Z_2}\right) - x_2E\left(\frac{Z_2}{Z_1Z_2}\right) \\ -x_1E\left(\frac{Z_1}{Z_1Z_2}\right) - x_2E\left(\frac{Z_2}{Z_1Z_2}\right) & E\left(\frac{Z_1}{Z_1Z_2}\right) + E\left(\frac{Z_2}{Z_1Z_2}\right) \end{pmatrix},
\end{aligned}$$

where $Z_i = \sum_{r=1}^{n_i}(1-\mathcal{M}_r)$ follows a Binomial distribution with mean $n_i = Nn_i/N = Nw_i$ and probability $(1-P(x_i))$, and

$$E\left(\frac{Z_i}{Z_iZ_j}\right) \approx \frac{1}{Nw_j(1-P(x_j))} + \frac{P(x_i)P(x_j)}{Nw_i(1-P(x_i))(Nw_j(1-P(x_j)))^2} + \frac{P(x_j)}{(Nw_j(1-P(x_j)))^2}$$

for $i \neq j$ (see Chapter 4.2 for more details).

To facilitate the analytical work in finding a design criterion, we consider the expected covariance between $y_{i(*)}$ and $y_{j(*)}$ that are imputed at $x_i = a$ and $x_j = b$ respectively,

given all the data of the experiment, i.e.

$$
\begin{aligned}
&E\left[Cov\left(y_{i(*)}, y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_i = a, x_j = b, i = 1, ..., N, j = 1, ..., N\right)\right] \\
&= \begin{pmatrix} 1 & a \end{pmatrix} E((X^T X)^{-1}_{obs}) \begin{pmatrix} 1 \\ b \end{pmatrix} \sigma^2 \\
&= \frac{\sigma^2}{(x_1 - x_2)^2} \begin{pmatrix} 1 & a \end{pmatrix} \begin{pmatrix} x_1^2 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2^2 E\left(\frac{Z_2}{Z_1 Z_2}\right) - b\left(x_1 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2 Z_2\right) \\ -x_1 E\left(\frac{Z_1}{Z_1 Z_2}\right) - x_2 E\left(\frac{Z_2}{Z_1 Z_2}\right) + b\left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) \end{pmatrix} \\
&= \frac{\sigma^2}{(x_1 - x_2)^2} \left( x_1^2 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2^2 E\left(\frac{Z_2}{Z_1 Z_2}\right) - b\left(x_1 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2 E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) \right. \\
&\qquad \left. - a\left(x_1 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2 E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) + ab\left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + E\left(\frac{Z_2}{Z_1 Z_2}\right)\right)\right) \\
&= \frac{\sigma^2}{(x_1 - x_2)^2} \left( x_1^2 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2^2 E\left(\frac{Z_2}{Z_1 Z_2}\right) - (a+b)\left(x_1 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2 E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) \right. \\
&\qquad \left. + ab\left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + E\left(\frac{Z_2}{Z_1 Z_2}\right)\right)\right).
\end{aligned}
$$

Setting $a = b = x_i = x_j$ in the above expression, we can obtain the expected variance of the imputed values given the data. For example, the variance of the imputed $y_{i(*)}$ in $G1$ which depends on support point $x_1$ is

$$
\begin{aligned}
&E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; i = j, x_i = x_j = x_1, i = 1, ...,\ n_1\right)\right] \\
&= E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_i = x_1, i = 1, ...,\ n_1\right)\right] \\
&= \begin{pmatrix} 1 & x_1 \end{pmatrix} E((X^T X)^{-1}_{obs}) \sigma^2 \begin{pmatrix} 1 \\ x_1 \end{pmatrix} + \sigma^2 \\
&= \frac{\sigma^2}{(x_1 - x_2)^2} \left( x_1^2 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2^2 E\left(\frac{Z_2}{Z_1 Z_2}\right) - 2x_1\left(x_1 E\left(\frac{Z_1}{Z_1 Z_2}\right) + x_2 E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) \right. \\
&\qquad \left. + x_1^2\left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + E\left(\frac{Z_2}{Z_1 Z_2}\right)\right)\right) + \sigma^2 \\
&= \frac{\sigma^2}{(x_1 - x_2)^2} \left( x_2^2 E\left(\frac{Z_2}{Z_1 Z_2}\right) - 2x_1 x_2 E\left(\frac{Z_2}{Z_1 Z_2}\right) + x_1^2 E\left(\frac{Z_2}{Z_1 Z_2}\right)\right) + \sigma^2 \\
&= \left(E\left(\frac{Z_2}{Z_1 Z_2}\right) + 1\right)\sigma^2,
\end{aligned}
$$

and the variance of $y_{i(*)}$ in $G2$ which depends on support point $x_2$ is

$$
\begin{aligned}
&E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; i = j, x_i = x_j = x_2, i = n_1 + 1, ..., N\right)\right] \\
&= E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_i = x_2, i = n_1 + 1, ..., N\right)\right] \\
&= \left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + 1\right)\sigma^2.
\end{aligned}
$$

We now find the expected covariance of two imputed values that are in the same repeated imputation given the information. For two imputed values of two subjects who are in the same groups, we have

$$\begin{aligned}
&E[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_i = x_j = x_1,\ i \neq j,\ i = 1, ..., n_1\right)] \\
&= \begin{pmatrix} 1 & x_1 \end{pmatrix} E((X^T X)^{-1}_{obs}) \begin{pmatrix} 1 \\ x_1 \end{pmatrix} \sigma^2 \\
&= E\left(\frac{Z_2}{Z_1 Z_2}\right) \sigma^2
\end{aligned}$$

for the imputed values at $x_1$; and

$$\begin{aligned}
&E[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_i = x_j = x_2,\ i \neq j,\ i = n_1 + 1, ..., N\right)] \\
&= \begin{pmatrix} 1 & x_2 \end{pmatrix} E((X^T X)^{-1}_{obs}) \begin{pmatrix} 1 \\ x_2 \end{pmatrix} \sigma^2 \\
&= E\left(\frac{Z_1}{Z_1 Z_2}\right) \sigma^2
\end{aligned}$$

for those at $x_2$. Since the experimental units are independent, it can be shown that given the observed data, there is no correlation between the imputed value of the subject who is in $G1$ and the imputed value of the subject who is in $G2$, i.e.

$$E[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_i = x_1, x_j = x_2,\ i \neq j,\ i = 1, ..., n_1,\ j = n_1 + 1, ..., N\right)] = 0.$$

On the other hand, considering the imputed values which are in different imputed sets, we have

$$E\left[Cov\left(y_{i(*)},\ y_{j(**)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; \forall i, j\right)\right] = 0$$

because given the data, the repeated imputations are independent.

Substituting these results into the common elements in the expression of $E(\boldsymbol{B}_t)$, we get

$$
\begin{aligned}
& E\left(E(\mathbb{C}_{(*)}\ \mathbb{C}_{(*)} \mid \boldsymbol{\mathcal{M}})\right) - E\left(E(\mathbb{C}_{(*)}\ \mathbb{C}_{(**)} \mid \boldsymbol{\mathcal{M}})\right) \\
&= \sum_{i=1}^{n_1} P(x_1) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_1\right)\right] \\
&\quad + 2 \sum_{\substack{i=1 \\ j>i}}^{n_1} P(x_1)^2 E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_1\right)\right] \\
&\quad + \sum_{i=n_1+1}^{N} P(x_2) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_2\right)\right] \\
&\quad + 2 \sum_{\substack{i=n_1+1 \\ j>i}}^{N} P(x_2)^2 E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_2\right)\right] \\
&= n_1 P(x_1) \left(E\left(\frac{Z_2}{Z_1 Z_2}\right) + 1\right) \sigma^2 + n_1(n_1 - 1) P(x_1)^2 E\left(\frac{Z_2}{Z_1 Z_2}\right) \sigma^2 \\
&\quad + n_2 P(x_2) \left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + 1\right) \sigma^2 + n_2(n_2 - 1) P(x_2)^2 E\left(\frac{Z_1}{Z_1 Z_2}\right) \sigma^2,
\end{aligned}
$$

$$
\begin{aligned}
& E\left(E(\mathbb{D}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}})\right)\} - E\left(E(\mathbb{D}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}})\right) \\
&= \sum_{i=1}^{n_1} x_1^2 P(x_1) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_1\right)\right] \\
&\quad + 2 \sum_{\substack{i=1 \\ i>j}}^{n_1} x_1^2 P(x_1)^2 E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_1\right)\right] \\
&\quad + \sum_{i=n_1+1}^{N} x_2^2 P(x_2) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_2\right)\right] \\
&\quad + 2 \sum_{\substack{i=n_1+1 \\ i>j}}^{N} x_2^2 P(x_2)^2 E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_2\right)\right] \\
&= n_1 x_1^2 P(x_1) \left(E\left(\frac{Z_2}{Z_1 Z_2}\right) + 1\right) \sigma^2 + n_1(n_1 - 1) x_1^2 P(x_1)^2 E\left(\frac{Z_2}{Z_1 Z_2}\right) \sigma^2 \\
&\quad + n_2 x_2^2 P(x_2) \left(E\left(\frac{Z_1}{Z_1 Z_2}\right) + 1\right) \sigma^2 + n_2(n_2 - 1) x_2^2 P(x_2)^2 E\left(\frac{Z_1}{Z_1 Z_2}\right) \sigma^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&E\left(E(\mathbb{C}_{(*)}\ \mathbb{D}_{(*)} \mid \boldsymbol{\mathcal{M}})\right\} - E\left(E(\mathbb{C}_{(*)}\ \mathbb{D}_{(**)} \mid \boldsymbol{\mathcal{M}})\right\} \\
&= \sum_{i=1}^{n_1} x_1 P(x_1) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_1\right)\right] \\
&\quad + 2\sum_{\substack{i=1 \\ i>j}}^{n_1} x_1 P(x_1)^2 E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_1\right)\right] \\
&\quad + \sum_{i=n_1+1}^{N} x_2 P(x_2) E\left[var\left(y_{i(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_2\right)\right] \\
&\quad + 2\sum_{\substack{i=n_1+1 \\ i>j}}^{N} x_2 P(x_2)^2 E\left[Cov\left(y_{i(*)},\ y_{j(*)} \mid \mathbf{x}, \mathbf{y}, \sigma^2, \boldsymbol{\mathcal{M}}; x_2\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= n_1 x_1 P(x_1)\left(E\left(\frac{Z_2}{Z_1 Z_2}\right)\sigma^2 + 1\right)\sigma^2 + n_1(n_1-1)x_1 P(x_1)^2 E\left(\frac{Z_2}{Z_1 Z_2}\right)\sigma^2 \\
&\quad + n_2 x_2 P(x_2)\left(E\left(\frac{Z_1}{Z_1 Z_2}\right)\sigma^2 + 1\right)\sigma^2 + n_2(n_2-1)x_2 P(x_2)^2 E\left(\frac{Z_1}{Z_1 Z_2}\right)\sigma^2,
\end{aligned}
$$

