Local Sensitivity Analysis and Bias Model Selection

Peng Yin

School of Mathematics and Statistics

University of Newcastle upon Tyne

A thesis submitted for the degree of

Doctor of Philosophy

22, January, 2014

Acknowledgements

I would like to thank my supervisor Dr. Jian Qing Shi for valuable guidance and advice, and for introducing me to the postgraduate study in the UK and inspiring me in the long but joyful doctoral journey. Thanks are due for his support, encouragement, and patience, which more than fulfilled the requirements of a supervisor role.

Thanks to my colleagues in the School of Mathematics and Statistics, and other peoples in the statistics group for sharing the knowledge and good times. I would also like to acknowledge the School of Mathematics and Statistics at Newcastle University for their financial support.

Thanks to Prof. Jane Hutton and Dr. Malcolm Farrow for being my thesis examiners and providing valuable suggestions and corrections.

I am grateful to my family, mum and girlfriend, for their endless love and support, and my best friends who have provided unwavering faith in me.

Abstract

Incomplete data analysis is often considered with other problems such as model uncertainty or non-identifiability. In this thesis I will use the idea of the local sensitivity analysis to address problems under both ignorable and non-ignorable missing data assumptions. One problem with ignorable missing data is the uncertainty for covariate density. At the mean time, the misspecification for the missing data mechanism may happen as well. Incomplete data biases are then caused by different sources and we aim to evaluate these biases and interpret them via bias parameters. Under non-ignorable missing data, the bias analysis can also be applied to analyse the difference from ignorability, and the missing data mechanism misspecification will be our primary interest in this case. Monte Carlo sensitivity analysis is proposed and developed to make bias model selection. This method combines the idea of conventional sensitivity analysis and Bayesian sensitivity analysis, with the imputation procedure and the bootstrap method used to simulate the incomplete dataset. The selection of bias models is based on the measure of the observation dataset and the simulated incomplete dataset by using K nearest neighbour distance. We further discuss the non-ignorable missing data problem under a selection model, with our developed sensitivity analysis method used to identify the bias parameters in the missing data mechanism. Finally, we discuss robust confidence intervals in meta-regression models with publication bias and missing confounder.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Problems of *model uncertainty* and *incomplete data* arise frequently in the statistical sciences. Most of the literature usually assumes that the model is correct and that we obtain observations on the variables that are described by that model: we have model certainty and complete data. In reality model certainty is always doubtful and incomplete sets of data are common.

Much of the theory and practice of statistics involves fitting parametric models for missing data, which comprises two components: one is for the complete data and the other is for the missing data mechanism (MDM). The former describes the probability distributions that fit the observations on the variables, while the latter characterizes the observation process by which some data may be missing or censored. We will first review the missing data types in Section 1.1. When we assume model certainty, many statistical techniques can be used to specify parametric models with missing data and we will review some of most popular methods in Section 1.2. Next the model uncertainty analysis will be addressed, and local bias analysis and sensitivity analysis will be discussed in Section 1.3. Since we will develop a novel approach of sensitivity analysis and make bias model selection by comparing the observed data set and a simulated data set, the choice of dissimilarities is important and the distance measures are reviewed in Section 1.4. Hypothesis testing is also considered and the procedure of permutation test is described. In Section 1.5, we will outline the structure of the thesis.

1.1 Missing Data Mechanism

Little and Rubin (2002) characterize the missing data mechanism into three types generally: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The first two missing types are usually considered as ignorable missing data while the MNAR is then named as non-ignorable missing data. Lu and Copas (2004) gave precise definitions of MAR and likelihood ignorable, and discussed the conditions when both are equivalent.

Suppose we have a n-dimensional random vector $Z = (z_1, \ldots, z_n)^T$ and a d-dimensional parameters of interest $\theta \in \Theta$. A working model $f(z;\theta)$ on $z \in \mathbb{R}^n$ can be assumed for inference. Suppose that the observation process of Z suffers from missing data and hence, we need to define a binary random vector $R = (r_1, \ldots, r_n)^T$ indicating the observational status of Z, where r_i takes the value 0 when the observation of z_i is missing and the value 1 when z_i is observed, $i = 1, \ldots, n$.

$$r_i = \begin{cases} 1, & z_i \text{ observed;} \\ 0, & z_i \text{ missing.} \end{cases} \tag{1.1}$$

The parameterization of the joint distribution of Z and R can always be fitted by the selection model form

$$f(z, r; \theta, \psi) = f(z;\theta) f(r|z, \psi),\ (\theta, \psi) \in \Theta \times \Psi, \tag{1.2}$$

with the parameters θ and ψ are assumed to be distinct (Rubin, 1976). The item $f(z;\theta)$ fits the probability density of the observations while the conditional density $f(r|z, \psi)$ characterizes the missingness process on the observations and thus specifies a model for the missing data mechanism. The pair of random variables (Z, R) induces an observable random variable Y, which is

$$Y = Y(Z, R) = (y_1, \ldots, y_n)^T. \tag{1.3}$$

where

$$y_i = \begin{cases} z_i, & \text{if}\, r_i\text{=1;} \\ \mathbb{R}, & \text{if}\, r_i\text{=0.} \end{cases} \quad i = 1, \ldots, n.$$

where the symbol $\mathbb{R}$ used in the vector argument means when $r = 0$ all we know is that the missing values are distributed at some points in $\mathbb{R} = (-\infty, -\infty)$. In this case, complete data Z can be separated into two components: the set of the observed values Z_{obs} and the set of the missing values Z_{mis}. [1] The density of incomplete data Y can be expressed as:

$$\begin{aligned} f(y;\theta,\psi) &= \int_{(y)} f(z;\theta) f(r|z;\psi) dz \\ &= f(z_{obs};\theta) \int f(z_{mis}|z_{obs};\theta) f(r|z;\psi) dz_{mis} \end{aligned} \qquad (1.4)$$

where (y) on the integration sign means the the marginal density is taken over the level set, i.e. $Y = Y(Z, R)$. Examples of level sets of $y(z, r)$ can be found in Copas and Eguchi (2005, p.463).

Thus Rubin's MAR condition can be expressed as follows. A MDM is said to be **MAR** if the conditional distribution $f(r|z;\psi)$ has the special form (Lu and Copas, 2004)

$$f(r|z;\psi) = h(y(z,r),\psi) \text{ for all } (z,r) \in Z \times R, \qquad (1.5)$$

where, for any fixed ψ and r, $h(.;\psi)$ is a function mapping real number field into [0,1]. Under MAR, the MDM depends on y only through the observed part of the sample $y = y(z, r)$.

Also it is well known that MCAR is a special case of MAR, where Z and R are statistically independent in the usual sense. Rubin (1976) and Little and Rubin (2002) distinguished between *missingness completely at random*, where the outcomes are independent of the mechanism governing missingness, and *missingness at random*, where there is dependence between, but only in the sense that missingness may depend on the observed, but not further on the unobserved measurements. Normally MAR (and MCAR) are named as ignorable in the likelihood setting. Lu and Copas (2004) give the definition of **likelihood ignorable** (LIG) to explain the meaning of it:

Definition 1. *A MDM is said to be LIG if the integral*

$$\int f(z_{mis}|z_{obs};\theta) f(r|z;\psi) dz_{mis} \qquad (1.6)$$

[1] The complete data Z can be separated in different ways for different purposes. For example, we use subscript here to denote a set of the observed values (i.e. Z_{obs}) or missing values (Z_{mis}). Also in Chapter 6, we use superscript to denote a set of variables which are always observed (Z^{obs}), or a set of variables with missing data (Z^{mis}).

is free of θ for almost all realizations of $(z, r) \in Z \times R$ and for all $(\theta, \psi) \in \Theta \times \Psi$.

And they stated that generally MAR is a necessary and sufficient condition for LIG for complete density family.

When neither MCAR nor MAR hold, we say the data are missing not at random or non-ignorable, which means that even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves. In this case, it is not always theoretically possible to characterize all parameters for this class of models given a certain choice of covariates, and this problem is termed as model non-identifiability.

1.2 Missing Data Methods

Over the last several decades a variety of models and methods are proposed to analyze incomplete data. Because standard techniques for regression models require fully observed information, one simple way to avoid the problem of missing data is to infer from the subjects that are completely observed. This method, known as a complete case (CC) analysis, is the technique most commonly used with missing values in the covariates and/or response, although it can be biased except the data are MCAR. Another ad hoc method of dealing with missing covariate data is to exclude those covariate variables with missingness from the analysis. But this procedure can lead to model misspecification (missing confounder) and is not recommended. Other approaches like maximum marginal distribution (MLE with EM algorithm), multiple imputation (MI), fully Bayesian (FB), and weighted estimating equations (WEEs) methods are getting popular for a wide variety of missing data problems, including missing covariate data in the linear regression model, generalized linear models (GLMs), survival analysis, as well as missing responses in the model of longitudinal data and meta analysis.

1.2.1 Complete Case Analysis

One simple way to avoid the missing data problem will be to use complete case analysis, excluding all units for which the outcome or any of the inputs are missing. This method has advantages such as simplicity, and comparability of univariate statistics, since these are all calculated on a common sample base of cases. However, this approach may suffer biased estimations as it discards incomplete cases and thus loss some information. Little and Rubin (2002) pointed out that the only unbiased situation is under MCAR assumption, then the complete case is just one effectively random subsample of the original dataset. But in other cases, the analysis without modification will cause seriously biased results and is not recommended.

1.2.2 Likelihood-Based Approach: EM Algorithm

The maximum likelihood method is one of the most popular methods for bias analysis with missing data. Many articles in the literature discuss missing responses and/or missing covariates under ignorable or non-ignorable assumption by this method. These include Little and Rubin (2002), Diggle and Kenward (1994), Ibrahim et al. (2005) and Molenberghs et al. (2008).

Chen and Ibrahim (2001) proposed semiparametric maximum likelihood estimators for identifiable regression coefficients. Under the same identity assumption and with a conditioning argument on MDM, Tang et al. (2003) made inferences based on a pseudo-likelihood function. Subsample ignorable likelihood for regression analysis with missing data has also been discussed by Little and Zhang (2011). Empirical likelihood based inference procedure has been proposed by Rao and Wang (2002) and Qin and Zhang (2007). There has also been some literature for likelihood based methods for establishing identifiability and asymptotic properties of estimators in missing covariate problems such as Robins and Rotnitzky (1995).

Let D_{obs} and D_{mis} denote the observed values and missing values respectively. The marginal probability density of D_{obs} is obtained by integrating out the missing data D_{mis}:

$$f(D_{obs}|\theta) = \int f(D_{obs}, D_{mis}|\theta) dD_{mis}.$$

We define the likelihood of θ based on data D_{obs} but *ignoring the missing-data mechanism* to be any function of θ proportional to $f(D_{obs}|\theta)$:

$$L(\theta|D_{obs}) \propto f(D_{obs}|\theta).$$

More generally, we can include in the model the distribution of a variable indicating whether each component of D is observed or missing. Similar to notation (1.1), we define an indicator R as follows

$$R = \begin{cases} 1, & D \text{ observed;} \\ 0, & D \text{ missing.} \end{cases} \tag{1.7}$$

We can treat R as a random variable and specify the joint distribution of R and D. The density of this distribution can be specified as the product of the densities of the distribution of D and the conditional distribution of R given D, that is,

$$f(D, R|\theta, \psi) = f(D|\theta)f(R|D, \psi).$$

The conditional distribution of R given D indexed by an unknown parameter ψ refers to the model of the missing-data mechanism we introduced. In some situations the distribution is known, and ψ is unnecessary. The actual observed data consist of the values of the variables (D_{obs}, R), and the distribution of the observed data is:

$$f(D_{obs}, R|\theta, \psi) = \int f(D_{obs}, D_{mis}|\theta)f(R|D_{obs}, D_{mis}, \psi)dD_{mis}.$$

The likelihood of θ and ψ is any function of θ and ψ proportional to the equation above:

$$L(\theta, \psi|D_{obs}, R) \propto f(D_{obs}, R|\theta, \psi).$$

And if missing data is *LIG*, then the distribution of observed data is:

$$f(D_{obs}, R|\theta, \psi) = f(R|D_{obs}, \psi)f(D_{obs}|\theta).$$

The Expectation-Maximization(EM) algorithm is a very general iterative algorithm for ML estimation in incomplete-data problems. In fact, the range of problems that can be attacked by EM is very broad and includes problems not usually considered to be ones arising from missing or incomplete data (e.g. variance components estimation, iteratively reweighted least squares). The algorithm is comprised of two steps: an Expectation step and a Maximization step. Specifically, let $\theta^{(i)}$ be the current estimate of the parameter θ. The *E step of EM* finds the expected loglikelihood if θ were $\theta^{(i)}$:

$$Q(\theta|\theta^{(i)}) = \int l(\theta|D)f(D_{mis}|D_{obs}, \theta = \theta^{(i)})dD_{mis}$$

where $l(\theta|D)$ is the log-likelihood of θ.

The *M step of EM* determines $\theta^{(i+1)}$ by maximizing this expected loglikelihood, and it has the following property:

$$Q(\theta^{(i+1)}|\theta^{(i)}) \geq Q(\theta, \theta^{(i)}), \ for \ all \ \theta.$$

The E step calculates the conditional average of the 'missing data' given the observed data conditional on the current parameter estimations, and then substitutes these expectations for the 'missing data'. The quotations around 'missing data' are there because the missing values themselves are not necessarily being substituted by EM, which is different from imputation procedure.

The M step is particularly simple to describe: perform maximum likelihood estimation of θ just as if there were no missing data, that is, as if they had been filled in. Thus the M step of EM uses the identical computational methods as ML estimation from $l(\theta|D)$. These two steps are then iterated until convergence happens. The stationary point is a global maximum and EM yields the unique maximum likelihood estimate of θ from $l(\theta, D_{obs})$ in well behaved problems (Schafer, 1997, pages 51-55), i.e. problems with not too many missing entries and not too many parameters.

1.2.3 Imputation Procedures

Imputation is another general and flexible method for handling missing data problems. There are many ways to make the fill-in, and we list some of the most popular below:

1. *Mean imputation*: where means from the responding units in the sample are substituted. The idea is to replace each missing value with the mean of the observed values for that variable. Let x_{ij} be the value of X for units j in variable $i, i = 1, \ldots, m, j = 1, \ldots, n$. Mean imputation substitutes the mean $\bar{x}_i$ of the n_i responding units for units that are sampled but that do not respond: $x_{ij}^{r=0} = \bar{x}_i^{r=1}$. However, this approach can distort the shape of distributions and then distort relationships between variables.

2. *Hot deck imputation*: can be broadly defined as a method where an imputed value is selected from an estimated distribution for each missing value, in contrast with mean imputation, where the mean of the distribution is substituted.

The simplest theory is obtained when imputed values can be selected from the values for the responding units by a probability sampling design. The hot deck with replacement selects is the most common one, but its estimator is only unbiased under unrealistic assumption that the probability of response is not related to the values of X. The nearest neighbour hot deck (Sander, 1983) and the sequential hot deck (Colledge et al., 1978) approach may be considered to improve the method.

3. *Regression imputation*: replaces missing values with predicted values from a regression of the missing item given items observed for the unit, usually calculated from units with both observed and missing variables present. One simple way is to fit a parametric regression model of variables with missingness against variables totally observed, based on observed samples only, then predict the variables with missingness by the regression model (Little and Rubin, 2002). There are many regression technique, such as stochastic regression imputation and Bayesian linear regression imputation. Generally, this method is model based imputation technique and is widely used in multiple imputation methods.

4. *Multiple imputation methods* (Rubin, 1978, 1987): impute more than one value for the missing items. This method is most widely used now and we have a detailed review below.

Multiple Imputation:

Multiple imputation was first proposed by Rubin (1978) and a comprehensive discussion can be found in Little and Rubin (2002), Schafer (1997) and Raghunathan et al. (2001). The method has valid inference on missing data problem, especially under ignorable missingness assumption and thus has a variety of applications. Single imputation introduced above has the advantage of allowing standard complete data methods of analysis, however, it is also difficult to reflect sampling variability under one model for nonresponse as pointed out by Little and Rubin (2002). While multiple imputation can overcome this problem as the method involve N complete data analyses to display variation in valid inferences across the models in dealing with uncertainty. The analysis of a multiply imputed data set is quite direct. Suppose $(\hat{\theta}_i, W_i), i = 1, \ldots, N$ are N complete-data estimates and their associated variance for an estimated θ respectively, calculated from N repeated imputations under one model. In order to make inferences for θ we average the results across the individual imputations:

$$\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i.$$

The variability associated with the estimate has two components: the within imputation variance:

$$W = \frac{1}{N}\sum \hat{W}_i,$$

and the between-imputation variance:

$$B = \frac{1}{N-1}\sum(\hat{\theta}_i - \bar{\theta})^2$$

thus the total variability is combined as

$$V = W + (1 + \frac{1}{N})B.$$

A rough 95% confidence interval can be obtained as $\bar{\theta} \pm 2V^{1/2}$, but a better calculation is to use the approximation of Student's t distribution:

$$(\theta - \bar{\theta})V^{-1/2} \sim t_\nu,$$

with the degrees of freedom,

$$\nu = (N-1)[1 + \frac{1}{N+1}\frac{W}{B}]^2.$$

Notice that when there is infinite number of imputations ($N = \infty$), the total variance V reduce to the sum of the two variance components, then the confidence interval is based on a normal distribution ($\nu = \infty$).

Rubin (1987) pointed out that the efficiency of the estimate based on N imputations of a proportion p of missing data is

$$(1 + \frac{p}{N})^{-1},$$

thus 3-10 imputations may be enough in practical examples.

As there are various imputation techniques which can be applied in practice, how to make a proper imputation strategy must be considered. MI procedure requires a mechanism and statistical assumptions to make valid inferences. The basic idea is sampling data from a conditional distribution of variables with missingness on variables without missingness. Take the missing covariates problem for example, assume response variables T is completely observed and covariate variables X is partially missing. R is the missingness indicator defined in equation (1.7). Then the imputation distribution is given as

$$f(x_{mis}|t, x_{obs}, R) \propto f(t|x, \theta)f(x)f(R|t, x, \psi).$$

Specially, when the missing data mechanism is assumed under ignorable missingness, the MDM need not to be specified in this case, and the above equation reduces as

$$f(x_{mis}|t, x_{obs}, R) \propto f(t|x, \theta)f(x).$$

1.3 Model Uncertainty and Sensitivity Analysis

An assessment of uncertainty due to incomplete data or model misspecification is a topic that has attracted many researchers for several decades, (see e.g Cornfield et al., 1959; Vemuri et al., 1969; Draper, 1995; Copas and Li, 1997), in which sensitivity analysis is one of the most commonly used approaches. It has been widely used in bias analysis for different areas, including: sensitivity analysis for publication bias in meta-analysis (Copas and Shi, 2000a,b) using the Heckman model (Heckman, 1979), sensitivity analysis for incomplete contingency tables by Molenberghs et al. (2001), local sensitivity analysis in Cook (1986), Copas and Eguchi (2001) and probabilistic sensitivity analysis in Oakley and O'Hagan (2004). Those discussions characterize the sensitivity analysis in different ways, but their aims are essentially the same: to examine the influence of individual uncertainty on model based inference. A different approach is to consider all possible sources of uncertainty by defining a prior density, and a Monte Carlo sensitivity analysis involves sampling '*bias parameters and then inverts the bias model to provide a distribution of bias-corrected estimates*' (Greenland, 2005, p.269). Also Draper (1995) evaluated the model uncertainty through Bayesian model averaging while Saha and Jones (2005) applied the bias analysis techniques to address non-identifiability issues.

