OPTIMAL DESIGN WHEN OUTCOME VALUES ARE NOT MISSING AT RANDOM

Kim May Lee, Robin Mitra and Stefanie Biedermann

University of Southampton, UK

Abstract: The presence of missing values complicates statistical analyses. In design of experiments, missing values are particularly problematic when constructing optimal designs, as it is not known which values are missing at the design stage. When data are missing at random it is possible to incorporate this information into the optimality criterion that is used to find designs; Imhof, Song and Wong (2002) develop such a framework. However, when data are not missing at random this framework can lead to inefficient designs. We investigate and address the specific challenges that not missing at random values present when finding optimal designs for linear regression models. We show that the optimality criteria will depend on model parameters that traditionally do not affect the design, such as regression coefficients and the residual variance. We also develop a framework that improves efficiency of designs over those found assuming values are missing at random.

Key words and phrases: Covariance matrix, information matrix, linear regression model, missing observations, not missing at random, optimal design.

1. Introduction

Missing values are a common problem in many fields. Their presence complicates statistical analysis, and appropriate methods are required to handle the missing data to ensure valid inferences. There is a wide variety of techniques present to handle missing values once the data are observed, but the objective in this paper is to focus on handling the missing data problem at the design stage of an experiment. By incorporating information about the missing data mechanism we may be able to design a more efficient experiment that allows more information to be obtained from the resulting data collected.

There has been work on finding optimal designs for experiments with potentially missing values in the literature. The majority of the contributions is concerned with robustness of designs to missing values; see for example Hedayat and John (1974), Ghosh (1979), Ortega-Azurduy, Tan and Berger (2008)

or Ahmad and Gilmour (2010). Herzberg and Andrews (1976) and Hackl (1995) introduce design criteria that account for the presence of missing responses for some special cases. Imhof, Song and Wong (2002) develop a framework that finds optimal designs by taking the expectation of the information matrix with respect to the missing data mechanism, which has been extended by Lee, Biedermann and Mitra (2017) to improve the approximation of the covariance matrix.

All these contributions in optimal design implicitly assume that the data are missing at random. This is where it is assumed that the process that generates the missing values, the missing data mechanism, depends on only observed variables. This is referred to as a missing at random (MAR) mechanism, defined by Rubin (1976). If on the other hand it is assumed that the missing data mechanism depends on unobserved variables, such as the missing values themselves, Rubin (1976) referred to this as a not missing at random (NMAR) mechanism. Typically NMAR problems are much more challenging to handle, as this is an untestable assumption. Learning about the exact form of the NMAR mechanism is not typically possible, and thus this often leads to biased inferences.

To our knowledge there has not been any explicit consideration of dealing with NMAR when finding optimal designs. We are thus opening up a whole new area of research. This article intends to explore this area, and to address the specific problems that NMAR causes in optimal design. We also propose a framework motivated by that of Imhof, Song and Wong (2002), but extend this to incorporate the possibility of NMAR using an approximation to the bias. By doing so we can mitigate the problems caused by NMAR and find more efficient designs.

We assume that analysts will make inferences using a linear regression model once the experiment has been performed, and will deal with the missing data using the complete cases. Complete case analysis is a widely used strategy, where any unit with missing data is discarded from the analysis. In the context of regression analysis, complete case analysis can be appropriate when the missing mechanism is MAR (Little, 1992). Under NMAR there are obvious problems that can occur and these will be noted and mitigated through our proposed optimal design framework.

The remainder of the article is organised as follows. Section 2 presents some

background behind the key elements of missing data and optimal design. Section 3 motivates the problems NMAR causes in optimal design. Section 4 presents an optimal design framework that takes NMAR into account and compares how it relates to the traditional MAR framework. Section 5 empirically evaluates the proposed framework to determine the benefits of using this approach. Section 6 evaluates our methodology in a real data scenario. Finally Section 7 ends with some concluding remarks.

2. Background

We first review the relevant background to dealing with missing data. We then present the key concepts in constructing optimal designs when a linear regression analysis model is used. Finally we review how the potential for missing data can be taken into account when finding optimal designs.

2.1 Missing data

Let $\mathbf{x}_i,\ i = 1, \ldots, n$, represent a set of explanatory variables for unit i in the experiment, and let y_i be the outcome for unit i once the experiment is performed. We assume that the analysts will make inferences by fitting a linear regression model to the data of the form,

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{2.1}$$

where $\mathbf{y} = (y_1, \ldots . y_n)$, $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients and $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$ is the error vector with residual variance σ^2. We also define a missing indicator, m_i, for each unit i, where $m_i = 1$ corresponds to y_i missing and $m_i = 0$ corresponds to y_i observed. We can then denote $y_{mis} = \{y_i : m_i = 1\}$ and $y_{obs} = \{y_i : m_i = 0\}$ as the missing and observed outcomes respectively. We assume the analyst is interesting in making inferences about the regression parameters, $\boldsymbol{\beta}$. Typically, inference regarding the parameters should be made using the joint likelihood for (y_i, m_i). One way this can be expressed is as

$$p(m_i|\mathbf{x}_i, y_i, \boldsymbol{\gamma})p(y_i|\mathbf{x}_i, \boldsymbol{\beta}) \tag{2.2}$$

which is known as the selection model framework (Little and Rubin, 2002) where the vector $\boldsymbol{\gamma}$ represents parameters characterising the model for m_i, also known as the missing data mechanism. We implicitly assume in this model that the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are distinct. It is commonly assumed that missing values

arise under MAR, which implies that $p(m_i|\mathbf{x}_i, y_i, \boldsymbol{\gamma}) = p(m_i|\mathbf{x}_i, \boldsymbol{\gamma})$. Under MAR we can see that (2.2) factorises so that inferences concerning $\boldsymbol{\beta}$ can be made using only $p(y_i|x_i, \boldsymbol{\beta})$. In this paper, we assume the analyst will base inferences on the complete cases, i.e. subset on those units where $m_i = 0$. Under MAR, estimates for $\boldsymbol{\beta}$ will be unbiased in this situation (Little, 1992). In this paper, we assume that the missing mechanism can be modelled using a logit link function. Specifically, under MAR,

$$p(m_i = 1|\mathbf{x}_i, \boldsymbol{\gamma}) = \frac{exp(\mathbf{x}_i'\boldsymbol{\gamma})}{1 + exp(\mathbf{x}_i'\boldsymbol{\gamma})}. \tag{2.3}$$

We denote the expression in (2.3) by $P(\mathbf{x}_i)$ for short, indicating that it is explicitly dependent on values of $\mathbf{x}_i$. A corresponding NMAR mechanism, which incorporates the (potentially missing) values of the response variable and includes (2.3) as a special case, is proposed in Section 3.