where $n_1 = Nw_1$, $n_2 = N(1-w_1)$, and

$$
E\left(\frac{Z_i}{Z_i Z_j}\right) \approx \frac{1}{Nw_j(1-P(x_j))} + \frac{P(x_i)P(x_j)}{Nw_i(1-P(x_i))(Nw_j(1-P(x_j)))^2} + \frac{P(x_j)}{(Nw_j(1-P(x_j)))^2}
$$

for $i \neq j$. Using these results of $E(\boldsymbol{B}_t)$, and the elements of $E(\bar{\boldsymbol{U}}_t)$ of the simple linear regression model, we can find an optimal design by minimising a function of $E(\mathcal{T}_t)$ over $\mathfrak{X}$, given the monotone MAR mechanism $P(x_i)$.

To illustrate this, consider $\mathfrak{X} = [0, u]$ where $u$ is a non-negative number, $N = 60$, and assume the inverse logit link function

$$
P(x_i) = \frac{exp(\gamma_0 + \gamma_1 x_i)}{1 + exp(\gamma_0 + \gamma_1 x_i)}
$$

has $\gamma_0 = -4.572$ and $\gamma_1 = 3.191$ for the MAR mechanism of an experiment. These values were chosen such that $P(x_i = 0) = 0.01$. Having fixed the lower bound as one of the support points, say $x_1$, in the optimisation problem of the corresponding optimality, we can find a two-point optimal design for a simple linear regression model by solving the design problem numerically with respect to $x_2$ and $w_2$, subjecting to $w_1 + w_2 = 1$.

Table 6.1 shows some examples of $A$-, $c$- and $D$-optimal designs (denoted by $\xi_A^*$, $\xi_c^*$ and $\xi_D^*$ respectively) that are found by the different design framework. Each row corresponds to the optimal settings, i.e. the second optimal point $x_2$ and the corresponding weight $w_2$ of $A$-, $c$- and $D$-optimal designs respectively. The values under the header multiple imputation correspond to those that are found by the design framework which assumes

Table 6.1: The optimal support points, $x_2$, and the corresponding weights, $w_2$, that are found by different design framework for different design criteria.

| | Multiple imputation | | | | | Complete case analysis | | | |
|---|---|---|---|---|---|---|---|---|---|
| $t =$ | 100 | 200 | 500 | $\infty$ | | 1st order | | 2nd order | |
| | $x_2$ | $x_2$ | $x_2$ | $x_2$ | $w_2(n_2)$ | $x_2$ | $w_2(n_2)$ | $x_2$ | $w_2(n_2)$ |
| $\xi_A^*$ | 1.495 | 1.496 | 1.496 | 1.497 | 0.454(28) | 1.515 | 0.454(28) | 1.485 | 0.461 (28) |
| $\xi_c^*$ | 1.579 | 1.580 | 1.581 | 1.581 | 0.622(38) | 1.601 | 0.621(38) | 1.572 | 0.623 (38) |
| $\xi_D^*$ | 1.364 | 1.365 | 1.365 | 1.366 | 0.503(31) | 1.377 | 0.500(30) | 1.354 | 0.506 (31) |

multiple imputation with $t = 100$, $t = 200$, $t = 500$ and $t = \infty$ respectively; the values under header complete case analysis are the optimal settings that are found by the framework which assumes the first order and the second order Taylor series approximation to $E\left(\frac{Z_i}{Z_iZ_j}\right)$ respectively in the complete case analysis. Comparing the optimal designs that are found by the different design framework for the same design criterion, we find that $x_2$ of the designs that assume multiple imputation lies between the $x_2$ of those designs that assume a complete case analysis, with the first and the second order approximation respectively. Nevertheless, $w_2$ of these optimal designs are very similar especially after rounding $Nw_2$ to integer values for implementing the experiment.

Concerning the role of $t$ on the optimal designs, we see that the impact of $t$ on $x_2$ is not significant when we compare the designs that assume multiple imputation and with different $t$ (in Table 6.1) for a specific design criterion. To explore further, we employ the same MAR mechanism in the design problems, and find the optimal designs for $5 < t \leq 1000$ and different $N$ using the framework that assumes multiple imputation. Figure 6.1 shows the values of $x_2$ of the $A$-, $c$- and $D$-optimal designs across the values of $t$ for the experiments that have $N = 30$, $N = 300$ and $N = 3000$ respectively. The average $w_2$ of these optimal designs are given in the legend of the plot. For each design criterion, we find that the average $w_2$ for the experiment with $N = 30$ is slightly greater than those of the experiments that have larger sample sizes; and the impact of sample size on $x_2$ becomes smaller when we compare $x_2$ of a design criterion for the experiment that has $N = 300$ with those for the experiment that has $N = 3000$. For an experiment with a given $N$, we see that as $t$ increases, the optimal support point of a design criterion approaches to $x_2$ of the optimal design that assumes $t = \infty$.

In the next section, we compare the performance of the designs that are presented in Table 6.1 through some simulation studies. We assess the empirical values of the total variance-covariance of $\bar{\boldsymbol{\beta}}$, and those of the $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ which is computed in complete case analysis.

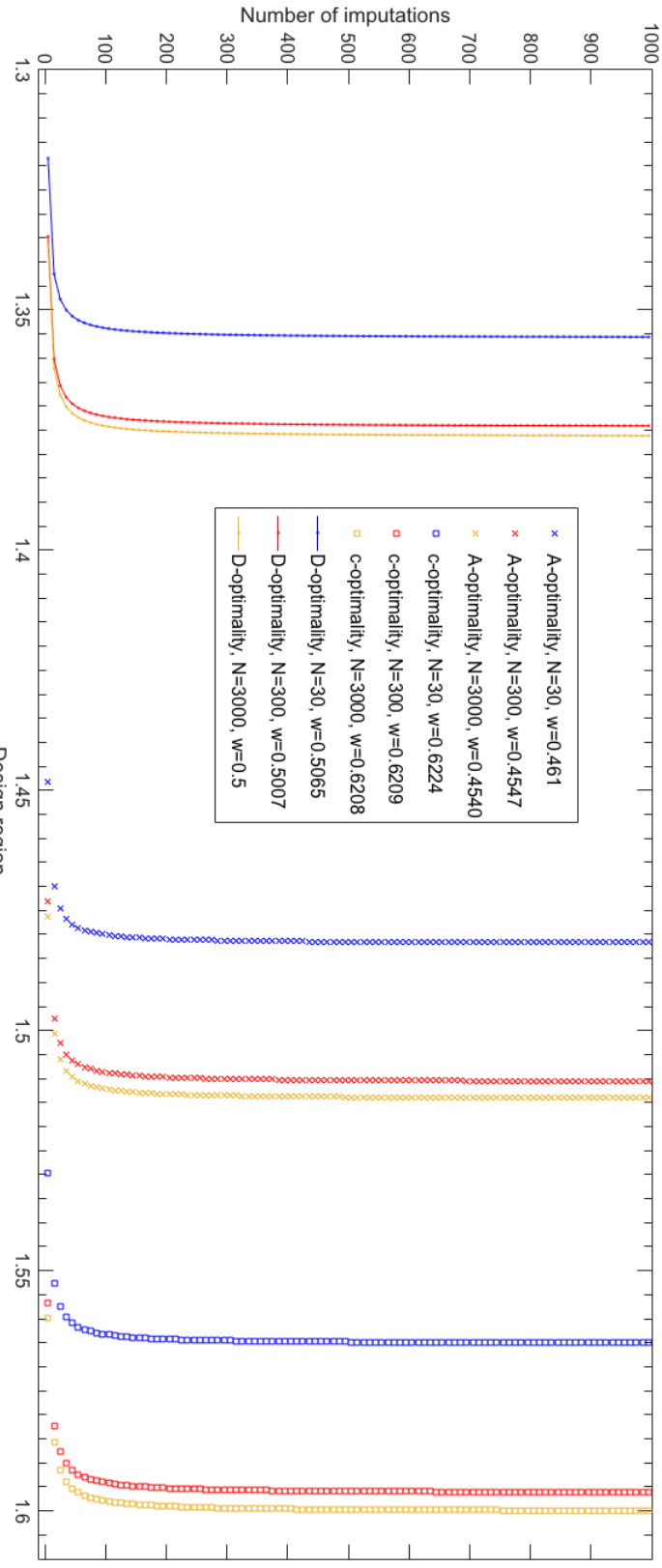



Figure 6.1: The second optimal support point of $A$-, $c$- and $D$-optimal designs with different $t$ and $N$.