1.3.1 Local Sensitivity Analysis

Copas and Eguchi (2005, 2001) discuss local model uncertainty when inference is based on incomplete data. We still use the notation defined in Section 1.1, denoting Z for complete data and Y for incomplete data. The data sampling distribution under complete data is denoted as $g_Z(z;\theta)$ and its marginal model as $g_Y(y;\theta)$. The working model f_Z under complete data (which is misspecified from g_Z) has the corresponding marginal distribution f_Y under incomplete data, inference based on f_Y (misspecified marginal model) θ_Y has a bias from inference from complete data θ_Z. The bias is named as *incomplete data bias* and the models for measuring the bias are called *bias models*. However, Lin et al. (2012) found under identifiable assumption, the working model f_Y is not always the same as the marginal model of f_Z, and extra misspecification occurs. Lin et al. (2012) extended Copas and Eguchi's work and discussed the so-called *marginal model bias* in missing confounder problem for GLMs with nonlinear link functions. The details of local sensitivity analysis for incomplete data will be discussed in Chapter 2 and 3.

1.3.2 Bias Model and Bayesian Sensitivity Analysis

Sensitivity analysis is mainly used to determine the statistical uncertainty issue in factorizing *models* or *parameter* errors. Good references about sensitivity analysis about modelling uncertainty include Saltelli et al. (2004), Saltelli et al. (2008) and Oakley and O'Hagan (2004), but in this thesis, we mainly focus on the sensitivity analysis with nuisance parameters in the missing data problem. Let D and R denote the observations vector and missingness indicator vector which takes 1 if data is observed or 0 otherwise. The complete data model can be factorized into an extrapolation model and an observed data model,

$$f(D,R|\theta) = f(D_{mis}|D_{obs},R,\theta_{mis})f(D_{obs},R|\theta_{obs}). \tag{1.8}$$

The observed data distribution $f(D_{obs},R|\theta_{obs})$ is identifiable and can be fitted by a parametric or nonparametric approach. However, the extrapolation distribution

$f(D_{mis}|D_{obs}, R, \theta_{mis})$ cannot be identified unless extra assumptions are made. Sensitivity of non-identifiable parameters should be considered carefully. Those parameters are therefore described as *sensitivity parameters* or *bias parameters* (Daniels and Hogan, 2008; Greenland, 2005), denoted by η. Local sensitivity analysis is based on derivatives of parameters of interest evaluated at some belief $\eta = \eta_0$ which helps us to understand the robustness of the practical model in a local area, but has limited value in understanding the consequences of global uncertainty about η. Global sensitivity analysis considers these more substantial changes individually without limitation (see e.g. Oakley and O'Hagan, 2004) although an unrealistically wide range is usually a troublesome problem without proper selection on the inputs. Bayesian techniques were then proposed to overcome the difficulty (McCandless et al., 2007, 2008; Gustafson et al., 2010, see e.g.), offering a route to sample smoothly via a prior distribution, and it weights possible scenarios rather than the conventional method which only reflects the investigator's plausible beliefs. Take one example in McCandless et al. (2007), let T be disease variable, X_1 as exposure and X_2, C denote the measured and unmeasured confounders respectively. They used the factorization $P(T, C|X_1, X_2) = P(T|X_1, C, X_2)P(C|X_1, X_2)$ and model the confounding effect of C using logistic regression models:

$$\text{logit}[P(T = 1|X_1, C, X_2)] = \theta_0 + \theta_1 X_1 + \theta_2 C + \theta_3 X_2,$$

$$\text{logit}[P(C = 1|X_1, X_2)] = \eta_0 + \eta_1 X_1 + \eta_2 X_2.$$

To interpret the parameter of interest $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)$, we need to specify a joint prior distribution of (θ, η) as

$$\begin{aligned} f(\theta|T, X_1, X_2, C) &= \int f(\theta|T, X_1, C, X_2, \eta) f(\eta|T, X_1, C, X_2) d\eta \\ &\propto \int P(T, X_1, C, X_2; \theta, \eta) f(\theta, \eta) d\eta. \end{aligned}$$

In principal, a prior distribution $f(\theta, \eta)$ from any standard parametric family can be used for Bayesian sensitivity analysis (BSA). In most literatures, priors are usually specified independently as

$$f(\theta, \eta) = f(\theta) f(\eta) \tag{1.9}$$

and the exponential family is always a popular choice. However, there is rarely discussions on testing the prior choice on the performance of interval estimators, since the sensitivity parameter is unknown. Depending on the specified prior distribution, the posterior average may be asymptotically biased and credible intervals may not have expected coverage probability, according to Gustafson (2005).

Monte Carlo sensitivity analysis (MCSA) is a type of Bayesian sensitivity analysis with modifications. Assuming that $f(\theta|\eta)$ is uniformly distributed and posterior distribution $f(\eta|D_{obs}, R)$ is close to the prior distribution $f(\eta)$, the MCSA procedure is to sample from

$$f(\theta|D_{obs}, R) = \int f(\theta|D_{obs}, R, \eta) f(\eta|D_{obs}, R) d\eta \approx \int f(\theta|D_{obs}, R, \eta) f(\eta) d\eta;$$

the details can be found in Greenland (2005). However, since

$$\begin{aligned} f(\eta|D_{obs}, R) &\propto \int f(D_{obs}, R|\eta, \theta_{obs}) f(\theta_{obs}|\eta) f(\eta) d\theta_{obs} \\ &= \int f(D_{obs}, R|\theta_{obs}) f(\theta_{obs}|\eta) f(\eta) d\theta_{obs} \\ &= f(D_{obs}, R|\eta) f(\eta), \end{aligned}$$

that means the posterior of the bias parameters is not equal to the prior i.e., $f(\eta|D_{obs}, R) \neq f(\eta)$; more discussion can be found in Daniels and Hogan (2008).

1.3.3 Missing Data Mechanism Bias

As well as the totally missing confounder problem, partially missing covariates issue is also very common. The literature analyse the partially missing data in partially linear models such as Liang et al. (2004) , GLMs such as Ibrahim and Lipsitz (1999), survival analysis such as Herring et al. (2004) and longitudinal data study such as Chen and Zhou (2011) etc.

The selection model $g_Z = f(Z)f(R|Z)$ and pattern mixture model $g_Z = f(Z|R)f(R)$ are two classes of models described by Little(1993,1994) for missing data problems. When the MAR assumption is plausible, the selection model formulation seems compelling because it leads to likelihood ignorable for complete density family. However, as pointed out by Little (1993), valid inference is based on knowledge of the missing data mechanism; if assumptions about the missing data mechanism are misspecified, extra uncertainty bias exists and we call it *missing data mechanism bias.* We will consider the model uncertainty under the three types of MDM respectively, and local bias analysis is conducted under identifiable assumption. And the incomplete data bias is separate, particularly as covariate bias, missing data mechanism bias and marginal model bias (for non-linear models) due to their bias sources. Bias analysis under non-ignorable missing data is particularly difficult and we can assume an ignorable working model, then the MDM bias actually measures the departure from non-ignorability. However, the true MDM model is unknown in practice and further sensitivity analysis is required.

Local sensitivity analysis for misspecified MDM will be discussed generally in Chapter 3 and the problems with non-ignorable missing data will be further disscussed in Chapter 5 and 6.

1.4 Dissimilarity

In Chapter 4, we will propose a new method for sensitivity analysis. One key step is to measure the similarity or dissimilarity between the observed data set and a simulated set.

A quantitative measure of closeness is named as *dissimilarity*, *distance* or *similarity* (a general term is *proximity*) (Everitt et al., 2011). Gower and Legendre (1986) summarized a list of similarity measures for binary data, and Gower (1971) proposed one general similarity measure to construct proximities for mixed mode data (with continuous and categorical):

$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} / \sum_{k=1}^{p} w_{ijk}$$

where s_{ijk} is the similarity between the ith and jth individual as measured by the kth variable, and w_{ijk} is typically one or zero depending on whether or not the comparison is considered valid. For binary and categorical variables with more than two categories, the component similarities, s_{ijk}, take the value one when the two individuals have the same value and zero otherwise. For continuous variables, Gower suggests using the similarity measure

$$s_{ijk} = 1 - |x_{ik} - x_{jk}| / R_k$$

where x_{ik} and x_{jk} are respectively the kth variable value of the p-dimensional observations for individuals i and j, and R_k is the range of observations for the kth variable. More suggested similarity measures can be found in Estabrook and Rodgers (1966), Legendre and Chodorowski (1977), Lerman (1987) and Ichino and Yaguchi (1994).

Dissimilarity measures or distance measures between individuals are typically calculated to describe the proximities for continuous variables, where a dissimilarity measure, d_{ij}, is termed a **distance measure** if it fulfills the *metric inequality*

$$d_{ij} + d_{im} \geq d_{jm}$$

for pairs of individuals ij, im and jm (Everitt et al., 2011). Also a series of measurement spaces have been proposed for deriving a dissimilarity matrix, such as *Euclidean distance*, *Minkowski distance* , *Canberra distance*, etc. See Table 1.4. More summary lists can be found in Gower (1985), Gower and Legendre (1986), Jajuga et al. (2003) and Everitt et al. (2011). The most commonly used distance is **Euclidean distance**

$$d_{ij} = [\sum_{k=1}^{p} (x_{ik} - x_{jk})^2]^{1/2},$$

which is a special case ($r = 2$) of the Minkowski metric

$$d_{ij} = [\sum_{k=1}^{p} (x_{ik} - x_{jk})^r]^{1/r}.$$

This distance can be interpreted as physical distance between two p-dimensional points $\boldsymbol{x}'_i = (x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{x}'_j = (x_{j1}, \ldots, x_{jp})$ in Euclidean space. It is commonly used to evaluate the proximity of objects in two or three dimensional space and it works well when a data set has ‘compact’ or ‘isolated’ clusters (Mao and Jain, 1996). Investigations of the relationships between dissimilarity matrices, distance matrices and Euclidean matrices are carried out in Gower and Legendre (1986) and Cailliez and Kuntz (1996). Another widely used distance is Mahalanobis distance, which is scaled space from the Euclidean norm but would reduce into Euclidean norm when covariance matrix shrinks into diagonal. It is given as

$$d_{ij} = [\sum_{k}^{p} (x_{i,k} - x_{j,k})^T S^{-1} (x_{i,k} - x_{j,k})^T]^{1/2}$$

Table 1.1: Dissimilarity measures for continuous data.

Measure	Formula
Euclidean distance	$d_{ij} = [\sum_{k=1}^{p} (x_{ik} - x_{jk})^2]^{1/2}$
Manhattan distance	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
Minkowski distance	$d_{ij} = (\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^r)^{1/r} \quad (r \geq 1)$
Canberra distance	$d_{ij} = \begin{cases} 0 & \text{for } x_{ik} = x_{jk}\text{=0;} \\ \sum_{k=1}^{p} \frac{\lvert x_{ik} - x_{jk} \rvert}{(\lvert x_{ik} \rvert + \lvert x_{jk} \rvert)} & \text{for } x_{ik} \neq 0 \text{ or } x_{jk} \neq 0 \end{cases}$
Pearson correlation	$d_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} (x_{ik} - \bar{x}_{i\cdot})(x_{jk} - \bar{x}_{j\cdot})/[\sum_{k=1}^{p} (x_{ik} - \bar{x}_{i\cdot})^2 \sum_{k=1}^{p} (x_{jk} - \bar{x}_{j\cdot})^2]^{1/2}$ where $\bar{x}_{i\cdot} = \sum_{k=1}^{p} x_{ik}/p$
Angular separation	$d_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk}/[\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2]^{1/2}$
Mahalanobis distance	$d_{ij} = [(x_i - x_j)^T S^{-1} (x_i - x_j)]^{1/2}$, S is covariance matrix

with S as covariance matrix.

Figure 1.1 (by Maesschalck et al., 2000) presents points with the same inter-cluster Euclidean and Mahalanobis distances from centre points by circles and ellipses respectively. The Euclidean distance spread evenly as circles while Mahalanobis distance as ellipses scaled by its covariance matrix, i.e. point 4 has the same distance as point 20 from centre under the Euclidean metric; but point 20 is farther than point 4 under the Mahalanobis metric. However, Mahalanobis space will reduce into Euclidean space if the covariance matrix is diagonal.

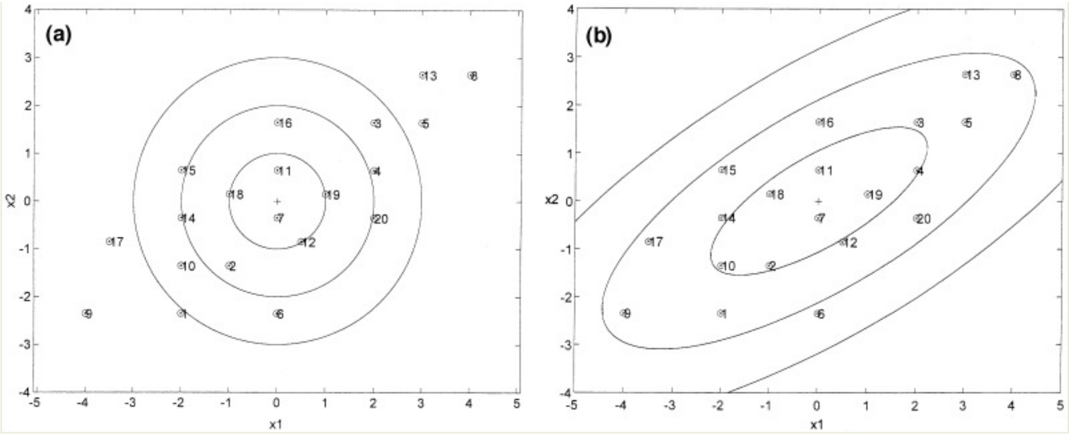


Figure 1.1: Euclidean and Mahalanobis distance illustration. (a) Plot of the simulated data for two variables x_1 and x_2 together with the circles representing equal Euclidean distances towards the centre point; (b) Plot of the simulated data for two variables x_1 and x_2 together with the ellipses representing equal Mahalanobis distances towards the centre point.

1.4.1 Nearest Neighbour Distance

A series of methodologies have been developed since people find the necessity of clustering observations into different groups, which include hierarchical and partitional approaches (hierarchical classification consists of a series of partitions while partitional methods produce only one). There is literature that summarizes these methodologies such as Jain and Dubes (1988), and Jain et al. (1999) and Everitt et al. (2011).

Hierarchical Clustering:

Everitt et al. (2011) pointed out that hierarchical clustering techniques may be subdivided into *agglomerative* methods, which begin with each pattern in a singleton cluster and merge clusters together, and *divisive* methods which separate the whole cluster (observations) into finer groupings. Most popular heuristic clustering criteria include *single linkage* (nearest neighbour), *complete linkage* (farthest neighbour) and *average linkage*. The single link was first introduced by Florek et al. (1951) and later by Sneath (1957) and Johnson (1967). It is also known as the nearest neighbour technique, but if not only the one closest individual defined as its neighbour, but kth nearest are chosen as neighbours, we call it kth Nearest Neighbour. Complete linkage is the opposite of single linkage, and the defining feature is that the distance between groups is that of the most distant pair of individuals. Average linkage - the distance between all pairs of individuals from each group or weighted average linkage

(McQuitty, 1966) works well in clustering, and these methods are compared in lots of studies including Milligan (1981), Cunningham and Ogilvie (1972), Blashfield (1976), Hubert (1974) and Duflou and Maenhaut (1990).

The single, complete, average linkage is illustrated by Figure 1.2:

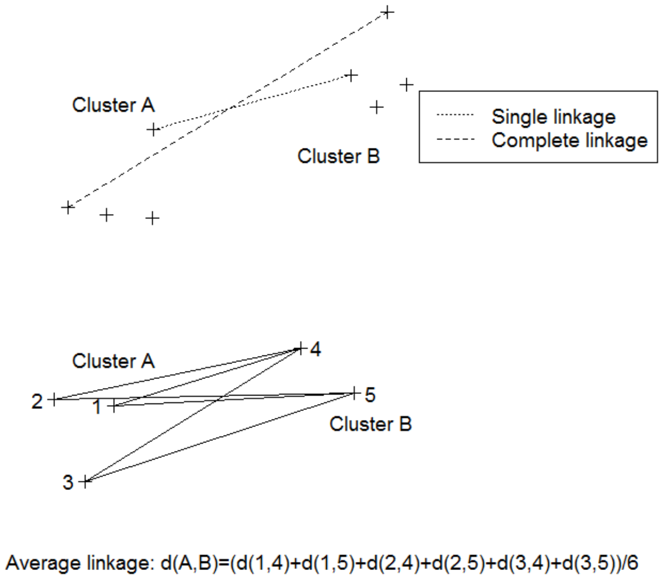


Figure 1.2: Examples of three inter-cluster distance measures: single, complete and average

1. Single linkage (Sneath, 1957): minimum distance between pair of objects, one in one cluster and one in the other.

2. Complete linkage (Sorensen, 1948): maximum distance between pair of objects, one in one cluster, one in the other.

3. Average linkage (Sokal and Michener, 1958): average distance between pair of objects, one in one cluster, one in the other.

Nearest Neighbour Clustering:

The kth Nearest Neighbour clustering procedure was proposed by Wong and Lane (1983), and it is designed to be strongly set-consistent based on density estimates. The earlier literature with discussion of density estimating in clustering procedures can be found in Bock (1979), Wishart (1969) and Ling (1972).

Let the observations $x_1, \dots x_n$ be independent, Wong and Lane (1983) estimate the density at a point x by $f_n(x)$ given by

$$f_n(x) = k/[nV_k(x)],$$

where $V_k(x)$ is the volume of the smallest sphere centred at x containing k sample observations. Then the relationship of 'neighbour' for two points is given by

Definition 2. *Two observations x_i and x_j are said to be K-neighbours if*

$$d^*(x_i, x_j) \leq d_k(x_i) \,\mathrm{or}\, d_k(x_j),$$

where d^* is the Euclidean metric and $d_k(x_i)$ is the kth nearest-neighbour distance to point x_i.

A distance matrix arises from these density estimates according to the following definition:

Definition 3. *The distance $d(x_i, x_j)$ between the observations x_i and x_j is*

$$\begin{aligned} d(x_i, x_j) &= \frac{1}{2}[\frac{1}{f_n(x_i)} + \frac{1}{f_n(x_j)}] \\ &= \begin{cases} \frac{n}{2k}[V_k(x_i) + V_k(x_j)] & \textit{if } x_i \textit{ and } x_j \textit{ are neighbours} \\ \infty & \textit{otherwise.} \end{cases} \end{aligned}$$

The kth nearest neighbor rule is considered the simplest and most intuitively appealing nonparametric classification procedure (Hall et al., 2008). However application of this method is inhibited by lack of knowledge about its properties, in particular, the parameter selection, and the absence of techniques for empirical choice of k, and the presence of noisy or irrelevant features. Much effort has been exerted in selecting or scaling features to improve classification. Wong and Lane (1983) suggested $k = 2log_2N$ to be effective for sample size N from 50 to 500 (see Wong and Schaack, 1982). And its increase should correspond to the increase in sample size. Hall et al.