If the MAR assumption is a not a reasonable assumption, so the missing mechanism is NMAR, then estimates for $\boldsymbol{\beta}$ based only on $p(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ will be biased (including those obtained using a complete case analysis). The presence of NMAR is an untestable assumption, and if it exists there is currently little that can be done to adjust for this, beyond assessing sensitivity of the results to different NMAR mechanisms (Little and Rubin, 2002). The problem arises in that it is not possible to fully determine $p(m_i|\mathbf{x}_i, y_i, \boldsymbol{\gamma})$ as it depends implicitly on the missing outcome values, y_{mis}. In Section 4 we propose a strategy that mitigates the effect NMAR has in finding designs and estimating regression coefficients.

2.2 Optimal design

In experimental design the goal is to choose values of $\mathbf{x}_i$ for each unit i that optimise a relevant criterion to obtain maximum information from the experiment. Typically, the optimality criterion minimises a function of the covariance matrix of the estimators. We assume that the regression coefficients, $\boldsymbol{\beta}$, will be estimated using maximum likelihood, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}$, with covariance matrix

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

We consider approximate designs, i.e. designs of the form

$$\xi = \begin{Bmatrix} \mathbf{x}_1^* & \cdots & \mathbf{x}_m^* \\ w_1 & \cdots & w_m \end{Bmatrix}, \quad 0 < w_i \leq 1, \quad \sum_{i=1}^{m} w_i = 1,$$

where $\mathbf{x}_1^*, \ldots, \mathbf{x}_m^*$ $(m \leq n)$ represent the distinct values of the explanatory variables and are referred to as the support points of the design, and the weights $w_1, \ldots, w_m$ represent the relative proportions of observations taken at the corresponding support points $\mathbf{x}_i^*$, $i = 1, \ldots, m$.

This approach is independent of sample size and avoids the problem of discrete optimisation and is thus widely used in finding optimal designs for experiments. Since nw_i, $i = 1, \ldots, m$, are not necessarily integer valued, in order to run such a design in practice, a rounding procedure can be applied; see, for example, Pukelsheim and Rieder (1992).

For an approximate design ξ, the Fisher information matrix for model (2.1) is

$$\boldsymbol{M}(\xi) = \sum_{i=1}^{m} \boldsymbol{f}(\mathbf{x}_i^*)\boldsymbol{f}^T(\mathbf{x}_i^*)\ w_i$$

where the vector $\boldsymbol{f}^T(\mathbf{x}_i^*)$ is a row in the design matrix $\boldsymbol{X}$ corresponding to $\mathbf{x}_i^*$, and its inverse, $\boldsymbol{M}^{-1}(\xi)$, is proportional to $\boldsymbol{var}\ (\hat{\boldsymbol{\beta}})$.

We consider the following optimality criteria, which are commonly used in the literature when the aim of the experiment is to estimate the parameter vector $\boldsymbol{\beta}$ with high precision.

- D-optimality: Minimise $|\boldsymbol{M}^{-1}(\xi)|$. A D-optimal design minimises the volume of a confidence ellipsoid for $\boldsymbol{\beta}$.

- A-optimality: Minimise trace$(\boldsymbol{M}^{-1}(\xi))$. An A-optimal design minimises the sum of the variances of the individual elements of $\hat{\boldsymbol{\beta}}$.

2.3 Optimal design for missing values

Now when certain values y_i may be missing we can take account of this through the missing data mechanism. Assuming MAR we denote $p(m_i = 1|\mathbf{x}_i, \boldsymbol{\gamma}) = P(\mathbf{x}_i)$. The Fisher information matrix containing the missing data indicators $\boldsymbol{\mathcal{M}} = \{m_1, m_2, \ldots, m_n\}$ is given by $\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})$ and we can take the expectation over these,

$$\begin{aligned} \boldsymbol{E}[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})] &= \boldsymbol{E}[\sum_{i=1}^{n} \boldsymbol{f}(\mathbf{x}_i)\boldsymbol{f}^T(\mathbf{x}_i)\ (1 - m_i)] \\ &= \sum_{i=1}^{n} \boldsymbol{f}(\mathbf{x}_i)\boldsymbol{f}^T(\mathbf{x}_i)[1 - P(\mathbf{x}_i)] \\ &= n\sum_{i=1}^{m} \boldsymbol{f}(\mathbf{x}_i^*)\boldsymbol{f}^T(\mathbf{x}_i^*)\ w_i[1 - P(\mathbf{x}_i^*)] \end{aligned} \tag{2.4}$$

which is equivalent to $\boldsymbol{M}(\xi)$ if the responses are fully observed. Imhof, Song and Wong (2002) proposed a general framework where $\boldsymbol{M}(\xi)$ is replaced by (2.4) in the respective optimality criterion. However this assumes that $E\{[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}\}$ is proportional to $E[\boldsymbol{var}\ (\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})]$ which may result in a rather crude approximation to the covariance matrix, in particular for small to moderate sample sizes. Lee, Biedermann and Mitra (2017) develop an improved approximation by considering the expectation of a 2nd order Taylor expansion of $[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}$ which also results in better designs. For large sample sizes, however, the two approaches will generate very similar designs.

Nevertheless both approaches are implicitly based on assuming MAR. If the potential for NMAR exists then this framework may lead to inefficient designs, with biased estimates, unless this is taken account of, regardless of which approach is used. In the next Section we determine what effect NMAR might have on the performance of designs found assuming MAR holds. We then consider how to best address the problem of NMAR in Section 4. In Sections 5 and 6 we present results that incorporate our findings from Section 4 to find designs and evaluate performance, which we use in conjunction with the Lee, Biedermann and Mitra (2017) approach. We also considered equivalent results that use the Imhof, Song and Wong (2002) approach, but the results were of a similar profile, due to the relatively large sample sizes considered, and so we do not present the results for brevity.

3. Effect of NMAR on optimal designs

If we have NMAR when constructing optimal designs then our missing data mechanism implicitly depends on the outcome variable. We consider one such

situation and modify the missing data mechanism in (2.3) to become,

$$p(m_i = 1|\mathbf{x}_i, y_i, \boldsymbol{\gamma}) = \frac{exp(\mathbf{x}_i'\gamma + \delta y_i)}{1 + exp(\mathbf{x}_i'\gamma + \delta y_i)} \tag{3.1}$$

for $i = 1, \ldots, n$.

We now illustrate what effect, if any, NMAR might have in the construction of optimal designs and their resulting performance. We focus on the simple linear regression model for the design region $\mathfrak{X} = [0, u]$ for some value $0 < u < \infty$. As such we have an analysis model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \tag{3.2}$$

for $i = 1, \ldots, n$. Without loss of generality we assume $\delta = 1$ which gives us a missing data mechanism

$$p(m_i = 1|x_i, y_i, \gamma_0, \gamma_1) = \frac{exp(\gamma_0 + \gamma_1 x_i + y_i)}{1 + exp(\gamma_0 + \gamma_1 x_i + y_i)}. \tag{3.3}$$

From the above two equations, it is immediately obvious that our design will depend on the regression coefficients β_0 and β_1. This can be seen by re-expressing (3.3) as

$$p(m_i = 1|x_i, y_i, \gamma_0, \gamma_1) = \frac{exp(\gamma_0^* + \gamma_1^* x_i + \epsilon_i)}{1 + exp(\gamma_0^* + \gamma_1^* x_i + \epsilon_i)} \tag{3.4}$$

where $\gamma_0^* = \gamma_0 + \beta_0$ and $\gamma_1^* = \gamma_1 + \beta_1$. For different values of β_0 and β_1 we can see γ_0^* and γ_1^* will change and hence our design will change. We assume that designs are constructed under some known fixed values of β_0 and β_1. In practice knowing these values is unrealistic, and finding their values is the goal of the experiment in the first place. However, it may be possible to assume that the analyst has some prior information about likely values of β_0 and β_1 that can be used. The resultant designs will thus be locally optimal. It is important to note that traditionally for linear models the optimal design is not sensitive to the specific values of the regression coefficient, even when missingness is MAR, so this is a specific complication that arises due to NMAR here.