## 6.6 Simulation study

We now illustrate the set-up of the simulation study. For each design that is given in Table 6.1, we simulate a response variable by the simple linear regression model

$$y_i = 1 + x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

i.e. $\beta_0 = \beta_1 = 1$, where $\sigma^2 = 1$ and $i = 1, \ldots, 60$, with $n_1$ responses at $x_1 = 0$ and $n_2$ responses at $x_2$ of the design. The missing values are introduced into the simulated $y_i$, by specifying a MAR mechanism through

$$P(x_i) = \frac{exp(-4.572 + 3.191x_i)}{1 + exp(-4.572 + 3.191x_i)}.$$

In each incomplete data set, we compute the sample estimate $\hat{\boldsymbol{\beta}}_{obs} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{obs}\boldsymbol{X}^T_{obs}\boldsymbol{y}_{obs}$ with variance-covariance $V_{\hat{\boldsymbol{\beta}}_{obs}}\sigma^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{obs}\sigma^2$ from the complete cases, i.e. $\boldsymbol{X}_{obs}$ and $\boldsymbol{y}_{obs}$ comprise only the information of the subjects who have observed responses. These correspond to the mean and the variance-covariance matrix of the posterior distribution of $\boldsymbol{\beta}^*$, i.e. $\boldsymbol{N}(\hat{\boldsymbol{\beta}}_{obs}, V_{\hat{\boldsymbol{\beta}}_{obs}}\sigma^2)$, for imputing the missing responses. Moreover, they also correspond to the least squares estimates $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ that are obtained in a complete case analysis.

To impute a missing response at the corresponding $x_i$, we draw a set of $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)^T$ and a value of $\epsilon_i$ from $N(\hat{\boldsymbol{\beta}}_{obs}, V_{\hat{\boldsymbol{\beta}}_{obs}}\sigma^2)$ and $N(0, \sigma^2)$ respectively, such that an imputed value of subject $i$ is

$$y_{i(*)} = \beta_0^* + \beta_1^* x_i + \epsilon_i.$$

After replacing the missing values in the incomplete data set, we compute the least square estimates $\hat{\boldsymbol{\beta}}^{(l)}$ with its covariance $\boldsymbol{cov}(\hat{\boldsymbol{\beta}}^{(l)})$ by treating all the imputed values and the observed responses as a complete-data set $l$. This imputation and least squares analysis are repeated for $l = 1, ..., t$, to obtain $\bar{\boldsymbol{\beta}} = \frac{\sum_{l=1}^{t}\hat{\boldsymbol{\beta}}^{(l)}}{t}$ and an empirical value of between imputation variance-covariance $\boldsymbol{B}_t$ that is computed using the sample estimates $\hat{\boldsymbol{\beta}}^{(l)}$, $l = 1, ..., t$.

By repeating the whole process for the same design on other sets of simulated responses and missing values, we find the empirical value of the expected total variance-covariance $E(\mathcal{T}_t)$ by computing the variances of $\bar{\boldsymbol{\beta}}$ across the simulated data; and the empirical value of the expected between imputation variance-covariance $E(\boldsymbol{B}_t)$ by averaging the sample estimates $\boldsymbol{B}_t$. Using these values, we can then find the empirical values of the expected within imputation variance-covariance, i.e. $E(\bar{\boldsymbol{U}}_t) = E(\mathcal{T}_t) - \left(1 + \frac{1}{t}\right)E(\boldsymbol{B}_t)$, to verify our assumption of $E(\bar{\boldsymbol{U}}_t)$ in the optimal design. To do that, we consider the relative difference

$$RD = 1 - \frac{(X^TX)^{-1}\sigma^2}{E(\mathcal{T}_t) - \left(1 + \frac{1}{t}\right)E(\boldsymbol{B}_t)},$$

Table 6.2: Simulation outputs of some two-point optimal designs with $t = 100$ and $N = 60$, averaged across 400 000 simulated sets.

| | $var(\bar{\beta}_0)$ | $var(\bar{\beta}_1)$ | $cov(\bar{\beta}_0, \bar{\beta}_1)$ |
|---|---|---|---|
| | $\xi_A^*$ with $x_2$=1.4948 | | |
| $E(\mathcal{T}_t)$ | 3.1575e-02 | 5.1636e-02 | -2.1180e-02 |
| $E(\boldsymbol{B}_t)$ | 3.3228e-04 | 2.1409e-02 | -2.2200e-04 |
| $E(\bar{\boldsymbol{U}}_t)$ | 3.1239e-02 | 3.0013e-02 | -2.0956e-02 |
| $(X^TX)^{-1}\sigma^2$ | 3.1250e-02 | 2.9969e-02 | -2.0906e-02 |
| $RD$ | -3.4014e-04 | 1.4698e-03 | 2.3757e-03 |
| | $\xi_c^*$ with $x_2$=1.5791 | | |
| $E(\mathcal{T}_t)$ | 4.5850e-02 | 4.7389e-02 | -2.9060e-02 |
| $E(\boldsymbol{B}_t)$ | 4.9014e-04 | 1.8324e-02 | -3.1124e-04 |
| $E(\bar{\boldsymbol{U}}_t)$ | 4.5355e-02 | 2.8882e-02 | -2.8745e-02 |
| $(X^TX)^{-1}\sigma^2$ | 4.5455e-02 | 2.8782e-02 | -2.8785e-02 |
| $RD$ | -2.1942e-03 | 3.4533e-03 | -1.3796e-03 |
| | $\xi_D^*$ with $x_2$= 1.3637 | | |
| $E(\mathcal{T}_t)$ | 3.4799e-02 | 5.0973e-02 | -2.5477e-02 |
| $E(\boldsymbol{B}_t)$ | 3.6755e-04 | 1.5006e-02 | -2.7020e-04 |
| $E(\bar{\boldsymbol{U}}_t)$ | 3.4428e-02 | 3.5817e-02 | -2.5204e-02 |
| $(X^TX)^{-1}\sigma^2$ | 3.4483e-02 | 3.5888e-02 | -2.5286e-02 |
| $RD$ | -1.5958e-03 | -2.0050e-03 | -3.2554e-03 |

where $(X^TX)$ is the full information matrix of the simple linear regression model, $\sigma^2 = 1$, $E(\mathcal{T}_t)$ and $E(\boldsymbol{B}_t)$ are the empirical values that are obtained in the simulation.

With 400 000 simulated data sets and $t = 100$ in a simulation, we find that $RD$ in general is very small. Table 6.2 shows the $RD$ of the $A$-, $c$-, and $D$-optimal designs that is found by the optimal design framework that considers multiple imputation with $t = 100$, i.e. the design settings in the first and the fifth column from the left in Table 6.1. With the small $t$ in this example, the $RD$ of the elements of the within imputation variance-covariance against the theoretical value of $(X^TX)^{-1}\sigma^2$ is less than 0.35%, implying that our assumption of $E(\bar{\boldsymbol{U}}_t) = (X^TX)^{-1}\sigma^2$ in the derivation of the optimal design framework is acceptable.

We now assess the performance of the designs that are given in Table 6.1. For each design criterion, we denote the optimal designs that are found by the design framework which assumes the first and the second order approximation in the complete case analysis by $\xi_{1st}$ and $\xi_{2nd}$ respectively; the optimal designs that are found by the design framework that considers multiple imputation by $\xi_{100}$, $\xi_{200}$, $\xi_{500}$, and $\xi_{\infty}$ respectively where the subscript indicates the $t$ that is chosen in the design problem. Table 6.3 shows the simulation outputs of these $A$-, $c$-, and $D$-optimal designs which are obtained by following the above described simulation procedures, with 400 000 number of simulated sets for each design. The rows of values that are under each design criterion correspond to the function of the covariance matrix of interest. For example, under $A$-optimality, the first column of values correspond to the empirical values of the trace of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$, where $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$

Table 6.3: The simulation output of various designs averaged across 400 000 replications.

| | Complete case analysis | Multiple imputation | | | | | |
|---|---|---|---|---|---|---|---|
| | | $t = 100$ | | $t = 200$ | | $t = 500$ | |
| | | Obj | Truth | Obj | Truth | Obj | Truth |
| | | | | *A*-optimality (e-02) | | | |
| $\xi_{1st}$ | 8.30901 | 8.32900 | 8.33204 | 8.31722 | 8.31926 | 8.30994 | 8.31329 |
| $\xi_{2nd}$ | **8.29355** | **8.31679** | **8.31439** | **8.30559** | **8.30360** | **8.29916** | **8.29830** |
| $\xi_{100}$ | 8.29824 | 8.31775 | 8.32110 | 8.30807 | 8.30816 | 8.30125 | 8.30265 |
| $\xi_{200}$ | 8.29972 | 8.31935 | 8.32148 | 8.30811 | 8.31093 | 8.30154 | 8.30422 |
| $\xi_{500}$ | 8.29990 | 8.31996 | 8.32074 | 8.30841 | 8.31196 | 8.30177 | 8.30431 |
| $\xi_{\infty}$ | 8.30063 | 8.31963 | 8.32133 | 8.30882 | 8.31127 | 8.30168 | 8.30495 |
| | | | | *c*-optimality (e-02) | | | |
| $\xi_{1st}$ | 4.72608 | 4.73673 | 4.74409 | 4.72622 | 4.73558 | 4.72107 | 4.72988 |
| $\xi_{2nd}$ | **4.72067** | **4.72815** | 4.73958 | **4.71887** | **4.72931** | **4.71325** | **4.72399** |
| $\xi_{100}$ | 4.72161 | 4.72892 | **4.73890** | 4.71928 | 4.73118 | 4.71405 | 4.72405 |
| $\xi_{200}$ | 4.72143 | 4.72890 | 4.74073 | 4.71943 | 4.73112 | 4.71395 | 4.72559 |
| $\xi_{500}$ | 4.72208 | 4.72931 | 4.74091 | 4.71984 | 4.73171 | 4.71389 | 4.72620 |
| $\xi_{\infty}$ | 4.72215 | 4.72920 | 4.74048 | 4.71960 | 4.73202 | 4.71401 | 4.72573 |
| | | | | *D*-optimality (e-03) | | | |
| $\xi_{1st}$ | 1.12079 | 1.12685 | 1.12601 | 1.12405 | 1.12355 | 1.12250 | 1.12199 |
| $\xi_{2nd}$ | **1.11933** | **1.12519** | **1.12468** | **1.12273** | **1.12199** | **1.12116** | **1.12035** |
| $\xi_{100}$ | 1.11936 | 1.12583 | 1.12472 | 1.12298 | **1.12199** | 1.12151 | 1.12041 |
| $\xi_{200}$ | 1.11982 | 1.12560 | 1.12531 | 1.12302 | 1.12243 | 1.12146 | 1.12070 |
| $\xi_{500}$ | 1.11966 | 1.12578 | 1.12503 | 1.12314 | 1.12219 | 1.12160 | 1.12082 |
| $\xi_{\infty}$ | 1.11965 | 1.12560 | 1.12495 | 1.12310 | 1.12224 | 1.12153 | 1.12071 |