(2008) detailed the way in which the value of k determines the misclassification error, and advised empirical choice of k to minimize the average error rate. They considered the Possion and Binomial models for training samples, and the kth nearest neighbour method locates the cluster position for each test sample. However, we find these choices are relatively conservative for the sensitivity analysis. In practice, a series of k may be considered, for example, different values of K are used in KNN regression and classification in R package 'caret' (from Jed Wing et al., 2013).

1.4.2 Permutation Test

Compared with the abundant discussion about cluster algorithm procedures, little research has investigated the properties of significance tests for distinguishing between the hypothesis H_0 of a 'homogeneous' population and an alternative H_1 involving 'clustering' or 'heterogeneity'. But fortunately Lee (1979) and Bock (1985) and most recently Auffermann et al. (2002) contributed to this area. The likelihood ratio (LR) and union-intersection (UI) criteria and a 'linear discrimination' statistics are shown in Lee (1979), and these tests are claimed to be equivalent. Meanwhile Bock (1985) considered four types of test statistics: the largest gap between observations, their mean distance (or similarity), the minimum with-in cluster sum of squares resulting from a k-mean algorithm and the resulting maximum F statistics. These tests are used to investigate the uniformity and unimodality hypothesis and alternatives. Although Bock (1985) provided theoretical discussion of the test measure, and a possible threshold is suggested with the measurement statistics distribution (asymptotically) estimated, the accuracy for the critical threshold and the power of the test still need to reconsidered. With the development of computing technology, the bootstrap algorithm (Efron and Tibshirani, 1993) was applied in testing fMRI data by Auffermann et al. (2002), where Fisher's linear discriminant function (Fisher, 1936) is chosen as the statistical measure.

Permutation Test and Bootstrap Test:

When we consider the two samples/clusters problem, Fisher's permutation test (Fisher, 1971) is popularly used. Our target is to test the null hypothesis H_0 of no difference between two groups $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$,

$$H_0 : \boldsymbol{X_1} = \boldsymbol{X_2}$$

The equality here means $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ assign equal probabilities to all sets, $Prob_{\boldsymbol{X_1}}\{A\} = Prob_{\boldsymbol{X_2}}\{A\}$ for any subset A of the common sample space of the x_1 and x_2. Normally the test statistic can be the mean difference, $\hat{d} = |\bar{X_1} - \bar{X_2}|$ (for scalar variables), and we expect that if the H_0 is not true, the value of $\hat{d}$ will be larger than if H_0 is true. To carry out the test, the **achieved significance level**(ASL) of the test is defined as the probability of observing at least that large a value $\hat{d}^*$ when the null hypothesis is true,

$$ASL = Prob_{H_0}\{\hat{d}^* \geq \hat{d}\}.$$

The smaller the value of ASL, the stronger the evidence against H_0. Fisher's permutation test is a clever way of calculating an ASL for the general null hypothesis $\boldsymbol{X_1} = \boldsymbol{X_2}$. First of all, we combine the two groups together as $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2})$, with sample size $N = n_1 + n_2$. We re-write the data frame as $D = (\boldsymbol{X}, \boldsymbol{R})$, where vector $\boldsymbol{R}$ indicates which group each observation belongs to. It consists of n_1 individuals from group 1 and n_2 individuals from group 2, there are $\binom{N}{n_1}$ possible R vectors, corresponding to all possible ways of partitioning N elements into two subsets of size n_1 and n_2. Permutation theory thus considers the permutations of x_1's and x_2's as equally likely if H_0 is true. In other words, let $\hat{d} = S(\boldsymbol{R}, \boldsymbol{X})$ for some function S, and for any one of the $\binom{N}{n_1}$ possible vectors $\boldsymbol{R}^*$, the corresponding test statistics

$$\hat{d}^* = \hat{d}(\boldsymbol{R}^*) = S(\boldsymbol{R}^*, \boldsymbol{X})$$

should be the same as $\hat{d}$ under H_0. The distribution that puts probability $1/\binom{N}{n_1}$ on each one of these $(\hat{d}^*)$ is called the permutation distribution of $\hat{d}$. The permutation ASL is defined to be the permutation probability that $\hat{d}^*$ exceeds $\hat{d}$,

$$\begin{aligned} ASL_{perm} &= Prob_{perm}\{\hat{d}^* \geq \hat{d}\} \\ &= \#\{\hat{d}^* \geq \hat{d}\}/\binom{N}{n_1} \end{aligned}$$

where $\#\{.\}$ denotes the cardinality of the set.

Bootstrap method can be applied to calculate the ASL, which can be done by

$$\widehat{ASL}_{perm} = \#\{\hat{d}^*(b) \geq \hat{d}\}/B$$

where $b = 1, \ldots, B$ and B is the replication number.

More generally, two quantities of carrying out a bootstrap hypothesis test are:

1. A test statistic $t(\boldsymbol{x})$.
2. A null distribution $\hat{F}_0$ for the data under H_0.

The empirical distribution $\hat{F}_0$ is a nonparametric estimate specified by the null hypothesis H_0 given $\boldsymbol{X}$.

Given these, we generate B bootstrap values of $t(\boldsymbol{x}^*)$ under $\hat{F}_0$ and estimate the achieved significance level by

$$\widehat{ASL}_{boot} = \#\{t(\boldsymbol{x}^{*b}) \geq t(\boldsymbol{x})\}/B$$

Bootstrap tests are useful in situations where the alternative hypothesis in not-well specified, and normally it requires a large B. The choice of test statistic $t(\boldsymbol{x})$ will determine the power of the test, that is , the chance that we reject H_0 when it is false.

Permutation algorithm is quite similar to bootstrap algorithm, and the main difference is that permutation sampling is carried out without replacement while bootstrap with replacement. And their efficiencies are about the same.

1.5 Structure of the Thesis

This thesis mainly focuses on the missing data problem with the model uncertainty issue, and the procedure of missingness can be separated into ignorable and non-ignorable assumptions. Local bias analysis is conducted using an ML method to assess the impact on the estimation of parameters of interest. We recognize that the statistical modelling assumption with parametric models is questioned as the lack of identifiablility or the lack of randomization, thus sensitivity analysis is applied to these problems.

The structure of the thesis is as follows. We first use incomplete data bias analysis to address the model uncertainty problems. In Chapter 2, we will discuss the covariate distribution misspecification for partially missing confounder problems. And in Chapter 3, the covariate distribution misspecification and missing data mechanism misspecification are both investigated. We use some examples to illustrate the uncertainty issue and the local bias analysis. In Chapter 4, we concentrate on measuring the uncertainty sources and propose a novel Monte Carlo sensitivity analysis method to make bias model selection (MC-BMS). Under ignorable missingness, the uncertainty about covariates distribution will be the primary concern. And in Chapter 5, we further apply the incomplete data bias analysis to non-ignorable missing data. And the missing data mechanism bias is calculated given covariate distribution, although it may be difficult to specify in practice. Further discussion based on the MC-BMS method for covariate density specification (based on pattern mixture model frame) will be given in Chapter 5 and discussion for missing data mechanism modelling (based on selection model frame) will be given in Chapter 6. We also discuss the other missing data problem for meta-analysis in Chapter 7, such as publication bias and missing confounder problems. And a robust confidence interval is proposed for meta regression models. Chapter 8 contains conclusions and suggestions for future work.

Chapter 2

Local Sensitivity Analysis for Missing Covariates Problems

2.1 Introduction

Copas and Eguchi (2005) discussed the model uncertainty issue with missing data by local bias analysis. They used a parametric model for inference when the data generating distribution is close to but not necessarily part of the considered parametric model. Bias is caused by the misspecified working model under incomplete data Y, and the bias is called *incomplete data bias* by Copas and Eguchi (2005). Lin et al. (2012) noticed that the actual working model may be a conditional model rather than the marginal model under incomplete data, and the so-called *marginal model bias* is measured under an identifiable local analysis assumption. We follow up their work and extend to partially missing data under ignorable assumption. The bias analysis is a useful tool for identifying the uncertainty parameters (termed as *bias parameters*) and analysing the model misspecifications, and we will apply it to missing data mechanism misspecification in Chapter 3 and non-ignorable missing data in Chapter 5.

We will introduce Copas and Eguchi's discussion about uncertainty analysis for missing data problems in Section 2.2, and interpret the incomplete data bias via bias parameters. One example about missing confounder problem will be discussed in Section 2.2.1. We further extend the inference to partially missing confounder problems, and argue that the model uncertainty issue is also important in this case due to the lack of identifiablility or the lack of randomization. The incomplete data bias analysis will be performed for a linear regression model under missing completely at random in Section 2.3 and missing at random assumption in Section 2.4. Then we will discuss the 'double misspecified' problems for generalized linear models in Section 2.5.

2.1.1 Missing Data Problem

Incomplete data is very common in epidemiology trials, and an example which illustrate some of the missing data problems is the case control studies to assess the link between alcohol consumption and breast cancer. A linear regression model may be assumed to examine the effect of alcohol use (denoted as variable X) towards breast cancer case (the log odds ratio is taken as the response variable T). Longnecker et al. (1988) reported significant association between the consumption of alcohol and the risk of breast cancer based on a meta-analysis of 16 published epidemiological studies. As agreed by these and later researchers, the estimation of parameter (denoted as θ_x) should be adjusted for the potential confounders (e.g. age, see Garland et al., 1999), which is denoted as C. The regression model is given as

$$t = \theta_0 + \theta_x x + \theta_c c + e \tag{2.1}$$

where $(\theta_0, \theta_x, \theta_c)$ are regression coefficients and $e \sim N(0, \sigma^2)$ brings t variation.

In practice, the confounder C is not always observed unfortunately and this analysis is likely to be influenced by missing the values and may lead to potential bias. This dissertation analyses the incomplete data biases for the missing data problems and also try to interpret the bias sources via bias parameters. The models we used for bias analysis is then named *bias models*.

2.2 Model Uncertainty and Incomplete-Data Bias

A statistical model is merely a parameterized family of probability distributions to which we believe the true distribution belongs (Amari, 1985). Given collected data, we specify a model $\{f(., \theta), \theta \in \Theta\}$ for inference about parameter θ, which is usually a vector and our interest may be part of it. We conceptually assume the observed data is from the true distribution, however in practice, data generating distribution, denoted as g, is not always equal to f. Also we should consider the influence of missing data.

Copas and Eguchi (2005) discussed the model uncertainty issue and incomplete-data bias analysis. They suggested a rather general asymptotic setting for exploring the link between local model uncertainty, defined in an appropriate way. For complete data Z and incomplete data Y, parametric models g_Z and g_Y specify the distribution of z and y respectively. In many cases, inference is based on a working model f_Z while z is in fact generated by a nearby distribution g_Z . Following Copas and Eguchi's discussion, to formulate distribution in a local neighbourhood of f_Z, let $u_Z(z;\theta)$ be any scalar function of z and parameter θ, standardized to have mean 0 and variance 1 under the model f_Z. Then for small values of ϵ, the sampling model

$$g_Z = g_Z(z;\theta,\epsilon,u_Z) = f_Z(z;\theta)\exp\{\epsilon u_Z(z;\theta)\} \tag{2.2}$$

is non-negative and integrates to 1 up to and including first-order terms in ϵ, and so identifies a distribution in the neighbourhood of f_Z. If $\epsilon = 0$ then $g_Z = f_Z$ meaning the working model is the correct model. Intuitively, ϵ can be thought of as the 'magnitude' of misspecification and u_Z can be thought of as the 'direction' of misspecification. If we fix ϵ and imagine θ and u_Z ranging over all possibilities, g_Z will cover all distribution within a 'tubular neighbourhood' of 'radius' ϵ around the working mode f_Z. And the distribution of $y = y(z)$ that is induced by g_Z is

$$\begin{aligned} g_Y &= g_Y(y;\theta,\epsilon,u_Z) \\ &= \int_{(y)} f_Z(z;\theta)\exp\{\epsilon u_Z(z;\theta)\}dz \\ &\approx f_Y(y;\theta)\exp\{\epsilon u_Y(y;\theta)\}, \end{aligned} \tag{2.3}$$

where $u_Y(y;\theta) = E_f\{u_Z(z;\theta)|y\}$ and f_Y is the corresponding working model of f_Z for incomplete data: $f_Y = \int_{(y)} f_Z dz$. The notation (y) on the intergration sign is interpreted in Section 1.1. These and later approximations are correct to first-order in terms of ϵ. We put these inferences into the following lemma:

Lemma 2.1. *The data sampling distribution under complete data (Z) is g_Z (Equation 2.2), which has the corresponding 'working model' f_Z. Correspondingly, the incomplete data (Y) distribution is marginal of g_Z denoted g_Y, which has the corresponding* *'working model' f_Y. The estimation of parameters is*

$$\theta_{gZ} = arg_\theta[E_g\{s_Z(z;\theta)\} = 0] \approx \theta + \epsilon I_Z^{-1} E_f\{u_Z(z;\theta)s_Z(z;\theta)\},$$

which is the limit of MLE when we use working model f_Z but the sampling model is g_Z. The estimation from Y is

$$\theta_{gY} = arg_\theta[E_g\{s_Y(y;\theta)\} = 0] \approx \theta + \epsilon I_Y^{-1} E_f\{u_Y(y;\theta)s_Y(y;\theta)\}.$$

Under the identifiability condition (see Lin et al., 2012), the incomplete-data bias b_θ is defined as the first-order approximation to the difference $\theta_{gY} - \theta_{gZ}$, which is given by

$$\theta_{gY} - \theta_{gZ} \approx b_\theta = \epsilon E_f[u_Z(z;\theta)\{I_Y^{-1}s_Y(y;\theta) - I_Z^{-1}s_Z(z;\theta)\}] \tag{2.4}$$

with I_Y, I_Z, s_Y, s_Z as information matrices and score vectors of f_Y, f_Z respectively.

Detailed proof for Lemma 2.1 is given in Appendix 2.7.1. Lemma 2.1 uses the first order approximation to estimate the bias, and thus require a local analysis assumption to make inference validly, which means that the misspecification quantity ϵ is small so that f_Z is in local neighbour of g_Z.

Notice that Copas and Eguchi's definition of **incomplete data bias** is given as $(\theta_{gY} - \theta_{gZ})$, which is the difference of estimators from incomplete data distribution g_Y and complete data distribution g_Z. In most literatures, the bias is commonly defined as the difference between the estimator and true value, that is $(\theta_{gY} - \theta_{true})$. Copas and Eguchi (2005) (page 470) argued that the difference of $(\theta_{gZ} - \theta_{true})$ is the difference of 'object of interest' θ^{INT} and 'object of inference' θ^{INF}. This is a fundamental problem on how to interpret θ. For example, if θ^{INT} is the mean of the population from which we are sampling, and object of inference θ^{INF} is the value of θ for which the model (noted as g_Z) is closest to the true distribution in the sense of Kullback-Leibler divergence. Royall and Tsou (2003) found $\theta^{INF} = \theta^{INT}$ for the model $N(\theta, \sigma^2)$ or Possion (θ) when the model fails, but not for log-normal distribution. They also argued that parametric inference about θ is meaningful only when $\theta^{INF} = \theta^{INT}$. Our discussion is based on this assumption, then the difference between θ_{gZ} and θ is not a bias but rather an artifact of the notations.

2.2.1 Linear Model with Missing Confounder

Assume we have an experiment design which contains response variable T, and independent covariates X and C. Variable X usually describes the treatments or therapies in clinical research. And C represents the confounder. We denote $\boldsymbol{Z} = (T, X, C)$ and $\boldsymbol{Y} = (T, X)$ as the complete and incomplete data respectively, with confounder C missing for all observations. We suppose $c \sim N(0, \sigma_c^2)$ here but the results can be extended to other distributions.

Complete data $\boldsymbol{Z} = (T, X, C)$ follows the distribution:

$$g_Z = f_{T|XC}(t, x, c) f_{XC}(x, c).$$

If X and C are assumed independent, the working model under $\boldsymbol{Z}$ is

$$f_Z = f_{T|XC}(t, x, c) f_X(x) f_C(c).$$

According to equation (2.3), incomplete data $\boldsymbol{Y} = (T, X)$ has distribution:

$$g_Y = f_Y \exp(\epsilon u_Y)$$

where f_Y is the working model under $\boldsymbol{Y}$

$$f_Y = f_{T|X}(t, x) f_X(x)$$

which is actually the marginal model of f_Z for the linear regression model if residuals are normally distributed, see the detailed discussion in Appendix 2.7.3. But if residuals are not normally distributed, the 'double misspecification' may be considered, see Lin et al. (2012). This case usually happens in non-linear models or generalized linear models, and we will discuss this issue in Section 2.5.

Assume that the response variable has a linear regression model:

$$t|(x, c) \sim N(\theta_0 + \theta_x x + \theta_c c, \sigma^2) \tag{2.5}$$

where σ^2 is the variance of t given x and c. If variable c is hidden, then the observable response distribution is an ordinary regression model without c

$$t|x \sim N(\theta_0 + \theta_x x, \sigma^2_{t|x}). \tag{2.6}$$

where $\sigma^2_{t|x}$ is the variance of t given x. Residuals are assumed to be i.i.d with covariates and it can be proved that $\sigma^2_{t|x} = \sigma^2 + \theta_c^2 \sigma_c^2$. Since x is a scalar variable, there is a bound to limit the quantity of incomplete data bias (Copas and Eguchi, 2005):

Lemma 2.2. *Incomplete Data Bias for Linear Model:*
The bias of parameter estimation b_{θ_x} (θ_x-component) for linear model between complete data model and incomplete data model is bounded by

$$\frac{b^2_{\theta_x}}{n\text{var}_f(\hat{\theta}_x)} \leq \text{corr}(t, c|x)^2 \text{corr}(x, c)^2 \tag{2.7}$$

where n is the sample size.

The first term on the right-hand side of inequality (2.7) is proportional to the partial correlation between t and c given x, which measures how much we lose since not observing the hidden variable c. The second term is the dependence between the treatments (X) and confounder (C) that is caused by the lack of randomization, which is a measure of non-ignorability in the design. The most troublesome confounder is one which is linearly correlated with treatment, see more discussion in Appendix 2.7.2.

Corollary 2.1. *When corr$(x, c) = 0$ and under the ignorable assumption we have $b_{\theta_x} = 0$.*

Below we will extend the uncertainty problems for partially missing confounder data problems.

2.3 Partially Missing Confounder under MCAR

2.3.1 Bias Models

In this section, we continue to discuss the missing data problem in (2.1). Now confounder C is partially missing with probability π, and suppose its missing type is missing completely at random, which indicates π is a constant.

We denote R as an indicator vector:

$$r = \begin{cases} 1, & c \text{ observed;} \\ 0, & c \text{ missing.} \end{cases} \tag{2.8}$$

The complete data set is $\boldsymbol{Z} = (T, X, C, R)$, with R Bernoulli distributed $R \sim B(1, \pi)$. And the corresponding incomplete data set is $\boldsymbol{Y} = (T, X, C^{(r)}, R)$, where

$$c^{(r)} = \begin{cases} c, & r = 1; \\ \mathbb{R}, & r = 0. \end{cases} \tag{2.9}$$

The symbol $\mathbb{R}$ used here means when $r = 0$ all we know is that c takes some value in $\mathbb{R} = (-\infty, \infty)$.