It is also of interest to consider the residual variance σ^2. It is not immediately obvious what effect, if any, σ^2 has on the efficiency of the design. As ϵ_i has zero mean, it may be the case that this term does not influence the design, but from a practical point of view the larger the value of σ^2 the greater the uncertainty

about the expected amount of missing data at any given point x_i within the design region $\mathfrak{X}$, and hence it is logical that this might influence what design we choose.

Let $u = 2$, so the design space is $\mathfrak{X} = [0, 2]$ in what follows. We first find the optimal two-point designs, under D- and A- optimality, assuming $(\gamma_0, \gamma_1, \beta_0, \beta_1)$ are known and equal to $(-5.572, 2.191, 1, 1)$ respectively, we also assume $\sigma^2 = 0$ when finding the designs. This is equivalent to setting $\epsilon_i = 0$ in (3.4) and assumes a MAR mechanism with parameters $(\gamma_0^*, \gamma_1^*) = (-4.572, 3.191)$. With these values we find the probability of missing at the end points of the design space, 0 and 2, are 0.01 and 0.859 respectively and is monotone increasing over this space. Thus the potential for missing data is not too extreme at any point in the design space, while still allowing the potential for missing data to have an impact on the performance of any given design. When the missing mechanism is monotone increasing, Lee, Biedermann and Mitra (2017) show that the lower bound of the design space will always be one of the support points in an optimal design. Thus in a two-point design it suffices to find the second support point, x_2^* and its weight w_2, as $w_1 = 1 - w_2$. These optimisation problems are subject to $w_1 + w_2 = 1$. Using the *fmincon* function in *Matlab*, we find an optimal design of $\{x_1^*, x_2^*; w_1, w_2\} = \{0, 1.3766; 0.5, 0.5\}$ under the D-optimality criterion and $\{x_1^*, x_2^*; w_1, w_2\} = \{0, 1.5147; 0.546, 0.454\}$ under the A-optimality criterion.

For each optimal design, we then simulate $n = 60$ (where $n_1 = nw_1$ and $n_2 = nw_2$ with integer rounding if necessary) observations from (3.2) using the support points, the values of β_0, β_1 above and under different σ^2. We make some outcome values missing using (3.3) with the values of γ_0, γ_1 above, as well as the simulated y_i values. Estimates of β_0, β_1 are obtained using the complete case data. This process is repeated 100,000 times to obtain measures of bias, and mean squared error for β_0 and β_1 respectively. We also present the determinant and trace of the variance-covariance and mean squared error matrix which correspond to the objective functions we are seeking to minimise under D- and A- optimality respectively.

Table 3.1 presents the performance of the two respective optimal designs under different missing data mechanisms and different values of σ^2. The outputs under NMAR correspond to the situations where ϵ_i in (3.4) has the corresponding

Table 3.1: Simulation outputs of A- and D-optimal designs across 100,000 simulated data sets under different missing data mechanisms.

	under NMAR			under MAR		
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
A-optimal design, $\{x_1^*, x_2^*; n_1, n_2\} = \{0, 1.5147; 33, 27\}$						
bias of $\hat{\beta}_0$	-0.00303	-0.0163	-0.0555	-7.82 $\times 10^{-5}$	-2.34 $\times 10^{-4}$	-3.91 $\times 10^{-4}$
bias of $\hat{\beta}_1$	-0.0855	-0.292	-0.538	2.56 $\times 10^{-4}$	7.69 $\times 10^{-4}$	0.00128
mse of $\hat{\beta}_0$	0.00765	0.0306	0.0701	0.00191	0.0172	0.0478
mse of $\hat{\beta}_1$	0.0198	0.130	0.379	0.00327	0.0294	0.0817
tr(mse)	0.0275	0.161	0.449	0.00518	0.0466	0.130
|mse|	1.29 $\times 10^{-4}$	0.00374	0.0263	4.65 $\times 10^{-6}$	3.77 $\times 10^{-4}$	0.00291
$var(\hat{\beta}_0)$	0.00764	0.0303	0.0670	0.00191	0.0172	0.0478
$var(\hat{\beta}_1)$	0.0125	0.0447	0.0891	0.00327	0.0294	0.0817
$tr(\boldsymbol{var}(\hat{\boldsymbol{\beta}}))$	0.0201	0.0750	0.156	0.00518	0.0466	0.130
$\lvert\boldsymbol{var}(\hat{\boldsymbol{\beta}})\rvert$	7.01 $\times 10^{-5}$	9.53 $\times 10^{-4}$	0.00401	4.65 $\times 10^{-6}$	3.77 $\times 10^{-4}$	0.00291
D-optimal design, $\{x_1^*, x_2^*; n_1, n_2\} = \{0, 1.3766; 30, 30\}$						
bias of $\hat{\beta}_0$	-0.00312	-0.0165	-0.0559	-1.26 $\times 10^{-4}$	-3.79 $\times 10^{-4}$	-6.31 $\times 10^{-4}$
bias of $\hat{\beta}_1$	-0.0761	-0.266	-0.501	2.43 $\times 10^{-4}$	7.29 $\times 10^{-4}$	0.00121
mse of $\hat{\beta}_0$	0.00840	0.0335	0.0766	0.00210	0.0189	0.0525
mse of $\hat{\beta}_1$	0.0180	0.116	0.345	0.00317	0.0285	0.0793
tr(mse)	0.0264	0.150	0.422	0.00527	0.0475	0.132
|mse|	1.17 $\times 10^{-4}$	0.00351	0.0258	4.37 $\times 10^{-6}$	3.54 $\times 10^{-4}$	0.00273
$var(\hat{\beta}_0)$	0.00839	0.0333	0.0735	0.00210	0.0189	0.0525
$var(\hat{\beta}_1)$	0.0123	0.0456	0.0945	0.00317	0.0285	0.0793
$tr(\boldsymbol{var}(\hat{\boldsymbol{\beta}}))$	0.0206	0.0789	0.168	0.00527	0.0475	0.132
$\lvert\boldsymbol{var}(\hat{\boldsymbol{\beta}})\rvert$	6.59 $\times 10^{-5}$	9.36 $\times 10^{-4}$	0.00410	4.37 $\times 10^{-6}$	3.54 $\times 10^{-4}$	0.00273

σ^2 whereas those under MAR correspond to the situations where ϵ_i in (3.4) has $\sigma^2 = 0$. In all cases, the responses are simulated with the corresponding values of σ^2. We see that the bias and the mean squared error increase as σ^2 increases in the simulation. Comparing the two different scenarios for the same σ^2, we find that the estimates obtained in the presence of a NMAR mechanism have more bias and larger mean squared errors than those obtained in the presence of the MAR mechanism. We also find a similar profile for the determinant and trace of the covariance and the mean squared error matrix.

