The Non-Ignorable Missing-Data Problem in Consumer Banking

by

I-Ding Wu

July 2008

A thesis presented for the degree of Doctor of Philosophy of the University of London and the Diploma of Imperial College

Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ
Abstract

The thesis aims to solve a specific missing-data problem in consumer banking. Application scoring and behaviour scoring are two of the main applications of statistics and probability modelling in consumer banking. In application scoring, a missing data problem occurs due to the selection of applicants by the bank. This has attracted much interest, and relevant discussion can be found under the topic of “reject inference”. On the contrary, a similar problem in behaviour scoring has not been widely explored. The problem we wish to solve in the present thesis is a missing data problem that results from selection in behaviour scoring.

We review the nature of the missing data problem and the existing solutions. Missing-data problems can be categorised into: MCAR, MAR, and MNAR problems. MCAR and MAR problems have attracted much attention; less discussion can be found on the MNAR problems. The problem we solve in this thesis is a MNAR problem.

Two of the best known solutions to MNAR problems are: the two-step method proposed by Heckman, and the EM algorithm proposed by Little and Rubin. We illustrate how these existing methods can be extended to solve our problem. The extensions of these existing methods are constrained by an inflexible assumption, i.e. each method assumes that an unrecorded variable has a specific distribution. We introduce solutions that remove this constraint so as to be able to use the empirical distribution. The thesis also presents solutions making use of updated MAR data, which are available in the case of behaviour scoring.
Acknowledgements

First of all, I would like to thank my supervisor, Professor David J. Hand, who has provided constant support and advise. I would also like to express my gratitude to my examiners, Professor Jonathan Crook and Dr. Nick Heard, for their helpful comments and suggestions.

I am certainly deeply indebted to my parents, sister, brother, and Billy Wu. Without support from any of them, I would not have achieved this stage.
Contents

1 Introduction 12

2 Missing data problems 18

  2.1 Problem statement ........................................... 19
    2.1.1 The aim ............................................. 19
    2.1.2 The problem ........................................... 19

  2.2 The general definition of the missing data mechanism .......... 21
    2.2.1 Missing data patterns ................................ 21
    2.2.2 Missing data indicators ................................ 22
    2.2.3 Missing data mechanisms ................................ 22

  2.3 The missing data mechanism of the data set available .......... 24
    2.3.1 Missing data patterns ................................ 24
    2.3.2 Missing data indicators ................................ 26
    2.3.3 Missing data mechanisms ................................ 26

  2.4 Further assumptions ........................................... 32
    2.4.1 Linear response functions ................................ 32
    2.4.2 Linear and deterministic previous assignment function 33

3 Existing adjustment methods 35

  3.1 EM algorithm for ignorable missingness ....................... 37
    3.1.1 ML method .......................................... 37
    3.1.2 EM algorithm .......................................... 38
    3.1.3 Application to the ignorable missing data problem in credit scoring ........................................ 40

  3.2 Imputation methods for ignorable missingness .................. 42
3.2.1 Single static imputation ........................................ 42
3.2.2 Single stochastic imputation .................................. 43
3.2.3 Multiple imputations ........................................... 45

3.3 Propensity score for ignorable missingness ..................... 49
3.3.1 Propensity score .............................................. 49
3.3.2 Application to the ignorable missing data problem in credit scoring ........................................... 51

3.4 Type II Tobit model .............................................. 52
3.4.1 Model description ............................................. 52
3.4.2 Non-ignorable missingness .................................... 54
3.4.3 Comparison with the non-ignorable missing data problem in credit scoring ........................................... 54

3.5 Heckman's two-step method for non-ignorable missingness ...... 55
3.5.1 Different explanatory variables ................................ 55
3.5.2 Heckman's two-step method .................................... 56
3.5.3 Application to the non-ignorable missing data problem in credit scoring ........................................... 58

3.6 EM algorithm for non-ignorable missingness ..................... 59
3.6.1 ML method ..................................................... 59
3.6.2 EM algorithm .................................................. 60
3.6.3 EM algorithm for Type II Tobit Model ......................... 61
3.6.4 Simplified EM algorithm for Type II Tobit Model .......... 63
3.6.5 Application to the non-ignorable missing data problem in credit scoring ........................................... 65

4 Simulation design .................................................. 68
4.1 Variables and parameters used .................................... 68
4.1.1 Explanatory variables ........................................ 69
4.1.2 Parameters in the response functions ......................... 69
4.1.3 Parameters in the previous assignment function ............ 72
4.2 Simulation details ................................................ 73
4.3 Evaluation ...................................................... 75
5 Key characteristics of the problem

5.1 The complete-case analysis is insufficient

5.2 The distribution of $Z$ varies from case to case

5.3 Newly collected data are MAR

6 Flexible adjustment approaches

6.1 An EM algorithm that uses an empirical density function to estimate the previous assignment function

6.1.1 The E step

6.1.2 The M step

6.2 Generate empirical samples from the newly collected data

6.2.1 Regression residuals obtained from applying the complete-case analysis to the newly collected data

6.2.2 Adequacy of using residuals as empirical samples

6.3 A semi-GEM algorithm that estimates the previous assignment function

6.3.1 The E step

6.3.2 The semi-generalised M step

6.4 Estimate the response functions

6.5 Simulation results

6.5.1 Effectiveness of the semi-GEM algorithm

6.5.2 Effectiveness of obtaining the empirical sample from the newly collected data set

6.6 When $Z$ is multivariate

7 Adaptive adjustment approaches

7.1 Ad-hoc Adaptive Method

7.2 Combined least squares regression approaches

7.2.1 Naive combined least squares regression approach

7.2.2 Adjusted combined least squares regression approach

7.2.3 Weighted combined least squares regression approach

8 Conclusions

8.1 Contributions
8.1.1 Drew attention to the non-ignorable missing data problem in behaviour scoring ................. 123
8.1.2 Described extensions from existing methods and pointed out the inadequacies ................... 124
8.1.3 Proposed adjustment methods so that the empirical distribution can be used ...................... 125
8.1.4 Proposed adaptive approaches that make use of the newly collected MAR data .................. 127

8.2 Future research ................................................. 129
8.2.1 Consider more than two actions ...................... 129
8.2.2 Relax the linearity assumption ....................... 129
## List of Figures

2.1 Examples of missing data patterns. .................................. 22  
2.2 Missing data pattern of a credit scoring data set $D_{cs}$. ..... 25  
2.3 Missing data patterns of credit scoring data sets $D_a$ and $D_b$. 25  

4.1 The histogram of $V_1$ used in our simulations. ............... 70  
4.2 The histogram of $V_2$ used in our simulations. ............... 70  
4.3 The histogram of $V_3$ used in our simulations. ............... 70  
4.4 The histogram of $V_4$ used in our simulations. ............... 71  
4.5 The histogram of $V_5$ used in our simulations. ............... 71  

5.1 The performance of the complete-case analysis. ............... 80  
5.2 The performance of the Heckman’s two-step method. ........... 82  

6.1 The performance of the proposed adjustment method that makes use of empirical samples of $Z$. .......................... 99  
6.2 The performance of the proposed adjustment method that makes use of empirical samples obtained from regression residuals. .................................................. 102  

7.1 The concept of dynamic updating approaches. .................. 106  
7.2 The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. ........................................... 110  
7.3 The performance of the ad-hoc adaptive approach and Heckman’s two-step method was applied to the previously collected data set. ........................................... 111
7.4 The performance of the naive combined least squares regression approach .................................. 115
7.5 The performance of the adjusted combined least squares regression approach .......................... 117
7.6 The performance of the weighted combined least squares regression approach extended from the naive combined least squares regression approach .............................................. 120
7.7 The performance of the weighted combined least squares regression approach extended from the adjusted combined least squares regression approach .............................................. 121

A-1 ........................................ 132
A-2 ........................................ 133
A-3 ........................................ 134
A-4 ........................................ 135
A-5 ........................................ 136
A-6 ........................................ 137
A-7 ........................................ 138
A-8 ........................................ 139
A-9 ........................................ 140
A-10 ....................................... 141
A-11 ....................................... 142
A-12 ....................................... 143
A-13 ....................................... 144
A-14 ....................................... 145
A-15 ....................................... 146
A-16 ....................................... 147
A-17 ....................................... 148
A-18 ....................................... 149
A-19 ....................................... 150
A-20 ....................................... 151
A-21 ....................................... 152
A-22 ....................................... 153
List of Tables

4.1 Parameters Values .................................. 73
Chapter 1

Introduction

In recent years, statistics and probability modelling have been widely applied to estimate financial risk. This has been mostly in the context of investment banking and consumer banking. In this thesis, we focus on the applications to consumer banking. The main consumer banking applications for which forecasting financial risk is relevant is credit scoring.

There are two main branches of credit scoring: application scoring and behaviour scoring. Application scoring assesses risk for new applicants at the point of application, and behaviour scoring assesses risk for existing customers. Application scoring enables financial service providers to decide how to treat new applicants, whereas behaviour scoring is used in deciding how to treat existing customers. For example, application scoring is used to decide whether to issue a credit card to a new applicant or not, and behaviour scoring is used to decide whether to increase the credit limit of an existing credit card holder or not. Similarly, when a new applicant asks for a loan, application scoring is used to decide whether to lend the money or not; and if an existing borrower defaults on a repayment, behaviour scoring is used to decide whether to collect the payment in-house or to sell off the debt to a third party.

We can think of accepting an applicant as assigning them one action, and
rejecting an applicant as assigning them another action. Likewise, we can think of increasing an existing customer’s credit limit to £600, £1,000, or £3,000 respectively as assigning to them three different actions. We can also regard collecting a debt in-house as assigning to the default borrower one action, and selling off the debt as assigning to them another action.

We can interpret the aim of credit scoring as estimating an applicant’s or an existing customer’s response to different actions [51] so that we can choose the optimal action to take. For example, to estimate the probability that a new applicant will default on their repayment if their applications are accepted; to estimate the probability that a customer will pay the full statement balance if their credit limits are increased to £600, £1,000 or £3,000; and to estimate the proportion of a debt that can be collected if the debt is collected in house or sold to a third party.

For simplicity, in this thesis we assume that there are only two possible actions: Action A and Action B. We also assume that each customer can only receive one of the available actions. This is because, for example, an applicant can only be either accepted or rejected but not both; a credit card holder’s credit limit can be increased to either £600, £1,000 or £3,000; and a debt can only be collected in-house or sold to a third party but not both. Therefore, we wish to estimate a customer’s responses to Actions A and B respectively so that they can be assigned the action expected to yield the most favourable outcome.

Typically, credit scoring models are based on a data set that contains information on previous customers. The problem is that each previous customer also received only one of the available actions. In this way, the responses to Action B are missing among customers who received Action A; similarly, the responses to Action A are missing among customers who received Action B. The data set used thus contains missing data. If one draws an inference about the response to Action A (or B) using only information on previous customers who received Action A (or B), the inference may be biased.
For example, a lending institution usually keeps the application details of all previous applicants, whether accepted or rejected. However, the firm can only record the responses (i.e. whether there is default on the repayment or not) of accepted applicants. If only the information on accepted applicants is used to estimate the default rate of a new applicant, such an estimation might be biased. This is because, unless previous applicants were randomly accepted or rejected, the default rate of an accepted applicant is likely to be lower than that of a rejected applicant. If previous applicants were not randomly accepted, the distribution of the default rate among accepted applicants is different from that among all applicants. Inferences drawn from accepted applicants are thus not representative of applicants overall. Consequently, it has been suggested that one should make use of the information on rejected applicants: this is the idea of “reject inference”. Reject inference has attracted much interest, and relevant discussion can be found in [6] [16] [17] [18] [33][47].

Similarly, in the example of increasing one’s credit card limit, the only available information about how a card holder will react if their credit limit is increased to £3,000 is that from those whose credit limit was indeed increased to £3,000. We do not know how those whose credit limit was increased to £600 would have behaved if their credit limit had been increased to £3,000. In this thesis, we consider a more general problem in behaviour scoring of which the accept/reject problem is a special case.

The accept/reject problem only takes into account the response to one action, i.e. acceptance, whereas we are interested in comparing responses to Actions A and B. Moreover, in the accept/reject problem, one can only observe responses from accepted but not rejected applicants; whereas in the problem we are concerned with, we can always observe the response (either to Action A or to Action B) from each previous customer. When the response to only one action is under consideration, the problem due to the missing data is insignificant if the proportion of missing data is small. Note that this is not the case if responses to more than one actions are taken into account (see Section 5.1 for further discussion). Nevertheless, approaches applicable
to the general problem discussed in this thesis are also applicable to the accept/reject problem.

Our aim is to use the possibly biased data, in which different customers have received each action, but not on a random assignment basis, to construct an optimal rule for assigning new customers to actions. Chapter 2 characterises this problem using mathematical structures and describes the conventional assumptions we made in order to focus on the missing-data problem. Whether using an incomplete data set without any adjustment will result in biased estimations or not depends on the missing-data mechanisms. In Section 2.2, we review the general definition of the missing-data mechanisms and discuss how to determine the type of missing-data mechanisms for the data sets considered in this thesis.

The missing-data mechanism can be used to classify missing-data problems into ignorable problems and non-ignorable problems. In Chapter 3, we review existing adjustment methods for both ignorable and non-ignorable missing data problems. Apart from reviewing existing methods, we also describe how these methods can be extended and applied to the problem we wish to solve. We show that, if applied to the credit scoring data sets we are concerned with, existing adjustment methods for ignorable missing data problems are either inapplicable or do not provide extra information. On the contrary, if the missing data problem is ignorable, using an incomplete data set without any adjustment is sufficient to derive unbiased estimations. From Chapter 4 on, we only consider the non-ignorable missing data problem in credit scoring.

Estimates derived directly from a data set with a non-ignorable missing data problem are biased. Several researchers have argued that, when response to only one action is taken into account, such a bias can be relatively small and applying an adjustment method is not worthwhile. In Section 5.1, we show that this is not the case when the interest lies in the response to more than one action. In Section 5.2, we point out that although it is worthwhile to adjust the bias in our case, extensions of existing methods are not always
applicable. The disadvantage of applying extensions of existing methods to the non-ignorable missing data problem in credit scoring is also discussed. In Section 5.3, we further point out that the missing data problem we wish to solve results from self-selection. Hence, although the newly collected data set\(^1\) is also incomplete, its missing data problem is ignorable. Such a data set can be useful when applying the adjustment methods we present in Chapters 6 and 7.

The non-ignorable missing data problem with which we are concerned results from an unrecorded variable. Each existing method assumes the unrecorded variable has a specific distribution. It is thus only reasonable to apply an extension of an existing method when one assumes the unrecorded variable has the same distribution. In Chapter 6, we propose adjustment methods that are applicable regardless of the type of distribution one assumes. Constructing an optimal decision rule using a biased data set requires information on how actions were assigned to previous customers (i.e. the previous assignment function). In Section 6.1 we introduce an EM algorithm that makes use of an empirical density function to estimate the previous assignment function. However, when the unrecorded variable \(Z\) is un-recordable, its density function is unavailable. Therefore in Section 6.2 we propose a method that generates empirical samples of \(Z\) using newly collected data. Further in Section 6.3 we introduce a semi-GEM algorithm that makes use of these empirical samples to estimate the previous assignment function. In Section 6.4 we describe how to construct an optimal decision rule using the estimated previous assignment function. Adjustment methods proposed in Chapter 6 assume that the unrecorded variable is univariate. In Section 6.6 we discuss that if the univariate assumption fails, these methods are inapplicable.

In Chapter 7, we propose adaptive models that avoid the univariate as-

\(^1\)We call customers who enter the system before the analysis starts previous customers, and those who enter after the analysis starts newly existing customers. The data set that contains information on previous customers is called the previously collected data set, and the one that contains information on newly existing customers the newly collected data set.
assumption and make use of both previously and newly collected data. Since the information in the newly collected data set increases whenever a new customer enters the system, adaptive methods can update the estimations continuously. Contrary to the previously collected data set, the missing data problem in the newly collected data set is ignorable, and estimations derived directly from this data set are unbiased. In Section 7.1, we propose an ad-hoc adaptive approach that combines the unbiased estimations derived from the newly collected data set with the biased estimations derived from the previously collected data set. In Section 7.2, we propose other adaptive models: combined least squares regression approaches, which combine both available data sets (one with an ignorable and the other with a non-ignorable missing data problem) before deriving estimations.

In this thesis, we review existing adjustment methods, point out the characteristics of the problem we wish to solve, and propose new adjustment methods. In order to evaluate extensions of existing methods, support our arguments and examine proposed methods, we make use of simulated data sets. Details regarding how these data sets were simulated can be found in Chapter 4.
Chapter 2

Missing data problems

In Section 2.1, we describe in detail the mathematical structure of the problem we wish to solve.

We wish to estimate the function on which optimal decisions are based using a credit scoring data set with missing values. A straightforward approach to derive estimations from an incomplete data set is to ignore data points with a missing value. This approach is called the complete-case analysis.

Whether applying the complete-case analysis will result in biased estimations depends on the missing data mechanism of the data set. In Section 2.2, we review the general definition of the missing data mechanism. In Section 2.3, we further discuss how to determine the type of the missing data mechanism for the credit scoring data sets with which we are concerned.

To focus on the missing data problem, in Section 2.4, we simplify the problem by making linearity assumptions for the response functions and the function on which previous assignments were based.
2.1 Problem statement

2.1.1 The aim

We assume that there are two available actions: Action A and Action B. For each customer, we wish to estimate their responses to Actions A and B respectively. In this way, the action expected to yield the most favourable result can be assigned to each individual. Set the variable \( R_a \) to represent the response to Action A and \( R_b \) the response to Action B. Without loss of generality, we assume that the larger the response the better. For a new customer, we wish to estimate their values of \( R_a \) and \( R_b \), compare the estimated responses, and assign the optimal action, which has a larger estimated response, to that customer.

Suppose that some recorded background information, i.e. the vector values of \( \mathcal{X} = \{x_1, \ldots, x_p\} \), are recorded for each customer. We wish to estimate the difference between \( R_a \) and \( R_b \) for each new customer according to their vector values of \( \mathcal{X} \). That is to estimate

\[
E(r_{ai} - r_{bi} | x_i),
\]

where \( x_i = \{x_{1i}, \ldots, x_{pi}\} \); \( r_{ai}, r_{bi} \), and \( x_{ji} \) represent the values of \( R_a \), \( R_b \), and \( X_j \) respectively of a customer \( i \).

Equation (2.1) is the function on which optimal decisions are based, and we will call this "the optimal decision function". The aim is to estimate the optimal decision function so that based on the estimation, a customer \( i \) will receive Action A if \( E(r_{ai} - r_{bi} | x_i) > 0 \), or Action B if \( E(r_{ai} - r_{bi} | x_i) \leq 0 \).

2.1.2 The problem

Estimation models are developed using a data set that consists of information on \( n \) previous customers. The data set records the values of \( \mathcal{X}, R_a \) and \( R_b \),
for these $n$ customers, if they are available. We assume the values of $X$ are fully available and recorded, but the values of $R_a$ and $R_b$ are not. Similar to new customers, previous customers only received one of the two actions. In this way, if a customer received Action A (or B), the value of $R_a$ (or $R_b$) but not $R_b$ (or $R_a$) of this customer is available and recorded. The values of $R_a$ and $R_b$ are never jointly observed and it will be difficult to derive a model for $E(r_{ai} - r_{bi}|x_i)$ directly.

Nevertheless, due to the linearity of the expectation operator, the optimal decision function (2.1) can be expressed as

$$E(r_{ai}|x_i) - E(r_{bi}|x_i).$$

From the original data set that records the values of $X$, $R_a$ and $R_b$ for previous customers, we can create another two data sets: one containing the values of $X$ and $R_a$ (if available) and the other containing the values of $X$ and $R_b$ (if available) for the same $n$ previous customers. Thereafter, one can derive a model for $E(r_{ai}|x_i)$ using one of the data sets and a model for $E(r_{bi}|x_i)$ using the other. However, both data sets remain incomplete.

We can split an incomplete data set into a complete-cases data set and an incomplete-cases data set. A complete-cases data set consists of data points without missing values, and an incomplete-cases data set consists of data points with missing values. In order to draw inferences from an incomplete data set, one might exclude data points with missing values and apply standard statistical methods to the complete-cases data set. Such a strategy is termed the complete-case analysis.

When one applies the complete case analysis to estimate $E(R_a|X)$ (or $E(R_b|X)$), only information on previous customers who received Action A (or Action B) is used. However, applying the complete-case analysis is not always appropriate. This is because the interest lies in making inferences about the entire target population, rather than the part of the target population from which one would observe the values of all relevant variables in the
analysis. For example, suppose that a previous customer was assigned Action A because the decision maker expected the value of $R_a$ to be larger than that of $R_b$ for that customer. In this way, the distribution of $R_a$ among previous customers who were assigned Action A is unlikely to be the same as the distribution of $R_a$ for all customers.

When it is possible to apply the complete-case analysis (e.g. the size of the complete-cases data set is large enough), whether applying such an analysis is appropriate or not depends on the missing data mechanism.

2.2 The general definition of the missing data mechanism

Rubin [42] first formalised the missing data mechanism by treating the missing data indicator as a random variable. In order to review the definition of each missing data mechanism, we shall first describe missing data patterns and missing data indicators.

2.2.1 Missing data patterns

Let $D = (d_{ij})$ denote a general data set, where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, q$. The data set $D$ records the values of $q$ variables $D_1, D_2, \ldots, D_q$, from $n$ independent cases (units) randomly drawn from the target population. Moreover, $d_{ij}$ represents the value of variable $D_j$ for case $i$.

If all of $d_{ij}$ are observed, the data set $D$ is complete. Using such a complete data set, one may apply standard statistical methods to draw valid inferences about the target population. However if the data set is incomplete, some adjustments may have to be made before applying standard methods.

The missing data patterns provide information about which values in a data set are observed, and which are missing. In Figure 2.1, we show some
2.2.2 Missing data indicators

Let $M = (m_{ij})$, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., q$, refers to a missing data indicator matrix for the general data set $D$. If $d_{ij}$ is missing, $m_{ij} = 1$, otherwise, $m_{ij} = 0$. The following matrices represent the missing data indicator matrices corresponding to the different missing data patterns in Figure 2.1:

$$
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 \\
0 & 1 & 1 & 1
\end{bmatrix}
\quad
\begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}
\quad
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{bmatrix}
$$

2.2.3 Missing data mechanisms

There are three types of missing data mechanisms: Missing Completely At Random, denoted MCAR, Missing At Random, denoted MAR, and Missing
Not At Random, denoted MNAR.

MCAR

The data in the general data set $D$ are defined as missing completely at random if the missingness is unrelated to the values (missing or observed) of the variables recorded in $D$. That is, if ([27], p12)

$$f(M|D) = f(M), \quad (2.3)$$

where $f(\cdot)$ denotes the probability mass function for $M$.

MAR

We can split the incomplete data set $D$ into \{${D}_{obs}, D_{mis}$\}, where $D_{obs}$ refers to the observed information, and $D_{mis}$ refers to the missing information. The data are defined as missing at random if, conditioned on the observed information $D_{obs}$, the missingness is independent of the unobserved $D_{mis}$. That is, if ([27], p12)

$$f(M|D) = f(M|D_{obs}). \quad (2.4)$$

When Equation (2.3) holds, Equation (2.4) also holds. Consequently, the MCAR can be considered as a special case of MAR.

MNAR

In the general data set $D$, the data are defined as missing not at random if, even conditioned on the observed information, the missingness depends on the unobserved one. That is, if ([27], p12)

$$f(M|D) \neq f(M|D_{obs}). \quad (2.5)$$
When the data are MCAR or MAR, the bias resulting from the missing data can be adjusted by making use of observed data, and the missing data problem is defined as ignorable. In contrast, when the data are MNAR, using observed data is insufficient to correct the bias, and the missing data problem is defined as non-ignorable. If the missing data problem of an incomplete data set is ignorable, applying the complete-case analysis (if it can be applied) will result in unbiased estimations. In contrast, if the missing data problem is non-ignorable, applying the complete-case analysis is inappropriate.

2.3 The missing data mechanism of the data set available

The previous section reviewed the general definition of each missing data mechanism. In this section, we discuss how to determine the type of the mechanism for data sets that contain the information on $X$ and $R_a$ and the information on $X$ and $R_b$ for the same $n$ previous customers.

2.3.1 Missing data patterns

Let $D_{ca}$ denote the original available data set that records the values of $X$, $R_a$ and $R_b$. The missing data pattern in this data set is shown in Figure 2.2. Since $R_a$ and $R_b$ are never jointly observed, such a data set does not provide any information on the correlation between $R_a$ and $R_b$. Using the data set $D_{ca}$ directly is equivalent to using data sets $D_a$ and $D_b$. The data set $D_a$ (or $D_b$) contains records of the values of $X$ and $R_a$ (or $R_b$), and both $D_a$ and $D_b$ contain information on the same set of previous customers. The missing data patterns of each of these two data sets are shown in Figure 2.3.
Figure 2.2: Missing data pattern of a credit scoring data set $D_{cs}$. Each row represents information on one previous customer, and each column represents a variable. Shaded blocks indicate observed values, and white ones indicate missing values.