'Obj' corresponds to the value of function of $E(\mathcal{T}_t) = (X^TX)^{-1}\sigma^2 + \left(1 + \frac{1}{t}\right) E(\boldsymbol{B}_t)$ where $E(\boldsymbol{B}_t)$ are the empirical values in the simulation; 'Truth' corresponds to the value of function of empirical $E(\mathcal{T}_t)$ in the simulation.

are obtained in complete case analysis; the values under the headers 'Obj' correspond to the trace of $E(\mathcal{T}_t) = (X^TX)^{-1}\sigma^2 + \left(1 + \frac{1}{t}\right) E(\boldsymbol{B}_t)$ where $E(\boldsymbol{B}_t)$ are the empirical values in the simulation for each design; and the values under the headers 'Truth' correspond to the trace of the empirical values of $E(\mathcal{T}_t)$ in the simulation. The numbers of repeated imputations employed in the simulation studies for each design candidate are given in the table.

Assessing the simulation outputs for each design criterion, we find that in general, the optimal design that is found by the design framework which assumes the second order approximation in the complete case analysis, i.e. $\xi_{2nd}$, has achieved the best performance. Among the considered designs under each design criterion, $\xi_{2nd}$ obtains the lowest value of the corresponding function of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$, where $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$ are computed in complete case analysis. Moreover, concerning the outputs of the multiple imputation with the corresponding $t$, $\xi_{2nd}$ also attains the minimum value (all cases under headers Obj and most cases under headers Truth) of the corresponding function of $E(\mathcal{T}_t)$. We

are not surprised with this finding as the $x_2$ that are found by this design framework are close to the $x_2$ that are found by the design framework that assumes multiple imputation. Consequently the numbers of missing responses (i.e. generated by the MAR mechanism that is dependent on the values of the support points) in the simulation studies of these optimal designs are very similar, yielding a similar performance in the analysis of the simulated data. Moreover, we find that the corresponding values that are computed with complete cases are smaller than those that are obtained in the multiple imputation. This finding verifies the proposition which states that for the linear regression model with missing values present only in the outcome variable and are missing at random, complete case analysis provides valid inferences (Little (1992)).

## 6.7 Conclusion

In this chapter, we have derived the optimal design framework for the linear regression model which assumes the implementation of multiple imputation for the incomplete data that contains responses missing at random. We assume that the imputation model has the same formulation of the linear regression model that is employed for making inferences. To tackle the uncertainty of not observing the experiments at the design stage of a study, we consider some iterative expectations and approximations to the elements of the total variance-covariance matrix of the estimators that are computed in the multiple imputation. Note that in this design framework, we have applied the approximation to some fractions of random variables (such as $\frac{Z_i}{Z_i Z_j}$) that we learnt in the design framework which assumes complete cases analysis (see Chapter 4), as the information of the subjects who have fully observed responses is involved in the construction of the imputation model. Employing a first order approximation to the fractions of random variables that are in the between imputation variance-covariance, we find that the optimal designs coincide with those that are found by the framework that is proposed by Imhof et al. (2002), which assumes complete cases analysis and with the first order approximation.

Nevertheless, we have compared the optimal designs that assume multiple imputation and the second order approximation, with those that are found by the design framework which assumes complete case analysis and the first order and the second order approximation respectively in the previous section. The simulation studies show that the optimal designs that assume complete case analysis and the second order approximation achieve the best performance when the empirical values of the corresponding design criterion are compared. Moreover, the simulation outputs agree with Little (1992) that a multiple imputation approach introduces unnecessary variability into the least squares estimates when a MAR mechanism presents only in the response variable of the linear regression model. Nevertheless, to the best of our knowledge, no literature has considered the features of multiple imputation at the design stage of an experiments. We

believe that this work can be served as a foundation for future research in this area, such as finding the optimal designs for the linear mixed models when a multiple imputation approach is considered at the design stage of an experiment.

# Chapter 7

# Conclusion

In this thesis, we have derived optimal design framework that accounts for the presence of responses missing at random and the features of some missing data analysis approaches. Here we summarise and discuss the main findings of this novel research, and propose some key ideas for future investigation on this research area.

## 7.1 Results and conclusions

In general, this project is a fusion of missing data analysis and design of experiments. This work is of great importance due to its applications in areas that involve expensive experiments and which often suffer from having great loss of information in the study. Employing the notion of design of experiments, we can find design settings that provide sufficient information for making conclusions with high precision. The novelty of our research is clear from the incorporation of different missing data analysis approaches into the optimal design framework, which covers the design and the analysis stage of a study. More specifically, this work provides experimental designs that are more robust to the presence of responses missing at random. We have considered three missing data analysis methods respectively in the optimal design framework for two classes of statistical models. The main challenge of this work is to find the approximation to the variance-covariance matrix of the estimators that are computed in the considered missing data analysis approaches given the monotone missing at random mechanism, before constructing an optimal design for implementing the experiment.

Assuming complete case analysis for the linear regression model, we have assessed the variance-covariance matrix of the least square estimates at the design stage of an experiment. When finding an expectation of this matrix for the construction of a design, complication arose due to the chances of having singular information matrix when all observations at one of the support points are missing. In Chapter 4 of this thesis, we

have considered multivariate Taylor series approximation to the elements of the matrix which consist of some fractions of random variables. We have proposed an optimal design framework that includes a framework proposed by Imhof et al. (2002) as a special case, where the authors implicitly assume the first order Taylor series approximation in their investigation. Moreover, we have illustrated some theoretical properties and results of using an approach based on the first order Taylor series approximation.

Overall, the assumption of a monotone MAR mechanism in the design framework for the linear regression model causes the support points to be shifted away from the design region that has high missing probability, and allocates more experimental units to the support points that are expected to have relatively more missing responses. This finding is more realistic than the design settings that assume full observations especially in its applications to a dose-response study (where dose level is the support point), as our optimal designs also provide an insight of scaling up the sample size of different groups unequally to account for the impact of missing responses on the power of the study.

On the other hand, for those studies that focus on the changes of an outcome variable over time, we consider some linear mixed models and an available case analysis for handling the longitudinal data that has monotone dropout patterns in Chapter 5. Following Ortega-Azurduy et al. (2008) who investigate the efficiency loss of $D$-optimal designs due to the presence of dropouts in a cohort, we have investigated the impact of dropouts on optimal cohort designs where different groups of subjects may have different monotone MAR mechanisms. Same as Ouwens et al. (2002) who investigate optimal cohort design with the assumption of fully observed repeated measurements and Ortega-Azurduy et al. (2008), we capture the within-subject correlations by a first order autoregressive process in the distribution of observational errors, and the random effect that is caused by the explanatory variable in the distribution of the random coefficients, in the formulation of the linear mixed models.

To this date, we believe that our design framework for the linear mixed models with the first order autoregressive covariance structure and the presence of dropouts are new contributions to the literature on design of experiments. This work constructs optimal cohort designs in the presence of different monotone dropout patterns for different groups, and the optimal cohort designs can be applied to the experiments that have the same baseline measurements on all subjects and the experiments that have different baseline measurements on different cohorts respectively. In other words, given some monotone MAR mechanisms, we can find the optimal time points of measuring the outcome variable on different cohorts and the corresponding weight of the groups for two different formulations of the linear mixed models. For some classes of the linear mixed models, we find that the optimal sets of time points are not close to the equally spaced time points. Moreover, this work gives the perception of measuring different groups of subjects at different sets of time points, which could be a new approach to conducting

longitudinal study in practice when the monotone dropout patterns of different groups are significantly different from one another.

Considering a more appealing missing data analysis approach than the complete case analysis and available case analysis, we have studied the role of multiple imputation in the optimal design framework for the linear regression model, i.e. find the support points for an explanatory variable and the corresponding weights assuming that missing responses are imputed repeatedly by the realisations of a plausible distribution. To the best of our knowledge, none of the literature has accounted for the features/impact of employing a multiple imputation approach at the design stage of an experiment. In this work, we have assumed the same formulation of linear regression model for both the imputation model and the analysis model. Inspecting Rubin's combining rule (Rubin (1987)) in Chapter 6 of this thesis, we have scrutinised the expectation of the total variance-covariance matrix of the least squares estimates that are obtained in the multiple imputation. The main challenge on finding the expectation for the matrix before constructing an optimal design lies in tackling the uncertainty of not knowing which subjects have missing responses and hence require imputation. Nevertheless, employing iterative expectations and the second order Taylor series approximation to the elements of the matrix, we find that the optimal designs for the experiment that assumes multiple imputation are similar to those that are found by the design framework which considers complete case analysis. In particular, the optimal designs that employ the first order Taylor series approximation to the total variance-covariance matrix (that is obtained in multiple imputation) coincide with those that assume complete case analysis and the first order Taylor series approximation. Moreover, comparing the performance of the optimal designs that are found by the different design framework for the same monotone MAR mechanism and the same sample size, we see that the optimal design that assumes complete case analysis and the second order Taylor series approximation outperformed the other two optimal designs (that assume complete case analysis and the first order approximation, and that assume multiple imputation and the second order approximation respectively) in the simulation studies. Furthermore for an optimal design, the variance of the least squares estimates that is computed in complete cases analysis is smaller than those that is obtained in the multiple imputation. This latter finding reaffirms the proposition from the design point of view that complete case analysis is sufficient for the linear regression model when missing values present only in the outcome variable of the study (Little (1992)).