Starting from the sampling model, we rewrite the density function g_Z as follows:

$$\begin{aligned} g_Z(z; \theta, \pi) &= f_Z \exp\{\epsilon u_z\} \\ &= f_{T|XC}(t|x, c; \theta) f_{XC}(x, c) h(r; \pi) \end{aligned}$$

where the missing data mechanism component is $h(r; \pi) = \pi^r (1 - \pi)^{1-r}$. And the working model (assuming X and C are independent) is given as :

$$f_Z = f_{T|XC}(t|x, c; \theta) f_X(x) f_C(c) h(r; \pi). \tag{2.10}$$

Then the misspecification of the model is caused by the association between observed variable X and missing variable C, represented by

$$\exp\{\epsilon u_z\} = \frac{f_{XC}(x, c)}{f_X(x) f_C(c)}. \tag{2.11}$$

As the misspecification is related to $[XC]$ only [1], we write u_Z as u_{XC} in the following. And b_{XC} represents the incomplete data bias b_θ caused by covariate density misspecification. Here the incomplete data bias can also be termed covariate bias according to the bias source.

[1] [.] is used throughout this thesis to denote a generic distribution

As covariate C is partially missing, we split all cases $\boldsymbol{Y}$ into two parts: complete cases and incomplete cases, $\boldsymbol{Y} = (\boldsymbol{Y_{cc}}, \boldsymbol{Y_{ic}})$. For complete cases (when $r = 1$)

$$f_{Y_{cc}} = f_Z = f_{T|XC}(t|x,c)f_X(x)f_C(c)h(r=1;\pi).$$

While for incomplete cases (when $r = 0$)

$$\begin{aligned} f_{Y_{ic}} &= \int_{(y)} f_Z dz \\ &= \int_{(y)} f(t,x,c)h(r=0;\pi)dz \\ &= \int_c f_{T|XC}(t|x,c)f_X(x)f_C(c)h(r=0;\pi)dc \\ &= f_X(x)h(r=0;\pi)\int_c f_{T|XC}(t|x,c)f_C(c)dc \\ &= f_{T|X}(t|x)f_X(x)h(r=0;\pi), \end{aligned}$$

where $f_{T|X}(t|x) = \int_c f_{T|XC}(t|x,c)f_C(c)dc$. Similarly to the discussion given in Section 2.2.1, the working model under incomplete data is $f_Y = \int_{(y)} f_Z dz$, the marginal model of complete data working model.

Then we write the models into one general form:

$$f_Y = f^r_{T|XC}(t|x,c;\theta)f_X(x)f^r_C(c)h(r;\pi) \tag{2.12}$$

where

$$f^r_{T|XC}(c) = \begin{cases} f_{T|XC}, & r{=}1; \\ f_{T|X}, & r{=}0. \end{cases} \tag{2.13}$$

and

$$f^r_C(c) = \begin{cases} f_C(c), & r{=}1; \\ 1, & r = 0. \end{cases} \tag{2.14}$$

Estimation of parameters θ from f_Y is calculated by maximizing the log-likelihood of (2.12), which is biased if covariate correlation is not equal to zero. The incomplete data bias analysis is conducted below.

2.3.2 Incomplete Data Bias

For complete data $\boldsymbol{Z} = (T, X, C, R)$, we use a linear fixed effect model to fit the data

$$t|(x,c) \sim N(\theta_0 + \theta_x^T x + \theta_c c, \sigma^2) \tag{2.15}$$

where σ^2 is the variance of error and variable x can be a vector. And with the incomplete data :

$$t|(x,c,r) \sim N(\theta_0 + \theta_x^T x + r\theta_c c, \sigma^2 + (1-r)\theta_c^2\sigma_c^2).$$

For complete cases, the incomplete data model is the same with the complete data model; while for incomplete cases, we assume $(t|x, r = 0) \sim N(\theta_0 + \theta_x^T x, \sigma^2 + \theta_c^2\sigma_c^2)$ which is similar to the totally missing confounder problem discussed in Section 2.2.1. Here we use the ML method to estimate the parameters $\theta = (\theta_0, \theta_x^T, \theta_c)$ and the *incomplete-data bias*. For complete data and incomplete data, the log-likelihood for the linear model is

$$\begin{aligned} l_Z(\theta; z) &= \log f(t|x,c;\theta) \\ &= \text{Cons} - \log h(r) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\frac{(t-\theta_0-\theta_x^T x - \theta_c c)^2}{\sigma^2} \end{aligned}$$

$$\begin{aligned} l_Y(\theta; y) &= \log f(t|x,c,r;\theta) \\ &= \text{Cons} - \log h(r) - \frac{1}{2}\log(\sigma^2 + (1-r)\theta_c^2\sigma_c^2) - \frac{1}{2}\frac{(t-\theta_0-\theta_x^T x - r\theta_c c)^2}{\sigma^2 + (1-r)\theta_c^2\sigma_c^2}. \end{aligned}$$

The above formulas have component $-\log h(r)$ which is constant under MCAR and thus can be ignored. Here, we assume σ^2 is given (it can be replaced by its estimation s^2 which can be obtained from each study). From log-likelihood l_Z and l_Y, the score functions under complete data and incomplete data are

$$\boldsymbol{s_Z}(z;\theta) = \begin{pmatrix} \frac{t-\theta_0-\theta_x^T x-\theta_c c}{\sigma^2} \\ \frac{(t-\theta_0-\theta_x^T x-\theta_c c)x}{\sigma^2} \\ \frac{(t-\theta_0-\theta_x^T x-\theta_c c)c}{\sigma^2} \end{pmatrix} \text{ and } \boldsymbol{s_Y}(y;\theta) = \begin{pmatrix} \frac{t-\theta_0-\theta_x^T x-r\theta_c c}{\sigma^2+(1-r)\theta_c^2\sigma_c^2} \\ \frac{(t-\theta_0-\theta_x^T x-r\theta_c c)x}{\sigma^2+(1-r)\theta_c^2\sigma_c^2} \\ \frac{(t-\theta_0-\theta_x^T x-r\theta_c c)rc}{\sigma^2+(1-r)\theta_c^2\sigma_c^2} \end{pmatrix}$$

respectively. Fisher information matrix for complete data is

$$\boldsymbol{I_Z} = \frac{1}{\sigma^2}\begin{pmatrix} 1 & \mu_x & \mu_c \\ \mu_x & \Sigma_x & E(xc) \\ \mu_c & E(xc) & \Sigma_c \end{pmatrix}, \tag{2.16}$$

where $\mu_x = E(x)$, $\mu_c = E(c)$, $\Sigma_x = E(xx^T)$, $\Sigma_c = E(c^2)$. Similarly,

$$\begin{aligned}\boldsymbol{I_Y} = &\frac{\pi}{\sigma^2}\begin{pmatrix} 1 & E(x|r=1) & E(c|r=1) \\ E(x|r=1) & E(xx^T|r=1) & E(xc|r=1) \\ E(c|r=1) & E(xc|r=1) & E(c^2|r=1) \end{pmatrix} \\ &+\frac{1-\pi}{\sigma_Y^2}\begin{pmatrix} 1 & E(x|r=0) & 0 \\ E(x|r=0) & E(xx^T|r=0) & 0 \\ 0 & 0 & 0 \end{pmatrix}\end{aligned} \tag{2.17}$$

where $\sigma_Y^2 = \sigma^2 + \theta_c^2\sigma_c^2$. Using Lemma 2.1, we have the incomplete data bias as

$$\theta_{gY} - \theta_{gZ} = \epsilon E_f[u_Z(z;\theta)\{\boldsymbol{I_Y}^{-1}\boldsymbol{s_Y}(y;\theta) - \boldsymbol{I_Z}^{-1}\boldsymbol{s_Z}(z;\theta)\}].$$

In this chapter, we only concentrate on the covaraite distribution misspecification, and the incomplete data bias is mainly generated by the correlation between X and C, so we call it *covariate bias* particularly, denoted by $\boldsymbol{b_{XC}}$. For simplifying notations, we define $\boldsymbol{v} = (1, x, rc)^T$, $\boldsymbol{v_1} = (1, x, c)^T$ and $\boldsymbol{v_0} = (1, x, 0)^T$. Since $E_{T|XC}(\boldsymbol{s_Z}) = 0$ for all x and c, thus

$$\begin{aligned}\boldsymbol{b_{XC}} &= \epsilon\boldsymbol{I_Y}^{-1}E_{f_Z}\{u_{XC}\boldsymbol{s_Y}\} - \epsilon\boldsymbol{I_Z}^{-1}E_{f_Z}\{u_{XC}\boldsymbol{s_Z}\} \\ &= \epsilon\boldsymbol{I_Y}^{-1}[\pi E\{u_{XC}\boldsymbol{s_{Y|r=1}}\} + (1-\pi)E\{u_{XC}\boldsymbol{s_{Y|r=0}}\}] \\ &\quad -\epsilon\boldsymbol{I_Z}^{-1}E_{XC}\{u_{XC}E_{T|XC}(\boldsymbol{s_Z})\} \\ &\approx \epsilon\boldsymbol{I_Y}^{-1}(1-\pi)E[u_{XC}\frac{c\theta_c}{\sigma_Y^2}\boldsymbol{v_0}] \\ &= \epsilon\theta_c(1-\pi)\frac{\boldsymbol{I_Y}^{-1}}{\sigma_Y}E(cu_{XC}\boldsymbol{v_0}).\end{aligned}$$

In this chapter we consider the uncertainty caused by missing covariate and thus the misspecification of $f(x,c)$. As shown in formula (2.11), if the misspecification quantity is small, then

$$\begin{aligned}
E_{f_Z}(\epsilon c u_{XC}\boldsymbol{v_0}) &= E_{f_Z}\{c\boldsymbol{v_0}\log\frac{f(x,c)}{f(x)f(c)}\} \\
&\approx E_{X,C}\{c\boldsymbol{v_0}\frac{f(x,c)-f(x)f(c)}{f(x)f(c)}\} \\
&= E_{XC}(c\boldsymbol{v_0}) - E_{X,C}(c\boldsymbol{v_0}) \\
&= (0,\mathrm{cov}(x,c),0)^T.
\end{aligned}$$

Here E_{XC} indicates the expectation under distribution $f(x,c)$, while $E_{X,C}$ indicates the expectation under independent distribution $f(x)f(c)$. So we have the incomplete data bias for θ_{gY} as

$$\boldsymbol{b_{XC}} = \frac{\theta_c(1-\pi)\boldsymbol{I_Y}^{-1}}{\sigma_Y^2}\begin{pmatrix} 0 \\ \mathrm{cov}(x,c) \\ 0 \end{pmatrix}. \tag{2.18}$$

If covariate correlation $\mathrm{corr}(x,c)=0$, the incomplete data bias $\boldsymbol{b_{XC}}=\mathbf{0}$ as we stated in Corollary 2.1. If we write the inverse of the Fisher information matrix $\boldsymbol{I_Y}$ as

$$\boldsymbol{I_Y}^{-1} = \begin{pmatrix} I^{\theta_0\theta_0} & I^{\theta_0\theta_x} & I^{\theta_0\theta_c} \\ I^{\theta_0\theta_x} & I^{\theta_x\theta_x} & I^{\theta_x\theta_c} \\ I^{\theta_0\theta_c} & I^{\theta_x\theta_c} & I^{\theta_c\theta_c} \end{pmatrix},$$

then we have the incomplete data bias for θ_x-component:

$$b_{\theta_x} \approx \theta_c(1-\pi)\frac{I^{\theta_x\theta_x}}{\sigma_Y^2}\mathrm{cov}(x,c). \tag{2.19}$$

If covariate C is totally missing ($\pi=0$), then $I^{\theta_x\theta_x}=\frac{\sigma_Y^2}{\sigma_x^2}$and

$$b_{\theta_x} = \theta_c\frac{I^{\theta_x\theta_x}}{\sigma_Y^2}\mathrm{cov}(x,c) = (I^{\theta_x\theta_x})^{1/2}\mathrm{corr}(t,c|x)\mathrm{corr}(x,c) \tag{2.20}$$

since

$$\mathrm{corr}^2(t,c|x) \approx \frac{\theta_c^2\sigma_c^2}{\sigma_Y^2}$$

for the regression model (2.15). It is easy to notice from equations (2.19) and (2.20) that the incomplete data bias for partially missing data is also impacted by the covariate correlation, but the size of the bias gets smaller than totally missing confounder problem as $(1-\pi) < 1$.

2.4 Partially Missing Confounder under MAR

Another missing data mechanism is termed as missing at random (MAR, Rubin, 1976). It means that the probability that a variable is observed/missing depends on the values of the other completely observed variables. This concept has been extensively studied, and effective computational methods for handling missing data under the MAR assumption have been developed, for example, using EM algorithm or Multiple Imputation. Good references include Tanner (1993), Schafer (1997), Kenward and Molenberghs (1998), and Little and Rubin (2002) among many others.

In this section, we extend the incomplete data bias analysis of the missing covariates problem to the missing at random assumption. Under the covariate distribution misspecification setting, the likelihood for working model f_Z with complete data is

$$L_Z \propto f_Z = f_{T|XC}(t|x,c) f_X(x) f_C(c) h(r|x). \tag{2.21}$$

Without loss generality the missing data mechanism is assumed to depend on variable X only, and $h(r|x) = h(r=1|x)^r h(r=0|x)^{1-r}$.

The likelihood for incomplete data $\boldsymbol{Y}$ is given as

$$L_Y \propto f_Y = f^r_{T|XC}(t|x,c) f_X(x) f^r_C(c) h(r|x) \tag{2.22}$$

with distribution $f_Y = \int_{(y)} f_Z dz$ as the marginal model of the complete data model. Since the missing data mechanism $h(r|x)$ does not depend on missing variable C, so the marginal model can 'ignore' this component, according to Definition 1. Thus there is no technical difficulty to calculate the incomplete data bias. Below we will discuss some special cases and explain the bias parameters. This can help us to understand the source of incomplete data bias.

2.4.1 Incomplete Data Bias

Under linear regression model (2.15), the estimation of parameters by maximising the log-likelihood from incomplete data model (2.22) is biased and the incomplete data bias is given as

$$\begin{aligned}\boldsymbol{b_{XC}} &= \epsilon \boldsymbol{I_Y}^{-1} E_{f_Z}\{u_{XC}\boldsymbol{s_Y}\} - \epsilon \boldsymbol{I_Z}^{-1} E_{f_Z}\{u_{XC}\boldsymbol{s_Z}\} \\ &= \epsilon \boldsymbol{I_Y}^{-1} E\{u_{XC} E_{R|X}(\boldsymbol{s_Y})\} \\ &= \epsilon E\{u_{XC}\boldsymbol{s_{Y|r=0}} h(r=0|x)\} \\ &= \epsilon \theta_c \frac{\boldsymbol{I_Y}^{-1}}{\sigma_Y^2} E(cu_{XC}\boldsymbol{v_0}h_x)\end{aligned}$$

Put it simply, we use h_x to represent the probability of missingness conditional on x: $h_x = h(r=0|x)$.

The difference with the MCAR section is that the weight h_x in the integral

$$E(cu_{XC}\boldsymbol{v_0}h_x) = \int cu_{XC}\boldsymbol{v_0}h_x f(x)f(c)dxdc$$

is no longer a constant probability. But the misspecification ϵu_{XC} has the same meaning, and the incomplete data bias can be written as

$$b_{XC} \approx \theta_c \frac{\boldsymbol{I_Y}^{-1}}{\sigma_Y^2}\{E_{XC}[c\boldsymbol{v_0}h_x] - E_{X,C}[c\boldsymbol{v_0}h_x]\}. \tag{2.23}$$

when we put the quantity of $\epsilon u_{XC} = \log\frac{f(x,c)}{f(x)f(c)}$ into the integral. Now we are interested in discovering the difference in bias models between MAR and MCAR, and interpreting the factors that influence the incomplete data bias.

Since the missing data mechanism depends on x only and can be expressed as

$$h_x = h(r=0|x) + h'(r=0|x)x + O(h''(r=0|x)) \tag{2.24}$$

by Taylor's series. It is easy to notice that the difference between MAR and MCAR is the existence of the first order of the missing data mechanism $h'(r=0|x)$ and we expect the incomplete data bias to be a function of it.

Specifically, if x is a scalar and follows a normal distribution $x \sim N(0, \sigma_x^2)$, we can express incomplete data bias under MAR for a linear regression model as follows:

Theorem 2.1. *For a linear regression model*

$$t|(x,c) \sim N(\theta_0 + \theta_x x + \theta_c c, \sigma^2),$$

covariate C is partially missing with probability conditional on X: $h_x = h(r=0|x)$. The incomplete data bias for estimation of parameters $\theta = (\theta_0, \theta_x, \theta_c)$ is given as

$$\boldsymbol{b_{XC}} = \frac{\sigma_x}{\sigma_Y}\text{corr}(t,c|x)\text{corr}(x,c)\boldsymbol{I_Y^{-1}}\begin{pmatrix} E(h'_x) \\ E(h_x) \\ 0 \end{pmatrix}. \tag{2.25}$$

The proof is given in Appendix 2.7.4. The item $E(h_x)$ is considered the average of missing probability, which can be approximated by the missing proportion of studies. The item $E(h'_x) = 0$ under MCAR, but does not equal to zero under MAR, which illustrates the complexity from MCAR. It is the expectation of the first derivative of the missing procedure, and can be calculated if a MDM model is specified. We show one example below.

Assume that the MDM is a logistic linear model:

$$\text{logit}(h(r=1|x;\psi)) = \psi_0 + \psi_1 x. \tag{2.26}$$

Then we can get a fairly precise approximation of incomplete data bias if we apply a skew normal distribution in the integral. In the expression (2.25), we need to evaluate

$E(h'_x)$:

$$\begin{array}{rcl} E(h'_x) & = & \int h'_x f(x)dx \\ & = & -\int h_x f'(x)dx \\ & \overset{x\sim N(0,\sigma_x^2)}{=} & -\int h_x \frac{-x}{\sigma_x^3}\phi(\frac{x}{\sigma_x})dx \\ & = & \frac{1}{\sigma_x^3}\int (1-h(r=1|x))x\phi(\frac{x}{\sigma_x})dx \\ & \overset{expit(x)\approx\Phi(vx)}{\approx} & \frac{1}{\sigma_x^3}\int x\phi(\frac{x}{\sigma_x})dx - \frac{1}{\sigma_x^3}\int \Phi(v(\psi_0+\psi_1 x))x\phi(\frac{x}{\sigma_x})dx \\ & = & -\delta_1\phi(\delta_0) = \lambda \end{array}$$

where we denote $\lambda = -\delta_1\phi(\delta_0) = E(h'_x)$. Here $\delta_0 = \frac{v\psi_0}{\sqrt{1+v^2\psi_1^2\sigma_x^2}}$, $\delta_1 = \frac{v\psi_1}{\sqrt{1+v^2\psi_1^2\sigma_x^2}}$, $\phi(.)$ and $\Phi(.)$ are the density function and cumulative distribution of the standard normal distribution. Also, we use the approximation of expit$(x) \approx \Phi(vx)$ with v as a constant $16\sqrt{3}/(15\pi)$. The cumulation under skew normal distribution is used in the last step (see Arnold and Beaver, 2000).