Figure 2.3: Missing data patterns of credit scoring data sets $D_a$ and $D_b$. Each row represents information on one previous customer, and each column represents a variable. Shaded blocks indicate observed values, and white ones indicate missing values.
2.3.2 Missing data indicators

Let $S = (s_i)$, where $i = 1, \ldots, n$, refer to the missing data indicators for both data sets $D_a$ and $D_b$. We have assumed $X$ to be fully observed and recorded, and only the response variable possibly missing. Thus we set $s_i = 0$ if a customer received Action A ($r_{ai}$ is observed and $r_{bi}$ is missing), and set $s_i = 1$ if a customer received Action B ($r_{bi}$ is observed and $r_{ai}$ is missing). The missing data indicator matrix corresponding to both $D_a$ and $D_b$ is thus

$$
\begin{bmatrix}
0 \\
1 \\
1 \\
0 \\
1
\end{bmatrix}
$$

Note that the missing data indicators of $D_b$ are identical to that of $D_a$. To avoid repetition, we will discuss the missing data mechanisms for $D_a$ only.

2.3.3 Missing data mechanisms

According to Equation (2.3), the values of the response in $D_a$ are MCAR if

$$f(S|X, R_a) = f(S). \quad (2.6)$$

Define $R_{a, obs}$ to be the observed values of the response to Action A, and $R_{a, mis}$ to be the unobserved values. According to Equations (2.4) and (2.5), the values of the response in $D_a$ are MAR if

$$f(S|X, R_a) = f(S|X, R_{a, obs}). \quad (2.7)$$

and are MNAR if

$$f(S|X, R_a) \neq f(S|X, R_{a, obs}). \quad (2.8)$$
Most of the time, recorded background information $X$ is used to make the decision about how to treat a customer. However, previous assignment decisions might have depended not only on recorded information but also on some criteria that are not recorded, e.g. the subjective opinion of a decision maker. We therefore assume that there is another set of covariates $Y = \{Y_1, \ldots, Y_q\}$, and that all variables on which previous assignments were based are included in $Y$. In this way, conditioned on the vector values of $Y$, the missing data indicator $S$ is independent of all of $X$ and $R_a$.

\[ S \perp X, R_a. \tag{2.9} \]

The missing data mechanism of $D_a$ can then be determined by the relationship between the variables in $Y$ and the variables recorded in $D_a$. We consider the following four cases.

2.3.3.1 Case I: $Y$ is independent of $X$ and $R_a$

According to (2.9), if all variables in $Y$ are independent of all of $X$ and $R_a$, the missing data indicator $S$ is also independent of $X$ and $R_a$,

\[ S \perp X, R_a. \tag{2.10} \]

In this case, Equation (2.6) holds, and the values of $R_a$ are MCAR. Moreover,

\[ f(R_a) = f(R_a|S = 0). \tag{2.11} \]

The distribution of the response among the whole target population is identical to that among respondents in $D_a$ (i.e. customers who received Action A). Applying the complete-case analysis will provide valid inferences about the target population.

Note that the definition does not imply the missingness to be random, but rather that the missingness is independent of all variables recorded in
the available data set. For example, suppose that a previous customer was
assigned Action A if they arrived at the branch before lunch time, and that
when a customer arrives is uncorrelated either to their recorded background
(i.e. variables in $X$) or to their response (i.e. $R_a$). Data in such a data set
are MCAR, and thus applying the complete-case analysis will result in unbi-
ased estimates for any quantities related to either the response, the recorded
background, or the relationship between the two for all customers. Although
the inferences drawn are unbiased, the power of the complete-case analysis is
diminished since the number of data points without missing values is never
larger than the total number of data points in an incomplete data set.

However, it is very unlikely that $Y$ is independent of all of $X$ and $R_a$. This
is because, as mentioned, recorded background information is usually taken
into consideration when making the decision about how to treat a customer.
Furthermore, it is also unlikely that variables used to make previous decisions
are uncorrelated to $R_a$. Consequently, it is rarely the case that the missing
data mechanism of the data set $D_a$ is MCAR.

2.3.3.2 Case II: $X$ can fully explain $Y$

If all variables used to make the previous assignment decisions can be ex-
plained by the recorded variables, Equation (2.9) implies

$$S|X \perp R_a.$$  \hspace{1cm} (2.12)

If this is the case, Equation (2.7) holds, and the values of response are
MAR. Moreover,

$$f(R_a|X) = f(R_a|X, S = 0).$$  \hspace{1cm} (2.13)

Within each subclass defined by the vector values of $X$, the distribution
form of $R_a$ among the target population is identical to that among respon-
dents. Therefore, one can use the respondent sample in $D_a$ to construct an
unbiased model for $E(R_a|X)$. 28
If conditioned on recorded information, any inference drawn from the respondent sample is valid. However, without such a constraint, the inference can be biased. For example, according to Equation (2.13), \( E(R_a|X) = E(R_a|X, S = 0) \), but the equation \( E(R_a) = E(R_a|S = 0) \) does not necessarily hold.

Suppose that whether a previous customer was assigned Action A or B depended only on their level of income so that the higher the income, the more likely the customer to receive Action A. If each customer’s income level is recorded, the values of the response are MAR. If the level of income is negatively correlated to the response (e.g. probability of default), the average response among customers who received Action A will be lower than that among all customers. In contrast, the average response among respondents with a certain level will be the same as that among all customers with the same income level. Applying the complete-case analysis will thus provide a valid inference for the response among all customers with a certain level of income.

It should be noted that if the previous assignment rule was deterministic, all customers with the same vector values of \( X \) would have been assigned the same action. The model for \( E(R_a|X) \) will be estimated from a sub-region of the full possible range of \( X \). Moreover, the model for \( E(R_a|X) \) and that for \( E(R_b|X) \) will be estimated from complementary regions. Estimating a model for either \( E(R_a|X) \) or \( E(R_b|X) \) thus involves extending estimates from a sub-region to a mutually exclusive one. Nevertheless, one can always extrapolate over the complementary region.

In contrast, if the previous assignment rule was probabilistic, the problem of estimating from sub-regions does not exist. The reason is that customers who are identical in \( X \) were randomly assigned to either action with the same probability \( \pi \), where \( 0 < \pi < 1 \). Therefore, models for both \( E(R_a|X) \) and \( E(R_b|X) \) are estimated from the full possible range of \( X \).
2.3.3.3 Case III: $\mathcal{X}$ cannot fully explain $\mathcal{Y} = \{\mathcal{X}, \mathcal{Z}\}$ and $\mathcal{Z}$ is independent of $\mathcal{R}_a$

When $\mathcal{X}$ cannot explain $\mathcal{Y}$ completely, we assume that there is a single variable $\mathcal{Z}$ so that $\{\mathcal{X}, \mathcal{Z}\}$ can fully explain $\mathcal{Y}$. In this way, (2.9) implies

$$S|\{\mathcal{X}, \mathcal{Z}\} \perp \mathcal{R}_a. \quad (2.14)$$

If $\mathcal{Z}$ is independent of $\mathcal{R}_a$, we have

$$S|\mathcal{X} \perp \mathcal{R}_a. \quad (2.15)$$

According to the above relationship, Equation (2.7) holds, and the values of the response are MAR. Similar to Case II, the complete-case analysis can be used to construct a model for $E(\mathcal{R}_a|\mathcal{X})$ without bias.

In this case, previous assignment decisions were made based on both $\mathcal{X}$ and $\mathcal{Z}$. Customers who were identical in $\mathcal{X}$ might be different in $\mathcal{Z}$. Therefore, even when the previous assignment rule was deterministic, customers who had the same vector $\mathcal{X}$ might have been assigned different actions. Consequently, the issue of estimating from sub-regions might not exist.

2.3.3.4 Case IV: $\mathcal{X}$ cannot fully explain $\mathcal{Y} = \{\mathcal{X}, \mathcal{Z}\}$ and $\mathcal{Z}$ is not independent of $\mathcal{R}_a$

We consider a case where $\mathcal{Y} = \{\mathcal{X}, \mathcal{Z}\}$ and $\mathcal{Z}$ is not independent of $\mathcal{R}_a$. Given the vector values of the recorded $\mathcal{X}$, the missing data indicator $S$ is still not independent of $\mathcal{R}_a$. In this case, conditioned on $\mathcal{X}$, the distribution

\footnote{Because $\mathcal{Z}$ is assumed to explain what cannot be explained by $\mathcal{X}$, $\mathcal{Z}$ is independent of $\mathcal{X}$. Since $\mathcal{Z}$ represents information that is not recorded, the distribution of $\mathcal{Z}$ is not given. For simplicity, we assume $\mathcal{Z}$ to be a single variable. However, such an assumption can result in loss of generality, and this will be further discussed in Section 6.6.}
of \( R_a \) among respondents is different from that among all customers

\[
f(R_a|\mathcal{X}) \neq f(R_a|\mathcal{X}, S = 0),
\]

and applying the complete-case analysis will result in a biased estimate for \( E(R_a|\mathcal{X}) \).

If the missing data mechanism is MNAR, regardless of whether conditioned on the observed information or not, inferences drawn from the complete-case analysis are biased. For example, suppose that whether a previous customer received Action A or B depended on both their level of income and the subjective opinion of the branch manager, which is not recorded. Suppose that among customers with the same income level, those more favoured by the branch manager were more likely to receive Action A. Moreover, if these favoured customers indeed have lower probability of default, then the responses are MNAR. Under this circumstance, the overall response among customers who received Action A is different from that among all customers. Moreover, the average response among respondents with a certain income level is also different from that among all customers with the same level of income.

Note that if \( Z \) is not independent of either \( R_a \) or \( R_b \), the complete-case analysis will result in a biased estimate for \( E(R_a - R_b|\mathcal{X}) \). To simplify the discussion, we will only take into account the circumstance where \( Z \) is not independent of both \( R_a \) and \( R_b \).

Note that the unrecorded variable is not necessarily un-recordable. If the unrecorded \( Z \) is recordable, one might obtain the distribution of \( Z \) from other sources. For example, the values of \( Z \) may be recorded for another group of customers. The availability of the distribution of \( Z \) can be helpful to adjust the bias resulting from the missing \( Z \), which we will further discuss in Chapter 6.
2.4 Further assumptions

2.4.1 Linear response functions

In order to focus on the missing data problem, we assume that the expected responses are linear functions of the predictor variables. (At the cost of increased complications, this assumption could be relaxed.)

- If the missing data problem is ignorable (e.g. Cases II and III), the responses can be considered as linear functions of the recorded \( \mathcal{X} \). That is

\[
\begin{align*}
    r_{ai} &= x_i \beta_{ra, x} + \varepsilon_{ai}, \\
    r_{bi} &= x_i \beta_{rb, x} + \varepsilon_{bi},
\end{align*}
\]  

(2.17) (2.18)

where \( x_i = [1 \ x_{i1} \ldots \ x_{ip}] \); \( x_{ji}, r_{ai}, \) and \( r_{bi} \) are the values of \( \mathcal{X}_j, \mathcal{R}_a \) and \( \mathcal{R}_b \) for a customer \( i \) respectively; \( \beta_{ra, x} \) and \( \beta_{rb, x} \) are \((1 + p)\) vectors of parameters; \( \varepsilon_{ai} \) and \( \varepsilon_{bi} \) are random errors that are assumed to follow normal distributions with means equal to zero and independent of \( \mathcal{X} \).

Because \( \varepsilon_a \) and \( \varepsilon_b \) have been assumed to be independent of \( \mathcal{X} \), the optimal decision function becomes

\[
E(r_{ai}|x_i) - E(r_{bi}|x_i) = x_i(\beta_{ra, x} - \beta_{rb, x}),
\]

(2.19)

where \( \beta_{ra, x} \) and \( \beta_{rb, x} \) are the parameters we wish to estimate.

- If the missingness is non-ignorable (e.g. Case IV), the responses are linear functions of both the recorded \( \mathcal{X} \) and the unrecorded \( \mathcal{Z} \). That is

\[
\begin{align*}
    r_{ai} &= x_i \beta_{ra, x} + z_i \beta_{ra, z} + \varepsilon_{ai}, \\
    r_{bi} &= x_i \beta_{rb, x} + z_i \beta_{rb, z} + \varepsilon_{bi},
\end{align*}
\]

(2.20) (2.21)

where \( z_i \) is the value of \( \mathcal{Z} \) of a customer \( i \); and \( \beta_{ra, z} \) and \( \beta_{rb, z} \) are single parameters.
Because all of $Z$, $\varepsilon_n$, and $\varepsilon_b$ have been assumed to be independent of $X$, the optimal decision function can be expressed as

$$E(r_{ai}|x_i) - E(r_{bi}|x_i) = x_i(\beta_{r_a,x} - \beta_{r_b,x}) + E(z_i|x_i)(\beta_{r_a,z} - \beta_{r_b,z}), \quad (2.22)$$

where $E(z_i|x_i) = E(Z)$ is the population mean of $Z$.

Note that the values of $Z$ for previous customers are not recorded, and thus the value of $E(Z)$ depends on the assumed distribution of $Z$. Because $X$ includes a constant term and $\beta_{r_a,x}$, $\beta_{r_a,z}$, $\beta_{r_b,x}$ and $\beta_{r_b,z}$ are not given, we can assume $Z$ to follow a distribution with $E(Z)$ and $Var(Z)$ equal to any constant value without loss of generality.

2.4.2 Linear and deterministic previous assignment function

We assume that whether a previous customer received Action A or B was based on a function called "the previous assignment function". In this thesis, we are interested in the case where the information about the previous assignment function is lost. To simplify the problem, we assume this function to be linear.

- If the missingness is ignorable (e.g. Cases II and III), we assume that the previous assignment function is a linear function of the recorded $X$, i.e.

$$y_i = x_i\beta_{g,x}. \quad (2.23)$$

- If the missingness is non-ignorable (e.g. Case IV), we assume that the previous assignment function is a linear function of both the recorded $X$ and the unrecorded $Z$, i.e.

$$y_i = x_i\beta_{g,x} + z_i. \quad (2.24)$$

Because the values of $y_i$, for $i = 1, ..., n$, and $\beta_{g,x}$ are not given, even if we set $Var(Z) = 1$, there is no loss of generality to further assume the coefficient for $Z$ to be 1.
We also assume that the previous assignment function is deterministic, thus there is no random error term. This assumption is generally applicable since most decisions made in consumer banking involve no randomness. We further assume that if \( y_i > c \), where \( c = 0 \) is a threshold, customer \( i \) was assigned Action A, otherwise he was assigned Action B.

\footnote{Because the values of \( g_i \), for \( i = 1, ..., n \), are not recorded, the vector values of \( \beta_{g,x} \) are unknown, and \( x \) includes a constant term, we can always assume \( c = 0 \) without loss of generality.}
Chapter 3

Existing adjustment methods

We described in the previous chapter how determining whether inferences drawn from the complete-case analysis are biased or not depends on whether the missing data problem is ignorable or non-ignorable. If the missing data problem is ignorable, estimates drawn from the complete-case analysis are unbiased. In contrast, if the missing data problem is non-ignorable, applying the complete-case analysis will result in biased estimates.

In the case of an ignorable missing data problem in general, although applying the complete-case analysis is expected to derive unbiased estimates, it might not always be possible to perform it. For example, consider a case where one wishes to discover the relationship between one response variable and \( p \) explanatory variables. If two or more explanatory variables are never jointly observed, there is no complete case (i.e. no data point without any missing values). In this case, the complete-case analysis cannot be implemented. In Section 3.1, we review the EM algorithm that can be applied to solve this problem.

Applying the EM algorithm to the ignorable missing data problem this thesis is concerned with is similar to applying the technique named single static imputation. In Section 3.2 we review a series of imputation techniques. We first review the concept of the single imputation technique. This
technique underestimates the variability of each data point. This disadvantage can be avoided by applying the single stochastic imputations technique. We also review the bootstrap resampling method which can be combined with the single stochastic imputation to derive variances for the estimates. Multiple stochastic imputation is another type of imputation method that one can apply. We review this method together with the data-augmentation algorithm, which generates data to be imputed.

If only the response variables is missing, and if the missing data problem is ignorable, one can apply the propensity score technique to draw inferences for the response variable. The propensity score technique has been widely used in epidemiology. In Section 3.3, we review this method and discuss how it can be applied to the ignorable missing data problem considered in this thesis.

The non-ignorable missing data problem discussed in this thesis is similar to the Type II Tobit problem. Several researchers have suggested solutions to this problem and we review two of the best known approaches: the two-step method proposed by Heckman (see Section 3.5) and the EM algorithm proposed by Little and Rubin (see Section 3.6). We also present how to extend these two approaches to solve the non-ignorable missing data problem on which this thesis focuses.

Although this thesis focuses on the missing data problem in behavior scoring, many characteristics of this problem are similar to those of the missing data problem in application scoring. Several researchers also have looked at the probit model (part of the Heckman's two-step method) and the maximum likelihood method (EM algorithm), and one can refer to [7], [4], and [12].

Note that we emphasize the case where the variable subject to being missing is continuous. Other methods have been suggested to solve a similar non-ignorable missing data problem when the partly missing variable (i.e. the response variable) is either binary [35] or categorical [36].
3.1 EM algorithm for ignorable missingness

3.1.1 ML method

Assume that the aim is to estimate parameters in a parametric model. Usually, this can be derived from the Maximum Likelihood (ML) method. As suggested by its name, this method obtains estimators by maximising a likelihood function. Let us formulate a model for a complete data set \( D \) with density \( f(D|\theta) \), indexed by an unknown vector of parameters \( \theta \). The likelihood function of \( \theta \) can be defined as any function of \( \theta \) proportional to \( f(D|\theta) \).

\[
L(\theta|D) \propto f(D|\theta). \quad (3.1)
\]

Because the likelihood (3.1) makes use of a complete data set, we term this likelihood a \textit{complete-data likelihood}. When the data set available is incomplete, \( D \) can be split into \( D_{\text{obs}} \) and \( D_{\text{mis}} \). The actual available information thus consists of the observed data \( D_{\text{obs}} \) and the missing data indicator matrix \( M \). We formulate the model for \( M \) using the probability distribution \( f(M|D, \psi) \), indexed by an unknown vector of parameters \( \psi \). A likelihood that concerns both the observed information and the missing data mechanism is termed an \textit{incomplete-data full likelihood}. This can be defined as any function of \( \theta \) and \( \psi \) proportional to \( f(D_{\text{obs}}, M|\theta, \psi) \).

\[
L(\theta, \psi|D_{\text{obs}}, M) \propto f(D_{\text{obs}}, M|\theta, \psi), \quad (3.2)
\]

where

\[
f(D_{\text{obs}}, M|\theta, \psi) = \int f(D_{\text{obs}}, D_{\text{mis}}, M|\theta, \psi) dD_{\text{mis}}
\]

\[
= \int f(D_{\text{obs}}, D_{\text{mis}}|\theta) f(M|D_{\text{obs}}, D_{\text{mis}}, \psi) dD_{\text{mis}}. \quad (3.3)
\]

Little and Rubin [27] showed that the inference for \( \theta \) based on the incomplete-data full likelihood (3.2) is identical to the one based on the \textit{incomplete-data}
likelihood

\[ \text{likelihood} \]

\[ L(\theta|D_{\text{obs}}) \propto f(D_{\text{obs}}|\theta), \quad (3.4) \]

which only makes use of \( D_{\text{obs}} \) and ignores the missing data mechanism, if both of the conditions below are true.

- The data are missing at random, i.e.

\[ f(M|D, \psi) = f(M|D_{\text{obs}}, \psi), \quad (3.5) \]

so that from Equation (3.3), the full likelihood (3.2) becomes

\[ L(\theta, \psi|D_{\text{obs}}, M) \propto f(D_{\text{obs}}|\theta)f(M|D_{\text{obs}}, \psi). \quad (3.6) \]

- The parameters \( \theta \) and \( \psi \) are distinct, so that from (3.4) and (3.6),

\[ L(\theta, \psi|D_{\text{obs}}, M) = L(\theta|D_{\text{obs}})L(\psi|M). \quad (3.7) \]

3.1.2 EM algorithm

Sometimes, explicit ML estimates can be difficult to obtain directly from incomplete data. In other words, ML estimates are sometimes easier to derive from the complete-data likelihood (3.1), than the incomplete-data likelihood (3.4). An intuitive solution to this problem is to:

1. based on estimates for the unknown parameters, impute the missing data and obtain an imputed data set;
2. from the imputed data set, derive a complete-data likelihood, e.g. (3.1);
3. based on the complete-data likelihood, re-estimate the unknown parameters by the standard ML method.

The above concept forms the EM algorithm. Each EM iteration consists of an E (Expectation) step, and a M (Maximisation) step.
E step at the $t^{th}$ iteration

The E step find the expected complete-data likelihood if $\theta$ were $\theta^{(t-1)}$: (see [27] p.168)

\[
Q(\theta|\theta^{(t-1)}) = \int f(\theta|D_{\text{obs}}, D_{\text{mis}}) f(D_{\text{mis}}|D_{\text{obs}}, \theta^{(t-1)}) dD_{\text{mis}}.
\] (3.8)

We interpret the above equation through a conventional way, and interpret the E step as imputing the missing values in the complete-data likelihood. The missing values are imputed using their expectation conditioned on the last updated parameter $\theta^{(t-1)}$. The observed and the imputed data form the expected complete-data likelihood

\[
Q(\theta|\theta^{(t-1)}) = E[L(\theta|D_{\text{obs}}, D_{\text{mis}})|\theta^{(t-1)}]
\] (3.9)

Note that in the EM algorithm, the E step imputes missing values in the complete-data likelihood, rather than those in the $D_{\text{mis}}$.

M step at the $t^{th}$ iteration

The M step obtains $\theta^{(t)}$ by maximising the expected complete-data likelihood (or log-likelihood) with respect to $\theta$

\[
Q(\theta^{(t)}|\theta^{(t-1)}) \geq Q(\theta|\theta^{(t-1)}), \text{ for all } \theta.
\] (3.10)

Dempster, Laird, and Rubin [8] proved that in each iteration the EM algorithm increases the likelihood $L(\theta|D_{\text{obs}})$. They also proved that if $L(\theta|D_{\text{obs}})$ is bounded, the sequence of $L(\theta^{(t)}|D_{\text{obs}})$ converges to a stationary value of $L(\theta|D_{\text{obs}})$.

The EM algorithm has been criticised for its possibly slow convergence and the fact that the M step is not always straightforward.
The convergence rate of the EM algorithm is linear to the proportion of observed information about $\theta$ in $L(\theta|D)$. Thus, the convergence rate can be painfully slow if the proportion of missing information is large. Algorithms that speed up the convergence can be found in [23] and [25].

When direct maximisation is difficult, some extensions of the EM can be applied. For example, the GEM (Generalised EM) algorithm, which replaces the original M step with a generalised one

$$Q(\theta^{(t)}|\theta^{(t-1)}) \geq Q(\theta|\theta^{(t-1)}), \text{ (not necessarily for all } \theta)$$  (3.11)

and the ECM algorithm, which replaces the original M step with two or more conditional maximisation steps.

3.1.3 Application to the ignorable missing data problem in credit scoring

Here we describe how the above EM algorithm concept can be applied to fit a least squares regression model for $R_a$ on $X = \{X_1, X_2, ..., X_p\}$, when some values of $R_a$ are missing. That is to use data set $D_a$ to estimate parameters in Equation (2.17). (Likewise, to use data set $D_b$ to estimate parameters in Equation (2.18).)

Because we have assumed $\varepsilon_{ai}$ to follow a normal distribution, this is equivalent to estimate parameters in:

$$r_{ai}|x_i, \beta_{r,x}, \sigma_a \sim N(x_i\beta_{r,x}, \sigma_a^2) \text{ for } i = 1, ..., n,$$  (3.12)

where $\sigma_a^2$ is the variance of $\varepsilon_a$.

In this case, a complete-data log-likelihood function of $\theta = \{\beta_{r,x}, \sigma_a\}$ can be any function of $\theta$ proportional to

$$\sum_{i=1}^{n} \log f(r_{ai}|x_i, \theta).$$  (3.13)
E step at the $t^{th}$ iteration

Because of the normality assumption in (3.12), the missing values in the log-likelihood function are the values of $R_a$ and $R_a^2$ among non-respondents.

If $r_{ai}$ is observed,

$$r_{ai}^{(t)} = r_{ai}$$

$$(r_{ai}^2)^{(t)} = r_{ai}^2.$$  

(3.14)  

(3.15)

If $r_{ai}$ is missing,

$$r_{ai}^{(t)} = x_i \beta_{ra,x}^{(t-1)}$$

$$(r_{ai}^2)^{(t)} = (x_i \beta_{ra,x}^{(t-1)})^2 + (\sigma_{a}^{(t-1)})^2.$$  

(3.16)  

(3.17)

M step at the $t^{th}$ iteration

1. regress $R_a^{(t)} = (r_{ai}^{(t)})$ on $X = (x_i)$, where $i = 1, ..., n$, to obtain $\beta_{ra,x}^{(t)}$;

2. set

$$(\sigma_{a}^{(t)})^2 = n^{(-1)} \sum_{i=1}^{n} [(r_{ai}^2)^{(t)} - 2r_{ai}^{(t)} x_i \beta_{ra,x}^{(t)} + (x_i \beta_{ra,x}^{(t)})^2].$$  

(3.18)

If the estimates converge at the $T^{th}$ iteration, estimators for $\beta_{ra,x}$ and $\sigma_{a}$ are $\beta_{ra,x}^{(T)}$ and $\sigma_{a}^{(T)}$ respectively. When the interest lies in $\beta_{ra,x}$ only but not $\sigma_{a}$, the above EM algorithm can be simplified so that only (3.16) in E step and the first M step are needed.