Focussing on the bias and the mean squared error of the estimates in the

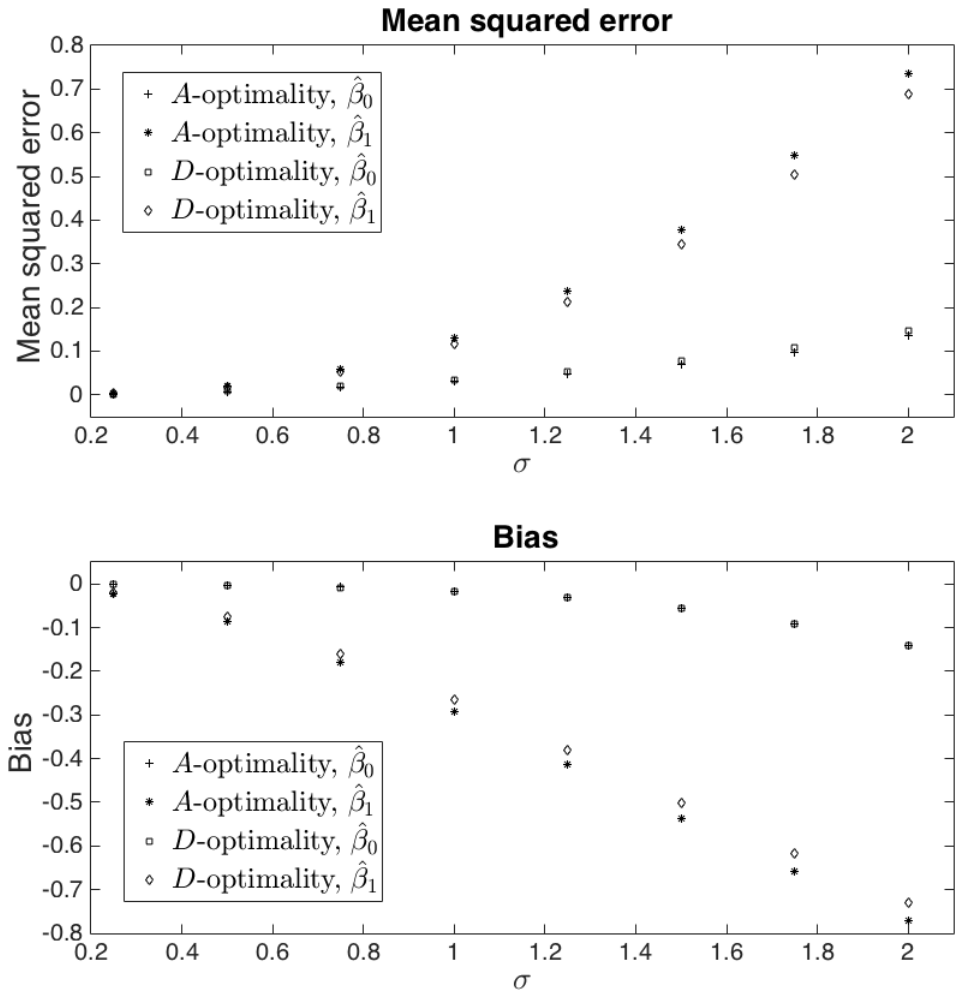

Figure 3.1: Mean squared error and bias of the estimates that are computed using the A- and the D-optimal design in the presence of NMAR mechanisms.

presence of a NMAR mechanism, in Figure 3.1 we plot how this varies with different values of σ^2 under the D- and A- optimal designs found above. We see that the mean squared error of each estimate increases with the values of σ^2 and the estimates are biased downward when σ^2 is large. These results show that σ^2 plays a role in affecting the performance of any design under NMAR. Again, this is a parameter that traditionally does not affect the optimal design found in linear regression when missingness is MAR. In the next section we investigate how we can take account of the effect of σ^2 in constructing optimal designs.

4. Optimal design under NMAR

In this section we first provide intuition behind why new theory needs to be developed in constructing optimal designs when NMAR is present. We then present details concerning our investigation into approximating the missing indicator probability, and finally we consider broadening the framework to include bias into the optimality criterion.

4.1 Incorporating NMAR into the design framework

Recall from Section 2.3 when missing data are present the optimality criterion would seek to minimise a function of $E\{[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}\}$ as this can be viewed as a surrogate for minimising the corresponding function of $E[\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})]$. Evaluating this expectation is not straightforward however and must be approximated. Imhof, Song and Wong (2002) approximate this by $\{E[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]\}^{-1}$ while Lee, Biedermann and Mitra (2017) first take a 2nd order Taylor expansion of $[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}$ and then take the expectation.

Regardless of which approach is taken, both approaches assume MAR, and the expectations involve taking expectations of the missing data indicators $E(m_i) = P(x_i)$ which are then components of the resulting optimality criterion. However, in order to account for NMAR when finding optimal designs, it is crucial to recognise the presence of the outcome in the missing mechanism. Hence rather than using the notation $P(x_i)$ we use $P(x_i, y_i)$ here, where $P(x_i, y_i) = E(m_i|x_i, y_i)$ is now random.

Clearly the presence of $P(x_i, y_i)$ in the approximation to $E\{[\boldsymbol{M}(\xi, \boldsymbol{\mathcal{M}})]^{-1}\}$ complicates the construction of optimal designs as we have not yet observed the outcome values y_i and so treat $P(x_i, y_i)$ as random. In order to proceed we replace $P(x_i, y_i)$ with its expected value $E[P(x_i, y_i)]$ where the expectation is

taken with respect to y_i. Evaluating this expectation is however not typically available in closed form and we investigate ways to approximate this expectation in Section 4.2.

Another key consideration is the potential for bias. Typically optimal design criteria focus on minimising a function of the covariance matrix, and this is because it assumes analysis of the resulting data will yield unbiased estimates. Minimising a function of the covariance matrix will then be equivalent to minimising the corresponding function of the mean squared error (MSE). However, this is not necessarily the case when NMAR is present, as estimates are likely to be biased and this is evident from the results presented in Section 3. Optimal design criteria should then incorporate the bias, or some approximation to the bias, in order to find designs with small MSE. This idea is discussed in more detail in Section 4.3.