Then the incomplete data bias can be approached by

$$\boldsymbol{b}_{XC} \approx \frac{\theta_c \boldsymbol{I_Y}^{-1}}{\sigma_Y^2}\text{cov}(x,c)\begin{pmatrix}\lambda\\ p\\ 0\end{pmatrix}. \tag{2.27}$$

As we can see, the bias depends on the correlation between covariates corr(x, c), the average missing proportion p and λ, where λ is the expectation of the first derivative of the MDM model and depends on the parameter ψ_1 in (2.26).

2.4.2 Simulation Study

Inference about the covariate bias is given in Theorem 2.1, but the covariate correlation corr(x, c) is not given in practice, and the approximation in equation (2.27) needs to be examined. So we conduct a simulation study to measure the sensitivity of the bias parameters towards the estimation of parameter of interest. Complete data is generated by the following linear regression model

$$t = \theta_0 + \theta_x x + \theta_c c + e, e \sim N(0, \sigma^2)$$

where covariates (X, C) are multivariate normal distributed. The true values are $\theta = (1, 1, 1)$, $\mu_x = \mu_c = 0$, $\sigma_x^2 = 2$ and $\sigma_c^2 = 1$, to make the confounder C about the same scale as X. The correlation $\operatorname{corr}(x, c)$, denoted as ρ, takes different values $\rho = (0, 0.2, 0.5, 0.7)$ for different studies, to represent no correlation, small correlation, medium correlation and strong correlation. We take the values of σ^2 from $U(0.16, 1)$ to add on t variation, and we set the maximum of σ^2 to be 1 such that the size of the ratio $\frac{\theta_c}{\sigma_Y^2} = \frac{\theta_c}{\theta_c^2 \sigma_c^2 + \sigma^2}$ is not too small (greater than 1/2). Here we repeat the simulation study 100 times to reduce the errors, each with 20 observations.

Variable C is designed to be censored by the missing at random assumption, and the logistic linear regression model (2.26) is chosen with $\psi_0 = 1$ and ψ_1 varied between (0, -3) representing different censoring strengths. When $\psi_1 = 0$, it will reduce to MCAR, otherwise MAR. When ψ is too large or too small, the missing data mechanism will be too extreme and out of our interest.

Figure 2.1 displays the censoring probability curve with different strength of ψ_1. When MDM is under MCAR (black line), the missing data probability is a constant, and $\lambda = 0$ in bias expression (2.27). As the gradient ψ_1 increases, the missing data mechanism is more unlike MCAR; see the grey and red line for example. The effect of ψ_1 on incomplete data bias is shown by simulation results presented in Table 2.1. For each fixed value of $\operatorname{corr}(x, c)$, the incomplete data bias increases with smaller ψ_1, which indicates the negative relation between the bias size and the parameter ψ_1. The evaluation of the bias works well when compared to simulation bias.

When there are no correlation between covariate variables, no bias exists under the missing at random assumption. In the other cases, the incomplete data bias exists due to the misspecification of covariate distribution. The relation between incomplete data bias and variables correlation can be illustrated by Figure 2.2. Figure 2.2 (a) indicates the MCAR problem ($\psi_1 = 0$), when the correlation of X and C is ignored incorrectly, and the bias exists. But if $\operatorname{corr}(x, c) = 0$, we can ignore the missing data mechanism specification, since the estimation of parameters θ will be independent from ψ under the ignorable missing data assumption, as shown in Figure 2.2 (b). Figure 2.2 (c) shows the case when the missingness indicator vector R is dependent on X, while X is also correlated with C. In this situation, the misspecification of

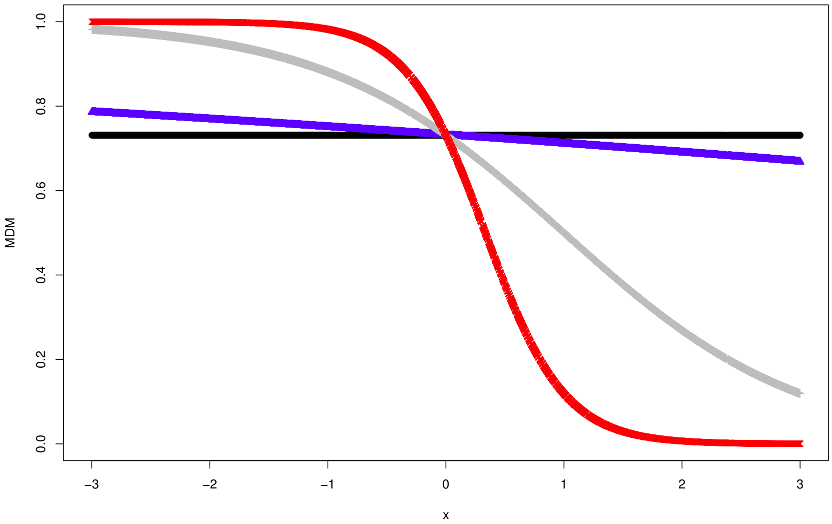
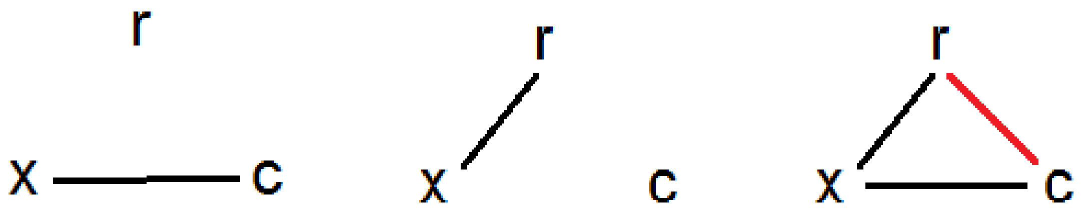


Figure 2.1: Curves $h(r = 1|x)$ for various MDM models. Black line: MCAR with $h(r = 1) = \text{expit}(1)$; Blue: MAR with $h(r = 1|x) = \text{expit}(1 - 0.1x)$; Grey: MAR with $h(r = 1|x) = \text{expit}(1 - x)$; Red: MAR with $h(r = 1|x) = \text{expit}(1 - 3x)$.



(a) MCAR; corr$(x, c) \neq 0$

(b) MAR; corr(x, c)=0

(c) MAR; corr$(x, c) \neq 0$

Figure 2.2: Picture of relationship between missingness indicator R with covariate variables X and C.

$f(x, c)$ causes bias, and the specification between R and X can not be ignored as it indirectly 'correlated' with missing values of C. And additional bias may be induced if we inference from a misspecified MDM model as we will show in the next chapter.

In all simulation studies, the missing proportion is ranged between 27% and 44% by given $\psi_0 = 1$. The estimation of parameter $\hat{\theta}$ is adjusted by the covariate bias(CB) and the 95% confidence interval is calculated as $(\hat{\theta} - CB) \pm 1.96\sqrt{\text{Var}(\hat{\theta})}$ with the reference distribution $\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$ assumed. The coverage probabilities (rates) are then calculated for the 100 replications. The simulation results show that the covariate bias is relatively large when there is medium or strong correlation between

Table 2.1: Covariate bias under MAR

		θ_0			θ_x			θ_c			MP
corr(x, c)	ψ_1	EB$(\hat{\theta} - \theta)$	CB	CR(%)	EB	CB	CR	EB	CB	CR	(%)
ρ=0	0	0.003	0	97	-0.001	0	96	-0.010	0	94	27
	-0.1	0.003	0	98	0.001	0	95	-0.019	0	95	27
	-0.3	0.002	0	97	0.001	0	96	-0.021	0	95	28
	-0.7	0.002	0	98	0.001	0	94	-0.017	0	97	33
	-1	0.005	0	98	0.002	0	96	-0.018	0	97	35
	-2	0.002	0	98	-0.001	0	96	-0.023	0	98	41
	-3	0.003	0	98	0.001	0	96	-0.022	0	98	44
ρ=0.2	0	0.001	-0.001	98	0.009	0.013	94	-0.015	-0.006	96	28
	-0.1	0.003	0.003	98	0.012	0.012	94	-0.016	-0.005	97	27
	-0.3	0.004	0.008	98	0.011	0.013	94	-0.020	-0.006	96	28
	-0.7	0.015	0.014	98	0.015	0.011	93	-0.022	0.002	97	33
	-1	0.019	0.021	97	0.017	0.016	96	-0.031	-0.006	96	35
	-2	0.023	0.027	98	0.020	0.019	94	-0.033	-0.007	98	41
	-3	0.026	0.030	98	0.021	0.021	96	-0.030	-0.009	99	43
ρ=0.5	0	-0.001	-0.017	97	0.031	0.013	90	-0.045	0.011	92	27
	-0.1	0.006	0.006	97	0.031	0.043	90	-0.044	-0.051	93	27
	-0.3	0.015	0.019	97	0.035	0.037	93	-0.049	-0.035	94	29
	-0.7	0.034	0.039	97	0.045	0.029	91	-0.054	0.009	95	33
	-1	0.039	0.059	98	0.047	0.054	91	-0.053	-0.058	94	35
	-2	0.053	0.069	97	0.056	0.049	92	-0.063	-0.036	96	41
	-3	0.052	0.082	97	0.052	0.066	91	0.056	-0.068	96	43
ρ=0.7	0	-0.001	-0.015	95	0.049	0.052	82	-0.079	-0.075	85	27
	-0.1	0.006	-0.047	95	0.054	-0.053	81	-0.090	0.221	85	27
	-0.3	0.021	0.033	97	0.062	0.078	80	-0.088	-0.110	87	28
	-0.7	0.038	0.062	95	0.068	0.079	85	-0.095	-0.066	90	33
	-1	0.045	0.087	95	0.073	0.096	83	-0.099	-0.143	88	35
	-2	0.068	0.121	94	0.082	0.122	85	-0.093	-0.180	90	41
	-3	0.066	0.144	90	0.086	0.131	84	-0.111	-0.202	89	44

MDM model $h(r = 1|x) = \text{expit}(1 - \psi_1 x)$. EB: Empirical bias $(\hat{\theta} - \theta)$; CB: Covariate bias approximation $\boldsymbol{b_{XC}}$; CR: coverage rate of adjusted estimator; MP: missing proportion.

X and C. Also the bias approximation works well comparing empirical bias with estimated covariate bias, and CR is around 95% as we expect when corr$(x, c) \leq 0.5$, and the case with stronger correlations may be beyond the local analysis assumption.

2.5 GLMs with Ignorable Missing Data

In this section, we consider the $f_{T|XC}$ to be a generalized linear model (GLM). There are quite a few literature discussing MLE for missing covariates in GLMs including Fuchs (1982), Little and Schluchter (1985), Ibrahim (1990), Ibrahim et al. (1999), etc. Quasi-likelihood approaches have been explored by Reilly and Pepe (1995), Lawless et al. (1999), and Tang et al. (2003).

Lin et al. (2012) argued that for nonlinear regression, if inference is to be based on incomplete data with local analysis, the estimate of the parameter of interest may bring additional *marginal model bias*.

We now consider a GLM with canonical form

$$f_{T|XC} = \exp\{\frac{t\pi_{xc} - b(\pi_{xc})}{a(\phi)} + d(t, \phi)\}$$

where $\pi_{xc} = \alpha + \theta x + \beta c$ and the conditional expectation satisfies

$$E_{f_{T|XC}}[t|x, c] = \xi(\pi_{xc}) = b'(\pi_{xc})$$

where $\xi(\pi_{xc})$ is the link function and $b'(\pi_{xc}) = \partial b(\pi_{xc})/\partial \pi_{xc}$.

It is well known that when the link function is linear, $\xi(\pi_{xc}) = \pi_{xc}$, then coefficient estimates are unbiased (Gail et al., 1984) under MAR; however, regressions with nonlinear link functions may lead to biased estimates, even in randomized experiments, if covariates are missing. Much literature has discussed this problem, including Dox (1972), Struthers and Kalbfleisch (1986) and Breslow and Lin (1995). Besides, Lin et al. (2012) pointed out that marginal model bias exists thus double misspecification should be considered. As we mentioned in the the previous sections, the working model under incomplete data $f_Y = \int_{(y)} f_Z dz$ is the marginal model of corresponding distribution under complete data, or rather under the assumption (Copas and Eguchi 2005, p464): *the components of θ which are fully identifiable from observations on y under model f_Z*. Lin's work is based on the consideration that when the working model is not the marginal model for incomplete data, which happens for nonlinear regression or GLM with nonlinear link function.

2.5.1 Incomplete Data Bias Analysis

Lin et al. (2012) proposed marginal model bias to measure the difference between the actual working model under incomplete data and the marginal distribution, and their discussion mainly focuses on the missing confounder problem. We follow their work and extend it to the missing covariate problem.

Recall the true distribution under complete data Z is

$$g_Z = f_{T|XC} f_{XC}(x, c) h(r|x)$$

with missing data mechanism $h(r|x)$ under MAR assumption. And the working model is

$$f_Z = f_{T|XC} f_X(x) f_C(c) h(r|x).$$

While under incomplete data Y, the marginal distribution is

$$\begin{aligned} g_Y &= \int_{(y)} g_Z dz \\ &= \int_{(y)} f_Z \exp(\epsilon_{XC} u_{XC}) dz \approx f_Y \exp(\epsilon_{XC} u_{XC|Y}) \\ &\approx f_Y^* \exp(\epsilon_{XC} u_{XC|Y}) \exp(\epsilon_M u_M). \end{aligned} \tag{2.28}$$

Here the misspecification $\exp(\epsilon_M u_M)$ is the ratio between the marginal model and the working model

$$\exp(\epsilon_M u_M) = \frac{f_Y}{f_Y^*}.$$

In this way, the working model f_Y^* is regarded as 'doubly misspecified' from true density g_Y with misspecification quantities separated into two parts:

$$\epsilon u_Y = \epsilon_{XC} u_{XC|Y} + \epsilon_M u_M.$$

For complete cases $(r = 1)$, $f_{Y_{cc}} = f_{T|XC} f_X f_C h(r|x)$ with

$$f_{T|XC} = \exp\{\frac{t\pi_{xc} - b(\pi_{xc})}{a(\phi)} + d(t, \phi)\}$$

and incomplete cases $(r = 0)$

$$\begin{aligned} f_{Y_{ic}} &= \int_{(y)} f_Z dz \\ &= \int_c f_{T|XC} f_C dc f_X h(r|x) \\ &= f_{T|X} f_X h(r|x) \end{aligned}$$

where

$$\begin{aligned}
f_{T|X} &= \int_c f_{T|XC} f_C dc \\
&= \exp\{\frac{t\pi_x - b(\pi_x)}{a(\phi)} + d(t,\phi)\}(1 + \frac{1}{2}\beta^2\sigma_c^2[(\frac{t-\xi(\pi_x)}{a(\phi)})^2 - \frac{\xi'(\pi_x)}{a(\phi)}] + O(\beta^4\sigma_c^4))
\end{aligned}$$

and let

$$f^*_{T|X} = \exp\{\frac{t\xi_x - b(\pi_x)}{a(\phi)} + d(t,\phi)\} \tag{2.29}$$

with link function

$$E_{f^*_{T|X}}[t|x] = \xi(\alpha + \theta x) = \xi(\pi_x),$$

then the working model under incomplete data can be written as

$$f^*_Y = f^r_{T|XC} f_X f^r_C h(r|x)$$

with

$$f^r_{T|XC} = \begin{cases} f_{T|XC}, & r = 1; \\ f^*_{T|X}, & r = 0. \end{cases}$$

and

$$f^r_C = \begin{cases} f(c), & r = 1; \\ 1, & r = 0. \end{cases}$$

If we let $\pi_{xc^r} = \alpha + \theta x + r\beta c$, then $f^r_{T|XC} = \exp\{\frac{t\pi_{xc^r} - b(\pi_{xc^r})}{a(\phi)} + d(t,\phi)\}$. Thus the incomplete data bias can also be decomposed into two components

$$\text{bias} = \boldsymbol{b_M} + \boldsymbol{b_{XC}}$$

with marginal bias given as

$$\begin{aligned}
\boldsymbol{b_M} &= \epsilon_M E_{f^*_Y}(u_M \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*) \\
&\approx \frac{\beta^2\sigma_c^2}{2a(\phi)} \boldsymbol{I_Y}^{*-1} E_X\{\xi''(\pi_x)\boldsymbol{v_0} h_x\}
\end{aligned}$$

and covariate bias as

$$\begin{aligned}
\boldsymbol{b_{XC}} &= \epsilon_{XC} E_{f_Z}(u_{XC} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*) - \epsilon_{XC} E_{f_Z}(u_{XC} \boldsymbol{I_Z}^{*-1} \boldsymbol{s_Z}^*) \\
&= \epsilon_{XC} \frac{\boldsymbol{I_Y}^{*-1}}{a(\phi)} E_{X,C}\{u_{XC} \boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)] h_x\} \\
&= \frac{\boldsymbol{I_Y}^{*-1}}{a(\phi)} [E_{XC}\{\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)] h_x\} - E_{X,C}\{\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)] h_x\}] \\
&\approx \frac{\beta \boldsymbol{I_Y}^{*-1}}{a(\phi)} [E_{XC}\{c \boldsymbol{v_0} \xi'(\pi_x) h_x\} - E_{X,C}\{c \boldsymbol{v_0} \xi'(\pi_x) h_x\}]
\end{aligned}$$

The detailed proof is given in appendix 2.7.5.

2.5.2 Simulation Study

Now we conduct a simulation study for a logistic regression model

$$f_{T|XC} = \pi_{xc}^t (1 - \pi_{xc})^{1-t}, \; \pi_{xc} = \frac{\exp\{\alpha + \theta x + \beta c\}}{1 + \exp\{\alpha + \theta x + \beta c\}}$$

with covariate distribution following a multivariate normal distribution. The mean and variance of covariates are given as 0 and 1 respectively, with covariate variables correlation ρ selected from (0, 0.3, 0.5) corresponding to no correlation, moderate and strong correlation. Variable c is designed to be partially observed. The missing data mechanism is under the ignorable assumption: $h(r = 1|x) = \text{expit}(\psi_0 + \psi_1 x)$. The true values of parameters are $(\alpha, \theta, \beta) = (1, 1, 1)$. We conduct 100 replications and each has sample size of 100.

The empirical bias is defined as the average difference between the MLEs for incomplete data $\hat{\theta}_Y$ and true value θ, which is approximately approached by the incomplete data bias as discussed. It contains two components: marginal bias b_M and covariate bias b_{XC}. We are particularly interested in the size and direction of marginal bias. As shown in Table 2.2, the marginal bias $\boldsymbol{b_M}$ is always negative for all the studies since the second derivation $\xi''(\pi_x) < 0$ for the logistic model, and it always exists even when $\rho = 0$ and MCAR. It seems independent from the covariate correlation ρ as we expected and mainly depends on the nonlinearity of the model and the missing data mechanism. The size of marginal model bias decreases when the value of parameter ψ_1 in the MDM model gets larger. The estimation of covariate bias is similar to what we found in linear model simulation, which exists when $\rho \neq 0$, and it is also affected by both correlation and missing data selection strength.