In fact, if the missing values are confined to the response variable, and if the missingness is ignorable, the non-respondents do not provide extra information about the unknown parameters. For example, if we set $\hat{\beta}_{ra,x}^{(0)} = \hat{\beta}_{ra,x}$, where $\hat{\beta}_{ra,x}$ is the least squares estimators derived from the respondent sample, the above iteration converge at once (see Section 3.2.1 for details). Thus, applying the above EM algorithm to the ignorable missing data problem
this thesis focuses on does not provide more information than applying the complete-case analysis.

3.2 Imputation methods for ignorable missingness

3.2.1 Single static imputation

Intuitively, if one can impute a proper value for each piece of missing data, the standard statistical tools can be used. The concept of imputation methods is to impute the missing data, and then perform analyses on the imputed data set.

Single static imputation consists of unconditional and conditional mean imputation. When the missing data mechanism is MCAR, the missing data can be imputed by the unconditional mean of observed values. Consider a case where $d_{ij}$, the value of the variable $j$ from unit $i$, is missing. The missing value can be imputed by the mean of all the observed values for the variable $j$. However, if the missingness is not completely at random, such imputation is inappropriate.

Take the ignorable missing data problem in a credit scoring data set as an example. If the missingness of responses depends on the values of $X$, applying unconditional mean imputation might not result in a plausible inference. But a valid inference can be obtained if we apply the regression imputation method. That is to impute each missing response $r_{ai}$ by

$$
\hat{r}_{ai} = x_i \hat{\beta}_{r_{ai}},
$$

(3.19)

where $x_i = [1 \ x_{i1} \ ... \ x_{ip}]$, $x_{ji}$ is the value of the $j^{th}$ variable in $X$ from the unit $i$, and $\hat{\beta}_{r_{ai}}$ is the least squares estimator obtained from the respondent sample.
After imputations, one may apply the standard least squares regression approach to the imputed data set and derive another estimator $\hat{\beta}_{r_a,x}$ for $\beta_{r_a,x}$. Note that such an approach is identical to applying the EM algorithm described in the previous section and setting $\beta_{r_a,x}^{(0)} = \hat{\beta}_{r_a,x}$.

It turns out that $\beta_{r_a,x}^{*} = \hat{\beta}_{r_a,x}$. This is because when setting $\beta_{r_a,x} = \hat{\beta}_{r_a,x}$, both terms in the following residual sum of squares function are minimised.

$$SS(\beta_{r_a,x}) = \sum_{i \in \Omega_{mis}} (r_{ai} - x_i \beta_{r_a,x})^2 + \sum_{i \in \Omega_{obs}} (r_{ai} - x_i \beta_{r_a,x})^2,$$

(3.20)

where $\Omega_{mis}$ denotes the sample of non-respondents and $\Omega_{obs}$ denotes the sample of respondents.

Since $\beta_{r_a,x}^{*} = \hat{\beta}_{r_a,x}$, single static imputation does not provide extra information to the ignorable missing data problem the thesis is concerned with.

From the above illustration, we can see clearly the main disadvantage of single imputation methods. Because the first term of Equation (3.20) equals to null when setting $\beta_{r_a,x} = \hat{\beta}_{r_a,x}$, one can see that the resulting residual sum of squares is underestimated. Such an underestimation is common among all single static imputation methods. In the following sub sections, we review the stochastic imputation technique that can resolve this issue.

### 3.2.2 Single stochastic imputation

Stochastic imputation methods impute the missing data by a draw from a predictive distribution, instead of imputing the missing data by either a conditional or an unconditional mean. For example, in a MCAR case, the missing $d_{ij}$ can be imputed by a draw from all the observed values of the $j^{th}$ variable.

Consider the credit scoring ignorable missing data problem, where the missingness depends on variables in $X$ only. A missing response can be replaced by a random draw from the responses of respondents with homoge-
neous vector values of $\mathcal{X}$. Furthermore, the regression imputation described in Section 3.2.1 becomes a stochastic imputation method by adding a residual term to each imputation, i.e. adjust Equation (3.19) into

$$r_{ai} = x_{i} \hat{\beta}_{r_{a},x} + \varepsilon_{ai},$$

(3.21)

where $\varepsilon_{ai}$ is randomly drawn from the residuals generated by the regression of $R_{a}$ on $\mathcal{X}$ among the respondents.

When a single stochastic imputation technique is applied, the variability of each imputed piece of data are already embedded in the imputed data set. Nevertheless, the uncertainty of estimates derived directly from the imputed data set is still underestimated. This is because each missing piece of data are imputed using a single value, each imputed value is treated as an observed one, and the uncertainty with respect to the missing data are ignored. As a result, more effort is required to derive the variances of estimates.

**Bootstrap re-sampling**

It has been suggested that re-sampling methods can be combined with single imputation methods so that appropriate estimates for variances can be obtained. Bootstrap re-sampling has been widely used, and thus we briefly look at how it can be applied. (Refer to [9] for details.)

Assume that the values of $p$ variables from $n$ independent units are recorded in a data set $D$. Suppose that we would like to make inferences for an unknown parameter $\theta$ based on $D$. However, some entries in $D$ are missing. The concept of applying the bootstrap re-sampling is as follows.

For $b = 1, \ldots, B$, where $B$ is called the bootstrap size:

- draw $n$ units from the original incomplete data set $D$ with replacement and obtain a bootstrap sample $D^{(b)}$;
- replace all missing values in $D^{(b)}$ using a single stochastic imputation
method, and obtain an imputed data set \( \hat{D}^{(b)} \);

- derive \( \hat{\theta}^{(b)} \) from \( \hat{D}^{(b)} \) using standard statistical methods.

The bootstrap estimate of \( \theta \) is obtained from:

\[
\hat{\theta}^{\text{(Boot)}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}. \tag{3.22}
\]

The bootstrap estimate of the variance of \( \hat{\theta}^{\text{(Boot)}} \) is derived by:

\[
\hat{\nu}^{\text{(Boot)}} = \frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\theta}^{(b)} - \hat{\theta}^{\text{(Boot)}} \right)^2. \tag{3.23}
\]

The algorithm that combines the single stochastic imputation with the bootstrap re-sampling is straightforward. However, Little and Rubin [27] pointed out that "in order to derive a valid inference, re-sampling methods require 200 or more different imputed data sets, with imputations based in each re-sampled data set, and transmitting this large set of re-sampled and imputed data sets to users may not be practical".

### 3.2.3 Multiple imputations

The method of *multiple stochastic imputations* was proposed by Rubin [44]. The idea is to create more than one imputed data sets independently, derive estimates from each imputed data set through standard statistics methods, and then combine the estimates.

Suppose that one wishes to draw inferences about \( \theta \) from an incomplete data set \( D \). First create more than one, say \( K \), imputed data sets, e.g. \( \hat{D}^{(1)}, \hat{D}^{(2)}, \ldots \hat{D}^{(K)} \), independently from \( D \). From each imputed data set \( \hat{D}^{(k)} \), obtain an estimator \( \hat{\theta}^{(k)} \) for \( \theta \) and its associated variance \( \hat{\nu}^{(k)} \). The combined
estimate is
\[ \hat{\theta}^{(MI)} = \frac{1}{K} \sum_{k=1}^{K} \hat{\theta}^{(k)} , \]  
(3.24)

which is a more efficient estimate than a single \( \hat{\theta}^{(k)} \).

The variability associated with the above estimate is
\[ \hat{\gamma}^{(MI)} = \hat{\gamma}^{(W)} + \left(1 + \frac{1}{K}\right) \hat{\gamma}^{(B)} , \]  
(3.25)

where \( 1 + \frac{1}{K} \) is an adjustment for finite \( K \), \( \hat{\gamma}^{(W)} \) is the within-imputation variance
\[ \hat{\gamma}^{(W)} = \frac{1}{K} \sum_{k=1}^{K} \hat{\gamma}^{(k)} , \]  
(3.26)

and \( \hat{\gamma}^{(B)} \) is the between-imputation variance
\[ \hat{\gamma}^{(B)} = \frac{1}{K - 1} \sum_{k=1}^{K} \left( \hat{\theta}^{(k)} - \hat{\theta}^{(MI)} \right)^2 . \]  
(3.27)

According to the definition for \( \hat{\gamma}^{(MI)} \), the proportion of missing information is approximately
\[ \lambda = \frac{\gamma}{1 + \gamma} , \]  
(3.28)

where
\[ \gamma = \frac{1 + \frac{1}{K} \hat{\gamma}^{(B)}}{\hat{\gamma}^{(MI)}} . \]  
(3.29)

Rubin [44] has shown that the efficiency of an estimate based on \( K \) imputed data sets relative to the estimate based on infinite sets is approximately \( \left(1 + \frac{\lambda}{K}\right)^{-\frac{1}{2}} \) (in units of standard deviations), where \( \lambda \) is defined in Equation (3.28). When 50% of the information is missing, an estimate based on \( K = 10 \) imputed data sets has a standard deviation that is about 2.5% wider than the estimate based on \( K = \infty \) imputed data sets. Therefore, unless the proportion of missing data are extraordinarily high, no more than 10 imputation sets are needed in practice.
Data-Augmentation algorithm

Several methods have been suggested to generate the imputed data for multiple stochastic imputations. We review one of the most popular methods, which makes use of the data-augmentation algorithm.

Suppose that the available data set consists of missing and observed data, e.g. \( D = \{ D_{\text{obs}}, D_{\text{mis}} \} \), and the aim is to impute the missing \( D_{\text{mis}} \) using the observed \( D_{\text{obs}} \). If there is a parametric model for the complete data, e.g. \( f(D_{\text{obs}}, D_{\text{mis}}|\theta) \) with a set of unknown parameters \( \theta \), drawing a set of random sample for \( D_{\text{mis}} \) requires the knowledge of not only \( D_{\text{obs}} \) but also \( \theta \). However, the value of \( \theta \) is unknown, and estimating it possibly requires the knowledge of both \( D_{\text{mis}} \) and \( D_{\text{obs}} \). Nevertheless, such a conflict can be resolved by the data augmentation algorithm [48].

The data augmentation algorithm is a Markov Chain Monte Carlo (MCMC) method. MCMC is a collection of techniques for drawing one or more values from a variable, which is typically multidimensional. This method is attractive when it is difficult to draw directly from the density of the target variable, but rather straightforward to draw each covariate in sequence. Given a starting value \( \theta^{(0)} \), the algorithm below defines a Markov Chain \( \{(D_{\text{mis}}^{(t)}, \theta^{(t)}), t = 1, 2, \ldots \} \).

At the \( t \)th iteration:

- **Imputation Step**: draw a random sample for \( D_{\text{mis}} \) from
  \[
  D_{\text{mis}}^{(t)} \sim f_D(D_{\text{mis}}|D_{\text{obs}}, \theta^{(t-1)}) .
  \]
  \[ (3.30) \]

- **Posterior Step**: draw a random sample for \( \theta \) from
  \[
  \theta^{(t)} \sim f(\theta|D_{\text{obs}}, D_{\text{mis}}^{(t)}) .
  \]
  \[ (3.31) \]

Under some general conditions, the chain will converge to the stationary distribution \( f(D_{\text{mis}}, \theta|D_{\text{obs}}) \). Thus, performing the above iteration until the
distribution of iterates converges will produce a proper draw for the missing data $D_{\text{mis}}$. However, because proper imputations must be independent, successive draws, which are correlated, cannot be used. For example, one should not take $D_{\text{mis}}^{(t)}$ to create an imputed data set, and $D_{\text{mis}}^{(t+1)}$ to create another. Instead, proper imputations can be created by taking every $N^{th}$ draw among the sequence, whereas the choice of $N$ has to be large enough so that the dependence among imputations is insignificant. Otherwise, proper imputations can be obtained through independently simulating $K$ (the number of imputed data sets to be created) sequences of draws with length $N$. Only the $N^{th}$ draw of each sequence is used as an imputation. The length $N$ also has to be large enough so that the $N^{th}$ draw of each sequence is substantially independent of the starting one. Regardless of which method, amongst the two mentioned here, is used, $N$ has to be large enough so that the distribution of the iterates converges to their stationary distribution. Schafer [46] has pointed out that because fewer burn-in iterations will be discarded, a single sequence method is more precise when only one parameter is of interest. However, if parallel computation is available, the multiple sequences method becomes more attractive because it requires less computation time. Moreover, various starting values can be used to assure the convergence.

Because MI (multiple imputations) requires fewer imputed data sets (e.g. 10 or less) than re-sampling methods do (e.g. 200 or more), the storage of imputed data sets is less of an issue for MI. However, if the data-augmentation algorithm is used, MI demands more computation. The theory behind MI is Bayesian and thus it can provide better inferences when the sample size is small. But if the size of the data set is large, re-sampling methods are favoured because they are less dependent on parametric modelling assumptions. Nevertheless, when the knowledge of the parameter model is sound, MI is favoured. More discussions regarding this issue can be found in the articles by Rubin [45], Fay [11], and Rao [37].

Nevertheless, if only estimates but not variances of estimates are taken
into consideration, all of these imputation methods provide no extra information to the ignorable missing data problem we wish to solve.

3.3 Propensity score for ignorable missingness

3.3.1 Propensity score

Propensity scores have been commonly used in epidemiology to compare outcomes among units that were not randomly assigned to the group treatment or the group control. Units in the group treatment receive treatments, and those in the group control do not. Suppose one wishes to evaluate the effectiveness of a new treatment, e.g. to evaluate whether the new treatment does increase the level of antibodies or not. A survey is conducted with $n$ patients non-randomly assigned the new treatment ($s = 0$) or placebo ($s = 1$). The assignment is based solely on a set of variables $X$, which are measured prior to the start of the trial. Note that since $X$ is measured before the treatment, the value of $X$ is unaffected by it. Because a patient can either receive the new treatment or a placebo, the available data set will take the form of Figure 2.2. In this example, the first six columns indicate variables in $X$, the last two columns indicate the level of antibodies measured after receiving the new treatment, and that measured after receiving a placebo respectively.

Because the distribution of $X$ is different between the treatment and the control group, if any variable in $X$ is correlated to the production of antibodies, these two groups are not comparable. For example, assume that age is one of the variables in $X$, and the younger the patient the more likely he/she is to receive a placebo. Suppose that younger people tend to produce more antibodies regardless of whether they receive treatments or not. If the levels of antibodies turn out to be higher among patients who received placebos, one should not draw the conclusion that the new treatment is ineffective.
Suppose that we match patients based on the value of $X$, i.e. we stratify so that patients within the same stratum are homogeneous in the vector values of $X$. The treatment and the control groups are then comparable within each stratum, and it is said that the treatment assignment is ignorable given $X$.

Rosenbaum and Rubin [40] have shown that if the treatment assignment is ignorable given $X$, it is also ignorable given any balancing score for $X$. A balancing score for $X$ is a summary or function, i.e. $b(X)$, of $X$ such that the treatment assignment and the covariates $X$ are independent given $b(X)$ [24]. The most complex and finest balancing score for $X$ is $X$ itself, and the simplest one is the propensity score:

$$e(x_i) = Pr(s_i = 0|x_i), \quad (3.32)$$

where the estimator $\hat{e}(x_i)$ of $e(x_i)$ can be derived from a logistic or a probit regression of $S$ on $X$.

Suppose that we match patients based on the value of their propensity score. It is then appropriate to compare, within each stratum, the outcomes (levels of antibodies) between the treatment and the control group. It should be stressed that the propensity scores $e(X)$, or any other balancing scores for $X$, only balance variables that are included in $X$. When the probabilities of receiving the new treatment depend on some variables that are not included in $X$, the propensity scores method will not provide valid inferences.

Given that the treatment is ignorable given $X$, the reason to match patients based on the value of $e(X)$ instead of directly on $X$ is the following. Suppose that the stratification is made so that only patients with the same vector values of $X$ are grouped in the same stratum. If the number of variables in $X$ is large, it is likely that there will be very few patients in each stratum. Thereafter, it is possible that there are some strata consisting only of patients who received the treatment. Therefore, there is no way to compare the survival rate between the treatment and the control group.

To overcome this problem, one would wish to match patients not with the
same but with similar vector values of $\mathcal{X}$. However, given the large number of variables in $\mathcal{X}$, it will be difficult to identify such a similarity. For this reason, matching on a single variable is preferred. Under such a circumstance, one can use the propensity score to stratify patients into, for example, $C$ strata so that the first stratum consists of patients with $0 < e(\mathcal{X}) < 1/C$, the second stratum consists of those with $1/C < e(\mathcal{X}) < 2/C$, etc. The outcomes between the treatment and the control group are then comparable within each stratum.

Moreover, the propensity scores can be used in conjunction with further model-based adjustment methods [41]. For example, one can use a log-linear model that relates patient outcomes to the treatment, the propensity score (or the index of a stratum $c = 1, \ldots, C$), and other key covariates from $\mathcal{X}$. In this way, the comparison can be made with emphasis on a few key variables and at the same time, while the less interesting variables in $\mathcal{X}$ are balanced.

There are several other matching methods, e.g. weighted Euclidean metric matching [43], optimal matching [38], and full matching [39]. A simulation study [15] suggested that, when the size of $\mathcal{X}$ is large, matching on the propensity scores outperforms others.

### 3.3.2 Application to the ignorable missing data problem in credit scoring

The propensity score can also be used to stratify previous customers in a credit scoring data set, if the missingness is ignorable given completely observed $\mathcal{X}$. In this way, customers with similar propensity score are expected to have homogeneous value of $\mathcal{R}_a$ (or $\mathcal{R}_b$).

However, propensity score matching is only applicable when the previous assignment function is probabilistic rather than deterministic. This is because, for example, when the estimator $\hat{e}(x_i)$ is derived from a probit analysis, this is equivalent to assuming that the previous assignment function to
be:
\[ g_i = x_i \beta_{g,x} + \varepsilon_{g_i}, \quad (3.33) \]
where \( \varepsilon_{g_i} \) is a random error and follows a normal distribution.

Note that if the previous assignment function is indeed deterministic, the following inequality
\[ 0 < Pr(s_i = a|x_i) < 1 \quad (3.34) \]
then suggests that previous assignments depend on variables other than those in \( X \). If this is the case, the missingness is no longer ignorable given \( X \), and matching with the propensity score \( e(x) \) will not provide valid inferences.

Consequently, due to our assumption about the deterministic assignment rule, the propensity score method is not applicable to the problem we wish to solve.

### 3.4 Type II Tobit model

The Type II Tobit model concerns a non-ignorable missing data problem, and has been widely discussed in econometrics. This model is very similar to the non-ignorable missing data problem we wish to solve. In this section, we describe the similarities and the differences between the Type II Tobit model and the problem we consider in the present thesis. Based on these, we can extend existing solutions to the Type II Tobit model to obtain a solution to our problem.

#### 3.4.1 Model description

The original Tobit Model [49] aims to estimate, from a random sample of size \( n \), the parameters in a standard regression model
\[ r_i = x_i \beta_{r,x} + u_i \text{ for } i = 1, ..., n, \quad (3.35) \]
where $x_i = [1 \ x_{i1} \ ... \ x_{ip} \ ]$; $x_{ji}$, $r_i$ and $u_{ri}$ are respectively the values of $X_j$, $R$ (e.g. expenditure on durable goods) and a random variable $U_r$ of the $i^{th}$ unit; and $\beta_{r,x}$ is a $(1 + p)$ vector of parameters.

It is assumed that, in the random sample, the vector values of $X$ are completely observed whereas those of $U_r$ are completely missing. Moreover, the original Tobit Model assumed $U_r$ to be independent of $X$ and

$$u_{ri} \sim N(0, \sigma_r^2) \text{ for } i = 1, ..., n.$$  \hspace{1cm} (3.36)

Note that because Equation (3.35) includes a constant term as a regressor, there is no loss of generality in assuming the mean of the completely unobserved variable $U_r$ to be zero.

Since there is no negative expenditure, the original Tobit Model assumes that only positive values of $R$ can be observed. The Type II Tobit Model is a more general model which assumes that whether the value of $R$ is observed or not depends on another covariate $G$. More specifically, $r_i$ is observed if $g_i > 0$, where $g_i$ is the value of $G$ of the unit $i$. In contrast, if $g_i \leq 0$, $r_i$ is missing. However, the value of $G$ is never observed, and is assumed to be linearly correlated to the completely observed $X$ and a completely missing variable $U_g$

$$g_i = x_i \beta_{g,x} + u_{gi} \text{ for } i = 1, ..., n,$$  \hspace{1cm} (3.37)

where $\beta_{g,x}$ is a $(1 + p) \times 1$ vector of parameters; $u_{gi}$ is the value of $U_g$ of unit $i$, and

$$u_{gi} \sim N(0, 1) \text{ for } i = 1, ..., n.$$  \hspace{1cm} (3.38)

The variable $U_g$ is also assumed to be independent of $X$. Note that there is also no loss of generality if we assume unit variance for $U_g$, since both values of $G$ and $U_g$ are completely unobserved, and the vector values of $\beta_g$ are not given.
3.4.2 Non-ignorable missingness

According to the above assumptions, we can obtain the following bivariate normal distribution

\[
\begin{pmatrix}
    r_i \\
    g_i
\end{pmatrix} | x_i, \theta, \psi \sim N \left( \begin{pmatrix}
    x_i \beta_{r,x} \\
    x_i \beta_{g,x}
\end{pmatrix}, \begin{pmatrix}
    \sigma^2_r & \rho \sigma_r \\
    \rho \sigma_r & 1
\end{pmatrix} \right), \text{ for } i = 1, \ldots, n.,
\]

where \( \theta = \{ \beta_{r,x}, \sigma_r \}, \psi = \{ \beta_{g,x}, \rho \} \).

Based on (3.39), Little and Rubin [27], p322 showed that the probability that \( r_i \) is missing, that is \( g_i \leq 0 \), can be calculated by

\[
Pr(g_i \leq 0 | r_i, x_i, \theta, \psi) = 1 - \Phi \left( \frac{x_i \beta_{g,x} + \rho \sigma_r^{-1} (r_i - x_i \beta_{r,x})}{\sqrt{1 - \rho^2}} \right), \tag{3.40}
\]

where \( \Phi(\cdot) \) is the cumulative distribution function of a standard normal distribution.

According to Equation (3.40), if \( \rho \neq 0 \) the data in Type II Tobit model are missing not at random. This is because the probability that \( r_i \) is missing depends on its own value, which is sometimes missing.

3.4.3 Comparison with the non-ignorable missing data problem in credit scoring

If we set

\[
\begin{align*}
    u_{gi} &= z_i, \tag{3.41} \\
    u_{ri} &= z_i \beta_{r,z} + \varepsilon_{ai}, \tag{3.42} \\
    \text{or } u_{ri} &= z_i \beta_{r,b} + \varepsilon_{bi}, \tag{3.43}
\end{align*}
\]

the response function and the previous assignment function in the Type II Tobit model (i.e. Equations (3.35) and (3.37)) become identical to those in
a credit scoring problem (i.e. Equations (2.20), (2.21) and (2.24)).

However, the Type II Tobit model assumes that both $U_r$ and $U_g$ follow normal distributions, whereas we only restrict the random errors $\varepsilon$, $\varepsilon_n$, and $\varepsilon_b$ to follow normal distributions. Thus solutions to the Type II Tobit model are not applicable unless we assume that the unrecorded variable $Z$ also follows a normal distribution.

3.5 Heckman’s two-step method for non-ignorable missingness

3.5.1 Different explanatory variables

Heckman[20][21][22] proposed a model to estimate the market wage $R$ for married women under the constraint that one can never observe the values of $R$ among non-working women. Since the market wages among non-working women are missing, the available data set is incomplete. Moreover, whether a woman $i$ decides to work or not depends on the value of $g_i$, which is the difference between her market wage and her asking wage. Only when $g_i > 0$, is a woman $i$ willing to take up employment thus allowing her market wage to be observed.

All the assumptions made by Heckman are the same as those made in the Type II Tobit Model except that the variables affecting the asking wages are assumed not to be exactly the same as those affecting the market wages. For example, a husband’s income might influence the asking wage of a woman, but not her market wage. Therefore, in the model proposed by Heckman, he assumes that $R$ is correlated to the completely observed covariates $X_r = \{X_{r1}, ..., X_{rp}\}$ whereas $G$ is correlated to another set of completely observed
covariates $\mathcal{X}_g = \{\mathcal{X}_{g1}, ..., \mathcal{X}_{gq}\}$. Moreover, he assumed that

$$r_i = x_{ri}\beta_{r,x} + u_{ri} \text{ for } i = 1, ..., n, \tag{3.44}$$
$$g_i = x_{gi}\beta_{g,x} + u_{gi} \text{ for } i = 1, ..., n, \tag{3.45}$$

$$\begin{pmatrix} r_i \\ g_i \end{pmatrix} \mid x_{ri}, \theta, x_{gi}, \psi \sim N \left( \begin{pmatrix} x_{ri}\beta_{r,x} \\ x_{gi}\beta_{g,x} \end{pmatrix}, \begin{pmatrix} \sigma_r^2 & \rho \sigma_r \\ \rho \sigma_r & 1 \end{pmatrix} \right), \text{ for } i = 1, ..., n, \tag{3.46}$$

where $\theta = \{\beta_{r,x}, \sigma_r\}$, $\psi = \{\beta_{g,x}, \rho\}$.