4.2 Evaluating the expectation of $P(x_i, y_i)$

To evaluate the expectation of $P(x_i, y_i)$ we consider the specific example of the NMAR mechanism (3.3) introduced in Section 3, where we use a logit link and where the model for y_i is given in (3.2). In principle however, the approach would work with any appropriate NMAR missing data mechanism. We can re-express

$$\begin{aligned} P(x_i, y_i) &= \frac{exp(\gamma_0 + \gamma_1 x_i + y_i)}{1 + exp(\gamma_0 + \gamma_1 x_i + y_i)} \\ &= \frac{exp(z_i)}{1 + exp(z_i)} \end{aligned} \tag{4.5}$$

where $z_i \sim N(\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i, \sigma^2)$ which implies $exp(z_i)$ has a Log-normal distribution with parameters given by the mean and variance of z_i, and $P(x_i, y_i) = \frac{exp(z_i)}{1+exp(z_i)}$ has a logit-normal distribution with parameters given by the mean and variance of z_i. Unfortunately, the mean of the logit normal distribution is not available in closed form and so we consider approaches to approximating the expected value of $P(x_i, y_i)$.

The simplest approach is to simply replace z_i with its expected value in (4.5), i.e. approximate

$$E[P(x_i, y_i)] \approx \frac{exp[E(z_i)]}{1 + exp[E(z_i)]}. \tag{4.6}$$

This is equivalent to the naive approach of finding an optimal design in Section 3 which assumes MAR, and we see that it does not perform well.

An improved approximation uses the fact that $E[exp(z_i)] = exp[\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i + \sigma^2/2]$ and taking a first order Taylor expansion of $P(x_i, y_i)$ as a function of $exp(z_i)$ about the mean of $exp(z_i)$ which results in,

$$\begin{aligned} E[P(x_i, y_i)] &\approx \frac{E[exp(z_i)]}{1 + E[exp(z_i)]} \\ &= \frac{exp[\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i + \sigma^2/2]}{1 + exp[\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i + \sigma^2/2]}, \end{aligned} \tag{4.7}$$

We also consider approximating expectation of $P(x_i, y_i)$, the mean of a logit-normal random variable, using numerical methods. Specifically, defining $P(x_i, y_i) = t_i$ for simplicity, we use the function *integral* in *Matlab* to evaluate

$$E\left[\frac{exp(z_i)}{1 + exp(z_i)}\right] = E(t_i) = \int_0^1 t_i \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{t_i(1 - t_i)} e^{-\frac{[logit(t_i) - \mu_i]^2}{2\sigma^2}} dt_i \tag{4.8}$$

We conducted simulation studies to empirically evaluate the performance of these different methods for approximating $E[P(x_i, y_i)]$. Specifically we generated data that followed a specific logit normal distribution, with parameters μ and σ corresponding to mean and variance of z_i above, and computed the estimated mean of this distribution using the different approximations. This was then repeated many times and estimates from the different methods were averaged over the replications and compared to the "true mean" obtained empirically by averaging the sample mean of observations over the replications. This process was then repeated for a range of different values of μ and σ. Our simulation studies show that approximations from (4.6) and (4.7) performed poorly compared to (4.8). In particular, the error from the approach given by (4.7) is only negligible when $\frac{\mu_i}{\sigma}$ is large. On the other hand, the approximation given by (4.8) gives us very small magnitude of absolute differences in the simulation for $-30 \leq \frac{\mu_i}{\sigma} \leq 30$. We also considered other alternatives to approximating the expected value, including a second order Taylor expansion about $exp(z_i)$ as well as first and second order Taylor expansions about z_i and also using the median of the logit normal distribution implied by $P(x_i, y_i)$ (which is available in closed form) as a surrogate for the expected value. However none of these performed as well as the numerical approximation considered. Hence, we use (4.8) in our design framework for the remainder of this article

4.3 Incorporating bias into the design criterion

When responses are not missing at random, it is unavoidable that estimates will be biased, as noted in Section 3. Hence, instead of simply considering $\boldsymbol{var}(\hat{\boldsymbol{\beta}})$, which is the typical approach in optimal design we consider broadening the framework to incorporate bias. Specifically we focus on now optimising a function of the mean squared error. Returning to the example of obtained regression coefficient estimates $\hat{\boldsymbol{\beta}}$, we can see the mean squared error incorporates both variance and bias,

$$m.s.e.\ (\hat{\boldsymbol{\beta}}) = \boldsymbol{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}}] = \boldsymbol{var}\ (\hat{\boldsymbol{\beta}}) + \left[\boldsymbol{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\right]\left[\boldsymbol{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\right]^{T}, \tag{4.9}$$

where $\boldsymbol{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$ measures the bias of the estimates. In the remainder of this article, we denote $\boldsymbol{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$ by $\Delta(\sigma, \xi)$, i.e. we assume the bias depends on σ as well as the design. Other more complex bias functions that depend on more parameters could also be considered but are not investigated further in this article.

Thus in order to find optimal designs in the presence of NMAR with good MSE properties, we need to include this bias into the criterion. This bias will not be known in advance of course and so in practice must be conjectured as must the functional relationship between the bias and (σ, ξ). In this article we numerically approximate the bias function by simulating the bias for a range of different pairs of values (σ, ξ). Each simulation step involves fitting the model and evaluating the bias for the given pair. We then fit a smooth function, e.g. a second order response surface or a LOESS function, to these simulated 'bias data', and use this function, $B(\sigma, \xi)$ say, as an approximation to the true bias.

We are thus now able to optimise a function of the MSE, or at least an approximation of this for finding an optimal design. As before we consider the same optimality criteria, A- and D- optimality, but these are now applied to the MSE matrix rather than the covariance matrix.

In the next section we evaluate how the approach of finding optimal designs based on the approximation given in (4.8) as well as the inclusion of a bias term performs in the presence of NMAR and specifically whether it offers any improvements over the optimal designs that assume MAR and thus assume $m.s.e.\ (\hat{\boldsymbol{\beta}}) = \boldsymbol{var}\ (\hat{\boldsymbol{\beta}})$.

5. Simulation study

As in Section 3, we set the design region $\mathfrak{X} = [0, 2]$ and sample size $n = 60$. For a given design we simulate a response variable by

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

for a given σ^2. We then introduce missing values into the observed y_i, $i = 1, \ldots, n$, using the NMAR mechanism given by (3.4), i.e. through specifying a missing data mechanism through the following logistic model,

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

with $\gamma_0 = -5.572$ and $\gamma_1 = 2.191$. We assume the analyst will fit a simple linear regression model to the complete case data, 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 restrict our optimal designs to the class of designs with two support points, i.e. $m = 2$. From the results given in Lee, Biedermann and Mitra (2017), 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. To find the second support point, x_2^*, we substitute the approximation to $E[P(x_i^*, y_i)]$ given by (4.8) with mean $-5.572+1+(2.191+1)x_i^*$ and a known value of σ^2, the value of x_1^* and $w_1 = 1-w_2$ into the mean squared error given in (4.9). The expected bias term in (4.9) is treated as being a function of x_2^* and σ^2, and is approximated numerically as described in Section 4.3. An optimal design is then found by minimising a function of this matrix with respect to x_2^* and w_2 in *Matlab* with the *fmincon* function.