## 7.2 Future work

We now suggest some key ideas for future research on this interdisciplinary project. In this project, the assumption of a monotone MAR mechanism plays a crucial role in the suggested optimal design framework. Hence, we are suggesting to explore the impact

of non-monotone MAR mechanisms on the optimal designs for the linear regression model and the linear mixed models respectively with the corresponding missing data analysis approaches that we have considered here. Moreover as mentioned in Chapter 5, an intermittent missing data pattern might be incorporated into the optimal design framework for the linear mixed models as an extension to the suggested framework here. The MAR mechanisms might account for the missingness that depend on the experimental conditions of different groups.

In practice, missing values are not necessarily being generated by a MAR mechanism. However only sensitivity analysis can be applied to test the assumptions of the missing data mechanism after observing the data. We believe that finding an optimal design that accounts for the impact of responses missing not at random (MNAR) will be beneficial for many experimental studies, especially in the clinical studies and the observational studies. Since MNAR in general is more complicated and more difficult to tackle, the challenges on finding some optimal designs in this circumstance may lie in specifying a suitable formulation of the missing data mechanism and in approximating the expected information matrix of the chosen statistical model at the design stage of an experiment.

On the other hand, since the optimal design framework that we has considered in this project is model based and is tailored for some missing data analysis methods, some optimality criteria could be developed to study multiple objectives of an experiment. These optimality criteria might include the impact of using some missing data analysis methods for finding a robust design that is adaptive to the data analysis method chosen by the data analyst after the implementation of the experiment. In addition to that, the optimal design framework for other models such as the generalised linear models and the generalised linear mixed models might be extended to account for the presence of missing responses based on our works. These models are useful in the context of modelling data that is not normally distributed. Besides, different approach of tackling the presence of missing data such as likelihood-based methods, and inverse probability weighting might be explored to find some robust designs.

Note that the suggested framework here provides only locally optimal designs for the linear regression models and the linear mixed models respectively as the missing data mechanisms and the structure of the covariance matrices (for the linear mixed models) are assumed to be known at the design stage of an experiment. These information can be estimated using data that is obtained from historical/pilot studies. A future investigation could include parameter robust designs, such as Bayesian design framework (Chaloner and Verdinelli (1995)). The probability that a design provides insufficient information for estimation due to the presence of missing responses may also be incorporated into the design criterion (see the idea of Müller (1995)).

In conclusion, we have shown that there is some statistical gain in accounting for the impact of missing responses and the features of some missing data analysis approaches

at the design stage of an experiment. In particular, the optimal design settings that are found by our design framework compromise between maximising information and minimising the proportion of missing information. We have also provided the novel research of design framework for the linear regression model that assumes multiple imputation. Even though this missing data analysis method is not beneficial for the situation that we have considered here, our finding could be served as a foundation to future research on design of experiments that considers a multiple imputation approach.

# Appendix A

In these appendices, we illustrate the proof of some materials that are presented in Chapter 4 of this thesis, i.e. the optimal design framework for general linear regression model with complete case analysis.

## A.1 Multivariate second order Taylor series approximation

Let $X$ and $Y$ be two independent random variables with expectations $\overline{X}$ and $\overline{Y}$, respectively. Define $F = X$ and $G = XY$, which have expected values $\overline{X}$ and $\overline{XY}$ respectively. Assume we want to expand $H(F, G) = F/G$ about the point $(\overline{X}, \overline{XY})$ that are non-zero values into a multivariate second order Taylor series. The partial derivatives evaluated at the point $(\overline{X}, \overline{XY})$ are

$$\frac{\partial\ H(\overline{X}, \overline{XY})}{\partial F} = \frac{1}{\overline{XY}}; \qquad \frac{\partial\ H(\overline{X}, \overline{XY})}{\partial G} = -\frac{\overline{X}}{(\overline{XY})^2};$$

$$\frac{\partial^2\ H(\overline{X}, \overline{XY})}{\partial F^2} = 0; \qquad \frac{\partial^2\ H(\overline{X}, \overline{XY})}{\partial G^2} = 2\frac{\overline{X}}{(\overline{XY})^3};$$

$$\frac{\partial^2\ H(\overline{X}, \overline{XY})}{\partial F\ \partial G} = -\frac{1}{(\overline{XY})^2}; \qquad \frac{\partial^2\ H(\overline{X}, \overline{XY})}{\partial G\ \partial F} = -\frac{1}{(\overline{XY})^2}.$$

Thus, a second-order Taylor series expansion for the function $H(F,G)$ expanded about the point $(\overline{X},\overline{XY})$ is

$$
\begin{aligned}
&H(F,G)\\
&\approx H(\overline{X},\overline{XY}) + (F-\overline{X})\tfrac{\partial\ H(\overline{X},\overline{XY})}{\partial F} + (G-\overline{XY})\tfrac{\partial\ H(\overline{X},\overline{XY})}{\partial G}\\
&\quad + \frac{1}{2}\left((F-\overline{X})^2\tfrac{\partial^2\ H(\overline{X},\overline{XY})}{\partial F^2} + (G-\overline{XY})^2\tfrac{\partial^2\ H(\overline{X},\overline{XY})}{\partial G^2} + 2(F-\overline{X})(G-\overline{XY})\tfrac{\partial^2\ H(\overline{X},\overline{XY})}{\partial G\ \partial F}\right)\\
&= \frac{\overline{X}}{\overline{XY}} + (F-\overline{X})\left(\frac{1}{\overline{XY}}\right) + (G-\overline{XY})\left(-\frac{\overline{X}}{(\overline{XY})^2}\right)\\
&\quad + \frac{1}{2}\left((F-\overline{X})^2(0) + (G-\overline{XY})^2\left(2\frac{\overline{X}}{(\overline{XY})^3}\right) + 2(F-\overline{X})(G-\overline{XY})\left(-\frac{1}{(\overline{XY})^2}\right)\right)\\
&= \frac{\overline{X}}{\overline{XY}} + (F-\overline{X})\left(\frac{1}{\overline{XY}}\right) - (G-\overline{XY})\left(\frac{\overline{X}}{(\overline{XY})^2}\right) + (G-\overline{XY})^2\left(\frac{\overline{X}}{(\overline{XY})^3}\right)\\
&\quad - (F-\overline{X})(G-\overline{XY})\left(\frac{1}{(\overline{XY})^2}\right).
\end{aligned}
$$

To construct an optimal design, the expected value of the approximated function expanded about the point $(\overline{X},\overline{XY})$ is required, i.e.

$$
\begin{aligned}
E\{H(F,G)\} &\approx \frac{\overline{X}}{\overline{XY}} + E\{(F-\overline{X})\}\left(\frac{1}{\overline{XY}}\right) - E\{(G-\overline{XY})\}\left(\frac{\overline{X}}{(\overline{XY})^2}\right)\\
&\quad + E\{(G-\overline{XY})^2\}\left(\frac{\overline{X}}{(\overline{XY})^3}\right) - E\{(F-\overline{X})(G-\overline{XY})\}\left(\frac{1}{(\overline{XY})^2}\right)\\
&= \frac{\overline{X}}{\overline{XY}} + E\{(G-\overline{XY})^2\}\left(\frac{\overline{X}}{(\overline{XY})^3}\right) - E\{(F-\overline{X})(G-\overline{XY})\}\left(\frac{1}{(\overline{XY})^2}\right)\\
&= \frac{\overline{X}}{\overline{XY}} + (E\{G^2\} - (\overline{XY})^2)\frac{\overline{X}}{(\overline{XY})^3} - (E\{FG\} - \overline{X}\ \overline{XY})\frac{1}{(\overline{XY})^2}\\
&= E\{G^2\}\frac{\overline{X}}{(\overline{XY})^3} - \frac{E\{FG\}}{(\overline{XY})^2} + \frac{\overline{X}}{\overline{XY}} \;=\; \frac{E\{G^2\}E\{F\}}{E\{G\}^3} - \frac{E\{FG\}}{E\{G\}^2} + \frac{E\{F\}}{E\{G\}}.
\end{aligned}
$$

### A.1.1 Approximating the elements of $\boldsymbol{E}\{[\boldsymbol{M}(\xi,\boldsymbol{\mathcal{M}})]^{-1}\}$ of the general linear model