### 3.5.2 Heckman’s two-step method

From Equation (3.44), Heckman showed that the regression function of $\mathcal{R}$ among the respondents, whose $r_i$ can be observed, is

$$E(r_i | r_{ri}, g_i > 0) = x_{ri}\beta_{r,x} + E(u_{ri} | g_i > 0)$$
$$= x_{ri}\beta_{r,x} + E(u_{ri} | u_{gi} > -x_{gi}\beta_{g,x}), \text{ for } i = 1, ..., n. \tag{3.47}$$

Based on the above bivariate normal distribution assumption, the last part of Equation (3.47) can be interpreted as

$$E(u_{ri} | u_{gi} > -x_{gi}\beta_{g,x}) = \rho \sigma_r E(u_{gi} | u_{gi} > -x_{gi}\beta_{g,x}), \text{ for } i = 1, ..., n. \tag{3.48}$$

Since $U_g$ has been assumed to follow a standard normal distribution,

$$E(u_{gi} | u_{gi} > -x_{gi}\beta_{g,x}) = \lambda(x_{gi}\beta_{g,x}), \text{ for } i = 1, ..., n, \tag{3.49}$$

where $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ is the inverse of Mill’s ratio, where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the probability density function and the cumulative distribution function of a standard normal distribution. Thus, Equation (3.47) becomes

$$E(r_i | r_{ri}, g_i > 0) = x_{ri}\beta_{r,x} + \rho \sigma_r \lambda(x_{gi}\beta_{g,x}), \text{ for } i = 1, ..., n. \tag{3.50}$$
Let \( \hat{u}_{gi} = \lambda(x_{gi}\hat{\beta}_{g,x}) \) denote the conditioned expected value of \( U_g \) for unit \( i \). Consider the case where the parameter \( \beta_{g,x} \) can be estimated so that the value of \( \hat{U}_g \) can be derived for each unit. In the respondent sample, regressing \( R \) on \( X_r \) and \( \hat{U}_g \) derives unbiased least squares estimators for \( \beta_{r,x} \) and \( \rho \sigma_r \) in Equation (3.50). Note that only the product \( \rho \sigma_r \) as a whole can be estimated, rather than \( \rho \) and \( \sigma_r \) separately.

Because \( G|x_i \sim N(x_i\beta_{g,x},1) \), Heckman suggested using probit analysis to estimate the value of \( \beta_{g,x} \). Probit analysis [13] has been commonly applied to the study of binary response relations, especially when the error term is assumed to follow a normal distribution. Therefore, Heckman’s two-step method is the following:

- Step 1: Apply probit analysis to the whole sample and obtain estimator \( \hat{\beta}_{g,x} \) of \( \beta_{g,x} \) in Equation (3.45), and calculate \( \hat{u}_{gi} = \lambda(x_{gi}\hat{\beta}_{g,x}) \) for each respondent \( i \).

- Step 2: In the respondent sample, regress \( R \) on \( X_r \) and \( \hat{U}_g \) to derive unbiased estimators \( \hat{\beta}_{r,x} \) of \( \beta_{r,x} \) and \( \hat{\delta} \) of \( \delta = \rho \sigma_r \) in Equation (3.50).

This method has been widely used for its simplicity. As shown above, the two-step method only involves probit analysis and least squares regression. However, Heckman’s two-step method is bounded by the normality assumption. Other similar approaches have been proposed, but all restricted to a specific type of distribution form (e.g. Olsen [34] proposed a similar method and assumed the distribution form of \( U_g \) to be uniform).

If the normality assumption is true, one can derive unbiased estimates from Heckman’s two-step method. Nevertheless, the two-step estimators can suffer from the collinearity problem. The collinearity problem occurs when predictors in a regression model are linearly correlated [50]. If this is the case, the variances of estimators will be inflated, hence the estimation becomes inefficient. Consider the second step of Heckman’s two-step method which regresses \( R \) on \( X_r \) and \( \hat{U}_g \) in the respondent sample. The collinearity problem appears if \( X_r \) and \( \hat{U}_g \) are highly correlated in the respondent sample.
Because \( \hat{u}_{gi} = \lambda(x_{gi}\hat{\beta}_{g,x}) \), Heckman avoids this problem by assuming \( \mathcal{X}_r \) to be not identical to \( \mathcal{X}_g \). Nevertheless, if variables in \( \mathcal{X}_r \) are highly correlated or even identical to those in \( \mathcal{X}_g \), \( \mathcal{X}_r \) will also be highly correlated to \( x_g\hat{\beta}_{g,x} \). Consequently, whether the collinearity problem exists relies on the non-linearity between \( x_g\hat{\beta}_{g,x} \) and \( \lambda(x_{gi}\hat{\beta}_{g,x}) \), which is unappealing [26]. However, if the normality assumption is true, estimators from such a computationally efficient method remains unbiased, and this is why Heckman’s two-step method has attracted much attention.

### 3.5.3 Application to the non-ignorable missing data problem in credit scoring

Heckman’s two-step method was originally proposed to estimate the response to only one action, whereas we are interested in responses to two actions. Here we describe how his method can be extended to solve our problem.

- **Step 1:** We have set \( s_i = 0 \) if customer \( i \) received Action A, and \( s_i = 1 \) if received Action B. In this way, the probit analysis can be performed to estimate \( \hat{\beta}_{g,x} \) for \( \beta_{g,x} \). Thereafter,
  - among customers who received Action A, set
    \[
    \hat{z}_i = \mathbb{E}(z_i | z_i > -x_i\hat{\beta}_{g,x}) = \lambda(x_i\hat{\beta}_{g,x});
    \tag{3.51}
    \]
  - among customers who received Action B, set
    \[
    \hat{z}_i = \mathbb{E}(z_i | z_i < -x_i\hat{\beta}_{g,x}) = -\lambda(-x_i\hat{\beta}_{g,x}).
    \tag{3.52}
    \]

- **Step 2:**
  - among customers who received Action A, regress \( R_a \) on \( \mathcal{X} \) and \( \hat{Z} \), yielding \( \hat{\beta}_{ra,x} \) and \( \hat{\beta}_{ra,z} \).
  - among customers who received Action B, regress \( R_b \) on \( \mathcal{X} \) and \( \hat{Z} \), yielding \( \hat{\beta}_{rb,x} \) and \( \hat{\beta}_{rb,z} \).
The estimation for the optimal decision function becomes
\[
\hat{E}(r_{ai}|x_i) - \hat{E}(r_{bi}|x_i) = x_i(\hat{\beta}_{ra,x} - \hat{\beta}_{rb,x}) + E(z_i|x_i)(\hat{\beta}_{ra,z} - \hat{\beta}_{rb,z}) \quad (3.53)
\]

Heckman's two-step method assumes that \( U \) follows the standard normal distribution and independent of all recorded variables. Applying Heckman's method to our problem implies \( E(z_i|x_i) = E(z_i) = 0 \). In this way, the estimation for the optimal decision function becomes
\[
\hat{E}(r_{ai}|x_i) - \hat{E}(r_{bi}|x_i) = x_i(\hat{\beta}_{ra,x} - \hat{\beta}_{rb,x}) \quad (3.54)
\]

A new customer \( i \) will be assigned Action A if \( \hat{E}(r_{ai}|x_i) - \hat{E}(r_{bi}|x_i) = x_i(\hat{\beta}_{ra,x} - \hat{\beta}_{rb,x}) > 0 \); otherwise will be assigned Action B.

Note that because the recorded variables in the previous assignment function are identical to those in both response functions, the variances of estimators from the two-step method are likely to be inflated. Moreover, if the unrecorded variable \( Z \) does not follow a normal distribution, the resulting estimation will be biased.

3.6 EM algorithm for non-ignorable missingness

3.6.1 ML method

As described in Section 3.1, when only an incomplete data set is available, the incomplete-data full likelihood, which takes into account the missing data mechanism, is proportional to
\[
f(D_{obs}, M|\theta, \psi) = \int f(D_{obs}, D_{mis}|\theta) f(M|D_{obs}, D_{mis}, \psi) dD_{mis}. \quad (3.55)
\]
However, according to the definition of MNAR, if the data are missing not at random

\[ f(M|D_{obs}, D_{mis}, \psi) \neq f(M|D_{obs}, \psi). \] (3.56)

This implies that we cannot always split Equation (3.55) (the joint distribution of the observed data and the missing data indicator matrix) into a marginal density function \( f(D_{obs}|\theta) \) for \( D_{obs} \), and a marginal probability mass function \( f(M|D_{obs}, \psi) \) for \( M \). If this is the case, \( f(D_{obs}, M|\theta, \psi) \) might not be proportional to \( f(D_{obs}|\theta) \). Therefore, the full likelihood \( L(\theta, \psi|D_{obs}, M) \) might not be proportional to the likelihood \( L(\theta|D_{obs}) \), which ignores the missing data mechanism. Consequently, if the data are missing not at random, one should make use of a more complicated likelihood function, i.e. \( L(\theta, \psi|D_{obs}, M) \).

### 3.6.2 EM algorithm

We have discussed that when it is difficult to derive ML estimates from a likelihood that only includes \( D_{obs} \), a complete-data likelihood in conjunction with the EM algorithm can be applied. The E step of the EM finds the conditional expectation for the missing values in the complete-data likelihood \( L(\theta|D) \). However, if the data are missing not at random, one should take into account the missing data mechanism when estimating the missing values. The EM algorithm for the MNAR problem is thus required to accommodate both the parameter \( \theta \) in \( f(D|\theta) \) and the parameter \( \psi \) in \( f(M|D, \psi) \). Consequently, in a MNAR case, one should use a complete-data full likelihood, which is proportional to the joint distribution of \( \theta \) and \( \psi \)

\[ L(D, M|\theta, \psi) \propto f(\theta, \psi|D, M). \] (3.57)

Based on the above complete-data full likelihood, the EM algorithm at the \( t^{th} \) iteration becomes
• E Step: using the last updated parameters $\theta^{(t-1)}$ and $\psi^{(t-1)}$, derive the expected likelihood function:

$$Q(\theta, \psi|\theta^{(t-1)}, \psi^{(t-1)}) = E[L(\theta, \psi|D_{\text{obs}}, D_{\text{mis}}, M)|\theta^{(t-1)}, \psi^{(t-1)}].$$ (3.58)

• M step: find the $\theta^{(t)}$ and $\psi^{(t)}$ that maximise $Q(\theta, \psi|\theta^{(t-1)}, \psi^{(t-1)})$, i.e.

$$Q(\theta^{(t)}, \psi^{(t)}|\theta^{(t-1)}, \psi^{(t-1)}) \geq Q(\theta, \psi|\theta^{(t-1)}, \psi^{(t-1)}) \quad \text{for all } \theta \text{ and } \psi,$$

or in the Generalised EM (GEM) algorithm, choose a $\theta^{(t)}$ and a $\psi^{(t)}$ such that

$$Q(\theta^{(t)}, \psi^{(t)}|\theta^{(t-1)}, \psi^{(t-1)}) \geq Q(\theta^{(t-1)}, \psi^{(t-1)}|\theta^{(t-1)}, \psi^{(t-1)}).$$ (3.60)

### 3.6.3 EM algorithm for Type II Tobit Model

Here we review the EM algorithm suggested by Little and Rubin [27], p323 that provides estimates for the Type II Tobit Model.

**E step at the $t^{th}$ iteration**

If $r_i$ is observed, $y_i > 0$,

$$g_i^{(t)} = x_i \beta_g^{(t-1)} + \lambda(x_i \beta_g^{(t-1)}),$$

$$g_i^{2(t)} = 1 + (x_i \beta_g^{(t-1)})^2 + x_i \beta_g^{(t-1)} \lambda(x_i \beta_g^{(t-1)}),$$

$$r_i^{(t)} = r_i,$$

$$r_i^{2(t)} = r_i^2,$$

$$r_i g_i^{(t)} = r_i x_i \beta_g^{(t-1)} + r_i \lambda(x_i \beta_g^{(t-1)}),$$

where $\lambda(\cdot)$ is the inverse of Mill’s ratio.
If \( r_i \) is missing, \( g_i \leq 0, \)

\[
\begin{align*}
   g_i^{(t)} &= x_i \beta_{r,x}^{(t-1)} - \lambda(-x_i \beta_{g,x}^{(t-1)}), \\
   (g_i^2)^{(t)} &= 1 + (x_i \beta_{r,x}^{(t-1)})^2 - x_i \beta_{g,x}^{(t-1)} \lambda(-x_i \beta_{g,x}^{(t-1)}), \\
   r_i^{(t)} &= x_i \beta_{r,x}^{(t-1)} - \rho^{(t-1)} \sigma_r^{(t-1)} \lambda(-x_i \beta_{g,x}^{(t-1)}), \\
   (r_i^2)^{(t)} &= (x_i \beta_{r,x}^{(t-1)})^2 + (\sigma_r^{(t-1)})^2 - \lambda(-x_i \beta_{g,x}^{(t-1)})(2 \mu_r^{(t-1)} - \rho^{(t-1)} \sigma_r^{(t-1)} x_i \beta_{g,x}^{(t-1)}) \\
   (r_i g_i)^{(t)} &= x_i \beta_{r,x}^{(t-1)}(x_i \beta_{g,x}^{(t-1)} - \lambda(-x_i \beta_{g,x}^{(t-1)}) + \rho^{(t-1)} \sigma_r^{(t-1)}). 
\end{align*}
\]

M step at the \( t^{th} \) iteration

The M step re-estimates the unknown parameters by

1. Regressing \( G^{(t)} = (g_i^{(t)}) \) on \( X = (x_i) \), where \( i = 1, ..., n \), yielding \( \beta_{g,x}^{(t)} \).
2. Regressing \( R^{(t)} = (r_i^{(t)}) \) on \( X = (x_i) \) and \( G^{(t)} = (g_i^{(t)}) \), where \( i = 1, ..., n \), yielding coefficients \( \beta_{r,x}^{(t)} \) for \( X \) and \( \delta^{(t)} \) for \( G \), and residual variance \( (\sigma_{r,g}^{(t)})^2 \).
3. Setting

\[
\begin{align*}
   \beta_{r,x}^{(t)} &= \beta_{r,x}^{(t)} + \delta^{(t)} \rho^{(t)}, \\
   (\sigma_r^{(t)})^2 &= (\sigma_{r,g}^{(t)})^2 + (\delta^{(t)})^2, \\
   \rho^{(t)} &= \frac{\delta^{(t)}}{\sigma_r^{(t)}}.
\end{align*}
\]

The simplicity of the above EM algorithm also relies on the normality assumption for both \( U_r \) and \( U_g \). If the actual distributions of \( U_r \) and \( U_g \) are not normal, the estimates will be biased. Criticisms have been made that if the normality assumption is incorrect, there is no way of distinguishing between the bias resulting from the non-ignorable missingness and that resulting from an incorrect distribution assumption. Discussion regarding the sensitivity of a model to un-testable assumptions can be found in Copas and Li [5].
Furthermore, as when applying EM to any incomplete data set, when the proportion of the missing information is large, the iterations will take a very long time to converge.

3.6.4 Simplified EM algorithm for Type II Tobit Model

We would like to point out that the EM algorithm proposed by Little and Rubin can be simplified.

- The E Steps can be simplified and only keep (3.61), (3.63), (3.66), and (3.68). This is because among the M steps, none of $g_i^2$, $r_1^2$, or $r_i g_i$ is required.

- We can also skip (3.72) and (3.73) in the third M step. Because after we simplify the E Steps, all parameters required in the remaining E steps are: $\beta_{g,x}$, $\beta_{r,x}$, and the product of $\rho$ and $\sigma$, but not $\rho$ and $\sigma$ individually.

- Because $g$ is a function of $X$ and $U_g$, and $U_g$ is assumed to be independent of all variables in $X$, we can merge the second M step and (3.71) into:

  Regress $R(t) = (r_i^{(t)})$ on $X = (x_i)$ and $U_g^{(t)} = (u_{gi}^{(t)})$, where $i = 1, \ldots, n$, yielding coefficients $\beta_{r,x}^{(t)}$ for $X$ and $\delta^{(t)}$ for $U_g$.

  (When $g_i > 0$, $u_{gi}^{(t)} = \lambda(x_i \beta_{g,x}^{(t-1)})$; when $g_i \leq 0$, $u_{gi}^{(t)} = -\lambda(-x_i \beta_{g,x}^{(t-1)})$.)

- According to Equation (3.73), we can replace all of $\rho^{(t)} \sigma_{r}^{(t)}$ by $\delta^{(t)}$.

Note that the E steps and the M steps for the missing values and the parameters in the previous assignment function are not affected by those in the response function. Therefore, we can split the simplified EM algorithm into two: one for the previous assignment function, and the other for the response function.
An EM algorithm for the previous assignment function

- E step at the $t^{th}$ iteration
  
  If $r_i$ is observed, $g_i > 0$,
  \begin{align}
  u_{gi}^{(t)} &= \lambda(x_i\beta_{g,x}^{(t-1)}), \\
  g_i^{(t)} &= x_i\beta_{g,x}^{(t-1)} + u_{gi}^{(t)}. \tag{3.74}
  \end{align}
  
  If $r_i$ is missing, $g_i \leq 0$,
  \begin{align}
  u_{gi}^{(t)} &= -\lambda(-x_i\beta_{g,x}^{(t-1)}), \\
  g_i^{(t)} &= x_i\beta_{g,x}^{(t-1)} + u_{gi}^{(t)}. \tag{3.76}
  \end{align}

- M step at the $t^{th}$ iteration
  
  Regress $G^{(t)} = (g_i^{(t)})$ on $X = (x_i)$, where $i = 1, ..., n$, yielding $\beta_{g,x}^{(t)}$.

An EM algorithm for the response function

- E step at the $t^{th}$ iteration
  
  If $r_i$ is observed, $g_i > 0$,
  \begin{equation}
  r_i^{(t)} = r_i. \tag{3.78}
  \end{equation}
  
  If $r_i$ is missing, $g_i \leq 0$,
  \begin{equation}
  r_i^{(t)} = x_i\beta_{r,x}^{(t-1)} + \hat{u}_{gi}\delta^{(t-1)}, \tag{3.79}
  \end{equation}
  
  where $\hat{u}_{gi}$ is the estimator for $u_{gi}$ derived from the EM algorithm for the previous assignment function.

- M step at the $t^{th}$ iteration
  
  Regress $R^{(t)} = (r_i^{(t)})$ on $X = (x_i)$ and $\hat{U}_g = (\hat{u}_{gi})$, where $i = 1, ..., n$, yielding $\beta_{r,x}^{(t)}$ and $\beta_{r,z}^{(t)}$. 

64
Note that if one takes estimator $\hat{\beta}_{g,x}$ estimated from the probit analysis as the initial guess for $\beta_{g,x}$, the EM algorithm for the previous assignment function converges at once. Moreover, using the estimated $\hat{u}_{gi}$ (for $i = 1, \ldots, n$) as observed values in an EM algorithm implies applying the EM method to an ignorable missing data problem, where only the response variable could be missing. As discussed in Section 3.1.2, the estimators obtained from such an EM algorithm are identical to the least squares estimators obtained from the respondent sample. Thus if one takes the estimators $\hat{\beta}_{r,x}$ and $\hat{\beta}_{r,z}$ derived from Heckman's method as initial guess, the EM algorithm for the response function also converges at once. Therefore, the EM algorithm suggested by Little and Rubin and Heckman's two-step method provide the same estimations for the Type II Tobit model. Nevertheless, the two-step method is more computationally efficient.

3.6.5 Application to the non-ignorable missing data problem in credit scoring

If we assume that the unrecorded $Z$ follows the standard normal distribution, we can perform the EM algorithm for the previous assignment function to obtain estimator $\hat{z}_i$ for each missing $z_i$. Thereafter, the EM algorithm for the response function can be extended into:

- **E step at the $t^{th}$ iteration**

  If customer $i$ was assigned Action A, $g_i > 0$,

  \[ r_{ai}^{(t)} = r_{ai}, \]  
  \[ r_{bi}^{(t)} = x_i' \beta_{r,x}^{(t-1)} + \hat{z}_i \beta_{r,z}^{(t-1)}. \]  

  \(^1\text{Replace all } u_{ai}^{(t)} \text{ by } z_i^{(t)}.\)
If customer $i$ was assigned Action B, $g_i \leq 0$,

\[
\begin{align*}
    r_{ai}^{(t)} &= x_i \beta_{ra,x}^{(t-1)} + z_i \beta_{ra,z}^{(t-1)}, \\
    r_{bi}^{(t)} &= r_{bi}.
\end{align*}
\]

(3.82) (3.83)

- M step at the $t^{th}$ iteration

- Regress $\mathcal{R}_a^{(t)} = (r_{ai}^{(t)})$ on $X = (x_i)$ and $\hat{Z} = (\hat{z}_i)$, where $i = 1, \ldots, n$, yielding $\beta_{ra,x}^{(t)}$ and $\beta_{ra,z}^{(t)}$.

- Regress $\mathcal{R}_b^{(t)} = (r_{bi}^{(t)})$ on $X = (x_i)$ and $\hat{Z} = (\hat{z}_i)$, where $i = 1, \ldots, n$, yielding $\beta_{rb,x}^{(t)}$ and $\beta_{rb,z}^{(t)}$.

If the above EM algorithm converges at the $T^{th}$ iteration, $\beta_{ra,x}^{(T)}$, $\beta_{ra,z}^{(T)}$, $\beta_{rb,x}^{(T)}$ and $\beta_{rb,z}^{(T)}$ are the estimators for $\beta_{ra,x}$, $\beta_{ra,z}$, $\beta_{rb,x}$ and $\beta_{rb,z}$ respectively. Since we assume $E(Z) = 0$ and $Z$ is independent of all recorded variables, the optimal decision function estimated from the above algorithm can be interpreted as

\[
\hat{E}(r_{ai}|x_i) - \hat{E}(r_{bi}|x_i) = x_i (\beta_{ra,x}^{(T)} - \beta_{rb,x}^{(T)}).
\]

(3.84)

According to the above estimation, a new customer $i$ should be assigned Action A if $\hat{E}(r_{ai}|x_i) - \hat{E}(r_{bi}|x_i) = x_i (\beta_{ra,x}^{(T)} - \beta_{rb,x}^{(T)}) > 0$, otherwise should be assigned Action B.

An EM algorithm can be computationally inefficient if large amount of data are missing. The above EM algorithm is thus likely to take a long time to converge if most previous customers received the same action. For example, if most customers received Action A, the E steps and the M steps for $\mathcal{R}_b$, $\beta_{rb,x}$ and $\beta_{rb,z}$ might not converge in a short time. Nevertheless, if we set estimators derived using extension from Heckman’s method to be the initial guess, the above EM algorithm converges at once. Therefore, when $Z$ is assumed to follow a normal distribution, the method extended from Heckman’s suggestion is preferred.
In this chapter, we reviewed existing adjustment methods for both ignorable and non-ignorable missing data problems. The EM algorithm and imputation methods for ignorable missing data problems were originally proposed to solve the problem when it is difficult to perform the complete-case analysis. However, such an issue does not exist in the problem considered in this thesis, and making use of incomplete cases (non-respondents) does not provide extra information. Moreover, the well-known propensity score method is inapplicable to our problem due to the deterministic assignment assumption. Nevertheless, applying the complete-case analysis to the ignorable missing data problem we are concerned with is sufficient to provide unbiased estimates.

Applying the complete-case analysis results in biased estimation when the data are missing not at random. Extensions of solutions to the Type II Tobit model are only applicable when the unrecorded \( Z \) indeed has a normal distribution. Albert and Chib [3] also proposed a Bayesian approach to solve a similar problem which make use of the tools of Gibbs Sampling [14]. Nevertheless, their proposed method is also restricted to the assumption. Other researchers have proposed solutions to problems that are similar to the Type II Tobit model but assumed that the unrecorded variable follows a distribution other than normal. For example, Olsen proposed a method that assume the unrecorded variable to follow a uniform distribution. However each existing solution was designed for a specific distribution. Due to the fact that the distribution of \( Z \) varies from case to case (see detailed discussion on Section 5.2), there is no single existing adjustment method that is applicable to all cases. Therefore, the following chapters focus on the non-ignorable missing data problem.
Chapter 4

Simulation design

We have discussed the fact that applying the complete-case analysis to the non-ignorable missing data problem on which we focus in this thesis will result in biased estimates. However, it is only appropriate to apply extensions from existing methods when one assumes that the unrecorded $Z$ follows a specific distribution that is consistent with the one the applied method assumes. In Sections 5.1 and 5.2 we will illustrate this insufficiency by applying the complete-case analysis or extensions of the existing methods to simulated data sets.