Table 5.2 presents the values of x_2^* given under the D- and A- optimality criteria for various different values of σ^2, the corresponding weight w_2 and the (rounded) number of replicates, n_2, of x_2^*. The first column from the left shows the optimal design found by the framework that assume a MAR mechanism and zero bias (see Section 3). We see that the optimal designs that account for the impact of NMAR have smaller x_2^* of both design criteria than the designs that assume the presence of MAR mechanism. The optimal weights of A-optimal designs remain constant in the considered cases whereas w_2 of the D-optimal

Table 5.2: The first column from the left shows the optimal designs that assume a MAR mechanism (4.6); the other columns show the optimal designs for NMAR mechanisms (4.5) with different σ^2. In all designs, $x_1^* = 0$, $n = 60$ and $w_1 = 1 - w_2$.

		MAR	σ =1	σ =1.5	σ =2
D-optimal	x_2^*	1.3766	0.9793	1.0202	1.1210
design	$w_2(n_2)$	0.5000(30)	0.3811 (23)	0.3194 (19)	0.2879 (17)
A-optimal	x_2^*	1.5147	1.0871	1.0617	1.0671
design	$w_2(n_2)$	0.4539(27)	0.4462 (27)	0.4508 (27)	0.4534 (27)

design decreases with σ^2 when responses are assumed to be NMAR. Figure 5.2 further illustrates the optimal designs that account for the impact of NMAR.

Now to illustrate the performance of these designs we repeatedly simulate an incomplete data set 200,000 times using each of the designs given in Table 5.2 and the models for the response and the missing data mechanism. For each design, we calculate the empirical bias and the mean squared error for β_0 and β_1, as well as the determinant and trace of the empirical mean squared error matrix for (β_0, β_1). Table 5.3 presents these results for various different values of σ.

When σ is large, we see that the biases and mean squared errors of the estimators are large, as expected. The designs that assume the presence of MAR have the largest biases and $m.s.e.$ $(\hat{\boldsymbol{\beta}})$ across the board. By taking NMAR into account at the design stage, we can mitigate some of its effects. For example, the A-optimal design for $\sigma = 1.5$ reduces the bias of $\hat{\beta}_1$ by more than 23% from -0.53864 to -0.41095, and a similar reduction applies to the trace of $m.s.e.$ $(\hat{\boldsymbol{\beta}})$. In general we see that the NMAR design with the conjectured value of σ performs best with respect to the relevant optimality criterion. However, it is also clear that NMAR designs with different conjectured values of σ also perform well and far better than the designs that assume MAR.

To complete the illustration, we also consider the potential problem of assuming the presence of NMAR when in fact a MAR assumption is reasonable. Specifically we evaluate the performance of the designs given in Table 5.2 when the missing mechanism is in fact MAR and given by,

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

with $\gamma_0 = -4.572$ and $\gamma_1 = 3.191$. The performance metrics considered as

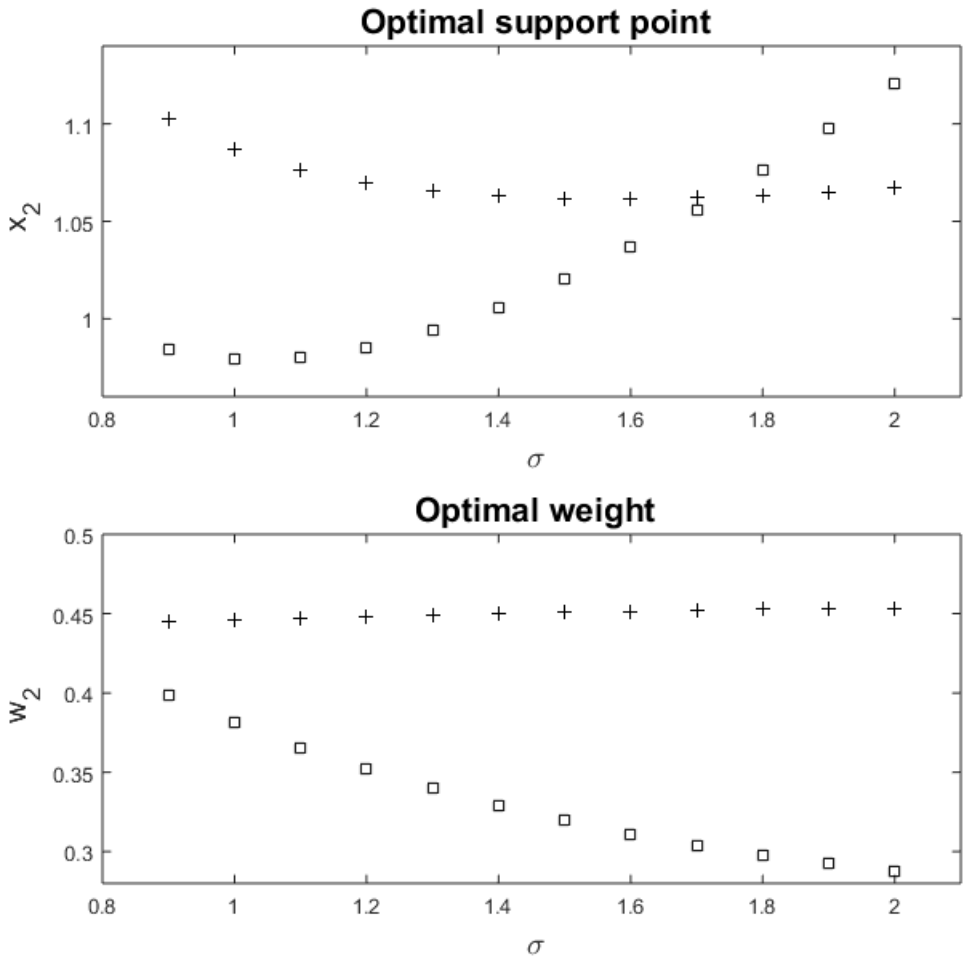

Figure 5.2: "+" correspond to A-optimal designs, "□" correspond to D-optimal designs in the presence of different NMAR mechanisms with $x_1 = 0$ and $w_1 = 1 - w_2$.

Table 5.3: Performance of various designs in the presence of NMAR mechanism over 200,000 simulated data sets.