Considering $F$ and $G$ are the products of independent binomial random variables present in the elements of $\boldsymbol{cov}(\hat{\boldsymbol{\beta}})$, the value of $\boldsymbol{E}\{[\boldsymbol{M}(\xi,\boldsymbol{\mathcal{M}})]^{-1}\}$ can be approximated. For instance, the matrix for a simple linear model contains $E\{Z_i/(Z_iZ_j)\}$ where $i \neq j$; for

a quadratic model, this matrix contains $E\{Z_iZ_j/(Z_iZ_jZ_k)\}$ where no pair of $i, j, k$ is equal. Rewriting $G = FZ$ where $Z$ is the extra independent variable, we have

$$E\{H(F,G)\} \approx \frac{E\{G^2\}E\{F\}}{E\{G\}^3} - \frac{E\{FG\}}{E\{G\}^2} + \frac{E\{F\}}{E\{G\}}$$

$$= \frac{E\{F^2\}E\{Z^2\}E\{F\}}{E\{F\}^3E\{Z\}^3} - \frac{E\{F^2\}E\{Z\}}{E\{F\}^2E\{Z\}^2} + \frac{E\{F\}}{E\{F\}E\{Z\}}$$

$$= \frac{E\{F^2\}E\{F\}\left(E\{Z^2\} - E\{Z\}^2\right)}{E\{F\}^3E\{Z\}^3} + \frac{1}{E\{Z\}}$$

$$= \frac{E\{F^2\}Var(Z)}{E\{F\}^2E\{Z\}^3} + \frac{1}{E\{Z\}}.$$

#### A.1.1.1 Example: Approximation of $E\{Z_i/(Z_iZ_j)\}$ for $i, j = 1, 2,\ i \neq j$

For $Z_i$ is a binomial random variable with mean $nw_i(1 - P(x_i))$ and variance $nw_i(1 - P(x_i))P(x_i)$, let $F = Z_i$ and $G = Z_iZ_j = FZ_j$. Using the above expression where the extra variable is $Z_j$ in this example, we have

$$E\left(\frac{Z_i}{Z_iZ_j}\right) \approx \frac{1}{E\{Z_j\}} + \frac{E\{Z_i^2\}Var(Z_j)}{(E\{Z_i\})^2(E\{Z_j\})^3}$$

$$= \frac{1}{nw_j(1 - P(x_j))} + \frac{(nw_i(1 - P(x_i))P(x_i) + (nw_i(1 - P(x_i)))^2)nw_j(1 - P(x_j))P(x_j)}{(nw_i(1 - P(x_i)))^2(nw_j(1 - P(x_j)))^3}$$

$$= \frac{1}{nw_j(1 - P(x_j))} + \frac{P(x_i)P(x_j)}{nw_i(1 - P(x_i))(nw_j(1 - P(x_j)))^2} + \frac{P(x_j)}{(nw_j(1 - P(x_j)))^2}$$

## A.2 Proof of Theorem 1

The details of this theorem is available in Section 4.2. We can prove that the $D$-optimal design has $p+1$ support points using the general equivalence theorem, by finding a contradiction. Assume $\xi^*$ has $p+2$ support points. Consider

$$g(x) := \frac{\boldsymbol{f}^T(x)\ \boldsymbol{M}^{-1}(\xi^*)\ \boldsymbol{f}(x)}{p+1} \leq \frac{1}{1-P(x)} := h(x)$$

where $g(x)$ is a polynomial of degree $2p$, which has to be less than $h(x)$ over the region $[l,u]\ \forall x$. We order the $p+2$ values for $x$ by size:

$$l \leq x_1 < x_2 < ... < x_{p+2} \leq u \tag{A.1}$$

such that the above equality is achieved. This implies $g(x_i)$ touches $h(x_i)$ and $g^{'}(x_i) = h^{'}(x_i)$ for $i = 2, 3, ..., p+1$. From (A.1), there are values $x_1^{'}, ..., x_{p+1}^{'}$ with $g^{'}(x_i^{'}) = h^{'}(x_i^{'})$ such that $x_1 < x_1^{'} < x_2 < x_2^{'} < x_3 < ... < x_{p+1} < x_{p+1}^{'} < x_{p+2}$ by the Mean Value Theorem.

Hence we have a total of $2p+1$ values where $g$ and $h$ have equal derivatives, and $g^{'}(x)$ is a polynomial of degree $2p-1$. Applying the Mean Value Theorem again to $g^{'}$ and $h^{'}$, there must be $2p$ values where $g^{''}$ and $h^{''}$ are equal. By repeating this process, we find that there must be 2 values where the $2p^{th}$ derivatives $g^{(2p)}$ and $h^{(2p)}$ are equal, and $g^{(2p)}(x)$ is a constant since $g$ is a polynomial of degree $2p$.

This is a contradiction since we assumed that $h^{(2p)}(x) = c$ has at most one solution in the real space for any constant $c$. The same contradiction occurs if we assume $\xi^*$ has more than $p+2$ support points. □

### A.2.1 Cauchy's Mean Value Theorem

If function $g$ and $h$ are both continuous on $[x_1, x_2]$ and differentiable on $(x_1, x_2)$, then there exists $x_1^{'}$ with $x_1 < x_1^{'} < x_2$ such that

$$[g(x_2) - g(x_1)]h^{'}(x_1^{'}) = [h(x_2) - h(x_1)]g^{'}(x_1^{'}).$$

In our situation, since $g(x_2) = h(x_2)$ and $g(x_1) = h(x_1)$, we have $h^{'}(x_1) = g^{'}(x_1)$. This applies to all the other $x_i^{'}$ and the higher order derivatives of $g$ and $h$.

## A.3 Proof of part (a) of Theorem 3

The details of this theorem is available in Section 4.3. We substitute $w_1 = 1 - w_2$ into the respective objective function and take the partial derivative with respect to $w_2$. For the $D$-objective function, we obtain

$$\begin{aligned}
&\frac{\partial}{\partial w_2}\left(E\left(\frac{Z_1}{Z_1Z_2}\right)E\left(\frac{Z_2}{Z_1Z_2}\right)\right)\\
&=\left(\frac{\partial}{\partial w_2}E\left(\frac{Z_1}{Z_1Z_2}\right)\right)E\left(\frac{Z_2}{Z_1Z_2}\right)+E\left(\frac{Z_1}{Z_1Z_2}\right)\left(\frac{\partial}{\partial w_2}E\left(\frac{Z_2}{Z_1Z_2}\right)\right)\\
&\approx\frac{(2\,w_2-1)\left({N'}^4{w_2}^4-2\,{N'}^4{w_2}^3-6\,P^2{N'}^2{w_2}^2-2\,P{N'}^3{w_2}^2\right)}{{N'}^6\,(-1+w_2)^4\,{w_2}^4}\\
&\quad+\frac{(2\,w_2-1)\left({N'}^4{w_2}^2+6\,P^2{N'}^2w_2+2\,P{N'}^3w_2+3\,P^4+3\,P^3N'\right)}{{N'}^6\,(-1+w_2)^4\,{w_2}^4}
\end{aligned}$$

and the optimal weight $w_2 = 1/2$ as $\frac{\partial}{\partial w_2}\left(E\left(\frac{Z_1}{Z_1Z_2}\right)E\left(\frac{Z_2}{Z_1Z_2}\right)\right) = 0$. On the other hand, the derivative of the $c$-objective function with respect to $w_2$ is

$$\begin{aligned}
&\frac{\partial}{\partial w_2}E\left(\frac{Z_1}{Z_1Z_2}\right)+\frac{\partial}{\partial w_2}E\left(\frac{Z_2}{Z_1Z_2}\right)\\
&\approx-\frac{(2\,w_2-1)\left(2\,PN'{w_2}^2-{N'}^2{w_2}^2-2\,PN'w_2+{N'}^2w_2+2\,P^2+2\,PN'\right)}{{N'}^3{w_2}^3\,(-1+w_2)^3},
\end{aligned}$$

which yields that the optimal weight is $w_2 = 1/2$ for this criterion as well. Hence, we have shown that the optimal weight for the experiment with MCAR mechanism is the same as the optimal weight for complete observations.

From expression (4.5), for both optimality criteria, the expressions above do not depend on the support points. Hence the objective functions in (4.8) and (4.9), respectively, are minimised with respect to $x_1$ and $x_2$ when the factor $1/(x_1 - x_2)^2$ is minimised. This is achieved by setting $x_1 = l$ and $x_2 = u$.

Taking partial derivatives in (4.10) with respect to $x_1$ and $x_2$, respectively, shows that regardless of the values of the expression in (4.5) the derivative with respect to $x_1$ is non-negative if $l \geq 0$ or $u \leq 0$. Hence the $A$-objective function is minimised when $x_1 = l$. Similarly, the derivative with respect to $x_2$ is non positive if $l \geq 0$ or $u \leq 0$. Hence the $A$-objective function is minimised when $x_2 = u$. □

## A.4 The covariance matrix for the Alzheimer's disease design example

For model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + N(0, \sigma^2),$$

the covariance matrix of the parameter estimates is

$$\begin{pmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & cov(\hat{\beta}_0, \hat{\beta}_2) \\ cov(\hat{\beta}_0, \hat{\beta}_1) & var(\hat{\beta}_1) & cov(\hat{\beta}_1, \hat{\beta}_2) \\ cov(\hat{\beta}_0, \hat{\beta}_2) & cov(\hat{\beta}_1, \hat{\beta}_2) & var(\hat{\beta}_2) \end{pmatrix} \propto [\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}$$

where

$$var(\hat{\beta}_0) = \frac{Z_2Z_3 + Z_2Z_4 + Z_4Z_3}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$var(\hat{\beta}_1) = \frac{(Z_2 + Z_4)(Z_1 + Z_3)}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$var(\hat{\beta}_2) = \frac{(Z_3 + Z_4)(Z_1 + Z_2)}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{(Z_2 + Z_4) Z_3}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

$$cov(\hat{\beta}_0, \hat{\beta}_2) = -\frac{(Z_3 + Z_4) Z_2}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4},$$

and

$$cov(\hat{\beta}_1, \hat{\beta}_2) = -\frac{Z_4Z_1 - Z_2Z_3}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4}.$$