2.6 Discussion

The purpose of addressing bias analysis into the sensitivity analysis is to discover and understand those uncertainty factors in the missing data problems. We usually eliminated the uncertainties in the working models (to be identifiable) to obtain valid inference, but the sensitivity of bias models should be considered in real trials and we assessed it via bias parameters, such as $\mathrm{corr}(x, c)$.

In this chapter, we discussed both linear and GLM regression models with missing covariate problems and misspecification of joint covariate density is considered. It is interesting to notice that in some occasions (when there is no relationship between missing confounder and other dependent variables), the missingness can be ignored and bring no bias towards the estimation of parameter θ although the variance becomes larger as pointed out by Copas and Eguchi (2005). However it can not be ignored in most of the cases. The covariate bias exists due to the lack of consideration over the bias parameters sensitivity.

For a generalized linear model, the working model f_Y^* may be double misspecified from true distribution g_Y with incomplete data Y. In this case, the additional marginal bias is generated and requires adjustment. These biases are then calculated and the sensitivity of MLEs on the bias parameters is presented in the simulation studies.

Beyond the examples we discussed, the local bias analysis can accommodate various response regression models and covariates densities, under the condition of identification. In this thesis, we use the maximum likelihood methods to estimate the parameters, but when the calculation of parameters becomes difficult we may use numerical computation methods such as Gaussian integral, MCEM, or Bayesian methods.

Table 2.2: Covariate bias for GLM under MAR

		Empirical Bias ($\hat{\theta}_Y - \theta$)			Marginal Bias			Covariate Bias			CR			MP
		b^α	b^θ	b^β	b^α_M	b^θ_M	b^β_M	b^α_{XC}	b^θ_{XC}	b^β_{XC}	α	θ	β	(%)
$\rho = 0$	ψ_1= -3	-0.110	-0.113	-0.024	-0.122	-0.130	-0.032	-0.001	-0.001	-0.001	92	95	94	40
	ψ_1= -2	-0.092	-0.091	-0.021	-0.117	-0.116	-0.033	0.001	-0.001	-0.001	95	92	94	35
	ψ_1= -1	-0.075	-0.071	-0.025	-0.088	-0.082	-0.031	-0.001	0.001	-0.001	95	95	92	30
	ψ_1= 0	-0.051	-0.048	-0.031	-0.061	-0.056	-0.028	0.001	0.001	-0.001	89	89	96	27
	ψ_1= 1	-0.031	-0.037	-0.006	-0.042	-0.054	-0.026	-0.001	0.001	-0.002	91	92	92	30
	ψ_1= 2	-0.026	-0.042	-0.015	-0.036	-0.053	-0.027	0.001	0.001	-0.002	92	93	95	35
	ψ_1= 3	-0.023	-0.036	-0.011	-0.035	-0.051	-0.028	0.001	0.001	0.002	93	95	93	39
$\rho = 0.3$	ψ_1= -3	-0.027	-0.004	-0.005	-0.131	-0.132	-0.022	0.085	0.121	0.012	93	94	94	39
	ψ_1= -2	-0.023	0.004	-0.006	-0.114	-0.113	-0.023	0.071	0.112	0.013	94	92	93	35
	ψ_1= -1	-0.032	0.016	-0.011	-0.091	-0.085	-0.021	0.047	0.092	0.006	96	92	93	31
	ψ_1= 0	-0.047	0.035	-0.014	-0.063	-0.059	-0.019	0.004	0.085	-0.003	92	95	96	27
	ψ_1= 1	-0.065	0.078	-0.027	-0.045	-0.058	-0.015	-0.027	0.126	-0.010	96	90	95	30
	ψ_1= 2	-0.076	0.109	-0.012	-0.037	-0.057	-0.016	-0.049	0.154	-0.009	91	90	90	34
	ψ_1= 3	-0.086	0.121	-0.017	-0.036	-0.053	-0.018	-0.058	0.158	-0.006	91	92	92	39
$\rho = 0.5$	ψ_1= -3	0.027	0.077	-0.013	-0.123	-0.131	-0.012	0.142	0.200	-0.006	92	97	93	39
	ψ_1= -2	0.019	0.076	-0.006	-0.112	-0.124	-0.011	0.123	0.191	-0.011	96	93	99	35
	ψ_1= -1	-0.003	0.076	-0.008	-0.089	-0.091	-0.011	0.073	0.158	-0.018	93	93	88	30
	ψ_1= 0	-0.048	0.112	-0.047	-0.065	-0.065	-0.009	0.012	0.160	-0.038	91	91	96	28
	ψ_1= 1	-0.075	0.183	-0.050	-0.046	-0.062	-0.005	-0.042	0.218	-0.054	94	93	96	31
	ψ_1= 2	-0.100	0.219	-0.042	-0.039	-0.061	-0.006	-0.074	0.247	-0.051	92	89	96	35
	ψ_1= 3	-0.122	0.251	-0.026	-0.038	-0.057	-0.008	-0.093	0.273	-0.042	90	84	91	39

MDM $h(r = 1|x) = \text{expit}(1 + \psi_1 x)$. CR: Coverage Rate (%). MP: Missing Proportion.

2.7 Appendix

2.7.1 Proof of Lemma 2.1

Suppose we fit the model $f_Z(z;\theta)$ to a random sample of n observations from g_Z, the log-likelihood and score function are

$$l(z;\theta) = \sum_{i=1}^{n} \log(g_Z(z_i;\theta)) = \sum_{i=1}^{n} \log(f_Z(z_i;\theta)) + \epsilon u_Z(z;\theta),$$

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^{n} s_Z(z_i;\theta) + \epsilon \frac{\partial u_Z(z_i;\theta)}{\partial \theta}. \tag{2.30}$$

As θ_Z is the MLE of model f_Z,

$$\sum_{i=1}^{n} s_Z(z_i;\hat{\theta}_Z) = 0$$

and a Taylor expansion leads to

$$\begin{aligned}\sum_{i=1}^{n} s_Z(z_i;\theta) &\approx \sum_{i=1}^{n} s_Z(z_i;\hat{\theta}_Z) + \sum_{i=1}^{n} \frac{\partial s_Z(z_i;\hat{\theta}_Z)}{\partial \theta}(\theta - \hat{\theta}_Z) \\ &= \sum_{i=1}^{n} \frac{\partial s_Z(z_i;\hat{\theta}_Z)}{\partial \theta}(\theta - \hat{\theta}_Z).\end{aligned}$$

The equation (2.30) becomes

$$\frac{\partial l}{\partial \theta} \approx \sum_{i=1}^{n} \frac{\partial s_Z(z_i;\hat{\theta}_Z)}{\partial \theta}(\theta - \hat{\theta}_Z) + \epsilon \frac{\partial u_Z(z_i;\theta)}{\partial \theta}. \tag{2.31}$$

As $n \to \infty$, the expression (2.31) tends to 0, and so we have

$$E_{f_Z}(\frac{\partial s_Z(z_i;\theta_{gZ})}{\partial \theta})(\theta - \hat{\theta}_Z) + \epsilon E_{f_Z}(\frac{\partial u_Z(z_i;\theta)}{\partial \theta}) \approx 0. \tag{2.32}$$

The first term in (2.32) is

$$E_{f_Z}(\frac{\partial s_Z(z_i;\hat{\theta}_Z)}{\partial\theta}) = -I_Z.$$

The second term is

$$E_{f_Z}(\frac{\partial u_Z}{\partial\theta}) = -E_{f_Z}(u_Z s_Z),$$

which can be easily derived by differentiating both sides of the following:

$$E_{f_Z}(u_Z) = \int_Z u_Z f_Z dz = 0.$$

Expression (2.32) then becomes:

$$-I_Z(\theta-\theta_{gZ}) - \epsilon E_{f_Z}(u_Z s_Z) \approx 0,$$

and this leads to

$$\theta_{gZ} = arg_\theta[E_g\{s_Z(z;\theta)\} = 0] \approx \theta + \epsilon I_Z^{-1} E_f\{u_Z(z;\theta)s_Z(z;\theta)\}.$$

in the sense of almost sure convergence. Similarly, if we are sampling from g_Y, the limiting value of $\hat{\theta}_Y$ is

$$\theta_{gY} = arg_\theta[E_{gY}\{s_Y(y;\theta)\} = 0] \approx \theta + \epsilon I_Y^{-1} E_f\{u_Y(y;\theta)s_Y(y;\theta)\}.$$

Thus, the incomplete data bias can be defined as:

$$\theta_{gY} - \theta_{gZ} \approx b_\theta = \epsilon E_f[u_Z(z;\theta)I_Y^{-1}s_Y(y;\theta) - I_Z^{-1}s_Z(z;\theta)]. \tag{2.33}$$

2.7.2 Proof of Lemma 2.2

For a linear regression model under complete data $Z = (T, X, C)$:

$$t = \theta_0 + \theta_x x + \theta_c c + e,\ e \sim N(0,\sigma^2).$$

We consider the scalar confounder C totally missing. And incomplete data $Y = (T, X)$:

$$t = \theta_0 + \theta_x x + e,\, e \sim N(0, \sigma^2 + \theta_c^2\sigma_c^2).$$

Assume $X \sim N(0, \sigma_x^2), C \sim N(0, \sigma_c^2)$, and denote ρ=corr(x, c), the log-likelihood under complete data Z and incomplete data Y are

$$l_Z = -\log(\sigma^2) - \frac{(t - \theta_0 - \theta_x x - \theta_c c)^2}{2\sigma^2}, \tag{2.34}$$

$$l_Y = -\log(\sigma^2 + \theta_c^2\sigma_c^2) - \frac{(t - \theta_0 - \theta_x x)^2}{2(\sigma^2 + \theta_c^2\sigma_c^2)} \tag{2.35}$$

respectively. The θ_x-component of the score function s_Z and s_Y are

$$\begin{aligned} s_Z &= \frac{(t - \theta_0 - \theta_x x - \theta_c c)x}{\sigma^2}, \\ s_Y &= \frac{(t - \theta_0 - \theta_x x)x}{(\sigma^2 + \theta_c^2\sigma_c^2)} \end{aligned}$$

respectively, and the information matrices (diagonal) of f_Z and f_Y for θ_x-component are

$$\begin{aligned} I_Z &= E_{f_Z}(-\partial l^2/\partial\theta_x\theta_x) = E(\frac{x^2}{\sigma^2}) = \frac{\sigma_x^2}{\sigma^2}, \\ I_Y &= E_{f_Y}(-\partial l^2/\partial\theta_x\theta_x) = E(\frac{x^2}{\sigma^2 + \theta_c^2\sigma_c^2}) = \frac{\sigma_x^2}{\sigma_Y^2} \end{aligned}$$

where $\sigma_Y^2 = \sigma^2 + \theta_c^2\sigma_c^2$. According to Lemma 2.1,

$$\begin{aligned} b &= \epsilon I_{T|X}^{-1} E_{f_Z}(s_{T|X} u_{XC}) \\ &= \epsilon(\frac{\sigma_Y^2}{\sigma_x^2}) E_{f_Z}(\frac{(t - \theta_0 - \theta_x x) x u_{XC}}{\sigma_Y^2}) \\ &= \epsilon(\frac{\sigma_Y^2}{\sigma_x^2}) E_{f_Z}(\frac{(\theta_c c + e) x u_{XC}}{\sigma_Y^2}) \\ &= \epsilon\theta_c\sigma_x^{-2} E_{f_Z}(c u_{XC} x) \end{aligned}$$

as we assume that e and (x, c) are independent, thus

$$E_{f_Z}\{e x u_{XC}\} = 0.$$

The size of the standardized bias is now

$$\begin{aligned} b^2 I_Y &= b^2(\frac{\sigma_x^2}{\sigma_Y^2}) = \frac{\epsilon^2\theta_c^2}{\sigma_Y^2}(E_{f_Z}(cu_{XC}x))\sigma_x^{-1}E_{f_Z}(cu_{XC}x) \\ &\leq \frac{\epsilon^2\theta_c^2}{\sigma_Y^2}E_{f_Z}(cx)^2\sigma_x^{-2}E_{f_Z}(u_{XC}^2). \end{aligned}$$

The equation is held when $u_{XC} = cdx$ for some constant vector d. While $\exp\{\epsilon u_{XC}\} = \frac{f(x,c)}{f(x)f(c)} = \frac{f(x)f(c|x)}{f(x)f(c)} = \frac{f(c|x)}{f(c)}$, so $f(c|x) = \exp(\epsilon c d x) f(c)$. This means that for small ϵ the conditional distribution of c given x is approximately

$$g_{C|X} \sim N(\epsilon\sigma_c^2 dx, \sigma_c^2). \tag{2.36}$$

So c becomes

$$c \approx \epsilon\sigma_c^2 dx + e_c,\ e_c \sim N(0, \sigma_c^2)$$

where e_c is independent of x. Here variable X is just scalar, in which case the correlation between c and x that is implied by distribution (2.36) is $\epsilon\sigma_c d\sigma_x$.

$$\begin{aligned} E_{f_Z}(cx)^2 &= E_{f_Z}\{(\epsilon\sigma_c^2 dx + e_c)^2 x^2\} \\ &= E_{f_Z}\{\epsilon^2\sigma_c^4(dx)^2x^2 + 2\epsilon\sigma_c^2 dx^2 e_c + e_c^2 x^2\} \\ &= \epsilon^2\sigma_c^4 d^2 3\sigma_x^4 + \sigma_x^2\sigma_c^2 \\ &\approx \sigma_x^2\sigma_c^2, \end{aligned}$$

and the approach is true when ϵ is supposed to be small. Similarly,

$$E_{f_Z}(c^2d^2x^2) \approx d^2\sigma_x^2\sigma_c^2.$$

The upper bound would be

$$\begin{aligned} &\frac{\epsilon^2\theta_c^2}{\sigma_Y^2}E_{f_Z}(cx)^2E_{f_Z}(cdx)^2\sigma_x^{-2} \\ &= \frac{\epsilon^2\theta_c^2}{\sigma_Y^2}d^2(\sigma_x^2\sigma_c^2)^2\sigma_x^{-2} \\ &= (\epsilon\sigma_c d\sigma_x)^2\frac{\theta_c^2\sigma_c^2}{\sigma_Y^2}. \end{aligned}$$

The correlation between t and c give x is

$$\mathrm{corr}(t, c|x)^2 = \frac{\theta_c^2 \sigma_c^2}{\sigma_Y^2}. \tag{2.37}$$

And

$$\mathrm{corr}(x, c) = \epsilon \sigma_c d \sigma_x. \tag{2.38}$$

So the size of the squared standardized bias is bounded by

$$b^2 I_Y \leq \mathrm{corr}^2(t, c|x)\mathrm{corr}^2(x, c). \tag{2.39}$$

If the sample size is n, then $I_Y^{-1} \approx n\mathrm{var}_f(\hat{\theta}_x)$, so

$$\frac{b^2}{n\mathrm{var}_f(\hat{\theta}_x)} \leq \mathrm{corr}^2(t, c|x)\mathrm{corr}^2(x, c). \tag{2.40}$$

When $\rho = 0$, the incomplete data bias equals to 0.

2.7.3 Proof of the Marginal Model

To prove f_Y as the marginal model of f_Z, we need to prove $f_{T|X} = \int f_{T|XC} f_C dc$. From Equation (2.5) and (2.6), we have

$$f_{T|XC} = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(t - \theta_0 - \theta_x x - \theta_c c)^2}{2\sigma^2}\}, \tag{2.41}$$

and

$$f_{T|X} = \frac{1}{\sqrt{2\pi(\sigma^2 + \theta_c^2 \sigma_c^2)}} \exp\{-\frac{(t - \theta_0 - \theta_x x)^2}{2(\sigma^2 + \theta_c^2 \sigma_c^2)}\}. \tag{2.42}$$

For the cases with missing data:

$$\begin{aligned}
& \int f_{T|XC}(t;x,c)f_C(c)dc \\
= & \int \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(t-\theta_0-\theta_x x-\theta_c c)^2}{2\sigma^2}\}\frac{1}{\sqrt{2\pi}\sigma_c}\exp\{-\frac{c^2}{2\sigma_c^2}\}dc \\
= & \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma_c}\int\exp\{-\frac{(t-\theta_0-\theta_x x)^2}{2\sigma_i^2}+\frac{2\theta_c c(t-\theta_0-\theta_x x)}{2\sigma^2}-\frac{\theta_c^2c^2}{2\sigma^2}-\frac{c^2}{2\sigma_c^2}\}dc \\
= & \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma_c}\exp\{-\frac{(t-\theta_0-\theta_x x)^2}{2\sigma^2}\}\int\exp\{\frac{2\theta_c c(t-\theta_0-\theta_x x)}{2\sigma^2}-\frac{\theta_c^2c^2}{2\sigma^2}-\frac{c^2}{2\sigma_c^2}\}dc \\
= & \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{(t-\theta_0-\theta_x x)^2}{2\sigma^2}\}\sqrt{\frac{\sigma^2}{\theta_c^2\sigma_c^2+\sigma^2}}\exp\{\frac{\theta_c^2\sigma_c^2(t-\theta_0-\theta_x x)^2}{2\sigma^2(\theta_c^2\sigma_c^2+\sigma^2)}\} \\
= & \frac{1}{\sqrt{2\pi(\sigma^2+\theta_c^2\sigma_c^2)}}\exp\{-\frac{(t-\theta_0-\theta_x x)^2}{2(\sigma^2+\theta_c^2\sigma_c^2)}\}.
\end{aligned}$$

That is $f_{T|X}=\int f_{T|XC}f_C dc$. The proof can be easily extended to multi-dimensional variables of x.