$\sigma^2 = 1$ in generating y_i and in the NMAR mechanism				
		D-optimal design that assumes		
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$	-0.015710	-0.015657	-0.015559	-0.015525
bias of $\hat{\beta}_1$	-0.26664	-0.18472	-0.19344	-0.21511
$m.s.e.$ $(\hat{\beta}_0)$	0.033581	0.027279	0.024665	0.023522
$m.s.e.$ $(\hat{\beta}_1)$	0.11689	0.11449	0.12077	0.12403
$tr(m.s.e.$ $(\hat{\boldsymbol{\beta}}))$	0.15047	0.14176	0.14544	0.14756
$\lvert m.s.e.$ $(\hat{\boldsymbol{\beta}})\rvert$	0.0035232	**0.0025149**	0.0025445	0.0026165
		A-optimal design that assumes		
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$	-0.015717	-0.015717	-0.015717	-0.015717
bias of $\hat{\beta}_1$	-0.29240	-0.20739	-0.20208	-0.20313
$m.s.e.$ $(\hat{\beta}_0)$	0.030604	0.030604	0.030604	0.030604
$m.s.e.$ $(\hat{\beta}_1)$	0.13022	0.10697	0.10728	0.10713
$tr(m.s.e.$ $(\hat{\boldsymbol{\beta}}))$	0.16083	**0.13758**	0.13788	0.13774
$\lvert m.s.e.$ $(\hat{\boldsymbol{\beta}})\rvert$	0.0037448	0.0026704	0.0026408	0.0026451
$\sigma^2 = 1.5^2$ in generating y_i and in the NMAR mechanism				
		D-optimal design that assumes		
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$	-0.054443	-0.054393	-0.054202	-0.054178
bias of $\hat{\beta}_1$	-0.50182	-0.38675	-0.39934	-0.42936
$m.s.e.$ $(\hat{\beta}_0)$	0.076555	0.062639	0.056827	0.054331
$m.s.e.$ $(\hat{\beta}_1)$	0.34630	0.32185	0.33703	0.34929
$tr(m.s.e.$ $(\hat{\boldsymbol{\beta}}))$	0.42285	0.38449	0.39386	0.40362
$\lvert m.s.e.$ $(\hat{\boldsymbol{\beta}})\rvert$	0.025828	0.018580	**0.018181**	0.018456
		A-optimal design that assumes		
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$	-0.054465	-0.054465	-0.054465	-0.054465
bias of $\hat{\beta}_1$	-0.53864	-0.41838	-0.41095	-0.41264
$m.s.e.$ $(\hat{\beta}_0)$	0.070012	0.070012	0.070012	0.070012
$m.s.e.$ $(\hat{\beta}_1)$	0.37910	0.31198	0.31145	0.31162
$tr(m.s.e.$ $(\hat{\boldsymbol{\beta}}))$	0.44912	0.38199	**0.38146**	0.38163
$\lvert m.s.e.$ $(\hat{\boldsymbol{\beta}})\rvert$	0.026319	0.020325	0.020139	0.020183

Table 5.4: Performance of various designs in the presence of a MAR mechanism, i.e. NMAR with $\sigma^2 = 0$. Responses y_i are generated with $\sigma^2 = 1.5^2$, and over 200,000 simulated data sets.

	D-optimal design that assumes			
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$ ($10^{-4}\times$)	3.7083	4.0648	5.8203	6.2709
bias of $\hat{\beta}_1$ ($10^{-4}\times$)	4.4560	2.9727	-5.9871	-8.3833
$m.s.e.$ ($\hat{\beta}_0$)	0.075687	0.061415	0.055479	0.052892
$m.s.e.$ ($\hat{\beta}_1$)	0.11455	0.19076	0.19898	0.18913
$tr(m.s.e.\ (\hat{\boldsymbol{\beta}}))$	0.19024	0.25218	0.25446	0.24202
$\lvert m.s.e.\ (\hat{\boldsymbol{\beta}})\rvert$	0.0056653	0.0078116	0.0081087	0.0077921
	A-optimal design that assumes			
	MAR	σ =1	σ =1.5	σ =2
bias of $\hat{\beta}_0$ ($10^{-4}\times$)	3.3924	3.3924	3.3924	3.3924
bias of $\hat{\beta}_1$ ($10^{-4}\times$)	2.4240	2.4140	3.2951	3.3618
$m.s.e.$ ($\hat{\beta}_0$)	0.068937	0.068937	0.068937	0.068937
$m.s.e.$ ($\hat{\beta}_1$)	0.11793	0.15299	0.15843	0.15727
$tr(m.s.e.\ (\hat{\boldsymbol{\beta}}))$	0.18687	0.22192	0.22736	0.22621
$\lvert m.s.e.\ (\hat{\boldsymbol{\beta}})\rvert$	0.0060607	0.0065411	0.0067209	0.0066843

before are empirical bias, and mean squared error for β_0 and β_1 as well as the determinant and trace of the empirical mean squared error matrix for (β_0, β_1).

Table 5.4 presents these results for MAR optimal designs and different NMAR optimal designs constructed assuming various values of σ^2. In this simulation, we have used a residual variance of $\sigma^2 = 1.5^2$ in generating the responses under each different design. We see that in this scenario the empirical biases are negligible, as expected. We will thus focus on the mean squared errors. The designs generated assuming MAR perform best as expected but there is evidence to suggest that the loss in assuming a positive value of σ is less severe than the one incurred when using the MAR design for NMAR data. In particular, from Table 5.3 we see the bias in $\hat{\beta}_1$ is smaller in all designs that assume NMAR than the MAR design and we would argue this is an important property analysts would value in any design.

6. Case study: Two-group A-optimal design for Alzheimer's Disease Trial

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. For illustration purposes, we only consider the experimental units in the placebo group and the donepezil-memantine treatment group, who were included in the primary intention-to-treat sample. The sample size in each group (n_1, n_2) respectively is 72 resulting in a total sample size of 144. Here we treat the rate of change of the primary outcome measure, SMMSE score (higher score indicates better cognitive function), as the response variable of a simple linear model,

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

where $x_i = 0$ for subject i in the placebo group and $x_i = 1$ for patient i in the treatment group, $i = 1, \ldots, 144$.

However, from the data set for the per-protocol analysis, we have 46 patients in the placebo group and 23 patients in the treatment group who have missing responses by the end of the study. Assuming that these responses are not missing at random, a logistic regression model is fitted to the missing data indicator, obtaining

$$\frac{exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_i)}{1 + exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_i)}$$

where $\hat{\gamma}_0 = 0.5705$ and $\hat{\gamma}_1 = -1.3269$. Using the observed responses, we fit a linear model to the data, obtaining $\hat{\beta}_0 = -0.10503$, $\hat{\beta}_1 = 0.04302$ and $\sigma^2 = 0.06143^2$. We then use these estimates to construct a NMAR mechanism, (4.5), where the logit-normal variable, t_i, has mean $\gamma_0 + \beta_0 + (\gamma_1 + \beta_1)x_i = 0.5705 - 0.10503 - (1.3269 - 0.04302)x_i$ and variance $\sigma^2 = 0.06143^2$. We use this information in (4.6) to approximate the expected NMAR mechanism, which is present in the elements of the approximation to $E[\boldsymbol{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{\mathcal{M}})]$ when finding optimal designs. See Section 2.3 and 4.1 for more details. Of course in practice NMAR is an untestable assumption and there is no guarantee that these conjectured mechanism corresponds to the true missing mechanism, the following analysis could have been repeated with different conjectured NMAR mechanisms.

In this scenario, the support points of an optimal design are given as $x_1^* = 0$ (placebo) and $x_2^* = 1$ (active treatment) since we are comparing two groups. We consider A-optimality. Hence we want to minimise $m.s.e.\ (\hat{\beta}_0) + m.s.e.\ (\hat{\beta}_1)$ for

Table 6.1: Fitted coefficients for the approximation function $B(\sigma, \xi)$ of $\Delta(\sigma, \xi)$ for $\hat{\beta}_0$ (first row) and $\hat{\beta}_1$ (second row), respectively.