Later, in Chapter 6, we will propose adjustment methods that are applicable regardless of the distribution one assumes for $Z$. In Chapter 7, we will introduce adaptive adjustment methods that make use of information from new customers. The suitability of these proposed methods will also be examined using the same sets of simulated data. In this chapter, we describe how these simulated data sets are obtained.

4.1 Variables and parameters used

Data sets used in this thesis are simulated based on the assumptions we made in Section 2.4. According to Section 2.4, the problem we wish to
solve concerns a set of recorded variables and an unrecorded one, response functions for Actions A and B, and the previous assignment function. In this section, we describe the variables and parameters used in the simulations.

4.1.1 Explanatory variables

According to Section 2.4, both the response functions and the previous assignment function contain a set of recorded explanatory variables and a single unrecorded variable. For simplicity, we assume that there is only one recorded explanatory variable $X_1$, whose values are drawn from the standard normal distribution.

The values of the unrecorded $Z$ are drawn with replacement from one of the following six different sample pools: a pool of samples drawn from the standard normal distribution, and five pools of samples drawn from five different standardised credit scoring variables $V_1, V_2, V_3, V_4$ and $V_5$. Each sample pool contains 4000 data points and the values of the five credit scoring variables are collected from a real credit scoring data set. The histograms of these credit scoring data points are shown in Figures 4.1, 4.2, 4.3, 4.4 and 4.5.

4.1.2 Parameters in the response functions

According to the assumptions in Section 2.4, we define the two response functions respectively as

$$r_{ai} = \beta_{ra,x_0} + x_1i\beta_{ra,x_1} + z_i\beta_{ra,z} + \varepsilon_{ai}, \quad (4.1)$$

$$r_{bi} = \beta_{rb,x_0} + x_1i\beta_{rb,x_1} + z_i\beta_{rb,z} + \varepsilon_{bi}. \quad (4.2)$$

Because we are interested in the difference between the two response vari-

---

1Standardised variables are used so that the effect from each variable on either previous assignments or responses is controlled by the respective coefficient.
Figure 4.1: The histogram of $V_1$ used in our simulations.

Figure 4.2: The histogram of $V_2$ used in our simulations.

Figure 4.3: The histogram of $V_3$ used in our simulations.
Figure 4.4: The histogram of $V_1$ used in our simulations.

Figure 4.5: The histogram of $V_5$ used in our simulations.
ables rather than in each response variable individually, we let \( \omega_i \) represent the difference between \( r_{ai} \) and \( r_{bi} \) so that

\[
\omega_i = r_{ai} - r_{bi} = (\beta_{r_{ai}x_0} - \beta_{r_{bi}x_0}) + x_{1i}(\beta_{r_{ai}x_1} - \beta_{r_{bi}x_1}) + z_i(\beta_{r_{ai}z} - \beta_{r_{bi}z}) + (\varepsilon_{ai} - \varepsilon_{bi})
\]

\[
= \beta_{\omega,x_0} + x_{1i}\beta_{\omega,x_1} + z_i\beta_{\omega,z} + \varepsilon_{wi}.
\]

(4.3)

This equation represents the optimal decision function. In order to generate most profit, one should assign Action A to a customer \( i \) if \( \omega_i > 0 \), and assign Action B otherwise.

For simplicity, we set \( \beta_{\omega,x_0} = 0 \) and \( \beta_{\omega,z} = 1 \). In this way, the parameter \( \beta_{\omega,x_1} \) determines the correlation between \( \omega \) and \( X_1 \), as well as the correlation between \( \omega \) and \( Z \). We allow \( \beta_{\omega,x_1} \) to vary over six different values, i.e. 1, 2, 3, 4, 5, and 6. The values of random error \( \varepsilon_{wi} \) are drawn from normal(0,0.1^2), so that the effect of \( \varepsilon_{\omega} \) on \( \omega \) is smaller than the effect of \( X_1 \) or \( Z \) on \( \omega \). We further set

\[
r_{ai} = \frac{1}{2} \omega_i, \tag{4.4}
\]

\[
r_{bi} = -\frac{1}{2} \omega_i. \tag{4.5}
\]

### 4.1.3 Parameters in the previous assignment function

According to the assumptions in Section 2.4, we define the previous assignment function to be

\[
g_i = \beta_{g,x_0} + x_{1i}\beta_{g,x_1} + z_i, \tag{4.6}
\]

and a previous customer \( i \) to have received Action A if \( g_i > 0 \), and Action B otherwise.

In the above equation, the parameter \( \beta_{g,x_1} \) determines the effect of the recorded variable \( X_1 \) and the unrecorded variable \( Z \) on the previous selection. Similarly to the values of \( \beta_{\omega,x_1} \), we allow the value of \( \beta_{g,x_1} \) to vary over 1, 2,
Table 4.1: Parameters Values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{w, x_1}$</td>
<td>1, 2, 3, 4, 5, 6.</td>
</tr>
<tr>
<td>$\beta_{g, x_1}$</td>
<td>1, 2, 3, 4, 5, 6.</td>
</tr>
<tr>
<td>$Pr(S = 0)$</td>
<td>0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.</td>
</tr>
</tbody>
</table>

3, 4, 5 and 6. After fixing $\beta_{g, x_1}$, the value of $\beta_{g, x_0}$ determines the proportion of previous customers that received Action A/B. The value of $\beta_{g, x_0}$ in each data set is set so that $Pr(S = 0)$, i.e. the proportion of previous customers who received Action A, varies over 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9.

The parameters values used to create different data sets are summarised in Table 4.1. The parameters used to construct the data sets are known with certainty. In this way, application of statistical methods to these simulated data sets illustrates clearly the level of accuracy of these methods.

In practice, the values of parameters $\beta_{g, x_1}$ and $\beta_{w, x_1}$ are unknown and need to be estimated. Thus we are less interested in any relationship between the performance of a given method and the values of these parameters. At the time an analysis is carried out, the value of $Pr(S = 0)$ is known. We will therefore only discuss in detail the simulation results in respect of different levels of $Pr(S = 0)$.

4.2 Simulation details

According to the previous section, there are 324 (6 different values of $\beta_{w, x_1}$ x 6 different values of $\beta_{g, x_1}$ x 9 different values of $Pr(S = 0)$) different sets of parameters. Each simulated data point $i$ consists of:

\begin{itemize}
  \item $\beta_{w, x_1}$ and $\beta_{g, x_1}$ that are greater than or equal to 1. This is because $\beta_{g, x_1} < 1$ and $\beta_{w, x_1} < 1$ imply that recorded information can only explain to a very limited extend how previous assignments were made and what an optimal decision should be. If this is the case, without information about the value of $\beta$ for each previous customer, it is impossible to obtain reasonably good estimates for either the previous assignment function or the optimal decision function.
\end{itemize}
• the value of $x_{hi}$ drawn from the standard normal distribution;

• the value of $z_i$ drawn from either the standard normal distribution, $V_1$, $V_2$, $V_3$, $V_4$, or $V_5$;

• the value of $r_{ai}$ and $r_{bi}$ calculated based on the value of $\beta_{a,x_1}$;

• the value of $g_i$ calculated based on the value of $\beta_{g,x_1}$ and $Pr(S = 0)$;

• the value of $s_i$, where $s_i = 0$ if $g_i > 0$ and $s_i = 1$ if $g_i \leq 0$.

Which values of $X_i$ and $Z$ are drawn is determined by a random number chosen. To ensure that the unique attributes of any given set of random numbers do not affect our conclusion, we apply each set of parameters to 60 different random data sets, yielding 19,440 ($324 \times 60$) separate data sets. Further, we take the following steps to guarantee that differences among parameters and methods, rather than differences among data sets, are the unambiguous causes of differences in outcome across our design:

• The same 60 sets of random numbers are used to generate the data sets for each set of designed parameters.

• Different methods are evaluated using the same 19,440 data sets.

Each simulated data set consists of 5200 data points: 4200 of the data points are used as training samples (4000 of them are considered as information on previous customers and the remaining 200 are information on newly existing customers), and 1000 of the data points are used as testing samples. When evaluating the performance of methods that only make use of the previously collected data set (e.g. the complete-case analysis, extensions of existing methods, and the adjustment methods proposed in Chapter 6), the 4000 data points are used. When evaluating methods that further make use of newly collected data (e.g. the method proposed in Section 6.2 that generates empirical sample pools, and the adaptive models proposed in Chapter 7), the 200 data points are also used. When applying statistical methods on
training samples to estimate the optimal decision function, only the values of $X_1$, $S$, and $R_a$ or $R_b$ (subject to the value of $S$) are used. In each testing sample, only the values of $X_1$ are used, and the details regarding how we use the testing sample to evaluate each method are described in the next section.

4.3 Evaluation

When evaluating the performance of a statistical method, we apply the method to each simulated training sample and obtain estimates for parameters in the optimal decision function. According to the estimated parameters $\hat{\beta}_{\omega,x_0}^{md}$ and $\hat{\beta}_{\omega,x_1}^{md}$ (where $md$ represents the method being evaluated), we can calculate the value of $\hat{E}_{\omega}^{md}(\omega|x_{1i})$ for each data point $i$ in the testing sample using the following equation:

$$\hat{E}_{\omega}^{md}(\omega|x_{1i}) = \hat{\beta}_{\omega,x_0}^{md} + x_{1i}\hat{\beta}_{\omega,x_1}^{md}.$$ (4.7)

The effectiveness of each method is judged by the difference between $\hat{E}_{\omega}^{md}(\omega|x_{1i})$ and $E(\omega|x_{1i})$, where

$$E(\omega|x_{1i}) = \beta_{\omega,x_0} + x_{1i}\beta_{\omega,x_1},$$ (4.8)

$\beta_{\omega,x_0}$ and $\beta_{\omega,x_1}$ are the actual values used in each simulated data set.

The difference between $\hat{E}_{\omega}^{md}(\omega|x_{1i})$ and $E(\omega|x_{1i})$ is measured using the weighted Euclidean distance between $\hat{E}_{\omega}^{md}(\omega|X_1)$ and $E(\omega|X_1)$ among the testing sample:

$$d_{md} = \sqrt{\frac{\sum_{i=1}^{1000} (\hat{E}_{\omega}^{md}(\omega|x_{1i}) - E(\omega|x_{1i}))^2}{\text{Var}(E(\omega|X_1))}}.$$ (4.9)

The smaller the weighted Euclidean distance, the better the method.  

---

Footnote: We use the weighted Euclidean distance instead of the ordinary Euclidean distance to
In order to examine the insufficiency of the complete-case analysis, we look at the value of $d^{cc}$ (where $cc$ represents the complete-case analysis). The values of $d^{cc}$ across different simulated data sets are shown and discussed in Section 5.1.

To examine the advantage of applying a given adjustment method rather than the complete-case analysis, we calculate

$$
\Delta^{adj} = d^{cc} - d^{adj},
$$

where $adj$ represents the given adjustment method.

The larger the value of $\Delta^{adj}$, the better the given adjustment method performs. When $\Delta^{adj} > 0$, the estimate for the optimal decision function derived from the given adjustment method is better than that derived from the complete-case analysis. On the contrary, when $\Delta^{adj} \leq 0$, applying the given adjustment method is not worthwhile.

The performance of the extension of Heckman's method, that of the method we propose in order to make use of empirical distributions (see Chapter 6), and that of the proposed adaptive methods (see Chapter 7) are evaluated by looking at the respective values of $\Delta^{adj}$ in all simulated data sets. The values of $\Delta^{adj}$ for these adjustment methods are shown and discussed in Sections 5.2, 6.5, and Chapter 7.

Note that when an adaptive method is applied, one updates the values of $\hat{\beta}_{w,z0}^{md}$ and $\hat{\beta}_{w,z1}^{md}$ whenever a newly existing customer's information becomes available (see Chapter 7 for details). Therefore, when evaluating an adaptive method, we recalculate the values of $d^{adj}$ and $\Delta^{adj}$ whenever the values of $\hat{\beta}_{w,z0}^{md}$ and $\hat{\beta}_{w,z1}^{md}$ are updated. In this way, one can observe the change of performance of an adaptive method as more and more information on newly existing customers become available.

---

measure the performance of each method. This is because the larger the value of $|\beta_{x,z1}|$, the larger the ordinary Euclidean distance without implying worse performance. On the other hand, the weighted Euclidean distance does not depend on the scale of measurement, and thus is preferred.
Chapter 5

Key characteristics of the problem

We discussed in Section 2.3 that when the responses are MNAR, applying the complete-case analysis will result in biased estimates. It has been pointed out that if the interest lies in responses to only one action and if the responses from most units can be observed, the bias is relatively small and can be ignored. Nevertheless this is not the case when responses to more than one action are taken into account. In Section 5.1, we discuss in detail that the bias results from the complete-case analysis should never be ignored in the non-ignorable missing data problem we are concerned with.

Each existing solution to the non-ignorable missing data problems makes a specific distribution assumption (e.g. normal distribution) for the unrecorded variable. However, we consider the unrecorded variable as information that cannot be explained by recorded ones. The variables which are recorded vary from case to case, and the unrecorded variable that affects the previous assignment also varies from case to case. Thus the assumed distribution for the unrecorded variable should also vary. In Section 5.2, we discuss the disadvantage of applying extensions from existing solutions to the problem focused on in this thesis.
Another key characteristic of the missing data problem we are interested in is that the missingness does not result from self-selection. For example, whether a previous customer received Action A or B was not decided by the customer themselves. For this reason, data collected after the start of analysis are MAR, and we discuss this in detail in Section 5.3.

5.1 The complete-case analysis is insufficient

In the literature on the non-ignorable missing data problem, most attention has been devoted to estimating responses to "one" action using an incomplete data set. For example, the original Tobit model has been used to estimate the expenditure on durable goods, when the amount of expenditure is observed only if it is positive (but not negative). Similarly, Heckman's method was originally proposed to estimate the wage rates of married women, whilst only the wage rate of working (but not non-working) women can be observed. In the case of application scoring, a lending institution wishes to estimate applicants' response to the acceptance, when only responses from accepted (but not rejected) applicants can be observed.

However if responses to only "one" action are taken into account, and if most responses can be observed, the bias results from the missing responses can be ignored. Little [26] argued that 'If the proportion of unselected cases is small, then the adjustment for selection is also small, and the fact that the difference between the means of the selected and unselected cases is poorly estimated is of little consequence.' Crook and Banasik [6] mentioned that 'when the rejection rate is not so large, the scope for improving a model parametrised only on those accepted appears to be very small.' Following such an argument, if most consumers did spend some money on durable goods, if most married women did work outside their family, and if most applicants were accepted, applying the complete-case analysis would lead to estimates that are close to the unbiased ones.

Nevertheless in the more general problem we are concerned with, re-
sponges to “two” actions are taken into consideration. If most previous customers received Action A (or B), applying the complete-case analysis is likely to provide an estimate for the response to Action A (or B) that is close to unbiased, whilst the estimate for the response to Action B (or A) might be extremely biased. Since the interest lies in the difference between these two response variables, estimates for responses to both actions are equally important. In this case, the complete-case analysis is insufficient. In this section, we examine this argument using the simulated data sets described in Chapter 4.

From Figure 5.1, we can compare the decision function estimated using the complete-case analysis with the actual optimal decision function when the values of $Z$ are drawn from either the standard normal distribution, or from $V_1$, $V_2$, $V_3$, $V_4$ or $V_5$ respectively. Figure 5.1 shows the relationship between the degree of bias and the proportion of the previous customers that received either action. When $Z$ follows a symmetric distribution, the bias is larger if customers for whom the values of $Z$ are on one of the tails received one action and the rest received the other. However, when the distribution of $Z$ is skewed to the right, in our simulated data sets, the bias is larger if those with values of $Z$ around the tail (the minority) received one action, and those with values of $Z$ around the mode (the majority) received the other.

In order to explain this phenomenon in more detail, we consider the following three cases where $Z$ follows a distribution that is skewed to the right:

1. 90%/10% of the previous customers received Action A/B.
2. 50%/50% of the previous customers received Action A/B.
3. 10%/90% of the previous customers received Action A/B.

Based on our simulation design, customers who received Action A tended to have larger values of $Z$. Consequently, in Case 1 and 2, the distribution forms of $Z$ among customers who received Action A, among previous
Figure 5.1: The performance of the complete-case analysis. The value of $d_{cc}$ (see Equation (4.9)), where "cc" represents the complete-case analysis, for different levels of $Pr(S = 0)$. Each graph represents average results among 2160 (6 different $\beta_{x,1} \times$ 6 different $\beta_{g,x1} \times$ 60 different random seeds) datasets.

(a) The values of $Z$ are drawn from a standard normal distribution.

(b) The values of $Z$ are drawn from V1.

(c) The values of $Z$ are drawn from V2.

(d) The values of $Z$ are drawn from V3.

(e) The values of $Z$ are drawn from V4.

(f) The values of $Z$ are drawn from V5.
customers who received Action B, and among all customers are similar, i.e. all are concentrated around the mode. In Case 3, although the distribution form of $Z$ among customers who received Action B is similar to the overall one, i.e. both are concentrated around the mode, the distribution form of $Z$ among previous customers who received Action A is significantly different, i.e. spread around the tail.

Hence, in Case 1 and 2, if the complete-case analysis is used, both the estimated response functions for Action A and B are similar to the unbiased ones. In these two cases, the estimation for the optimal decision function derived from the complete-case analysis are thus relatively close to the unbiased one. However, in Case 3, the estimate for the response function for Action A could be extremely different from the unbiased one, as could the estimate for the optimal decision function.

In practice, if previous customers were not evenly assigned to the available actions, and if the distribution form of the unknown variable is skewed, it is more likely that those with a value of $Z$ around the tail would have received one action, and those with a value of $Z$ around the mode would have received the other. Thus, regardless of whether $Z$ is symmetric or skewed, the bias resulting from the complete-case analysis is likely to be more severe when the majority of the previous customers received the same action. As a result, if one wish to compare responses to two actions, regardless of how previous customers were assigned, the complete-case analysis is insufficient and a proper adjustment method is needed.

5.2 The distribution of $Z$ varies from case to case

According to the definition in Chapter 2, $Z$ represents information that cannot be explained by the recorded $X$, and that is correlated to the response variables. Under this definition, the distribution of $Z$ can vary from case to
Figure 5.2: The performance of the Heckman's two-step method. The value of $\Delta^{\text{heck}}$ (see Equation 4.10), where "heck" represents Heckman's two-step method, for different levels of $Pr(S = 0)$. Each graph represents average results among 2160 (6 different $\beta_{z,x_1} \times 6$ different $\beta_{y,x_1} \times 60$ different random seeds) data sets.
As has been pointed out, the performance of an adjustment method for non-ignorable missing data problems depends on whether the underlying distribution assumption for the unrecorded variable is consistent with the actual distribution form. Since each existing solution has been designed for a specific distribution form of \( Z \), there is no existing solution that is applicable to all cases in credit scoring. In this section, we use the simulated data sets described in Chapter 4 to confirm such an argument. We apply the extension of Heckman’s two-step method \(^1\), which assumes that \( Z \) follows a normal distribution, to six groups of simulated data sets: one with \( Z \) following a normal distribution, and the others with \( Z \) following a skewed distribution (\( V_1 \), \( V_2 \), \( V_3 \), \( V_4 \) or \( V_5 \)).

We illustrate advantages/disadvantages of applying this adjustment method using the values of \( \Delta^{\text{heck}} \). The larger the value of \( \Delta^{\text{heck}} \) implies the more effective it is to apply the extension of Heckman’s method than to apply the complete-case analysis. Moreover, negative \( \Delta^{\text{heck}} \) suggests that the adjustment method generates even more bias. The values of \( \Delta^{\text{heck}} \) among simulated cases are shown in Figure 5.2.

In order to examine the performance of an adjustment method in more detail, we should take into account the insufficiency of the complete-case analysis. We thus compare the value of \( d^{cc} \) (see Figure 5.1) with the corresponding \( \Delta^{\text{heck}} \). From Figures 5.1 and 5.2, we can observe that an adjustment method that assumes \( Z \sim N(0,1) \) is effective when the distribution form of \( Z \) is indeed normal or is close to normal (e.g. when the values of \( Z \) are drawn from \( V_1 \)). On the contrary, when the actual distribution of \( Z \) is skewed, such an adjustment method is not as helpful and can be harmful.

For example, when the values of \( Z \) are drawn from a relatively skewed distribution (\( V_2 \), \( V_3 \), \( V_4 \) or \( V_5 \)), and when only less than 50% of the customers

\(^1\)As discussed in Section 3.6, estimations derived from Little and Rubin’s EM algorithm can be identical to that from Heckman’s method. We thus only apply the extension of Heckman’s method.
received Action A, the values of the $d^{cc}$ are significantly high whilst the values of $\Delta^{heck}$ are not. This implies that applying Heckman's method is insufficient especially when applying the complete-case analysis results in significant bias and when an adjustment method is needed most.

As discussed in the previous section, when the distribution form of $Z$ is skewed, and when customers were not evenly assigned to both actions, it is more likely that customers with a value of $Z$ around the mode received one action and those with a value of $Z$ around the tail received the other. This implies that it is more likely that $Pr(S = 0) < 0.5$ in our simulation design. Thus, when the distribution of $Z$ is skewed, it is more likely that neither the complete-case analysis nor any adjustment method that makes a normality assumption is adequate.

From simulation results, we can confirm that it is ineffective to apply an adjustment method that assumes that $Z$ follows a distribution which is significantly different from the actual one. The distribution of $Z$ can vary from case to case, adjustment methods that are only applicable to a certain type of distribution are thus insufficient. In Chapter 6, we will introduce adjustment methods that allow one to assume any distribution for $Z$.

5.3 Newly collected data are MAR

In the literature, MNAR problems often arise from self-selection. For example, the decision whether a married woman works outside her family is made by herself rather than by any econometricians. Whether to spend money on durable goods is also decided by consumers themselves rather than by any econometricians. In most cases, any variable that was unrecorded and affected the decision in the past is still unrecorded and affects the decision after an analysis is carried out. Although updated information, e.g. the newly collected information about married women and consumers, is available, the MNAR problem still exists within newly collected data sets.
In contrast, in the case we are concerned with, the MNAR problem does not result from self-selection. For example, whether to increase a given customer's credit limit or whether to sell off the debt from a given customer is decided by a decision maker in the bank rather than by the customer themselves. Once an analysis starts, a decision maker is restricted to use only variables that can be observed at that time they construct a decision function. In this way, any variable that caused the MNAR problem in the previously collected data set will no longer affect the assignment decision. Whether a response in the newly collected data set is observed or not depends only on recorded variables, the data are MAR rather than MNAR, and the missingness is ignorable.

In the following chapters, we will propose approaches that make use of the newly collected data set to adjust the bias that results from non-ignorable missingness in the previously collected data set.
Chapter 6

Flexible adjustment approaches

The non-ignorable missing data problem on which we focus results from an unrecorded variable $Z$. We have mentioned that this unrecorded variable is not necessarily un-recordable. If $Z$ is recordable, one might be able to obtain the empirical density function of $Z$ from other sources. In this case, it is only reasonable to apply an adjustment method that assumes that $Z$ follows the empirical distribution. Although several adjustment methods have been proposed for different distributions of $Z$, each existing method is designed for one specific distribution, and it is very likely that none of the existing methods assume that $Z$ follows a distribution that is identical to the empirical one. Therefore in this chapter, we introduce adjustment methods that enable one to assume any type of distribution for $Z$ or even without assuming a specific distribution form for $Z$.

In general, using a data set with a non-ignorable missing data problem to construct an optimal decision rule involves the following two steps:

1. estimate the parameters in the previous assignment function so as to obtain an estimator $\hat{Z}$ for the unrecorded $Z$ of each previous customer;

2. make use of the estimated $\hat{Z}$ to obtain estimators for parameters in the response functions.
In Section 6.1, we illustrate an EM algorithm that makes use of the empirical density function to derive an estimator for $\beta_{y,x}$ and the expected value of $Z$ for each previous customer. However when $Z$ is never recorded, the empirical density function of $Z$ is not available. If this is the case, we suggest making use of information contained in the newly collected data set to generate empirical samples of $Z$. (The details are described in Section 6.2.) Using the generated empirical samples, one may estimate the density function of $Z$ so that the method proposed in Section 6.1 can be applied. Nevertheless, we introduce another method: a semi-GEM algorithm, which is more straightforward to implement. In Section 6.3 we detail how the semi-GEM algorithm makes use of empirical samples to estimate the previous assignment function. Later in Section 6.4, we illustrate how to derive estimates for the response functions using $\hat{Z}$ derived from the methods proposed in Sections 6.1 and 6.3.

The performance of the proposed method is examined using the simulated data sets. The simulation results are shown and discussed in Section 6.5. Note that the methods proposed in this chapter are constrained by a univariate assumption for $Z$. The disadvantage of this constraint is further discussed in Section 6.6.