2.7.4 Proof of Theorem 2.1

According to equation (2.23), the incomplete data bias under MAR is

$$b_{XC}\approx\theta_c\frac{\boldsymbol{I_Y}^{-1}}{\sigma_Y^2}\{E_{XC}[c\boldsymbol{v_0}h_x]-E_{X,C}[c\boldsymbol{v_0}h_x]\}.$$

The missingness relies on observed covariate variable x, and it is assumed to be normal distributed $x\sim N(0,\sigma_x^2)$. Then the item

$$\begin{aligned}
& E_{XC}[c\boldsymbol{v_0}h_x]-E_{X,C}[c\boldsymbol{v_0}h_x] \\
= & E_X[\boldsymbol{v_0}h_x\{E_{C|X}(c)-E_C(c)\}] \\
= & \rho\frac{\sigma_c}{\sigma_x}E_X[x\boldsymbol{v_0}h_x)]
\end{aligned}$$

Let $h'(r=0|x)=\frac{\partial h(r=0|x)}{\partial x}$ as the first derivative for $h(r=0|x)$. And denote $\phi(.)$, $\varPhi(.)$ as the standard normal Probability density function and Cumulative distribution function. The component

$$\begin{aligned}E(xh_x) &= \int xh_x f_x dx \\ &= -\int h'_x \int_{-\infty}^{x} xf_x dx \\ &= \sigma_x \int \phi(\frac{x}{\sigma_x})h'_x dx \\ &= \sigma_x^2 E_X(h'_x)\end{aligned}$$

and

$$\begin{aligned}E(x^2h_x) &= \int x^2 h_x f_x dx \\ &= \sigma_x^2 h_{x\to\infty} - \int h'_x \int_{-\infty}^{x} x^2 f_x dx \\ &= \sigma_x^2 h_{x\to\infty} - \int h'_x[\sigma_x x\phi(\frac{x}{\sigma_x}) + \sigma_x^2 \Phi(\frac{x}{\sigma_x})]dx \\ &= \sigma_x^2 h_{x\to\infty} + \sigma_x \int h'_x x f_x dx - \sigma_x^2 \int h'_x \Phi(\frac{x}{\sigma_x})dx \\ &= \sigma_x^2 h_{x\to\infty} - \sigma_x^2 \int h''_x \int_{-\infty}^{x} xf_x dxdx - \sigma_x^2[h_{x\to\infty} - \frac{1}{\sigma_x}\int h_x \phi(\frac{x}{\sigma_x})dx] \\ &= \sigma_x^2[E_X(h_x) + \int h''_x f_x dx] = \sigma_x^2[E_X(h_x) + E_X(h''_x)].\end{aligned}$$

If the second derivative $h''(r=0|x)$ of MDM is small, then the bias has an approximation

$$b_{XC} = \frac{\theta_c \boldsymbol{I_Y^{-1}}}{\sigma_Y^2}\text{cov}(x,c)\begin{pmatrix} E_X(h'_x) \\ E_X(h_x) \\ 0 \end{pmatrix}.$$

2.7.5 Proof of Incomplete Data Bias for GLMs under MAR

For the models under complete data:

$$g_Z = f_{T|XC} f_{XC} h(r|x);$$

$$f_Z = f_{T|XC} f_X f_C h(r|x)$$

and the corresponding marginal model under incomplete data is g_Y, f_Y which is shown in equations (5.3). The actual working model is

$$f_Y^* = f_{T|XC}^r f_X f_C^r h(r|x).$$

Lin et al. (2012) give the marginal model bias $\boldsymbol{b_M} = \epsilon_M E_{f_Y^*}(u_M \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*)$. Two misspecification functions (MF) are involved between the true distribution g_Z and working model f_Y^*.

$$\begin{aligned} \text{MF } 1: \quad & \exp(\epsilon_{XC} u_{XC}) = \frac{g_Z}{f_Z};\ \exp(\epsilon_{XC} u_{XC|Y}) = \frac{g_Y}{f_Y}; \\ \text{MF } 2: \quad & \exp(\epsilon_M u_M) = \frac{f_Y}{f_Y^*}; \end{aligned}$$

where $u_{XC|Y} = E_{f_Z^*}(u_{XC}|Y)$.
Let $l_{T|XC} = \log(f_{T|XC})$, and $l_{T|X}^* = \log(f_{T|X}^*)$, so that we have

$$\int_t f_{T|XC} dt = 1, \frac{\partial f_{T|XC}}{\partial \pi_{xc}} = l'_{T|XC} f_{T|XC},$$

and

$$\int_t f_{T|X}^* dt = 1, \frac{\partial f_{T|X}^*}{\partial \pi_x} = l_{T|X}^{*'} f_{T|X}^*,$$

where $l'_{T|XC} = \frac{\partial l_{T|XC}}{\partial \pi_{xc}}$ and $l_{T|X}^{*'} = \frac{\partial l_{T|X}^*}{\partial \pi_x}$. So the score function under incomplete data is

$$\begin{aligned} \boldsymbol{s_Y}^* &= l_{T|XC^r}^{*'} \begin{pmatrix} 1 \\ x \\ c^r \end{pmatrix} \\ &= \frac{t - \xi(\pi_{xc^r})}{a(\phi)} \begin{pmatrix} 1 \\ x \\ c^r \end{pmatrix} \end{aligned}$$

where

$$l_{T|XC^r}^{*'} = \begin{cases} l_{T|XC}, & r = 1; \\ l_{T|X}^*, & r = 0. \end{cases}$$

We are fitting f_Y^* to a random sample of n observations from g_Y, and the limiting value of the MLE $\hat{\theta}_Y$ as $n \to \infty$ is

$$\begin{aligned} \theta_{gY} &= arg_\theta E_g[\boldsymbol{s_Y}^* = 0] \\ &\approx \theta + \epsilon_{XC} E_{f_Y^*}[u_{XC|Y} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*] + \epsilon^* E_{f_Y^*}[u_Y^* \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*]. \end{aligned}$$

Here we have

$$\epsilon_{XC} E_{f_Y^*}[u_{XC|Y} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*] \approx \epsilon_{XC} E_{f_Z}[u_{XC} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*]$$

as

$$\begin{aligned} & \epsilon_{XC} E_{f_Y^*}[u_{XC|Y} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*] \\ \approx\; & \epsilon_{XC} \int_{(y)} u_{XC|Y} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^* f_Y^* \exp(1 - \epsilon_M u_M) dy \\ \approx\; & \epsilon_{XC} \int_{(y)} u_{XC|Y} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^* f_Y^* dy - O(\epsilon_{XC} \epsilon_M) \\ \approx\; & \epsilon_{XC} E_{f_Z^*}[u_{XC} \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*]. \end{aligned}$$

Then the incomplete data bias is

$$\begin{aligned} b &\approx \theta_{gY} - \theta_{gZ} \\ &= \epsilon_{XC} E_{f_Z}[u_{XC}\{\boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^* - \boldsymbol{I_Z}^{*-1} \boldsymbol{s_Z}^*\}] + \epsilon_M E_{f_Y^*}\{u_M \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*\} \end{aligned}$$

with

$$\boldsymbol{b_{XC}} = \epsilon_{XC} E_{f_Z}[u_{XC}\{\boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^* - \boldsymbol{I_Z}^{*-1} \boldsymbol{s_Z}^*\}];$$

$$\boldsymbol{b_M} = \epsilon_M E_{f_Y^*}\{u_M \boldsymbol{I_Y}^{*-1} \boldsymbol{s_Y}^*\}.$$

First for misspecification problem of covariate distribution, covariate bias b_{XC} is :

$$\begin{aligned}
\boldsymbol{b_{XC}} &= \epsilon_{XC} E_{f_Z}[u_{XC}\{\boldsymbol{I_Y}^{*-1}\boldsymbol{s_Y}^{*} - \boldsymbol{I_Z}^{*-1}\boldsymbol{s_Z}^{*}\}] \\
&= \epsilon_{XC}\boldsymbol{I_Y}^{*-1} E_{f_Z}[u_{XC}\boldsymbol{s_Y}^{*}] \\
&= \epsilon_{XC}\boldsymbol{I_Y}^{*-1} E_{f_Z}[u_{XC} l^{*'}_{T|XC^r} \begin{pmatrix} 1 \\ x \\ c^r \end{pmatrix}] \\
&= \epsilon_{XC}\boldsymbol{I_Y}^{*-1} E_{f_Z}\{u_{XC}[\frac{t-\xi(\pi_{xc^r})}{a(\phi)}] \begin{pmatrix} 1 \\ x \\ c^r \end{pmatrix}\} \\
&= \frac{\boldsymbol{I_Y}^{*-1}}{a(\phi)} E_{XC}\{\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)]h_x\} - \frac{\boldsymbol{I_Y}^{*-1}}{a(\phi)} E_{X,C}\{\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)]h_x\} \\
&\approx \beta\frac{\boldsymbol{I_Y}^{*-1}}{a(\phi)} E_{XC}\{c\boldsymbol{v_0}\xi'(\pi_x)h_x\} - \beta\frac{\boldsymbol{I_Y}^{*-1}}{a(\phi)} E_{X,C}\{c\boldsymbol{v_0}\xi'(\pi_x)h_x\}
\end{aligned}$$

since $E_{T|XC}(l'_{T|XC}) = 0$.

And

$$\begin{aligned}
\boldsymbol{b_M} &= \epsilon_M E_{f_Y^*}\{u_M \boldsymbol{I_Y}^{*-1}\boldsymbol{s_Y}^{*}\} \\
&= E_{f_Y^*}\{\log\frac{f_Y^*}{f_Y^*}\boldsymbol{I_Y}^{*-1}\boldsymbol{s_Y}^{*}\} \\
&= E_{f_Y^*}\{\log\frac{f_{T|X}}{f^*_{T|X}}\boldsymbol{I_Y}^{*-1}\boldsymbol{s_{Y|r=0}}^{*}h_x\} \\
&\approx \boldsymbol{I_Y}^{*-1} E\{\frac{1}{2}\beta^2\sigma_c^2[(l'_{T|X})^2 + l''_{T|X}]l'_{T|X}\boldsymbol{v_0}h_x\} \\
&= \frac{1}{2}\beta^2\sigma_c^2\boldsymbol{I_Y}^{*-1} E_{f_x}\{E_{f^*_{T|X}}[l^{*'}_{T|X}]^3\boldsymbol{v_0}h_x\} \\
&= \frac{\beta^2\sigma_c^2}{2a(\phi)}\boldsymbol{I_Y}^{*-1} E_{f_x}\{\xi''(\pi_x)\boldsymbol{v_0}h_x\}.
\end{aligned}$$

Chapter 3

Local Sensitivity Analysis for Misspecified Missing Data Mechanism

3.1 Introduction

The majority of the literature handles the missing data problem through a selection model (Little and Rubin, 2002) $f(D, R; \theta, \psi) = f(D; \theta)f(R|D; \psi)$, and inference is based on probability distribution $f(D; \theta)$ that fit the observations on the variables and the observation process, or rather missing data mechanism (MDM) $f(R|D; \psi)$. The assumptions of *ignorable* and *parametric* modelling are the most frequently referenced, for example, a logistic linear model. This literature includes Scott and Wild (2002) with missing data in a case-control study discussed, Chen et al. (2010) considering missing response and missing covariate problems with longitudinal study and Ibrahim et al. (1999) with generalized linear regression models. Lu and Copas (2004) stated: 'A closely related, but logically distinct, concept is ignorability'. Rubin (1976) pointed out that, when making sampling distribution inferences about the parameter of the data θ, it is appropriate to ignore the process that causes missing data if the missing data are 'missing at random'. However, we should be aware that those inferences may have problems if models are misspecified, and an uncertainty analysis is then necessary to be considered in practice.

3.1.1 Uncertainty Problems for MDM

Basically, the uncertainty problem is not commonly discussed under ignorable missingness as the parameters involved are identifiable and thus 'can' be estimated. However, the inference with the 'ignorable' assumption is based on belief in the correctly specified distributions and missing data mechanism, or rather the trust of the working model. But knowledge is often limited as a result of the lack of randomization or lack of observations, e.g., or because assumptions are not proposed properly (failure of trial design). In these cases, the conventional analysis may encounter problems, and it may be better to discuss the sensitivity analysis for the potential uncertainties.

When we reconsider the missing covariates problem, we see that the true covariate distribution and missing data mechanism are actually unknown in practice. And our usual model assumptions may be questioned in some occasions. One example is US Federal Highway Administration Data in 2001 as we will discuss in Section 3.4. Data are collected in each state of the USA, and a linear regression model is assumed by Weisberg (2005) to explain the fuel consumption against four covariate variables: Federal-aid highway miles, personal Income, Drivers number and state gasoline tax. An incomplete data set will be designed artificially by dropping some of the income values with an ignorable missing data model. The covariate distribution is required to make the bias adjustment for MLEs, or perform multiple imputation for the 'missing' values. It can usually be specified as a parametric model based on the observations in complete cases since $f(D_{mis}|D_{obs}, R) = f(D_{mis}|D_{obs})$ under ignorable missing data assumption (see e.g. Rubin, 1987; Molenberghs et al., 2008). However, we notice for some trials especially with small sample sizes, that approach may not be valid for identifying a parametric conditional model on the complete cases. In this case, the missing data mechanism modelling should be considered carefully since misspecification of MDM may bring in additional bias, which is named as *missing data mechanism bias* in this thesis. In this way 'ignorable' missingness assumption can no longer be 'ignored'.

The rest of the chapter is arranged as follows. We will discuss the MDM misspecification problem generally in Section 3.2, and use the incomplete data bias analysis to assess the influence of uncertainty in Section 3.3. Then we will consider three examples. Section 3.4 will discuss the fuel consumption data example (with incomplete data designed under MAR), and local bias analysis is performed comparing with other methods without concern of uncertainty issue. We further consider a study in Section

3.5.1 when the missing data mechanism is assumed as MCAR but the true model is under MAR. The missing data mechanism bias will be induced. And Section 3.5.2 will discuss complex missing data mechanisms, and working from a logistic linear model as usual may result in bias. Both covariate bias and missing data mechanism bias are investigated by simulation studies.

3.2 Bias Models with Misspecified MDM

Now we recall missing covariate problems. Complete data $\boldsymbol{Z}$ contains (T, X, C, R) and incomplete data is $\boldsymbol{Y} = (T, X, C^{(r)}, R)$ with C partially missing, and R is the indicator vector of missingness.
The data generating model under $\boldsymbol{Z}$ is:

$$g_Z \quad = \quad f_{T|XC}(t|x, c; \theta) f_{XC}(x, c) h(r|t, x, c). \tag{3.1}$$

The regression model $f_{T|XC}$ can be any regression model and $h(r|t, x, c)$ represents the missing data mechanism, which is allowed to be ignorable or non-ignorable. The working model (assuming MDM model is $h_1(r|t, x)$) is :

$$f_Z = f_{T|XC}(t|x, c; \theta) f_{XC}(x, c) h_1(r|t, x) \tag{3.2}$$

with the misspecification of MDM as:

$$\exp\{\epsilon_R u_R\} = \frac{h(r|t, x, c)}{h_1(r|t, x)}. \tag{3.3}$$

Here $h_1(r|t, x)$ is supposed to be identifiable and usually selected as a parametric model (e.g. logistic linear model under MAR). When we further consider the covariate density uncertainty, we can assume that X and C are independent:

$$f_Z^* = f_{T|XC}(t|x, c; \theta) f_X(x) f_C(c) h_1(r|t, x) \tag{3.4}$$

with misspecification of covariate distribution:

$$\exp\{\epsilon_{XC} u_{XC}\} = \frac{f_{XC}(x, c)}{f_X(x) f_C(c)}. \tag{3.5}$$

Thus we can write another general form for g_Z

$$g_Z = f_Z \exp(\epsilon_R u_R) = f_Z^* \exp(\epsilon_{XC} u_{XC}) \exp(\epsilon_R u_R).$$

Here $(\epsilon_{XC}, \epsilon_R)$ can be thought of as the 'magnitude' of misspecifications, and (u_{XC}, u_R) can be thought of as the 'direction' of misspecifications. The two misspecifications $\exp\{\epsilon_{XC} u_{XC}\}$ and $\exp\{\epsilon_R u_R\}$ correspond to two uncertainty issues: the uncertainty of covariates distribution and the uncertainty of MDM.

We first consider the uncertainty of MDM based on the true distribution for $[XC]$: $g_Z \to f_Z$, and this step actually can transpose the non-ignorable missingness into ignorable missingness, the non-identifiable issue into identifiable. Next we will measure the uncertainty of covariates distribution under identifiable model f_Z^* as previously discussed. The illustration graph of the bias models is given in Figure 3.1. The arrows from g_Z to f_Z^* indicate the model misspecifications, i.e. MDM misspecification (u_R) and covariate distribution misspecification (u_{XC}). The arrows from complete data to incomplete data (e.g. $g_Z \to g_Y$) indicate that the corresponding model for Y is the marginal density of Z. Also the marginal distribution misspecification as discussed under GLMs (see Section 2.5) is illustrated by $f_Y^* \to f_Y^{**}$. Those different type of biases will be discussed in the next section.

$$\begin{tikzcd}
\mathrm{g_Z} \arrow[r, "\mathrm{u_R}"] \arrow[d, "\text{marginal}"] & \mathrm{f_Z} \arrow[r, "\mathrm{u_{XC}}"] \arrow[d] & \mathrm{f_Z^*} \arrow[d] & \\
\mathrm{g_Y} \arrow[r, "\mathrm{u_{R|Y}}"] & \mathrm{f_Y} \arrow[r, "\mathrm{u_{XC|Y}}"] & \mathrm{f_Y^*} \arrow[r, "\mathrm{u_M}"] & \mathrm{f_Y^{**}}
\end{tikzcd}$$

Figure 3.1: Bias models with misspecifications.

For incomplete data $\boldsymbol{Y} = (T, X, C^{(r)}, R)$, the data generating distribution is

$$\begin{aligned} g_Y &= \int_{(y)} g_Z dz \\ &= \int_{(y)} f_Z \exp(\epsilon_R u_R) dz \approx \int_{(y)} f_Z + f_Z \epsilon_R u_R dz \\ &= f_Y + \int_{(y)} f_Z \epsilon_R u_R dz \overset{u_{R|Y} = E_{f_Z}(u_R)}{\approx} f_Y \exp(\epsilon_R u_{R|Y}) \end{aligned}$$

$$\begin{aligned} \text{or } \; g_Y &\overset{f_Z = f_Z^* \exp(\epsilon_{XC} u_{XC})}{\approx} \int_{(y)} f_Z^* \exp(\epsilon_{XC} u_{XC}) dZ + \int_{(y)} f_Z \epsilon_R u_R dz \\ &\approx \int_{(y)} f_Z^* (1 + \epsilon_{XC} u_{XC}) dZ + f_Y \epsilon_R u_{R|Y} \\ &\overset{u_{XC|Y} = E_{f_Z^*}(u_{XC})}{=} f_Y^* + f_Y^* \epsilon_{XC} u_{XC|Y} + f_Y \epsilon_R u_{R|Y} \\ &\approx f_Y^* \exp(\epsilon_{XC} u_{XC|Y}) \exp(\epsilon_R u_{R|Y}) \end{aligned}$$

where

$$f_Y^* = \int_{(y)} f_Z^* dz$$

is the marginal model of f_Z^* on $\boldsymbol{Y}$ and

$$\begin{aligned} f_Y &= \int_{(y)} f_Z dy \\ &= \int_{(y)} f_Z^* \exp(\epsilon_{XC} u_{XC}) dz \\ &\approx \int_{(y)} f_Z^* (1 + \epsilon_{XC} u_{XC}) dz \\ &\approx f_Y^* \exp(\epsilon_{XC} u_{XC|Y}) \end{aligned}$$

is the marginal model of f_Z on $\boldsymbol{Y}$. And $u_{XC|Y} = E_{f_{Z^*}}(u_{XC}|\boldsymbol{Y})$; $u_{R|Y} = E_{f_Z}(u_R|\boldsymbol{Y})$. The actually working model under incomplete data is f_Y^* given by

$$f_Y^* = f_{T|XC}^r(t|x, c^{(r)}; \theta) f_X(x) f_C^r(c) h_1(r|t, x) \tag{3.6}$$

with

$$f_C^r(c) = \begin{cases} f_C(c), & r{=}1; \\ 1, & r = 0. \end{cases}$$

and

$$f_{T|XC}^r(c) = \begin{cases} f_{T|XC}, & r{=}1; \\ f_{T|X}, & r{=}0. \end{cases}$$

If we further consider the marginal misspecification from f_Z^* to f_Y^*, the bias analysis will be more complicated and we will have discussion later.

3.3 Incomplete Data Bias

We denote b_{XC} and b_R as the incomplete data bias components caused by misspecified covariates association and misspecified missing data mechanism.