After some simplification, we have the trace of $[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}$,

$$var(\hat{\beta}_0) + var(\hat{\beta}_1) + var(\hat{\beta}_2) = \frac{Z_1Z_2 + Z_1Z_3 + 2\,Z_1Z_4 + 3\,Z_2Z_3 + 2\,Z_2Z_4 + 2\,Z_3Z_4}{Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4}$$

where $Z_k = \sum_{i \in G_k}(1 - \mathcal{M}_i)$ is the sum of the response indicators in Group $G_k, k = 1, \ldots, 4$. Using Taylor second order linearisation where

$$E\left(\frac{F}{G}\right) \approx \frac{E\{G^2\}E\{F\}}{(E\{G\})^3} - \frac{E\{FG\}}{(E\{G\})^2} + \frac{E\{F\}}{E\{G\}},$$

the $A$-objective function can now be found by taking the expectation with respect to the binomial random variables where

$$E(Z_k) = nw_k(1 - P(x_k^*)) \qquad \text{and} \qquad E(Z_k^2) = nw_k(1 - P(x_k^*))P(x_k^*) + (E(Z_k))^2$$

for $k = 1, 2, 3, 4$. The expressions $F$ and $G$ are given by

$$F = Z_1Z_2 + Z_1Z_3 + 2\,Z_1Z_4 + 3\,Z_2Z_3 + 2\,Z_2Z_4 + 2\,Z_3Z_4$$

and

$$G = Z_1Z_2Z_3 + Z_1Z_2Z_4 + Z_1Z_3Z_4 + Z_2Z_3Z_4.$$

# Appendix B

## B.1 Diagonal elements of between imputation variance-covariance

Recall that

$$\boldsymbol{B}_t = \frac{1}{(t-1)} \begin{pmatrix} \sum_{l=1}^{t} \left(\hat{\beta}_0^{(l)} - \bar{\beta}_0\right)^2 & \sum_{l=1}^{t} \left(\hat{\beta}_0^{(l)} - \bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)} - \bar{\beta}_1\right) \\ \sum_{l=1}^{t} \left(\hat{\beta}_0^{(l)} - \bar{\beta}_0\right)\left(\hat{\beta}_1^{(l)} - \bar{\beta}_1\right) & \sum_{l=1}^{t} \left(\hat{\beta}_1^{(l)} - \bar{\beta}_1\right)^2 \end{pmatrix}.$$

We now illustrate the analytical work for expressing the diagonal elements of $\boldsymbol{B}_t$ in terms of the individual summation of the products of estimates (see Chapter 6.2). We first expand the squares in the summation of the diagonal elements in $\boldsymbol{B}_t$:

$$\left(\hat{\beta}^{(l)} - \frac{\sum_{l=1}^{t} \hat{\beta}^{(l)}}{t}\right)^2 = \hat{\beta}^{2(l)} + \left(\frac{\sum_{l=1}^{t} \hat{\beta}^{(l)}}{t}\right)^2 - 2\hat{\beta}^{(l)} \frac{\sum_{l=1}^{t} \hat{\beta}^{(l)}}{t},$$

where $\hat{\beta}$ with an appropriate subscript corresponds to $\hat{\beta}_0$ or $\hat{\beta}_1$.

Summing this expression from $l = 1, ..., t$ yields

$$
\begin{aligned}
&\sum_{l=1}^{t}\left(\hat{\beta}^{(l)} - \frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t}\right)^2 \\
&= \hat{\beta}^{2(1)} + \left(\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t}\right)^2 - 2\hat{\beta}^{(1)}\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t} + ... + \hat{\beta}^{2(t)} + \left(\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t}\right)^2 - 2\hat{\beta}^{(t)}\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t} \\
&= \sum_{l=1}^{t}\hat{\beta}^{2(l)} + \sum_{l=1}^{t}\left(\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t}\right)^2 - \sum_{l=1}^{t}2\hat{\beta}^{(l)}\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t} \\
&= \sum_{l=1}^{t}\hat{\beta}^{2(l)} + t\left(\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t}\right)^2 - 2\frac{\left(\sum_{l=1}^{t}\hat{\beta}^{(l)}\right)^2}{t} \\
&= \sum_{l=1}^{t}\hat{\beta}^{2(l)} + 2\sum_{l=1,q>l}^{t}\hat{\beta}^{(l)}\hat{\beta}^{(q)} - 2\sum_{l=1,q>l}^{t}\hat{\beta}^{(l)}\hat{\beta}^{(q)} - \frac{\left(\sum_{l=1}^{t}\hat{\beta}^{(l)}\right)^2}{t} \\
&= \sum_{l=1}^{t}\hat{\beta}^{2(l)} - \frac{\left(\sum_{l=1}^{t}\hat{\beta}^{(l)}\right)^2}{t} \qquad \text{(B.1)} \\
&= \sum_{l=1}^{t}\hat{\beta}^{2(l)} - \frac{1}{t}\left(\sum_{l=1}^{t}\hat{\beta}^{2(l)} + 2\sum_{l=1,q>l}^{t}\hat{\beta}^{(l)}\hat{\beta}^{(q)}\right) \\
&= \left(1 - \frac{1}{t}\right)\sum_{l=1}^{t}\hat{\beta}^{2(l)} - \frac{2}{t}\sum_{l=1,q>l}^{t}\hat{\beta}^{(l)}\hat{\beta}^{(q)}
\end{aligned}
$$

where the products of two summations in B.1 can be expanded as follows:

$$
\left(\sum_{l=1}^{t}\hat{\beta}^{(l)}\right)^2 = \begin{pmatrix}\hat{\beta}^{(1)}\\ \hat{\beta}^{(2)}\\ \vdots \\ \hat{\beta}^{(t)}\end{pmatrix}\begin{pmatrix}\hat{\beta}^{(1)} & \hat{\beta}^{(2)} & \dots & \hat{\beta}^{(t)}\end{pmatrix} = \begin{pmatrix}\hat{\beta}^{(1)}\hat{\beta}^{(1)} & \hat{\beta}^{(1)}\hat{\beta}^{(2)} & \dots & \hat{\beta}^{(1)}\hat{\beta}^{(t)} \\ \hat{\beta}^{(2)}\hat{\beta}^{(1)} & \hat{\beta}^{(2)}\hat{\beta}^{(2)} & \dots & \hat{\beta}^{(2)}\hat{\beta}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}^{(t)}\hat{\beta}^{(1)} & \hat{\beta}^{(t)}\hat{\beta}^{(2)} & \dots & \hat{\beta}^{(t)}\hat{\beta}^{(t)}\end{pmatrix}
$$

$$
= \sum_{l=1}^{t}\left(\hat{\beta}^{(l)}\right)^2 + 2\underbrace{\sum_{l=1,q>l}^{t}\hat{\beta}^{(l)}\hat{\beta}^{(q)}}_{\frac{1}{2}t(t-1)\ terms}
$$

where $l$ and $q$ denote the different imputed sets. Dividing through $t - 1$, we have

$$
\frac{1}{t-1}\sum_{l=1}^{t}\left(\hat{\beta}^{(l)} - \bar{\beta}\right)^2 = \frac{1}{t}\sum_{l=1}^{t}\left(\hat{\beta}^{(l)}\right)^2 - \frac{2}{t(t-1)}\sum_{l=1,q>l}^{t}\hat{\beta}^{(l)}\hat{\beta}^{(q)}.
$$

## B.2 Expectation of diagonal elements of between imputation variance-covariance

Following the above illustrated expression, the expected value of the diagonal elements of $\mathcal{B}_t$ has the following form:

$$
\begin{aligned}
& E\left(\frac{1}{t-1}\left(\hat{\beta}^{(l)}-\frac{\sum_{l=1}^{t}\hat{\beta}^{(l)}}{t}\right)^2\right) \\
&= \frac{1}{t}\sum_{l=1}^{t} E\left(\hat{\beta}^{(l)}\hat{\beta}^{(l)}\right) - \frac{2}{t(t-1)}\sum_{l=1,q>l}^{t} E\left(\hat{\beta}^{(l)}\hat{\beta}^{(q)}\right) \\
&= \frac{1}{t}t\ E\left(\hat{\beta}^{(*)}\hat{\beta}^{(*)}\right) - \frac{2}{t(t-1)}\ \frac{t\ (t-1)}{2} E\left(\hat{\beta}^{(*)}\hat{\beta}^{(**)}\right) \\
&= E\left(\hat{\beta}^{(*)}\hat{\beta}^{(*)}\right) - E\left(\hat{\beta}^{(*)}\hat{\beta}^{(**)}\right),
\end{aligned}
$$

where $\hat{\beta}$ with an appropriate subscript corresponds to $\hat{\beta}_0$ or $\hat{\beta}_1$, and superscript $(*)$ and $(**)$ correspond to two different imputation. In the third line, the summation becomes a multiplier as the individual expected value obtained in the iterative expectation is the same for all repeated imputation.

# References


Afifi, A. and Elashoff, R. (1966). Missing observations in multivariate statistics i. review of the literature. *Journal of the American Statistical Association*, 61(315):595–604.

Ahmad, T. and Gilmour, S. G. (2010). Robustness of subset response surface designs to missing observations. *Journal of Statistical Planning and Inference*, 140(1):92–103.

Akhtar, M. and Prescott, P. (1986). Response surface designs robust to missing observations. *Communications in Statistics-Simulation and Computation*, 15(2):345–363.

Alshurafa, M., Briel, M., Akl, E. A., Haines, T., Moayyedi, P., Gentles, S. J., Rios, L., Tran, C., Bhatnagar, N., Lamontagne, F., et al. (2012). Inconsistent definitions for intention-to-treat in relation to missing outcome data: systematic review of the methods literature. *PloS one*, 7(11):e49163.

Altman, D. G. (2009). Missing outcomes in randomized trials: addressing the dilemma. *Open Medicine*, 3(2):e51.

Andrews, D. F. and Herzberg, A. M. (1979). The robustness and optimality of response surface designs. *Journal of Statistical Planning and Inference*, 3(3):249–257.