6.1 An EM algorithm that uses an empirical density function to estimate the previous assignment function

From each previous customer $i$, we can observe their values of $x$, and to which action they were assigned, i.e. $s_i = 0$ or $s_i = 1$. If we assume that the unrecorded variable $Z$ follows a normal distribution, part of the EM algorithm suggested by Little and Rubin can be applied to estimate parameters in the previous assignment function (see Section 3.6.4). From this we can derive the expected value of $Z$ for each previous customer.
If the actual distribution of $Z$ is not normal, the resulting estimates will be very different from the unbiased ones. Nevertheless, the concept of the EM algorithm suggested by Little and Rubin can be extended so that one can estimate the previous assignment function and consequently the missing $Z$ using the empirical density function.

6.1.1 The E step

In the previous assignment function, the values of $g_i$ and $z_i$ of each previous customer $i$ are missing. The E step thus replaces these missing values with their expectations. Within the algorithm proposed by Little and Rubin, the expectations are calculated based on the assumption that $Z$ follows the standard normal distribution. We suggest relaxing this constraint and interpreting the E step as follows.

At the $t$th iteration,

- if customer $i$ received Action A, $s_i = 0$,

$$z_i^{(t)} = E(z_i | x_i, s_i = 0) = E(z_i | z_i > -x_i \beta_{g,x}^{(t-1)})$$

$$= \frac{\int_{-\infty}^{\infty} z f(z) dz}{\int_{-\infty}^{\infty} f(z) dz}$$

$$g_i^{(t)} = E(g_i | x_i, z_i^{(t)})$$

$$= x_i \beta_{g,x}^{(t-1)} + z_i^{(t)}; \quad (6.2)$$

- if customer $i$ received Action B, $s_i = 1$,

$$z_i^{(t)} = E(z_i | x_i, s_i = 1) = E(z_i | z_i \leq -x_i \beta_{g,x}^{(t-1)})$$

$$= \frac{\int_{-\infty}^{\infty} z f(z) dz}{\int_{-\infty}^{\infty} f(z) dz}$$

$$g_i^{(t)} = E(g_i | x_i, z_i^{(t)})$$

$$= x_i \beta_{g,x}^{(t-1)} + z_i^{(t)}; \quad (6.4)$$
where $f(\cdot)$ is empirical density function of $Z$ and can be any form of pdf.

### 6.1.2 The M step

The M step then makes use of the estimated values to derive estimators for parameters in the previous assignment function. In the M step proposed by Little and Rubin, the estimator for $\beta_{g,x}$ is obtained from regressing the estimated $G$ on $X$. Such an approach is equivalent to maximising the following likelihood function, subject to the vector value of $\beta_{g,x}$,

$$l(\beta_{g,x}) = \sum_{i} \log f(g_{i}^{(t)} - x_{i} \beta_{g,x}),$$ \hfill (6.5)

when $f(\cdot)$ is the pdf of the standard normal distribution.

Here we propose a more general M step that removes the normality constraint on $f(\cdot)$. That is, at the $t^{th}$ iteration,

$$\beta_{g,x}^{(t)} = \arg_{\beta_{g,x}} \sum_{i} \log f(g_{i}^{(t)} - x_{i} \beta_{g,x}),$$ \hfill (6.6)

where $f(\cdot)$ is the empirical density function of $Z$ and can be any form of pdf.

If the above E step and M step converge at the $T^{th}$ iteration, set $\hat{z}_{i} = z_{i}^{(T)}$. Therefore, the estimator $\hat{z}_{i}$ can be used in the method proposed in Section 6.4 to derive the response functions.

### 6.2 Generate empirical samples from the newly collected data

We have introduced an EM algorithm that estimates the previous assignment function using the empirical density function of $Z$. In this section, we
consider a case where the pdf of $Z$ is not given. We propose an approach to obtain empirical samples of $Z$ from the newly collected data set.

The main difference between the previously collected data and the newly collected data results from the different assignment rules. Whether a previous customer received Action A or B depended on both recorded and unrecorded information. In contrast, whether we assign Action A or B to a newly existing customer\(^1\) depends only on recorded information.

The fact that the decision rule applied to newly existing customers depends only on recorded information is appealing. This enables one to generate from the newly collected data set empirical samples of $Z$. In this section, we describe how such a sample pool can be generated.

### 6.2.1 Regression residuals obtained from applying the complete-case analysis to the newly collected data

Among newly existing customers who received Action A, we regress $R_a$ on $X$ and obtain a complete-case estimator $\hat{\beta}_{ra,x}$ for $\beta_{ra,x}$. Likewise, among newly existing customers who received Action B, regress $R_b$ on $X$ and obtain an estimator $\hat{\beta}_{rb,x}$ for $\beta_{rb,x}$.

Define $\xi_a$ and $\xi_b$ to be residuals from the regressions. According to the response function we assumed for Action A and B (see Equations (2.20) and (2.21)),

- if customer $i$ received Action A,

$$
\xi_{ai} = r_{ai} - x_i \hat{\beta}_{ra,x} i ;
$$

\(^1\)We define a newly existing customer to be a customer who entered the system after the analysis starts and has been assigned an action, and thus is different from a new customer who has not been assigned any action.
• if customer $i$ received Action B,

$$\xi_{bi} = r_{bi} - x_i \beta_{rb,x}.$$  \hspace{1cm} (6.8)

Let $sd(\xi_a)$ represents the standard deviation of $\xi_a$s, and $sd(\xi_b)$ represents that of $\xi_b$s. Set

$$\xi'_a = \frac{\xi_a}{sd(\xi_a)},$$  \hspace{1cm} (6.9)

and

$$\xi'_b = \frac{\xi_b}{sd(\xi_b)}.$$  \hspace{1cm} (6.10)

We can take $\xi' = \{\xi'_a, \xi'_b\}$ as empirical samples. In the following section, we explain the adequacy of using these residuals as empirical sample of $Z$.

### 6.2.2 Adequacy of using residuals as empirical samples

Due to the fact that newly collected data are MAR, $\hat{\beta}_{ra,x}$ and $\hat{\beta}_{rb,x}$ are unbiased estimators for $\beta_{ra,x}$ and $\beta_{rb,x}$ respectively. When the number of newly existing customers who received Action A is large,

$$\hat{\beta}_{ra,x} \approx \beta_{ra,x},$$  \hspace{1cm} (6.11)

and

$$\xi_{ai} \approx z_i \beta_{ra,z} + \varepsilon_{ai};$$  \hspace{1cm} (6.12)

In contrast, when the number of newly existing customers who received Action B is large,

$$\hat{\beta}_{rb,x} \approx \beta_{rb,x},$$  \hspace{1cm} (6.13)

and

$$\xi_{bi} \approx z_i \beta_{rb,z} + \varepsilon_{bi};$$  \hspace{1cm} (6.14)

If we assume variables in $X$ and $Z$ can explain the major part of the behaviour of $R_a$ and $R_b$, we can ignore $\varepsilon_a$ and $\varepsilon_b$ in Equations (6.12) and
so that

\[ \xi_{ni} \approx z_i \beta_{ra,z} \]  
\[ \xi_{bi} \approx z_i \beta_{rb,z} \]  

(6.15)  
(6.16)

In this way,

\[ sd(\xi_i) \approx sd(Z) \beta_{ra,z} \]  
\[ sd(\xi_i) \approx sd(Z) \beta_{rb,z} \]  

(6.17)  
(6.18)

where \( sd(Z) \) is the standard deviation of the unobserved variable \( Z \).

Thus, we have

\[ \xi'_a \approx \frac{z_i \beta_{ra,z}}{sd(Z) \beta_{ra,z}} \approx \frac{z_i}{sd(Z)} \]  
\[ \xi'_b \approx \frac{z_i \beta_{rb,z}}{sd(Z) \beta_{rb,z}} \approx \frac{z_i}{sd(Z)} \]  

(6.19)  
(6.20)

The values of the scaled residuals are therefore approximate to the values of standardised \( Z \). Using the sample pool that consists of \( \xi'_a \)’s and \( \xi'_b \)’s will result in biased estimators of coefficients for \( Z \), but estimators of coefficients for \( X \) remain unbiased.

Nevertheless, we are not interested in estimating

\[ E(r_{ai} - r_{bi} | x_i, z_i) = x_i(\beta_{ra,x} - \beta_{ra,x}) + z_i(\beta_{ra,z} - \beta_{rb,z}). \]  

(6.21)

Instead, we wish to estimate

\[ E(r_{ai} - r_{bi} | x_i) = x_i(\beta_{ra,x} - \beta_{ra,x}) + E(z_i | x_i)(\beta_{ra,z} - \beta_{rb,z}). \]  

(6.22)

Since we use regression residuals as empirical samples, \( E(z_i | x_i) = E(z_i) = 0 \),
Equation (6.22) becomes

\[ E(r_{ai} - r_{bi}|x_i) = x_i(\beta_{ra,x} - \beta_{rb,x}), \]

where only coefficients for recorded variables are taken into consideration.

Therefore, taking \( \xi'_a \)'s and \( \xi'_b \)'s as empirical samples of \( Z \) is adequate.

Note that the value of \( \xi'_{ai} \) (or \( \xi'_{bi} \)) is only close to \( z_i/sd(Z) \) when:

- the variance of \( \varepsilon_a \) (or \( \varepsilon_b \)) is relatively small, and
- the estimator \( \hat{\beta}_{ra,x} \) or \( \hat{\beta}_{rb,x} \) is close to the respective actual value.

The first criterion holds when variables in \( X \) and \( Z \) can explain most of the behaviour of the response variables. The second criterion holds when applying the complete-case analysis to the newly collected data sets is sufficient to obtain good estimators for either \( \beta_{ra,x} \) or \( \beta_{rb,x} \) but not necessarily both.

If most newly existing customers received Action A, using the newly collected data will only result in a good estimate for \( \beta_{ra,x} \) but not for \( \beta_{rb,x} \). In this way, it remains difficult to obtain good estimators for all parameters in the optimal decision function (e.g. \( E(r_{ai} - r_{bi}|x_i) \)). Nevertheless, one can make use of the empirical samples formed by \( \xi'_a \)'s and \( \xi'_b \)'s to estimate the pdf of \( Z \) so that the EM algorithm introduced in the previous section can be applied. From this we can derive good estimates for both \( \beta_{ra,x} \) and \( \beta_{rb,x} \), and consequently good estimators for all parameters in the optimal decision function. Therefore, when most newly existing customers received the same action, it is especially useful to apply the method we propose in this section.
6.3 A semi-GEM algorithm that estimates the previous assignment function

The EM algorithm introduced in Section 6.1 is only applicable if the density function of $Z$ is given. After obtaining empirical samples using the method described in Section 6.2, one can certainly estimate the density function of $Z$ and apply the EM algorithm introduced in Section 6.1. However, in this section we propose a semi-GEM algorithm so that the empirical samples can be used directly rather than as a means of estimating the density function. Moreover, the integrations and the differentiations in the EM algorithm that make use of the density function can be replaced by more straightforward calculations.

6.3.1 The E step

The concept of the E step here is identical to the one proposed in Section 6.1. That is to obtain the expectation for the missing $z_i$, and consequently the missing $g_i$. According to Equations (6.2) and (6.4), we can see that once the expectation for $z_i$ is obtained, it is straightforward to derive the expected value for $g_i$. However, deriving the expectation for $z_i$ using Equations (6.1) and (6.3) requires the pdf of $Z$.

Assume that from the newly collected data, one applies the method proposed in Section 6.2 and obtains $m$ empirical samples, i.e. $\Xi_1, \Xi_2, ..., \Xi_m$. We can obtain the conditional expectation for each missing $z_i$ using these empirical samples via the following E steps.

At the $t^{th}$ iteration,
• if customer $i$ received Action A, $s_i = 0$,

$$z_i^{(t)} = E(z_i|x_i, s_i = 0) = E(z_i|z_i > -x_i\beta_{g,x}^{(t-1)})$$

$$= \frac{1}{\sum_{j=1}^{m} I(\xi_j > -x_i\beta_{g,x}^{(t-1)})} \sum_{j=1}^{m} \xi_j I(\xi_j > -x_i\beta_{g,x}^{(t-1)}), \quad (6.24)$$

$$y_i^{(t)} = E(y_i|x_i, z_i^{(t)})$$

$$= x_i\beta_{g,x}^{(t-1)} + z_i^{(t)}; \quad (6.25)$$

• if customer $i$ received Action B, $s_i = 1$,

$$z_i^{(t)} = E(z_i|x_i, s_i = 1) = E(z_i|z_i \leq -x_i\beta_{g,x}^{(t-1)})$$

$$= \frac{1}{\sum_{j=1}^{m} I(\xi_j \leq -x_i\beta_{g,x}^{(t-1)})} \sum_{j=1}^{m} \xi_j I(\xi_j \leq -x_i\beta_{g,x}^{(t-1)}), \quad (6.26)$$

$$y_i^{(t)} = E(y_i|x_i, z_i^{(t)})$$

$$= x_i\beta_{g,x}^{(t-1)} + z_i^{(t)}; \quad (6.27)$$

where $I(\xi_j > -x_i\beta_{g,x}^{(t-1)}) = 1$ and $I(\xi_j \leq -x_i\beta_{g,x}^{(t-1)}) = 0$ if $\xi_j > -x_i\beta_{g,x}^{(t-1)}$; otherwise $I(\xi_j > -x_i\beta_{g,x}^{(t-1)}) = 0$ and $I(\xi_j \leq -x_i\beta_{g,x}^{(t-1)}) = 1$ if $\xi_j \leq -x_i\beta_{g,x}^{(t-1)}$.

### 6.3.2 The semi-generalised M step

In the EM algorithm that makes use of the empirical density function, the M step allows us to obtain an estimator for $\beta_{g,x}$ using the maximum likelihood method. Since the pdf of $Z$ is unknown, we cannot obtain the actual underlying likelihood function, much less maximise it. We therefore consider obtaining an estimator for $\beta_{g,x}$, but not necessarily the ML estimator.

If we think of $Z$ as a regression residual in the previous assignment function (which is consistent with the idea of obtaining the empirical samples from regression residuals), regardless of the distribution form of $Z$, the least squares estimator of $\beta_{g,x}$ can be obtained from regressing $G$ on $X$. 

95
We thus propose that at the $t^{th}$ iteration,

- regress $\mathcal{G}(t) = (g_i^{(t)})$ on $X = (x_i)$, where $i = 1, ..., n$, yielding $\beta_{g,x}^{(t)}$.

Suppose that the proposed algorithm converges at the $T^{th}$ iteration, set $\hat{z}_i = z_i^{(T)}$, and $\hat{z}_i$ to be the estimated expectation for each missing $z_i$.

When $Z$ indeed follows a normal distribution, least squares estimators do maximise the underlying likelihood. This is not necessarily true if $Z$ follows any other type of distribution form. However if least squares estimators obtained at each iteration do improve the unknown likelihood, $\beta_{g,x}^{(t)}$ (for $t = 1, 2, ...$) is likely to converge to the ML estimator for $\beta_{g,x}$, and the algorithm we propose here can be regarded as a Generalised EM algorithm. Since the actual density function of $Z$ is not given, the exact form of the underlying likelihood is unknown, and there is no way to prove whether least squares estimators do improve the likelihood at each iteration.

Let us assume that $\beta_{g,x}^{(t)\text{ML}}$ represents the ML estimator for $\beta_{g,x}$ at the $t^{th}$ iteration, $\beta_{g,x}^{(t)\text{LS}}$ and $\beta_{g,x}^{(t-1)\text{LS}}$ represent least squares estimators at the $t^{th}$ and the $(t - 1)^{th}$ iteration respectively. We can argue that, when the same set of data (observed and estimated) is used, least squares estimators are likely to be relatively close to the ML estimators. In this way, it is more likely that $\beta_{g,x}^{(t)\text{LS}}$ is closer to $\beta_{g,x}^{(t)\text{ML}}$ than $\beta_{g,x}^{(t-1)\text{LS}}$ is to $\beta_{g,x}^{(t)\text{ML}}$. At the $t^{th}$ iteration, $\beta_{g,x}^{(t)\text{LS}}$ is more likely to result in a higher likelihood than $\beta_{g,x}^{(t-1)\text{LS}}$ is. Consequently, least squares estimators are likely to improve the unknown likelihood at each iteration, and we can consider the proposed algorithm as a semi-GEM algorithm.

Note that applying the above semi-GEM algorithm avoids estimating the pdf from empirical samples, replaces the integrations in E steps by summations and replaces the differentiations in M steps by least-squares regression. Although this proposition is not theoretically sound, we expect this semi-GEM algorithm to outperform extensions of existing methods that assume $Z$ to follow a particular distribution. The performance of this algorithm is examined using simulated data sets, and the results are shown in Section 6.5.
6.4 Estimate the response functions

According to the second step of Heckman's method, once the expected value of $Z$ is obtained for each previous customer, response functions can be estimated by applying least squares regression to respondent samples. Among respondent samples, the response functions of $R_a$ and $R_b$ are respectively:

$$E(r_a|x_i, s_i = 0) = x_i \beta_{ra,x} + E(z_i|x_i, s_i = 0) \beta_{ra,z}, \quad (6.28)$$

$$E(r_b|x_i, s_i = 1) = x_i \beta_{rb,x} + E(z_i|x_i, s_i = 1) \beta_{rb,z}. \quad (6.29)$$

Since in both Sections 6.1 and 6.3, the estimator $\hat{z}_i$ for each missing $z_i$ is obtained from $E(z_i|x_i, s_i)$, we can replace $E(z_i|x_i, s_i = 0)$ in Equation (6.28) and $E(z_i|x_i, s_i = 1)$ in Equation (6.29) by $\hat{z}_i$. In this way, the response functions can be estimated using the following steps.

- Among customers who received Action A, regress $R_a$ on $X$ and $\hat{Z}$ to derive estimators $\hat{\beta}_{ra,x}$ and $\hat{\beta}_{ra,z}$.
- Among customers who received Action B, regress $R_b$ on $X$ and $\hat{Z}$ to derive estimators $\hat{\beta}_{rb,x}$ and $\hat{\beta}_{rb,z}$.

The estimated optimal decision function becomes:

$$\hat{E}(r_a|x_i) - \hat{E}(r_b|x_i) = x_i(\hat{\beta}_{ra,x} - \hat{\beta}_{rb,x}) + E(z|x_i)(\hat{\beta}_{ra,z} - \hat{\beta}_{rb,z}). \quad (6.30)$$

Because $Z$ is assumed to be independent of $X$, $E(z|x_i) = E(Z)$ is the population mean of $Z$. If empirical samples generated from regression residuals are used, then $E(Z) = 0$. If the EM algorithm proposed in Section 6.1 is used, the empirical density function of $Z$ is used.
6.5 Simulation results

In this chapter, we first introduced an EM algorithm to estimate the previous assignment function using the empirical density function of $Z$. When the empirical density function is not available, we suggested obtaining empirical samples from the newly collected data sets. We further proposed a semi-GEM algorithm that can make use of the empirical samples directly. Nevertheless, the semi-GEM algorithm is not theoretically sound, and thus we wish to examine the performance of this algorithm using the simulated data sets.

Whether it is worthwhile to generate empirical samples from the newly collected data and apply the semi-GEM algorithm depends not only on the quality of the proposed algorithm but also on the quality of the generated samples. We thus further examine the quality of the empirical samples using the simulated data sets.

6.5.1 Effectiveness of the semi-GEM algorithm

Applying the semi-GEM algorithm requires empirical samples drawn from $Z$. In order to examine the performance of this algorithm, in each simulated case, we use the data pool from which the values of $Z$ were originally drawn (see Chapter 4) and draw another 4,000 data points with replacement. These 4000 data points are used as empirical samples in the semi-GEM algorithm to estimate the previous assignment function and obtain the estimator $\hat{z}_i$ for each previous customer $i$. Thereafter, we apply the method proposed in Section 6.4 and use the estimated $\hat{z}$ to estimate the response functions and consequently the optimal decision function. The accuracy of the estimations is evaluated using the value of $\Delta_{\text{emp}}$, which represents the improvement resulting from applying the semi-GEM algorithm over applying the complete-case analysis. The results are shown in Figure 6.1.

When evaluating the performance of an adjustment method, we should also look at Figure 5.1, which shows the inadequacies of the complete-case
(a) The values of \( Z \) are drawn from a standard normal distribution.

(b) The values of \( Z \) are drawn from \( V_1 \).

(c) The values of \( Z \) are drawn from \( V_2 \).

(d) The values of \( Z \) are drawn from \( V_3 \).

(e) The values of \( Z \) are drawn from \( V_4 \).

(f) The values of \( Z \) are drawn from \( V_5 \).

Figure 6.1: The performance of the proposed adjustment method that makes use of empirical samples of \( Z \). The value of \( \Delta^{\text{emp}} \) (see Equation 4.10), where "emp" represents the adjustment method that makes use of empirical samples of \( Z \), for different levels of \( Pr(S = 0) \). Each graph represents average results among 2160 (6 different \( \beta_{\omega,r_1} \times 6 \) different \( \beta_{\beta,r_1} \times 60 \) different random seeds) data sets.
analysis. In order to examine whether the proposed adjustment method, which makes use of empirical samples, performs better than extensions of existing methods, which assume $Z$ to follow a specific distribution, we also take Figure 5.2 into consideration.

From Figures 5.1, 5.2 and 6.1, we can see that when the actual distribution of $Z$ is indeed normal or close to normal (e.g. $V_1$), an extension of Heckman’s method performs better than the semi-GEM algorithm we proposed. This is because the number of empirical samples used in our simulations is 4000 and it is unlikely that using such a sample pool will outperform when a density function is used directly.

Nevertheless, when the distribution of $Z$ is relatively skewed and the complete-case analysis is significantly ineffective (e.g. when $Pr(S = 0) < 0.5$ in our simulation cases), the method that makes use of empirical samples performs significantly better than the method that assumes $Z \sim normal$. Therefore among our simulated cases, although the semi-GEM algorithm is not theoretically perfect, it does perform significantly better when the actual distribution of $Z$ is significantly different from the specific distribution assumed by an existing method.

6.5.2 Effectiveness of obtaining the empirical sample from the newly collected data set

In Section 6.2, we introduced an approach to obtain empirical samples from the newly collected data set. In this section, we make use of the simulated data sets to examine whether a sample pool that consists of regression residuals is as useful as samples drawn directly from the actual distribution of $Z$.

In each simulated case, we have generated 200 data points as information on 200 newly existing customers (see Chapter 4 for details). We assume that 100 customers have been assigned Action A, and another 100 customers have been assigned Action B. By regressing $R_a$ ($R_b$) on $X$ among customers
who received Action A (B), one obtains a sample pool that consists of scaled residuals $\xi'_a$ and $\xi'_b$. The semi-GEM algorithm is then carried out using these $\xi'_a$ and $\xi'_b$ as empirical samples to estimate the previous assignment function and obtain estimator $\hat{z}_i$ for each previous customer $i$. Afterwards, the estimators $\hat{z}$s are used to estimate the response functions and the optimal decision function. The quality of these estimates can be judged from Figure 6.2.

In order to compare with the samples drawn directly from the actual distribution of $Z$, we compare Figure 6.2 with 6.1. We can see that these two figures are very similar. This implies that within our simulated cases, sample pools that consist of regression residuals are nearly as useful as those that consist of samples drawn directly from the actual distribution of $Z$.

Note that in our simulation design, the variances of random errors are relatively small ($Var(\xi_a) = Var(\xi_b) = 0.05^2$). In this way, most of the behaviour of $R_a$ and $R_b$ can be explained by the recorded variables in $X$ and the unrecorded variable $Z$. If this is not the case, the distribution form of the residual $\xi_a$ (or $\xi_b$) will be significantly different from that of the scaled unknown variable $Z\beta_{r_a,x}$ (or $Z\beta_{r_b,x}$), and empirical samples obtained from regression residuals might not be as useful.

Also note that in our simulated data sets, we assume that there is only one recorded variable. Thus, 200 data points are adequate to obtain a good estimator for $\beta_{r_a,x}$. If the number of recorded variables is large, information on more newly existing customers is required to derive good estimators for $\beta_{r_a,x}$ or $\beta_{r_b,x}$ so as to obtain informative samples.

### 6.6 When $Z$ is multivariate

Note that applying the adjustment methods proposed in this chapter is inappropriate if

- $Z$ is multivariate, and
Figure 6.2: The performance of the proposed adjustment method that makes use of empirical samples obtained from regression residuals. The value of $\Delta^{res}$ (see Equation 4.10), where “res” represents the proposed adjustment method that makes use of empirical sample pools obtained from residuals, for different levels of $Pr(S = 0)$. Each graph represents average results among 2160 (6 different $\beta_{x,r_1} \times 6$ different $\beta_{y,r_1} \times 60$ different random seeds) data sets.
more than one variable in \( Z \) significantly affects the previous assignment function and any of the response functions.

In this section, we explain the reason why this is so.

The methods proposed in this chapter estimate the previous assignment function and obtain the estimator \( \hat{z}_i \) for each previous customer \( i \). If \( Z \) is multivariate, e.g. \( Z = \{ Z_1, Z_2 \} \), the previous assignment function becomes

\[
g_i = x_i \beta_{g,x} + z_{1i} \beta_{g,z_1} + z_{2i} \beta_{g,z_2}. \tag{6.31}
\]

According to Equation (6.31), obtaining an estimator \( \hat{z}_i \) for each \( z_i \) implies obtaining \( E(z_{1i} \beta_{g,z_1} + z_{2i} \beta_{g,z_2}) \) for each previous customer \( i \).