$\hat{\lambda}_0$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
-2.5282×10^{-5}	1.8727×10^{-6}	-1.2511×10^{-3}	-1.4028×10^{-8}	8.7490×10^{-7}	-0.6023
-2.9306×10^{-5}	-4.1954×10^{-7}	1.6884×10^{-3}	2.9213×10^{-9}	3.1693×10^{-6}	0.2919

a future experiment. The optimisation problem is now in one variable, i.e. w_2, with the condition $w_1 + w_2 = 1$.

In this illustration, we conduct simulation studies on designs that have $n_2 = 37, 38, ..., 107$ in each design, with $\sigma = 0.04,\ 0.05, ...,\ 0.09$ in each case, to obtain empirical biases for $\hat{\beta}_0$ and $\hat{\beta}_1$. Fitting a second order response surface to these observed biases and values of n_2 and σ, we approximate bias as a function of n_2 and σ in the following way,

$$B(\sigma, \xi) = \hat{\lambda}_0 + \hat{\lambda}_1 n_2 + \hat{\lambda}_2 \sigma + \hat{\lambda}_3 {n_2}^2 + \hat{\lambda}_4 n_2 \sigma + \hat{\lambda}_5 \sigma^2$$

for each estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ (see Table 6.1).

Using this information, we can find the A-optimal design by using the *fmincon* numerical method in *Matlab* as was done in Section 5. The optimal design resulted in $w_2 = 0.34365$, i.e. $n_2 = 144 \times w_2 = 49.486 = 49$ subjects in the treatment group, giving $n_1 = 95$ subjects in the placebo group. We then conduct a simulation study comparing this design with other design candidates using the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\sigma}^2$ in generating responses (both observed and missing). Table 6.2 shows the performance of these designs in the simulation. Specifically we repeatedly simulate incomplete data under the various designs and compute the trace of the mean squared error matrix obtained from each design. The simulation study shows that the A-optimal design that accounts for NMAR and bias in the experiment performs better than all other designs considered and in particular is better than the original design that assumes equal sample size for both groups. There is about a 9% $(1 - 3.2919/3.6155) \times 100\%$ efficiency loss if we use the equal sample size design instead of the optimal design. This indicates that in this real application there is also the potential for obtaining estimates with smaller mean squared error if the proposed design is used rather than conventional designs.

7. Discussion and remarks

Table 6.2: Performance of various designs where n_2 is the sample size of the treatment group and $n_1 = 144 - n_2$ for each design.

n_2	52	51	50	49	72
$tr(m.s.e.\ (\hat{\boldsymbol{\beta}}))(\times 10^{-4})$	3.2950	3.2927	3.2934	**3.2919**	3.6155

We have opened up a new area of research, showing that the effects of NMAR mechanisms on estimation can be mitigated through a clever choice of experimental design.

We first assess the effect of the value of σ^2 under a NMAR mechanism on the performance of optimal designs, which have been generated under the assumption of MAR (i.e. $\sigma^2 = 0$). We see dramatic increases in empirical biases and mean squared errors. Having established the need for further investigation, we then propose a novel approximation to the information matrix taking NMAR into account. Unlike in the MAR case, or indeed the classical case of no missing data, our designs depend on the linear parameters of the mean model and on the value of σ^2. The optimal designs found through our new approach considerably outperform the naive designs generated assuming MAR when NMAR is present.

To the best of our knowledge this is the first paper dealing with optimal design of experiments in a NMAR framework. There are thus many open problems left to investigate. We present and illustrate our methodology through linear regression models. However, a similar approximation to (4.7) can be found for nonlinear models with normally distributed errors, and extensions to generalised linear models are also possible in our framework.

The designs we find are locally optimal in the sense that they depend on the unknown model parameters. Our numerical investigation shows that even when the value of σ^2 is misspecified at the design stage, the designs assuming NMAR with an incorrect σ^2 perform still better than the MAR design when the missing data mechanism is NMAR. For the other parameters, we assume here that good information can be elicited from the experimenter. If this is not the case, parameter robust design criteria, such as Bayesian or standardised maximin criteria (see, e.g., Chaloner and Verdinelli, 1995, and Dette, 1997, respectively), need to be developed for our approach, which is a topic for future research.

There are a plethora of possible methods to handle the problem of missing values, in addition to complete case analysis considered in this article. Other common approaches include multiple imputation, methods based on the EM algorithm, Hot Deck methods, plus many others. We do not investigate these here, as our approach focuses on the design aspect of the problem, rather than the specific method to deal with the missing data. For each new method for handling the missing values, the design framework would change to reflect the type of analysis being performed on the data and would need to be derived carefully mathematically. It would be interesting to investigate this further in future research to see whether the benefits seen here could be similarly observed when other methods are used to handle the missing data.

Our approach sheds new light on NMAR problems, and shows there is scope for optimal experimental design as a tool to reduce biases and mean squared errors in such scenarios.

Acknowledgement

The first author's research is funded by the Institute for Life Sciences at the University of Southampton.

References

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.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science* **10(3)**, 273-304

Dette, H. (1997). Designing experiments with respect to standardized optimality criteria. *J. Roy. Statist. Soc. Ser. B* **59**, 97-110.

Ghosh, S. (1979). On robustness of designs against incomplete data. Sankhyā: *The Indian Journal of Statistics, Series B*, 204-208.

Hackl, P. (1995). Optimal design for experiments with potentially failing trials. In *Proc. of MODA4: Advances in Model-Oriented Data Analysis* (Edited by C. P. Kitsos and W. G. Müller), 117-124. Physica Verlag, Heidelberg.

Hedayat, A. and John, P. W. M. (1974). Resistant and susceptible BIB designs. *Ann. Statist. 2* **1**, 148–158.

Herzberg, A. M. and Andrews, D. F. (1976). Some considerations in the optimal design of experiments in non-optimal situations. *J. Roy. Statist. Soc. Ser. B* **38**, 284-289.

Howard, R., McShane, R., Lindesay, J., Ritchie, C., Baldwin, A., Barber, R., ... and Phillips, P. (2012). Donepezil and memantine for moderate-to-severe Alzheimer's disease. *New England Journal of Medicine* **366(10)**, 893-903.

Imhof, L. A and Song, D. and Wong, W. K. (2002). Optimal design of experiments with possibly failing trials. *Statistica Sinica* **12**, 1145-1155.

Lee, K.M., Biedermann, S. and Mitra, R. (2017). Optimal design for experiments with possibly incomplete observations. *Under revision*.

Little, R. J. A. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association* **87**, 1227-1237.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data (Second Edition)*. Wiley-Interscience.

Ortega-Azurduy, S. A., Tan, F. E. S. and Berger, M. P. F. (2008). The effect of dropout on the efficiency of D-optimal designs of linear mixed models. *Statist. Med.* **27**, 2601-2617.

Pukelsheim, F. and Rieder, S. (1992). Efficient rounding of approximate designs. *Biometrika*, **79(4)**, 763-770.

Rubin, D.B (1976). Inference and missing data. *Biometrika* **63**, 581-592.