Antille, G., Dette, H., and Weinberg, A. (2003). A note on optimal designs in weighted polynomial regression for the classical efficiency functions. *Journal of Statistical Planning and Inference*, 113(1):285–292.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum experimental designs, with SAS*, volume 34. Oxford University Press Oxford.

Baek, I., Zhu, W., Wu, X., and Wong, W. K. (2006). Bayesian optimal designs for a quantal dose-response study with potentially missing observations. *Journal of biopharmaceutical statistics*, 16(5):679–693.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Berger, M. P. and Wong, W.-K. (2009). *An introduction to optimal designs for social and biomedical research*, volume 83. John Wiley & Sons.

Bickel, P. and Herzberg, A. M. (1979). Robustness of design against autocorrelation in time i: Asymptotic theory, optimality for location and linear regression. *The Annals of Statistics*, pages 77–95.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4):651–675.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.

Chang, F. and Jiang, B. (2007). An algebraic construction of minimally-supported d-optimal designs for weighted polynomial regression. *Statistica Sinica*, 17(3):1005.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, pages 586–602.

Coley, N., Gardette, V., Cantet, C., Gillette-Guyonnet, S., Nourhashemi, F., Vellas, B., and Andrieu, S. (2011). How should we deal with missing data in clinical trials involving alzheimer's disease patients? *Current Alzheimer Research*, 8(4):421–433.

Cramer, H. (1946). Mathematical methods of statistics. 1946. *Department of Mathematical SU*.

De la Garza, A. (1954). Spacing of information in polynomial regression. *The Annals of Mathematical Statistics*, 25(1):123–130.

Dette, H., Kunert, J., and Pepelyshev, A. (2008). Exact optimal designs for weighted least squares analysis with correlated errors. *Statistica Sinica*, 18(1):135–154.

Dette, H., Pepelyshev, A., and Zhigljavsky, A. (2014). 'nearly'universally optimal designs for models with correlated observations. *Computational Statistics & Data Analysis*, 71:1103–1112.

Dette, H., Pepelyshev, A., Zhigljavsky, A., et al. (2013). Optimal design for linear models with correlated observations. *The Annals of Statistics*, 41(1):143–176.

Dette, H. and Trampisch, M. (2010). A general approach to d-optimal designs for weighted univariate polynomial regression models. *Journal of the Korean Statistical Society*, 39(1):1–26.

Elfving, G. (1952). Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2):255–262.

Fedorov, V. and Nachtsheim, C. (1995). Optimal designs for time-dependent responses. In *MODA4Advances in Model-Oriented Data Analysis*, pages 3–13. Springer.

Fedorov, V. V. (1972). *Theory of optimal experiments*. Elsevier.

Fedorov, V. V., Gagnon, R. C., and Leonov, S. L. (2002). Design of experiments with unknown parameters in variance. *Applied Stochastic Models in Business and Industry*, 18(3):207–218.

Fedorov, V. V. and Hackl, P. (1997). *Model-Oriented Design of Experiments*. Springer-Verlag New York, 1 edition.

Gardette, V., Coley, N., Toulza, O., and Andrieu, S. (2006). Attrition in geriatric research: how important is it and how should it be dealt with? *The journal of nutrition, health & aging*, 11(3):265–271.

Ghosh, S. (1979). On robustness of designs against incomplete data. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 204–208.

Ghosh, S. (1980). On robustness of optimal balanced resolution v plans. *Journal of Statistical Planning and Inference*, 4(3):313–319.

Ghosh, S. (1982). Robustness of bibd against the unavailability of data. *Journal of Statistical Planning and Inference*, 6(1):29–32.

Gonzalez, J. M. and Eltinge, J. L. (2007). Multiple matrix sampling: A review. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 3069–75.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in clinical research*, 2(3):109.

Hackl, P. (1995). *Optimal design for experiments with potentially failing trials*. Springer.

Harman, R. and Štulajter, F. (2010). Optimal prediction designs in finite discrete spectrum linear regression models. *Metrika*, 72(2):281–294.

Hedayat, A., John, P., et al. (1974). Resistant and susceptible bib designs. *The Annals of Statistics*, 2(1):148–158.

Herzberg, A. M. and Andrews, D. F. (1976). Some considerations in the optimal design of experiments in non-optimal situations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 284–289.

Herzberg, A. M., Prescott, P., and Akhtar, M. (1987). Equi-information robust designs: Which designs are possible? *Canadian Journal of Statistics*, 15(1):71–76.

Hoel, P. G. (1958). Efficiency problems in polynomial estimation. *The Annals of Mathematical Statistics*, pages 1134–1145.

Howard, R., McShane, R., Lindesay, J., Ritchie, C., Baldwin, A., Barber, R., Burns, A., Dening, T., Findlay, D., Holmes, C., et al. (2012). Donepezil and memantine for moderate-to-severe alzheimer's disease. *New England Journal of Medicine*, 366(10):893–903.

Hu, J., Zhu, W., Su, Y., and Wong, W. K. (2010). Controlled optimal design program for the logit dose response model. *Journal of Statistical Software*, 35(i06).

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):173–190.

Imhof, L. A., Song, D., and Wong, W. K. (2002). Optimal design of experiments with possibly failing trials. *Statistica Sinica*, 12(4):1145–1156.

Imhof, L. A., Song, D., and Wong, W. K. (2004). Optimal design of experiments with anticipated pattern of missing observations. *Journal of theoretical biology*, 228(2):251–260.

John, P. W. (1976). Robustness of balanced incomplete block designs. *The Annals of Statistics*, pages 960–962.

John, R. S. and Draper, N. R. (1975). D-optimality for regression designs: a review. *Technometrics*, 17(1):15–23.

Kageyama, S. (1980). Robustness of connected balanced block designs. *Annals of the Institute of Statistical Mathematics*, 32(1):255–261.

Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16(3):199–218.

Kiefer, J. (1958). On the nonrandomized optimality and randomized nonoptimality of symmetrical designs. *The Annals of Mathematical Statistics*, pages 675–699.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 272–319.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.*, 2(5):849–879.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12(363-366):234.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, T. D. and Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, 7(4):199–204.

Low, J., Lewis, S., and Prescott, P. (1999). Assessing robustness of crossover designsto subjects dropping out. *Statistics and Computing*, 9(3):219–227.

Mackinnon, A. (2008). Statistical treatment of withdrawal in trials of anti-dementia drugs. *The Lancet*, 372(9647):1382–1383.

Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in medicine*, 30(6):627–641.

Mitra, R. and Reiter, J. P. (2012). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical methods in medical research*, page 0962280212445945.

Most, B. M. (1975). Resistance of balanced incomplete block designs. *The Annals of Statistics*, pages 1149–1162.

Müller, C. H. (1995). Breakdown points for designed experiments. *Journal of statistical planning and inference*, 45(3):413–427.

Näther, W. (1985). Exact designs for regression models with correlated errors. *Statistics*, 16(4):479–484.

Ortega-Azurduy, S., Tan, F., and Berger, M. (2008). The effect of dropout on the efficiency of d-optimal designs of linear mixed models. *Statistics in medicine*, 27(14):2601–2617.

Ouwens, M. J., Tan, P. E., and Berger, M. P. (2002). Maximin *d*-optimal designs for longitudinal mixed effects models. *Biometrics*, 58(4):735–741.

Pukelsheim, F. and Rieder, S. (1992). Efficient rounding of approximate designs. *Biometrika*, 79(4):763–770.

Raghunathan, T. E. and Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429):54–63.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. new york: J.

Sacks, J. and Ylvisaker, D. (1966). Designs for regression problems with correlated errors. *The Annals of Mathematical Statistics*, pages 66–89.

Sacks, J. and Ylvisaker, D. (1968). Designs for regression problems with correlated errors: many parameters. *The Annals of Mathematical Statistics*, pages 49–69.

Salim, A., Mackinnon, A., Christensen, H., and Griffiths, K. (2008). Comparison of data analysis strategies for intent-to-treat analysis in pre-test–post-test designs with substantial dropout rates. *Psychiatry research*, 160(3):335–345.

Schmelter, T. (2007a). Considerations on group-wise identical designs for linear mixed models. *Journal of Statistical Planning and Inference*, 137(12):4003–4010.

Schmelter, T. (2007b). The optimality of single-group designs for certain mixed models. *Metrika*, 65(2):183–193.

Silvey, S. D. (1980). *Optimal design*, volume 7. Springer.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, pages 1–85.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338:b2393.

Tan, F. E. and Berger, M. P. (1999). Optimal allocation of time points for the random effects model. *Communications in Statistics-Simulation and Computation*, 28(2):517–540.

Tekle, F. B., Tan, F. E., and Berger, M. P. (2008). D-optimal cohort designs for linear mixed-effects models. *Statistics in medicine*, 27(14):2586–2600.

Tekle, F. B., Tan, F. E., and Berger, M. P. (2011). Too many cohorts and repeated measurements are a waste of resources. *Journal of clinical epidemiology*, 64(12):1383–1390.

Thomas, R. G., Berg, J. D., Sano, M., and Thal, L. (2000). Analysis of longitudinal data in an alzheimer's disease clinical trial. *Statistics in medicine*, 19(11-12):1433–1440.

Wald, A. (1943). On the efficient design of statistical investigations. *The annals of mathematical statistics*, 14(2):134–140.

White, I. R., Horton, N. J., Carpenter, J., Pocock, S. J., et al. (2011). Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ*, 342.

Whittle, P. (1973). Some general points in the theory of optimal experimental design. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 123–130.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, 1(2):129–142.

Zhigljavsky, A., Dette, H., and Pepelyshev, A. (2010). A new approach to optimal design for linear models with correlated observations. *Journal of the American Statistical Association*, 105(491).