If \( Z = \{ Z_1, Z_2 \} \), the response functions for \( R_a \) and \( R_b \) become:

\[
\begin{align*}
r_{ai} & = x_i \beta_{ra,x} + z_{1i} \beta_{ra,z_1} + z_{2i} \beta_{ra,z_2}, \tag{6.32} \\
rb_i & = x_i \beta_{rb,x} + z_{1i} \beta_{rb,z_1} + z_{2i} \beta_{rb,z_2}. \tag{6.33}
\end{align*}
\]

In this way, assuming \( Z \) to be univariate and using the estimators \( \hat{z}_i \) to estimate the response functions is only appropriate when the following equations hold

\[
\begin{align*}
r_{ai} & = x_i \beta_{ra,x} + z_{1i} \beta_{ra,z_1} + z_{2i} \beta_{ra,z_2} \\
& = x_i \beta_{ra,x} + (z_{1i} \beta_{g,z_1} + z_{2i} \beta_{g,z_2}) \beta_{ra,z}, \tag{6.34} \\
r_{bi} & = x_i \beta_{rb,x} + z_{1i} \beta_{rb,z_1} + z_{2i} \beta_{rb,z_2} \\
& = x_i \beta_{rb,x} + (z_{1i} \beta_{g,z_1} + z_{2i} \beta_{g,z_2}) \beta_{rb,z}. \tag{6.35}
\end{align*}
\]

Equations (6.34) and (6.35) imply

\[
\frac{\beta_{g,z_1}}{\beta_{g,z_2}} = \frac{\beta_{ra,z_1}}{\beta_{ra,z_2}} = \frac{\beta_{tb,z_1}}{\beta_{tb,z_2}}, \tag{6.36}
\]

103
and this is rarely the case in practice. Thus if $\mathcal{Z}$ is not univariate and more than one variable in $\mathcal{Z}$ is highly correlated to the previous assignment decisions and any of the response variables, it is not appropriate to apply the methods proposed in this chapter or any other methods that make the univariate assumption.

Moreover, consider the case when the empirical samples are obtained from regression residuals. Similarly to the above argument, if the unrecorded $\mathcal{Z}$ is multivariate, one can hardly obtain a good estimator $\hat{z}_i$ for each previous customer $i$, much less good estimates for the response functions.

Consequently, unless one has strong evidence that either $\mathcal{Z}$ is univariate or that only one variable in $\mathcal{Z}$ significantly affects the previous assignment function and all the response functions, we do not suggest applying any method that makes the univariate assumption. In the next chapter, we will introduce other adjustment methods that are not affected by whether $\mathcal{Z}$ is univariate or multivariate.
Chapter 7

Adaptive adjustment approaches

Consider a case where whether a previous customer received Action A or B depends on both recorded and unrecorded information. In this case, data in the previously collected data set are MNAR, applying the complete-case analysis will result in biased estimates, and adjustment methods are required.

On the other hand, a newly existing customer received either action according only to information that is recorded. In this way, the newly collected data are MAR, and one can obtain unbiased estimates using the complete-case analysis without further adjustment. Although unbiased estimates can be obtained using newly collected data only, one should not ignore the previously collected data. This is because when the number of newly existing customers is relatively small, estimators derived from newly collected data might not be as close to the actual values as those derived from previously collected data. It is thus worthwhile to make use of both the previously collected and newly collected data sets.

In this chapter, we introduce two adaptive adjustment approaches to achieve this: an ad-hoc adaptive approach (Section 7.1) and a combined least-squares regression approach (Section 7.2). To avoid duplication, we
show how these two approaches can be applied to estimate the parameters in the response function of $\mathcal{R}_a$ only.

Note that the adaptive adjustment approaches described in this chapter can be performed continuously. Whenever a new customer enters the system, more information is recorded in the newly collected data set, and the adaptive procedures can be carried out continuously based on both the fixed previously collected data set and the updated newly collected data set. In this way, once a newly existing customer enters the system, we can update the estimators for parameters in the response function and thus those in the optimal decision function. The next customer is then assigned to either action according to the estimated optimal decision function. Figure 7.1 illustrates this process. If adaptive approaches are applied, estimators can be updated whenever a new customer enters the system. Therefore, we examine the quality of each updated estimator when evaluating adaptive methods.

In order to evaluate the performance of the proposed adaptive methods, according to the simulation design described in Chapter 4, we simulate another 200 data points based on the assumed response functions and these data points are treated as information on 200 new customers. To simplify the problem, we assume responses can be observed immediately after the assignment decisions are made.
the simulation, we assumed one new customer arriving at a time, each of them being assigned to either action according to the latest estimated optimal decision function, and the corresponding response being observed right after the assignment. In each simulated case, whenever information on a new customer becomes available, we update not only the estimations according to each adaptive method but also the value of $\Delta$ (see Equation (4.10)) to evaluate the accuracy of each estimator.

7.1 Ad-hoc Adaptive Method

The concept of the ad-hoc adaptive method is to combine estimates derived from previously collected data with that derived from newly collected data. When only the previously collected data set is used, we can apply one of the adjustment approaches in the previous chapters to obtain estimator $\hat{\beta}_{r,a,x}^{(pre)}$ for $\beta_{r,a,x}$ in the response function. When only newly collected data is used, we can apply the complete-case analysis and obtain an unbiased estimator $\hat{\beta}_{r,a,x}^{(new)}$.

Assume that there are $p$ recorded variables, and we set

$$
\hat{\beta}_{r,a,x}^{(pre)} = [ \hat{\beta}_{r,a,x_0}^{(pre)} \hat{\beta}_{r,a,x_1}^{(pre)} \ldots \hat{\beta}_{r,a,x_p}^{(pre)} ],
$$

$$
\hat{\beta}_{r,a,x}^{(new)} = [ \hat{\beta}_{r,a,x_0}^{(new)} \hat{\beta}_{r,a,x_1}^{(new)} \ldots \hat{\beta}_{r,a,x_p}^{(new)} ].
$$

In order to combine estimator $\hat{\beta}_{r,a,x}^{(pre)}$ with $\hat{\beta}_{r,a,x}^{(new)}$, we set

$$
\hat{\beta}_{r,a,x}^{(adhoc)} = w_j^{(pre)} \hat{\beta}_{r,a,x_j}^{(pre)} + w_j^{(new)} \hat{\beta}_{r,a,x_j}^{(new)},
$$

(7.1)

where $w_j^{(pre)}$ and $w_j^{(new)}$ are weights on the estimated $\hat{\beta}_{r,a,x_j}^{(pre)}$ and $\hat{\beta}_{r,a,x_j}^{(new)}$ respectively.

Regardless of how we estimate $\hat{\beta}_{r,a,x_j}^{(pre)}$, because the distribution form of $Z$ is unknown, we prefer $\hat{\beta}_{r,a,x_j}^{(new)}$ to $\hat{\beta}_{r,a,x_j}^{(pre)}$. For this reason, we wish to increase the weight on $\hat{\beta}_{r,a,x_j}^{(new)}$ at a higher rate, and thus choose the following functions
that contain exponential terms for \( w_j^{(\text{new})} \) and \( w_j^{(\text{pre})} \):

\[
\begin{align*}
\frac{w_j^{(\text{new})}}{w_j^{(\text{pre})}} &= \frac{\exp\left(\frac{\text{Var}(\hat{\beta}_r^{(\text{pre})})}{\text{Var}(\hat{\beta}_r^{(\text{new})})}\right)}{1 + \exp\left(\frac{\text{Var}(\hat{\beta}_r^{(\text{pre})})}{\text{Var}(\hat{\beta}_r^{(\text{new})})}\right)}, \\
(7.2) \\
&= 1 - w_j^{(\text{new})}, \\
(7.3)
\end{align*}
\]

for \( j = 0, 1, \ldots, p \).

The more reliable the estimator \( \hat{\beta}_r^{(\text{new})} \) becomes, the larger the weight the ad-hoc adaptive approach assigns to \( \hat{\beta}_r^{(\text{new})} \). The variance of each estimated parameter is used to evaluate the reliability. The smaller the variance of \( \hat{\beta}_r^{(\text{new})} \), the larger the value of \( w_j^{(\text{new})} \) (for \( j = 0, 1, \ldots, p \)).

The unbiased estimator \( \hat{\beta}_r^{(\text{new})} \) is derived from applying least squares regression to the respondent sample. Therefore, \( \text{Var}(\hat{\beta}_r^{(\text{new})}) \) can be obtained using standard approach. However the estimator \( \hat{\beta}_r^{(\text{pre})} \) is derived from either least squares regression or any adjustment method. A more general approach should be applied to derive \( \text{Var}(\hat{\beta}_r^{(\text{pre})}) \). Efron and Tibshirani [10] have shown that bootstrap size 50 to 100 is sufficient for variance estimation. Therefore, regardless of the method applied to derive \( \hat{\beta}_r^{(\text{pre})} \), we obtain \( \text{Var}(\hat{\beta}_r^{(\text{pre})}) \) using the bootstrap method with bootstrap size 50.

Since the same set of previously collected data is used, the values of \( \hat{\beta}_r^{(\text{pre})} \) and \( \text{Var}(\hat{\beta}_r^{(\text{pre})}) \) are fixed. However the information contained in the newly collected data set increases over time, and the values of \( \hat{\beta}_r^{(\text{new})} \) and \( \text{Var}(\hat{\beta}_r^{(\text{new})}) \) shall vary whenever information on a newly existing customer who received Action A becomes available. Thus we suggest performing the following steps whenever information on a new respondent becomes available.

For \( j = 0, 1, \ldots, p \)

1. update the values of \( \hat{\beta}_r^{(\text{new})} \) and \( \text{Var}(\hat{\beta}_r^{(\text{new})}) \) according to the information in the newly collected data set;

2. update the values of \( w_j^{(\text{new})} \) and \( w_j^{(\text{pre})} \) according to Equations (7.2) and
3. update the values of $\hat{\beta}_{R_a,x_j}$ according to Equation (7.1);

4. according to the last updated $\hat{\beta}_{R_a,x}^{(adhoc)}$ and $\hat{\beta}_{R_b,x}^{(adhoc)}$ (which can be estimated using a similar approach), assign to the next customer to Action A if $x_i(\hat{\beta}_{R_a,x}^{(adhoc)} - \hat{\beta}_{R_b,x}^{(adhoc)}) > 0$, otherwise assign to this customer Action B.

Note that least squares regression cannot be carried out until the sample size is greater than $p + 1$, where $p$ is the number of variables in $X$. Thus we can only perform steps 1, 2, and 3 above after the size of new respondents is greater than $p + 1$. Moreover, before we start to calculate $\hat{\beta}_{R_a,x}^{(new)}$, we set $\hat{\beta}_{R_a,x}^{(adhoc)} = \hat{\beta}_{R_a,x}^{(pre)}$.

After the first $\hat{\beta}_{R_a,x}^{(new)}$ is calculated, whenever information on a new respondent becomes available, we update the corresponding estimators (e.g. $\hat{\beta}_{R_a,x_j}^{(new)}$), variances (e.g. $Var(\hat{\beta}_{R_a,x_j}^{(new)})$), and the estimate for the optimal decision function. The assignment decision made for the next new customer will then be based on both the previously collected and the newly collected data sets.

In order to evaluate the performance of the ad-hoc adaptive method, we apply it to the simulated cases described in Chapter 4. Because $p = 1$ in our simulations, we did not start to calculate the value of $\hat{\beta}_{R_a,x}^{(new)}$ and its variance until the number of newly existing customers who received Action A reaches 10. Likewise, the value of $\hat{\beta}_{R_b,x}^{(new)}$ and its variance are not calculated when fewer than 10 newly existing customers received Action B.

We first apply either the complete-case analysis or Heckman’s method to each previously collected data set. Figures 7.2 and 7.3 show the performance of the ad-hoc adaptive approach where the complete-case analysis and Heckman’s method are respectively applied to the previously collected data set.

From these two figures, we can see that when the ad-hoc adaptive ap-
(a) The values of \( Z \) are drawn from a standard normal distribution.

(b) The values of \( Z \) are drawn from \( V_1 \).

(c) The values of \( Z \) are drawn from \( V_2 \).

(d) The values of \( Z \) are drawn from \( V_3 \).

(e) The values of \( Z \) are drawn from \( V_4 \).

(f) The values of \( Z \) are drawn from \( V_5 \).

Figure 7.2: The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. The value of \( \Delta^{ad-hoc} \) before the size of newly collected data set reaches 200. In this figure, "ad-hoc" represents the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. Each graph represents average results among 19,440 (9 different \( Pr(S = 0) \) \( \times \) 6 different \( \beta_{w,r} \) \( \times \) 6 different \( \beta_{g,r} \) \( \times \) 60 different random seeds) data sets. More details can be found in Figures A-1, A-5, A-3, A-4, A-2 and A-6 in the Appendix.
Figure 7.3: The performance of the ad-hoc adaptive approach and Heckman's two-step method was applied to the previously collected data set. The value of $\Delta^{ad-hoc}$ before the size of newly collected data set reaches 200. In this figure, "ad-hoc" represents the ad-hoc adaptive approach and Heckman's two-step method was applied to the previously collected data set. Each graph represents average results among 19,440 (9 different $Pr(S = 0) \times 6$ different $\beta_{w,x} \times 6$ different $\beta_{g,z} \times 60$ different random seeds) data sets. More details can be found in Figures A-7, A-11, A-8, A-9, A-10 and A-12 in the Appendix.
proach is applied, the quality of the estimated optimal decision function improves most of the time when the size of newly collected sample increases. The exception can be found when the distribution of $\mathcal{Z}$ is normal or close to normal and Heckman's method is applied to derive $\hat{\beta}_{r_a,x}^{(pre)}$, $\hat{\beta}_{r_b,x}^{(pre)}$, $\text{Var}(\hat{\beta}_{r_a,x}^{(pre)})$ and $\text{Var}(\hat{\beta}_{r_b,x}^{(pre)})$. This is because when a proper adjustment method is applied to the previously collected data, both the estimators $\hat{\beta}_{r_a,x}^{(pre)}$ and $\hat{\beta}_{r_b,x}^{(pre)}$ are already close to the actual values. In contrast, when the size of newly collected data set is small, the estimators $\hat{\beta}_{r_a,x}^{(new)}$ and $\hat{\beta}_{r_b,x}^{(new)}$ can be significantly different from the actual values. Nevertheless, regardless of how poor the estimators $\hat{\beta}_{r_a,x}^{(new)}$ and $\hat{\beta}_{r_b,x}^{(new)}$ are, the weight $w_j^{(new)}$ is always greater than 0.5. Therefore when the ad-hoc adaptive approach is applied, if one is confident about the adjustment method applied to the previously collected data set, the adaptive procedures should not be carried out until the size of the newly collected data set is relatively large.

Note that there is only one recorded variable in our simulation design. It is thus sufficient to start the adaptive procedure once the number of new respondents exceeds 10. However in consumer banking, usually a large number of parameters are recorded and used to make assignment decisions. In this way, the ad-hoc adaptive approach cannot be used until information on a large number of new customers becomes available.

7.2 Combined least squares regression approaches

The main disadvantage of the ad-hoc adaptive approach is that the estimated optimal decision function cannot be updated immediately. Moreover when the number of newly existing customers is relatively small, the weights assigned to the estimates derived from the newly collected data set are too

---

2This is because when $\frac{\text{Var}(\hat{\beta}_{r_a,x}^{(pre)})}{\text{Var}(\hat{\beta}_{r_a,x}^{(new)})} \rightarrow 0$, $w_j^{(new)} \rightarrow 0.5$. The larger the value of $\frac{\text{Var}(\hat{\beta}_{r_a,x}^{(pre)})}{\text{Var}(\hat{\beta}_{r_a,x}^{(new)})}$, the larger the weight $w_j^{(new)}$. The weight $w_j^{(new)}$ is thus always greater than 0.5.
large. In this section, we propose adaptive methods that can avoid these disadvantages.

7.2.1 Naive combined least squares regression approach

Set $X_{\text{a,new}}$ to be a $n_a \times (1 + p)$ matrix that contains the vector values of $X = [1 \ \mathcal{X}_1 \ldots \mathcal{X}_p]$ of all $n_a$ customers who received Action A in the newly collected data set. We also set $R_{\text{a,new}}$ to be a vector of size $n_a$ that contains values of $R_a$ of all newly existing customers who received Action A.

Since the missingness of $R_a$ is ignorable given $X$ in the newly collected data set, the complete-case analysis can be applied to derive an unbiased estimator for $\beta_{\text{a,x}}$ in the response function. Therefore, the unbiased estimator $\hat{\beta}_{\text{a,x}}$ can be obtained from applying least squares regression to the respondent sample:

$$\hat{\beta}_{\text{a,x}} = (X_{\text{a,new}}^T X_{\text{a,new}})^{-1} X_{\text{a,new}}^T R_{\text{a,new}}.$$  \hspace{1cm} (7.4)

If we ignore the effect of $Z$ on previous assignments, the respondent sample in both the newly and the previously collected data set can be treated as if generated by the same regression process. In this way, the regression parameter can be estimated using:

$$\hat{\beta}_{\text{a,x}} = (X_{\text{a,pre}}^T X_{\text{a,pre}} + X_{\text{a,new}}^T X_{\text{a,new}})^{-1} (X_{\text{a,pre}}^T R_{\text{a,pre}} + X_{\text{a,new}}^T R_{\text{a,new}}),$$  \hspace{1cm} (7.5)

where $X_{\text{a,pre}}$ is a $m_a \times (1 + p)$ matrix that contains the vector values of $X$, and $R_{\text{a,pre}}$ is a vector of size $m_a$ that contains values of $R_a$ for all $m_a$ previous customers who received Action A.

This approach is thus equivalent to combining the previously collected data set (without adjustment) with the newly collected one and then deriving least squares estimators from the combined data set.

The performance of the naive combined least squares regression approach
is examined using the simulated data sets described in Chapter 4, and the
simulation results are shown in Figure 7.4. We can see that when this adap-
tive approach is applied, the estimated optimal decision function is indeed
updated as soon as information on the first newly existing customer becomes
available. Moreover, because responses in the newly collected data set are
MAR whilst those on the previously collected data set are MNAR, the es-
timations derived from this adaptive approach get closer and closer to the
actual values when the size of newly collected data set increases.

Nevertheless if we compare Figure 7.4 with 7.2, we can see that estimates
derived from the ad-hoc adaptive approach improve at a higher speed, and
this is due to the exponential term in weights (see Equations (7.3) and (7.2)).
In Section 7.2.3, we will introduce an extension from the naive combined least
squares regression approach that overcomes this disadvantage.

7.2.2 Adjusted combined least squares regression ap-
proach

Applying the complete-case analysis to the previously collected data set will
result in biased estimates because the missingness of \( R_a \) is not ignorable
given \( X \). However, the missingness is ignorable given both \( X \) and \( Z \). If
one regresses \( R_a \) on not only \( X \) but also \( Z \) in the respondent sample, the
estimated parameters will be unbiased. We thus adjust Equation (7.5) into:

\[
\hat{\beta}_{ra,q} = (Q^T_{a,pre} Q_{a,pre} + Q^T_{a,new} Q_{a,new})^{-1}(Q^T_{a,pre} R_{a,pre} + Q^T_{a,new} R_{a,new}), \tag{7.6}
\]

where \( \hat{\beta}_{ra,q} = [\hat{\beta}_{ra,x_0} \hat{\beta}_{ra,x_1} ... \hat{\beta}_{ra,x_p} \hat{\beta}_{ra,z}] \), \( Q_{a,pre} \) is a \( m_a \times (p+2) \) matrix
that contains the vector values of \( X = [1 \ X_1 \ ... \ X_p \ Z] \) of \( m_a \) previous
customers who received Action A and \( Q_{a,new} \) is a \( n_a \times (p+2) \) matrix.

When applying Equation (7.6), one takes into consideration the effect of
unrecorded \( Z \) on the previous assignment decisions. The adjusted combined
least squares regression approach is thus recommended when one wishes to
Figure 7.4: The performance of the naive combined least squares regression approach. The value of $\Delta_{nls}$ before the size of newly collected data set reaches 200. In this figure, "nls" represents the naive combined least squares regression approach. Each graph represents average results among 10,440 (9 different $Pr(S = 0) \times 6$ different $\beta_{x,x} \times 6$ different $\beta_{y,x} \times 60$ different random seeds) data sets. More details can be found in Figures A-13, A-14, A-15, A-16, A-17 and A-18 in the Appendix.
make use of an adjustment method to correct the previously collected data set.

In this adaptive approach, the values of $Z$ in $Q_{a, pre}$ are replaced by their expectation estimated by applying a given adjustment method (e.g. Heckman's method or the methods we proposed in Chapter 6) to the previously collected data set. Since newly existing customers have been assigned to either action based on recorded variables in $\mathcal{X}$ only, the expected values of $Z$ for each newly existing customer are their unconditional mean. Thus the values of $Z$ in $Q_{a, new}$ are replaced by $E(Z)$. In this way, one can think of applying this approach as combining the adjusted previously collected data set with the newly collected one and then deriving least squares estimators from the combined data set.

We use the simulated data sets described in Chapter 4 to examine the performance of the adjusted combined least squares regression approach. The simulation results are shown in Figure 7.5, and the values of $Z$ among previous customers are estimated using Heckman's method. If we compare Figures 7.5 with 7.4, the adjusted combined least squares regression approach outperforms the naive one in most cases. The exceptions occur when the distribution of $Z$ is significantly different from a normal distribution (e.g. when the values of $Z$ are drawn from $V5$). In these cases, applying Heckman's method to the previously collected data set results in more bias and the adaptive approach begins with very poor estimates. Therefore, the adjusted combined least squares regression approach is preferred to the naive one only when a proper adjustment method is applied to the previously collected data.

From Figures 7.5 and 7.3, we can compare the combined least squares regression approach with the ad-hoc adaptive approach. We can see that when the actual distribution of $Z$ is normal or close to normal, the adjusted combined least squares regression approach is preferred to the ad-hoc adaptive approach. This is because when the distribution of $Z$ is normal or close to normal, applying Heckman's method provides estimators that are as good as if they were derived directly from a MAR data set. After adjustment, the
(a) The values of $Z$ are drawn from a standard normal distribution.

(b) The values of $Z$ are drawn from $V_1$.

(c) The values of $Z$ are drawn from $V_2$.

(d) The values of $Z$ are drawn from $V_3$.

(e) The values of $Z$ are drawn from $V_4$.

(f) The values of $Z$ are drawn from $V_5$.

Figure 7.5: The performance of the adjusted combined least squares regression approach. The value of $Δ_{als}$ before the size of newly collected data set reaches 200. In this figure, “als” represents the adjusted combined least squares regression approach and Heckman’s two-step method was applied to estimate the missing $Z$ for previous customers. Each graph represents average results among 19,440 (9 different $Pr(S = 0) \times 6$ different $β_{ω,x} \times 6$ different $β_{g,x} \times 60$ different random seeds) data sets. More details can be found in Figures A-19, A-20, A-21, A-22, A-23 and A-24 in the Appendix.
previously collected data set is as informative as a newly collected data set with the same sample size. Adaptive estimates should thus be affected by both previously and newly collected data sets proportional to their sample sizes. The adjusted combined least squares regression approach does update the estimation gradually according to the size of the newly collected data set. The ad-hoc adaptive approach assigns more than half of the weight to $\hat{\beta}_{r,w}^{(\text{new})}$ and $\hat{\beta}_{r,w}^{(\text{new})}$ regardless of the number of newly existing customers. In this way, when the adjustment method applied to the previously collected data set is proper, and when the size of the newly collected data set is relatively small, the adjusted combined least squares regression approach is likely to outperform the ad-hoc adaptive approach.

In contrast, in cases where the distribution of $Z$ is relatively skewed, the ad-hoc adaptive approach is preferred. This is because in these cases, applying Heckman’s method to the previously collected data set will result in estimates that are significantly different from the actual values. Thus, it is preferred to increase the relative “weight” on the newly collected data set at a higher speed. Due to the exponential terms in the weights (see Equations (7.3) and (7.2)), the ad-hoc adaptive approach increases the relative weights on the unbiased estimates at a higher speed and therefore outperforms the adjusted combined least squares regression approach.

7.2.3 Weighted combined least squares regression approach

When the previously collected data set is either unadjusted or adjusted using an inappropriate adjustment approach, it is less reliable than the newly collected one. In this case, it is reasonable to assign less “weight” to each previously collected data point so that the relative “weight” assigned to the newly collected data set increases at a higher speed.

A straightforward way to adjust the weight one assigns to less reliable
data points is to adjust Equations (7.5) and (7.6) into:

\[
\hat{\beta}_{a,x} = (w^2 X_{a,pre}^T X_{a,pre} + X_{a,new}^T X_{a,new})^{-1}(w^2 X_{a,pre}^T R_{a,pre} + X_{a,new}^T R_{a,new}),
\]

(7.7)

\[
\hat{\beta}_{a,q} = (w^2 Q_{a,pre}^T Q_{a,pre} + Q_{a,new}^T Q_{a,new})^{-1}(w^2 Q_{a,pre}^T R_{a,pre} + Q_{a,new}^T R_{a,new}).
\]

(7.8)

Applying this approach is identical to combining the previously collected data set (adjusted or not) with the newly collected one and then applying the weighted least squares regression. We assign weight \( w \) to each previously collected data point and weight 1 to each newly collected data point. Note that the choice of \( w (0 < w \leq 1) \) is subjective. The less confident one is about the method applied to the previously collected data set, the smaller the value one should choose for \( w \).

Using the simulated data sets described in Chapter 4, we evaluate the performance of the weighted combined least squares regression approach (set \( w = 0.1 \)) extended from the naive combined least squares regression. The results are shown in Figure 7.6. Let us compare Figure 7.4 with 7.6. We can observe that when less weight is assigned to each previously collected data point, the estimations are indeed improved at a higher speed. Since the choice of \( w \) is small in our simulations, the relative weight one assigns to the newly collected data set increases at a speed even higher than that when the ad-hoc adaptive approach is applied. Thus if the previously collected data set is used without adjustment, the weighted combined approach outperforms the ad-hoc adaptive approach (see Figures 7.2 and 7.6).

Figure 7.7 displays the performance of the weighted combined least squares regression approach \( (w = 0.1) \) extended from the adjusted combined least squares regression approach (which obtained the expected values of \( Z \) for each previous customer using the extension of Heckman’s method). Let us compare Figure 7.7 with Figures 7.3 and 7.5. Where the actual distribution of \( Z \) is normal or close to normal, the weighted combined least squares regression approach performs worst. In contrast, where the actual distribution
Figure 7.6: The performance of the weighted combined least square regression approach extended from the naive combined least squares regression approach. The value of $\Delta wls_n$ before the size of newly collected data set reaches 200. In this figure, "wls" represents the weighted combined least square regression approach extended from the naive combined least squares regression approach with weight $w = 0.1$. Each graph represents average results among 19,440 (9 different $Pr(S = 0) \times 6$ different $\beta_{x,x} \times 6$ different $\beta_{y,x} \times 60$ different random seeds) data sets. More details can be found in Figures A-25, A-26, A-27, A-28, A-29 and A-30 in the Appendix.
Figure 7.7: The performance of the weighted combined least square regression approach extended from the adjusted combined least squares regression approach. The value of $\Delta^{wls_a}$ before the size of newly collected data set reaches 200. In this figure, "wls_a" represents the weighted combined least square regression approach extended from the adjusted combined least squares regression approach (and Heckman's two-step method was applied to estimate the missing $Z$ for previous customers) with weight $w = 0.1$. Each graph represents average results among 19,440 (9 different $Pr(S = 0) \times 6$ different $\beta_{w,t}$ $\times 6$ different $\beta_{g,t}$ $\times 60$ different random seeds) data sets. More details can be found in Figures A-31, A-32, A-33, A-34, A-35 and A-36 in the Appendix.
of $Z$ is relatively skewed, the weighted combined least squares regression approach performs the best.

As has been mentioned, when the adjustment method applied to the previously collected data is appropriate, one should slowly increase the relative weight on the newly collected data set. In contrast, if one is uncertain about the adjustment method, the relative weight assigned to the newly collected data set should be increased at a higher speed. We chose $w = 0.1$ in our simulations, which implies that the relative weight on the newly collected data set would increase at a speed even higher than when the ad-hoc adaptive method is used. Therefore, when the actual distribution of $Z$ is similar to the one assumed in Heckman’s method, the values of $\Delta^{\text{tls}}$ are lower than those of $\Delta^{\text{ad-hoc}}$ and $\Delta^{\text{aft}}$. In contrast, when the assumed distribution form is different from the actual one, the values of $\Delta^{\text{tls}}$ are higher.

Among all the adaptive approaches we have proposed, the weighted combined least squares regression approach is the most general and the most flexible one. We recommend applying this adaptive approach so that one can

- take into account the unrecorded variable $Z$ among previously existing customers (if an adjustment method is available to be applied to the previously collected data);
- begin to update the estimates as soon as information on the first newly existing customer becomes available;
- adjust the value of $w \ (0 < w \leq 1)$ according to experts’ experience, so that one chooses a larger value for $w$ the more confident one is about the previously collected data set (adjusted or not).
Chapter 8

Conclusions

In the previous chapters, we described the problem we wish to solve, reviewed existing solutions, and proposed new ones. In this final chapter, we summarise the main contributions of this thesis.

In order to focus on the missing data problem, we made some conventional assumptions, which are not generally applicable. Therefore we further point out how these assumptions can be relaxed and propose directions for future research.

8.1 Contributions

8.1.1 Drew attention to the non-ignorable missing data problem in behaviour scoring

The missing data problem in application scoring has attracted much interest, and relevant discussion can be found under the topic of “reject inference”. However, less attention has been paid to a similar problem in behaviour scoring. In this thesis, we draw attention to this problem.

Application scoring considers the accept/reject problem and only takes
into account responses to one action, i.e. acceptance. Several researchers have pointed out that when only the responses to one action are under consideration, the problem due to the missing data is insignificant if the proportion of missing data is small. However, when the same action is assigned to most previous customers, a small proportion of missing data for one action necessarily implies a large proportion missing for some other action. One cannot, then, simply ignore the problem.

We argued in Section 5.1 that when most previous customers received the same action, applying the complete-case analysis is likely to provide a good estimate for one response function but not both, and the estimate for the optimal decision function is unlikely to be satisfactory. This argument was examined using simulated data sets. From the simulation results, we further found that the bias resulting from the complete-case analysis is likely to be more severe when the majority of the previous customers received the same action. Consequently, the complete-case analysis is insufficient to solve the non-ignorable missing data problem in behaviour scoring, and one should apply adjustment methods.

8.1.2 Described extensions from existing methods and pointed out the inadequacies

We reviewed two of the best known solutions to MNAR problems: the two-step method proposed by Heckman, and the EM algorithm proposed by Little and Rubin. We also illustrated how these solutions can be extended to solve the problem that considers two response variables.

In Sections 3.5, we pointed out that the variances of estimators are likely to be inflated if the two-step method is applied to our problem. This is because, in consumer banking, assignment decisions for a given customer are usually based on variables that are expected to be correlated to their response. Thus the recorded variables in the previous assignment function are likely to be identical to those in the response functions. In Section 3.6,
we noted that if most previous customers received the same action, the EM algorithm can be computationally inefficient. However if one sets Heckman’s estimators as initial values, the EM algorithm converges at once. Thus when the unrecorded variable Z is assumed to have a normal distribution, we suggest applying Heckman’s two-step method.

Other researchers have proposed adjustment methods that assume that Z follows other specific distributions. However, each existing method for non-ignorable missing data problem is constrained by a specific distribution assumption. If this assumption fails, the estimates are biased. In Section 5.2, we argued that in credit scoring, Z represents information that cannot be explained by the observed variables, and the distribution of Z can vary from case to case. Simulations were carried out to show that existing methods are effective only when the distribution of Z is indeed follows the distribution assumed by the method applied. In contrast, when the actual distribution of Z is different, existing methods might generate even more bias. Consequently, existing methods are inadequate to solve the non-ignorable missing data problem in behaviour scoring.

8.1.3 Proposed adjustment methods so that the empirical distribution can be used

Each existing method assumes the unrecorded variable follows a specific distribution. It is thus only appropriate to apply extensions from existing methods when one makes the same distribution form assumption. For example, it is only appropriate to apply an adjustment method that assumes $Z \sim \text{uniform}$ when one is confident to assume $Z \sim \text{uniform}$. Thus in Chapter 6, we proposed adjustment methods that removes this constraint so that the empirical distribution of Z can be used.

We first proposed an EM algorithm extended from the EM algorithm suggested by Little and Rubin. We re-interpreted the E step in a more general way so that the constraint on the distribution of Z can be relaxed.
Moreover, the M step proposed by Little and Rubin is replaced by an actual likelihood maximisation. In this way, once the density function of $Z$ is given, the EM algorithm proposed in Section 6.1 can be used to solve our problem.

It is likely that the density function of $Z$ is not available. Thus we proposed in Section 6.2 an approach to obtain empirical samples from the newly collected data set. We have shown that such samples are able to provide as much information as samples drawn directly from the actual distribution of $Z$.

In Section 6.3, we proposed a semi-GEM algorithm so that empirical samples can be used directly. In this semi-GEM algorithm, the E step replaces the missing values by their expectation calculated from the sample pool, and the generalised M step obtains estimators using least squares regression. When applying the semi-GEM algorithm, the pdf of $Z$ is assumed to be unknown, and there is no way of proving whether the least squares estimators improve the likelihood at each iteration. Nevertheless, we argued that the least squares estimators are likely to be relatively close to the ML estimators and thus are likely to improve the unknown likelihood. We have supported this argument by comparing the semi-GEM algorithm with existing methods using the simulated data sets.

In our simulated cases, we found that when $Z$ follows a distribution that is indeed normal or close to normal, the extension of Heckman's method (which assumes $Z \sim \text{normal}$) performs better than the proposed semi-GEM algorithm. Nevertheless, when the actual distribution of $Z$ is significantly different from a normal distribution and when the complete-case analysis is ineffective, the semi-GEM algorithm performs significantly better than the method that assumes $Z \sim \text{normal}$. Therefore, it can be more useful to apply the semi-GEM algorithm than to apply extensions of existing methods that assume $Z$ to have a specific distribution.

Note that the methods proposed in Chapter 6 are restricted by a univariate assumption. We have shown that if $Z$ is not univariate, and if more than one variable in $Z$ is highly correlated to previous assignment decisions and
response variables, applying methods that make the univariate assumption is inappropriate. Therefore, the adjustment approaches proposed in Chapter 6 are not applicable universally and should be applied as appropriate.

8.1.4 Proposed adaptive approaches that make use of the newly collected MAR data

We have pointed out that in the problem we wish to solve, the missing data problem does not result from self-selection. The reason is that assignment decisions are made by a decision maker in the bank rather than by customers themselves. After the start of the analysis, a decision maker is restricted to use only recorded variables. Therefore, any unrecorded variable that caused the MNAR problem in the previously collected data set will no longer affect the assignment decision. Whether a response in the newly collected data set is observed or missing depends only on recorded variables, the data are MAR, and the missingness is ignorable.

Due to the ignorable missingness, unbiased estimates can be obtained using newly collected data only. Nevertheless, we argued that one should not ignore the previously collected data. This is because when the size of the newly collected data set is relatively small, the estimates derived from this data set might not be as close to the actual values as those derived from the previously collected data set. Thus in Chapter 7, we proposed adaptive approaches that make use of both the previously collected MNAR data and the newly collected MAR data.

All adaptive approaches we have introduced can update the estimates continuously. Whenever a new customer's response becomes available, the newly collected data set is updated, and the adaptive procedures can be carried out continuously to update the estimates according to both the fixed previously collected data set and the updated newly collected data set. In this way, updated estimates will be derived from all available information.

We first introduced the ad-hoc adaptive method, which combines esti-
orators (e.g. $\hat{\beta}^{(\text{pre})}$) derived from the previously collected data set with estimators (e.g. $\hat{\beta}^{(\text{new})}$) derived from the newly collected one. Estimators are combined in the way that the smaller the variances of $\hat{\beta}^{(\text{new})}$, the larger the weight assigned to $\hat{\beta}^{(\text{new})}$. Variance for $\hat{\beta}^{(\text{pre})}$ is obtained using the bootstrap method and that for $\hat{\beta}^{(\text{new})}$ is obtained using the standard approach. However, the ad-hoc adaptive approach suffers from two disadvantages:

- The weight $w^{(\text{new})}$ on $\hat{\beta}^{(\text{new})}$ is always larger than 0.5 regardless of the size of the newly collected data set. When the number of newly existing customers is relatively small, the ad-hoc adaptive approach tends to put too much weight on $\hat{\beta}^{(\text{new})}$ and results in a combined estimator that is worse than $\hat{\beta}^{(\text{pre})}$, which is derived solely from the previously collected MNAR data.

- The ad-hoc adaptive approach cannot be used until the number of newly collected respondent samples exceeds $p+1$, where $p$ is the number of recorded variables, which is usually very large in consumer banking.

In order to overcome these two disadvantages, we further introduced a series of combined least squares regression approaches. Instead of combining the estimators directly, combined least squares regression approaches combine two available data sets (previously collected and newly collected) before deriving estimators. The naive combined least squares regression approach combines two available data sets directly without adjustment; the adjusted combined least squares regression approach combined the adjusted previously collected data set with the newly collected one; and the weighted combined least squares regression approach applies weighted least squares regression to the combined data set.

Among all the proposed adaptive approaches, we most highly recommend the weighted combined least squares regression approach. This approach not only avoids all the disadvantages that occur in the ad-hoc adaptive approach, but also provides flexibility to apply adjustment methods to the MNAR data and the flexibility to adjust the weights on the MNAR data according to
expert experience.

8.2 Future research

8.2.1 Consider more than two actions

In this thesis, we wish to draw attention to the missing data problem in behaviour scoring, where more than one action is taken into consideration. To begin with, we simplify the problem and assume there are only two actions available. However, it will be practical to relax this conventional assumption and take into account more than two actions.

We have shown that the non-ignorable missingness results from the unrecorded variable $Z$, and that the key to solve this non-ignorable missing data problem is to derive the expectation of $Z$ for each previous customer. Note that methods proposed for the two actions case remain applicable when previous customers were assigned to more than two actions. For example, consider a case where Actions A, B, and C are available. Methods reviewed or proposed in this thesis can be applied to estimating the values of $Z$ for each previous customer who received Action A (by considering that previous customers were assigned either Action A or non Action A), likewise for customers who received Action B or C. This implies that the values of the unrecorded $Z$ have to be estimated using more than one individual module, and this is not preferable. Nevertheless, using one single module to estimate the values of $Z$ when more than two actions are available has been widely discussed, and related research can be found in [1] [2] [3] [19] [28] [29] [30] [31] [32].

8.2.2 Relax the linearity assumption

For simplicity reasons, we assumed the responses are linear functions of the predictor variables. Nevertheless, the linearity assumption is not generally
applicable, and it can be helpful to relax this constraint. For example, logistic regression is commonly used in application scoring, and one may wish to extend the proposed methods to cases where responses are logistic functions of explanatory variables.
Appendix
Figure A-1: The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. (When Zs are drawn from a normal distribution.)
Figure A-2: The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. (When Zs are drawn from V1.)
Figure A-3: The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. (When Zs are drawn from $V_2$.)
Among previous customers $Pr(S = 0) = 0.1$

Among previous customers $Pr(S = 0) = 0.2$

Among previous customers $Pr(S = 0) = 0.3$

Among previous customers $Pr(S = 0) = 0.4$

Among previous customers $Pr(S = 0) = 0.5$

Among previous customers $Pr(S = 0) = 0.6$

Among previous customers $Pr(S = 0) = 0.7$

Among previous customers $Pr(S = 0) = 0.8$

Among previous customers $Pr(S = 0) = 0.9$

Figure A-4: The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. (When $Z$s are drawn from $V_3$.)
Figure A-5: The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. (When Zs are drawn from V.)
Figure A-6: The performance of the ad-hoc adaptive approach and the complete-case analysis was applied to the previously collected data set. (When $Z$s are drawn from $V_5$.)

(a) Among previous customers $Pr(S = 0) = 0.1$

(b) Among previous customers $Pr(S = 0) = 0.2$

(c) Among previous customers $Pr(S = 0) = 0.3$

(d) Among previous customers $Pr(S = 0) = 0.4$

(e) Among previous customers $Pr(S = 0) = 0.5$

(f) Among previous customers $Pr(S = 0) = 0.6$

(g) Among previous customers $Pr(S = 0) = 0.7$

(h) Among previous customers $Pr(S = 0) = 0.8$

(i) Among previous customers $Pr(S = 0) = 0.9$
Figure A-7: The performance of the ad-hoc adaptive approach and Heckman’s two-step method was applied to the previously collected data set. (When Zs are drawn from a normal distribution.)
(a) Among previous customers $Pr(S = 0) = 0.1$
(b) Among previous customers $Pr(S = 0) = 0.2$
(c) Among previous customers $Pr(S = 0) = 0.3$
(d) Among previous customers $Pr(S = 0) = 0.4$
(e) Among previous customers $Pr(S = 0) = 0.5$
(f) Among previous customers $Pr(S = 0) = 0.6$
(g) Among previous customers $Pr(S = 0) = 0.7$
(h) Among previous customers $Pr(S = 0) = 0.8$
(i) Among previous customers $Pr(S = 0) = 0.9$

Figure A-8: The performance of the ad-hoc adaptive approach and Heckman's two-step method was applied to the previously collected data set. (When $Z$s are drawn from $V_1$.)
Among previous customers $Pr(S = 0) = 0.1$

Among previous customers $Pr(S = 0) = 0.2$

Among previous customers $Pr(S = 0) = 0.3$

Among previous customers $Pr(S = 0) = 0.4$

Among previous customers $Pr(S = 0) = 0.5$

Among previous customers $Pr(S = 0) = 0.6$

Among previous customers $Pr(S = 0) = 0.7$

Among previous customers $Pr(S = 0) = 0.8$

Among previous customers $Pr(S = 0) = 0.9$

Figure A-9: The performance of the ad-hoc adaptive approach and Heckman’s two-step method was applied to the previously collected data set. (When $Zs$ are drawn from $V_2$.)
Figure A-10: The performance of the ad-hoc adaptive approach and Heckman’s two-step method was applied to the previously collected data set. (When $Zs$ are drawn from $V_3$.)
Figure A-11: The performance of the ad-hoc adaptive approach and Heckman’s two-step method was applied to the previously collected data set. (When Zs are drawn from $V_4$.)
Figure A-12: The performance of the ad-hoc adaptive approach and Heckman’s two-step method was applied to the previously collected data set. (When $Z$s are drawn from $V_5$.)
Figure A-13: The performance of the naive combined least squares regression approach. (When Zs are drawn from a normal distribution.)
Figure A-14: The performance of the naive combined least squares regression approach. (When $Z$s are drawn from $V_1$.)
Figure A-15: The performance of the naive combined least squares regression approach. (When $Z$s are drawn from $V_2$.)
Figure A-16: The performance of the naive combined least squares regression approach. (When $Z$s are drawn from $V_3$.)
Figure A-17: The performance of the naive combined least squares regression approach. (When Zs are drawn from $V_4$.)
Among previous customers $Pr(S = 0) = 0.1$

Among previous customers $Pr(S = 0) = 0.2$

Among previous customers $Pr(S = 0) = 0.3$

Among previous customers $Pr(S = 0) = 0.4$

Among previous customers $Pr(S = 0) = 0.5$

Among previous customers $Pr(S = 0) = 0.6$

Among previous customers $Pr(S = 0) = 0.7$

Among previous customers $Pr(S = 0) = 0.8$

Among previous customers $Pr(S = 0) = 0.9$

Figure A-18: The performance of the naive combined least squares regression approach. (When $Z$s are drawn from $V_0$.)
Figure A-19: The performance of the adjusted combined least squares regression approach. (When $Z$s are drawn from a normal distribution.)
Figure A-20: The performance of the adjusted combined least squares regression approach. (When Zs are drawn from $V_1$.)
Figure A-21: The performance of the adjusted combined least squares regression approach. (When Zs are drawn from $V_2$.)
(a) Among previous customers $Pr(S = 0) = 0.1$
(b) Among previous customers $Pr(S = 0) = 0.2$
(c) Among previous customers $Pr(S = 0) = 0.3$
(d) Among previous customers $Pr(S = 0) = 0.4$
(e) Among previous customers $Pr(S = 0) = 0.5$
(f) Among previous customers $Pr(S = 0) = 0.6$
(g) Among previous customers $Pr(S = 0) = 0.7$
(h) Among previous customers $Pr(S = 0) = 0.8$
(i) Among previous customers $Pr(S = 0) = 0.9$

Figure A-22: The performance of the adjusted combined least squares regression approach. (When $Z$s are drawn from $V_3$.)
Figure A-23: The performance of the adjusted combined least squares regression approach. (When $Z$s are drawn from $V_1$.)
Figure A-24: The performance of the adjusted combined least squares regression approach. (When Zs are drawn from $V_5$.)

(a) Among previous customers $Pr(S = 0) = 0.1$

(b) Among previous customers $Pr(S = 0) = 0.2$

(c) Among previous customers $Pr(S = 0) = 0.3$

(d) Among previous customers $Pr(S = 0) = 0.4$

(e) Among previous customers $Pr(S = 0) = 0.5$

(f) Among previous customers $Pr(S = 0) = 0.6$

(g) Among previous customers $Pr(S = 0) = 0.7$

(h) Among previous customers $Pr(S = 0) = 0.8$

(i) Among previous customers $Pr(S = 0) = 0.9$
Figure A-25: The performance of the weighted combined least square regression approach extended from the naive combined least squares regression approach. (When Zs are drawn from a normal distribution.)
(a) Among previous customers $Pr(S = 0) = 0.1$

(b) Among previous customers $Pr(S = 0) = 0.2$

(c) Among previous customers $Pr(S = 0) = 0.3$

(d) Among previous customers $Pr(S = 0) = 0.4$

(e) Among previous customers $Pr(S = 0) = 0.5$

(f) Among previous customers $Pr(S = 0) = 0.6$

(g) Among previous customers $Pr(S = 0) = 0.7$

(h) Among previous customers $Pr(S = 0) = 0.8$

(i) Among previous customers $Pr(S = 0) = 0.9$

Figure A-26: The performance of the weighted combined least square regression approach extended from the naive combined least squares regression approach. (When $Zs$ are drawn from $V_1$.)
Figure A-27: The performance of the weighted combined least square regression approach extended from the naive combined least squares regression approach. (When $Z$s are drawn from $V_2$.)
Figure A-28: The performance of the weighted combined least square regression approach extended from the naive combined least squares regression approach. (When Zs are drawn from V₃.)
Figure A-29: The performance of the weighted combined least square regression approach extended from the naive combined least squares regression approach. (When Zs are drawn from $V_4$.)
Figure A-30: The performance of the weighted combined least square regression approach extended from the naive combined least squares regression approach. (When Zs are drawn from $V_5$.)

(a) Among previous customers $Pr(S = 0) = 0.1$
(b) Among previous customers $Pr(S = 0) = 0.2$
(c) Among previous customers $Pr(S = 0) = 0.3$
(d) Among previous customers $Pr(S = 0) = 0.4$
(e) Among previous customers $Pr(S = 0) = 0.5$
(f) Among previous customers $Pr(S = 0) = 0.6$
(g) Among previous customers $Pr(S = 0) = 0.7$
(h) Among previous customers $Pr(S = 0) = 0.8$
(i) Among previous customers $Pr(S = 0) = 0.9$
Figure A-31: The performance of the weighted combined least square regression approach extended from the adjusted combined least squares regression approach. (When $Z$s are drawn from a normal distribution.)
Figure A-32: The performance of the weighted combined least square regression approach extended from the adjusted combined least squares regression approach. (When Zs are drawn from $V_1$.)
(a) Among previous customers $Pr(S = 0) = 0.1$

(b) Among previous customers $Pr(S = 0) = 0.2$

(c) Among previous customers $Pr(S = 0) = 0.3$

(d) Among previous customers $Pr(S = 0) = 0.4$

(e) Among previous customers $Pr(S = 0) = 0.5$

(f) Among previous customers $Pr(S = 0) = 0.6$

(g) Among previous customers $Pr(S = 0) = 0.7$

(h) Among previous customers $Pr(S = 0) = 0.8$

(i) Among previous customers $Pr(S = 0) = 0.9$

Figure A-33: The performance of the weighted combined least square regression approach extended from the adjusted combined least squares regression approach. (When $Zs$ are drawn from $V_2$.)
(a) Among previous customers $Pr(S = 0) = 0.1$

(b) Among previous customers $Pr(S = 0) = 0.2$

(c) Among previous customers $Pr(S = 0) = 0.3$

(d) Among previous customers $Pr(S = 0) = 0.4$

(e) Among previous customers $Pr(S = 0) = 0.5$

(f) Among previous customers $Pr(S = 0) = 0.6$

(g) Among previous customers $Pr(S = 0) = 0.7$

(h) Among previous customers $Pr(S = 0) = 0.8$

(i) Among previous customers $Pr(S = 0) = 0.9$

Figure A-34: The performance of the weighted combined least square regression approach extended from the adjusted combined least squares regression approach. (When $Z$s are drawn from $V_3$.)
Figure A-35: The performance of the weighted combined least square regression approach extended from the adjusted combined least squares regression approach. (When Zs are drawn from $V_t$.)
(a) Among previous customers $Pr(S = 0) = 0.1$
(b) Among previous customers $Pr(S = 0) = 0.2$
(c) Among previous customers $Pr(S = 0) = 0.3$
(d) Among previous customers $Pr(S = 0) = 0.4$
(e) Among previous customers $Pr(S = 0) = 0.5$
(f) Among previous customers $Pr(S = 0) = 0.6$
(g) Among previous customers $Pr(S = 0) = 0.7$
(h) Among previous customers $Pr(S = 0) = 0.8$
(i) Among previous customers $Pr(S = 0) = 0.9$

Figure A-36: The performance of the weighted combined least square regression approach extended from the adjusted combined least squares regression approach. (When $Z$s are drawn from $V_5$.)
Bibliography