The following theorem gives formula on how to calculate the bias.

Theorem 3.1. *The data generating distribution for complete data Z is noted as*

$$g_Z = g_Z(z;\theta,\epsilon_R,\epsilon_{XC},u_R,u_{XC}) = f_Z^*(z;\theta)\exp\{\epsilon_{XC}u_{XC}\}\exp\{\epsilon_R u_R\}$$

where f_Z^* is the working model, and the limiting value of MLE is denoted θ_{gZ}. Correspondingly, the sampling distribution under incomplete data Y is g_Y which is the marginal model of g_Z:

$$g_Y = f_Y^*(y;\theta)\exp\{\epsilon_R u_{R|Y}\}\exp\{\epsilon_{XC}u_{XC|Y}\}$$

where $u_{XC|Y} = E_{f_Z^*}(u_{XC}(z;\theta)|Y)$ and $u_{R|Y} = E_{f_Z}(u_R(z;\theta)|Y)$. So we use the model $f_Y^*(y;\theta)$ to fit the observations sampling from g_Y, the limiting value of MLE under Y is denoted θ_{gY}. Using Lemma 2.1, the incomplete data bias b_θ under the identifiability condition is given by

$$\begin{aligned} b_\theta &\approx \theta_{g_Y} - \theta_{g_Z} \\ &= \epsilon_{XC}{I_Y^*}^{-1}E_{f_Z^*}[u_{XC}s_Y^*] - \epsilon_{XC}{I_Z^*}^{-1}E_{f_Z^*}[u_{XC}s_Z^*] + \epsilon_R{I_Y^*}^{-1}E_{f_Z}(u_R s_Y^*) - \epsilon_R{I_Z^*}^{-1}E_{f_Z}(u_R s_Z^*) \end{aligned}$$

with s_Y^* and I_Y^* as score function and information matrix under model f_Y^*, while s_Z^* and I_Z^* are under f_Z^*.

The proof of theorem is given in Appendix 3.7.1.

In theorem 3.1, g_Z is non-negative and integrates to 1 up to and including first-order terms in $(\epsilon_R u_R, \epsilon_{XC} u_{XC})$ and it is distributed in the neighbourhood of f_Z^*.

Correspondingly, g_Y has a distribution in the neighbourhood of f_Y^*,

$$g_Y \approx f_Y^* \exp(\epsilon_R u_{R|Y}) \exp(\epsilon_{XC} u_{XC|Y}).$$

Based on the argument given around Figure 3.1, two types of biases need to be considered. One is caused by the misspecified MDM and the other is caused by the misspecified distribution for $[XC]$. Thus first two terms in the bias expression is described as covariate bias:

$$b_{XC} = \epsilon_{XC} {I_Y^*}^{-1} E_{f_Z^*}[u_{XC} s_Y^*] - \epsilon_{XC} {I_Z^*}^{-1} E_{f_Z^*}[u_{XC} s_Z^*],$$

and the last two terms as MDM bias:

$$b_R = \epsilon_R {I_Y^*}^{-1} E_{f_Z}(u_R s_Y^*) - \epsilon_R {I_Z^*}^{-1} E_{f_Z}(u_R s_Z^*).$$

The bias b_{XC} is mainly caused by the correlation of observed covariates X and missing covariate C, while the bias b_R is mainly caused by the non-identifiability of missing data mechanism.

When we consider the non-ignorable missing data problem, the working model may be wrongly assumed to be MAR. In this case MDM bias will be calculated differently. This problem will be discussed in Chapter 5.

For nonlinear model or GLM model, we need to consider the marginal model bias as well; see Figure 3.1 ($f_Y^* \to f_Y^{**}$). The discussion is similar to that in Section 2.5.

3.4 Fuel Consumption Data Example

US Federal Highway Administration published fuel consumption data over 50 United States and the District of Columbia in 2001; it was analysed in Weisberg (2005) (Chapter 1, page 15). The aim of the research is to understand the effect on fuel consumption (T) with Federal-aid highway miles (X_1), personal Income (X_2), Drivers number (X_3) and state gasoline tax (X_4). Summarized variables after using transformation and standardization are listed in Table 3.1. The linear regression model with parameter $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)^T$ estimated using complete data is

$$\hat{t} = 154.19 + 18.55x_1 - 6.14x_2 + 0.47x_3 - 4.23x_4.$$

Table 3.1: Variables in fuel consumption data

FuelC	Gasoline sold for road use, thousands of gallons
State	State name
Pop	2001 population age 16 and over
Miles	Miles of Federal-aid highway miles in the state
Drives	Number of licensed drivers in the state
Fuel (T)	1000 × FuelC/Pop
logMiles ($X1$)	Base-two logarithm of Miles
Income ($X2$)	Per person personal income for the year 2000, in thousands of dollars
Dlic ($X3$)	1000 × Drivers/Pop
Tax($X4$)	Gasoline state tax rate, cents per gallon

Source: Highway Statistics 2001. http://www.fhwa.dot.gov/ohim/hs01/index.htm.

We find the relevance of regression relationship between fuel and the other variables in Figure 3.2, also the correlation between some covariates. The correlation matrix is listed in Table 3.2.

Table 3.2: Correlation of covariates in fuel consumption data

	Logmiles (X_1)	Income (X_2)	Dlic (X_3)	Tax (X_4)
Logmiles (X_1)	1.000	-0.296	0.031	-0.044
Income (X_2)	-0.296	1.000	-0.176	-0.011
Dlic (X_3)	0.031	-0.176	1.000	-0.086
Tax (X_4)	-0.044	-0.011	-0.086	1.000

Bias analysis for incomplete data:

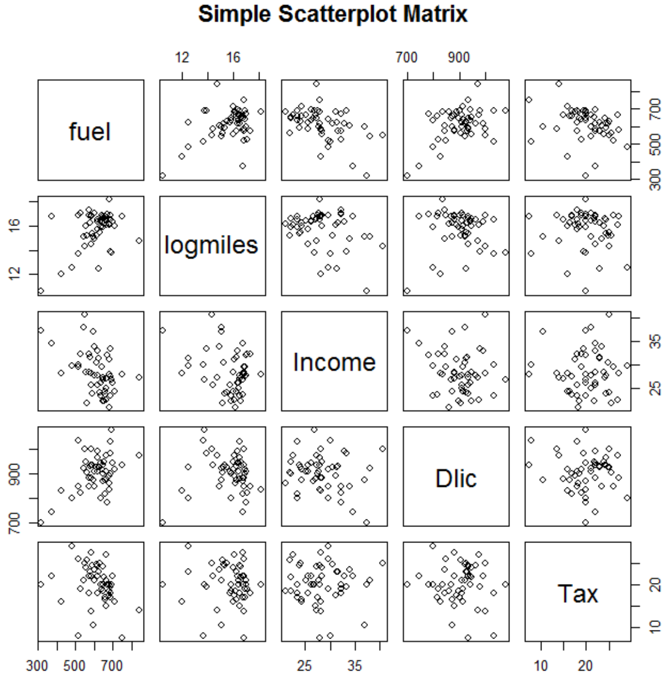


Figure 3.2: Scatterplot matrix for fuel consumption data

We design the incomplete dataset by letting Income (X_2) be missing with probability $h(r=0|x_1) = 1 - \text{expit}(1 + 0.5(x_1 - \bar{x}_1) - (x_1 - \bar{x}_1)^2)$ where $\bar{x}$ is the average $E(x)$ and r is the indicator for missingness of Income, which equals 1 when data is observed or 0 otherwise. Then we obtain incomplete data $Y = (T, X_1, X_2^{(r)}, X_3, X_4)$ [1]. To make it identifiable, we assume $\text{corr}(x_2, x_i) = 0$, for $i = 1, 3, 4$. Using incomplete data Y parameter θ can be estimated by ML method (denoted as θ_{gY}):

$$L \propto f_Y = f(t|x_1, x_2^{(r)}, x_3, x_4) f(x_1, x_3, x_4) f(x_2^{(r)}) h(r|x_1).$$

This is the marginal model of $f_Z = f(t|x_1, x_2, x_3, x_4) f(x_1, x_3, x_4) f(x_2) h(r|x_1)$ on complete data.

Bias analysis for misspecified models with missing data is conducted and bias of $\hat{\theta}_{gY}$ is given as:

$$\text{Bias}(\hat{\theta}_{gY}) = \theta_2 \frac{\boldsymbol{I_Y}^{-1}}{\sigma_Y^2} E\{x_2 \boldsymbol{v_0} h(r=0|x_1)\} \tag{3.7}$$

with $\boldsymbol{v_0} = (1, x_1, 0, x_3, x_4)^T$. Given the variable correlations listed in Table 3.2, the ad-

[1] Raw data and a simulated incomplete data are presented in Appendix 3.7.2

justed estimation $\hat{\theta}$ can be calculated by using this formula and the results are shown in Table 4.3. Given the MDM model, the incomplete data bias $(\theta_{gZ} - \theta_{gY})$ measures the misspeficiation of covariate distribution. As noticed, the covariate bias for θ_1 is significantly affected by the correlation between covariates; the other estimators are slightly biased, because of the small correlations between those variables.

Besides, we conduct complete case (CC) analysis and multiple imputation (MI) analysis with Bayesian linear regression imputation ('mice' in software R) method (van Buuren and Groothuis-Oudshoorn, 2011). The complete case estimation is seriously biased, which interprets the influence of personal Income toward fuel consumption in the wrong direction. Estimation based on multiple imputation method with MAR assumption works better than complete case analysis. But the bias is also large. These two methods provide the results without concern on model uncertainty, and reflect the necessary to consider the potential misspecifications.

Table 3.3: Simulation study result

		θ_0	θ_1	θ_2	θ_3	θ_4
	$\hat{\theta}_{gZ}$	154.193	18.545	-6.135	0.472	-4.228
	SE	194.906	6.472	2.194	0.129	2.030
$\hat{\theta}_{gY}$	Given MDM	149.762	22.581	-6.404	0.484	-4.802
	1.MCAR	147.029	24.212	-7.299	0.480	-5.203
	2.Logit Linear	148.535	22.788	-7.306	0.473	4.088
	3.GAM(Nonp)	148.557	22.312	-7.326	0.471	-5.246
Others	MI	155.176	11.808	-8.300	0.393	-5.474
	CC	164.041	-1.247	-6.998	0.331	-7.689

Note: $\hat{\theta}_{gZ}$ is calculated based on complete dataset, with the standard error 'SE'. $\hat{\theta}_{gY}$ is the estimation without adjustment of incomplete data bias. Three fitting models for MDM are considered: MCAR, logistic linear model and generalized additive model [2] (Hastie and Tibshirani, 1990) with nonparametric method; adjusted estimations are given respectively. Multiple imputation (MI) is performed under MAR with 'mice' (i.e. Bayesian linear regression method) in R software and CC is complete case analysis. Simulation study is repeated 100 times.

The generalized additive model is similar to generalized linear model, where an exponential family distribution is specified for the response variable T (for example normal, binomial or Poisson distributions) along with a link function ξ relating the expected value of T to the predictor variables via a structure such as

$$\xi(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

where $x_i, i = 1 \ldots m$ are predictor variables and the 'smooth functions' $f_i(x_i)$ may be specified parametrically (e.g. polynomial) or non-parametrically.

Covariate distribution uncertainty:

We noticed that the calculation of incomplete data bias in formula (3.7) requires the specification of the conditional distribution $f(x_2|x_1, x_3, x_4)$. And it can usually be specified on complete cases under ignorable missingness:

$$f(x_2|x_1, x_3, x_4) = f(x_2|x_1, x_3, x_4, r = 1), \tag{3.8}$$

but it is not always certainly going to have the density 'f' calculated accurately because of a lack of randomization, especially in a trial with small sample size (Stubbendick and Ibrahim, 2003). For example, when we fit the conditional density as a normal distribution [3]

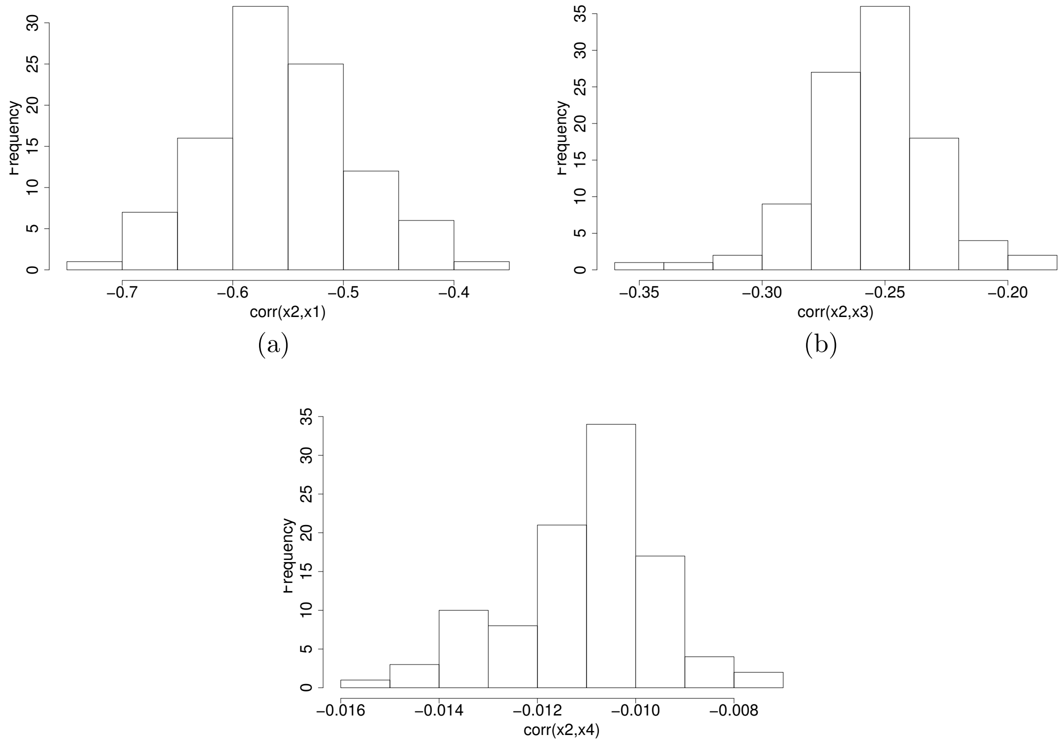
$$x_2|(x_1, x_3, x_4) \sim N(\gamma_0 + \gamma_1 x_1 + \gamma_3 x_3 + \gamma_4 x_4, \tau^2)$$

The estimators of parameter $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_3, \gamma_4, \tau^2)$ under complete cases by distribution $f(x_2|x_1, x_3, x_4, r = 1; \boldsymbol{\gamma})$ are actually 'biased' from under all cases $f(x_2|x_1, x_3, x_4; \boldsymbol{\gamma})$; as shown in Table 3.4. And correspondingly the crucial variable correlations such as corr(x_1, x_2) are not precisely estimated, e.g. $\widehat{\text{corr}}(x_1, x_2)$ =-0.552 based on modelling strategy (3.8). See Figure 3.3 for estimations from 100 repeated studies. Although the estimators for γ is not seriously biased (which may because the normal distribution is a good fit, see footnote) as shown in Table 3.4, the evaluation of the covariate correlations have moderate bias. In this case, the sensitivity analysis is necessary to be realistically considered to evaluate those bias parameters.

Table 3.4: Parameter estimation for conditional covariate density

		γ_0	γ_1	γ_2	γ_3
all cases	MLE	28.403	-0.875	-0.010	-0.037
	SE	0.604	0.411	0.008	0.135
	p-value	0.000	0.038	0.222	0.783
complete cases	MLE	28.066	-0.724	-0.014	-0.039
	SE	0.967	1.125	0.014	0.189
	p-value	0.000	0.456	0.351	0.591

[3] Since x_2 is the personal income, normal distribution seems a reasonable assumption, and the p-values of Shapiro Wilk normality test for regression residuals based on all cases and complete cases are 0.373 and 0.989 respectively.



(a)

(b)

(c)

Figure 3.3: Estimation of correlations.

Missing data mechanism uncertainty:

As indicated by Theorem 3.1, once the uncertainty issue exists in specifying covariate distributions, the missing data mechanism modelling is required. Since true missing data mechanism $h(r|x_1)$ is unknown in missing data problems, we consider different MDM working models, for example: 1) MCAR; 2) logistic linear model; and 3) generalized additive model (Hastie and Tibshirani, 1990) with non-parametric method. Simulation results are listed in Table 3.3. It shows the existance of MDM bias, although not serious in this example, and we should always bear this problem in mind. We also found that the non-parametric fitting is one of best considerations for complex missing data mechanism models (e.g. logistic quadratic model used in this simulation study) and fitting by MCAR and logistic linear model will be misspecified. The benefit of using non-parametric model will also be discovered in the non-ignorable examples.

3.5 Numerical Results under Misspecified MDM

3.5.1 An Example: MAR-MCAR

The misspecification problem of covariate density has been discussed in Chapter 2, and in this section we will address the MDM misspecification for a specific example when the true MDM is under MAR but is wrongly supposed as MCAR.

Assume a linear model

$$t|(x,c) \sim N(\alpha + \theta x + \beta c, \sigma^2)$$

with C partly missing. Covariate X has a Bernoulli distribution $X \sim B(1, p_x)$:

$$x = \begin{cases} 1, & p_x; \\ 0, & 1\text{-}p_x. \end{cases}$$

with density function as $f(x) = p_x^x(1-p_x)^{1-x}$. Then $E(x) = p_x$ and $\sigma_x^2 = p_x(1-p_x)$. The assumption of missingness for C is MAR and it depends on X through a logistic form:

$$\text{logit}\{h(r=1|x)\} = \psi_0 + \psi_1 x.$$

Under MCAR assumption, we have $h(r=1|x=1) = h(r=1|x=0)$, but when the true MDM is MAR: $h(r=1|x=1) = \text{expit}(\psi_0+\psi_1)$ while $h(r=1|x=0) = \text{expit}(\psi_0)$. We denote $\pi_1 = h(r=1|x=1)$ and $\pi_0 = h(r=1|x=0)$, thus $\psi_1 = \text{logit}(\pi_1) - \text{logit}(\pi_0)$, which describes the difference of observed probability on different values of x, and it also reflects the departure from MCAR to MAR. We still use the notation $h_x = h(r=0|x)$ for simiplicity. Using formula (2.23), we have

$$\begin{aligned} & E_{XC}(ch_x) - E_{X,C}(ch_x) \\ = \; & E_X[h_x E_{C|X}(c)] - E_C(c)E_X(h_x) \\ = \; & E_X[h_x \rho \frac{\sigma_c}{\sigma_x}(x - E(x))] \\ = \; & \rho \frac{\sigma_c}{\sigma_x}\{E_X(xh_x) - E(x)E_X(h_x)\} \\ = \; & \rho \frac{\sigma_c}{\sigma_x}[p_x\pi_1 - p_x\pi] \\ = \; & \text{cov}(x,c)(\pi_1 - \pi_0), \end{aligned}$$

where the marginal density $E(h_x)$ is calculated as:

$$\begin{aligned}\pi = E(h_x) &= p_x h(r=0|x=1) + (1-p_x)h(r=0|x=0)\\ &= p_x\pi_1 + (1-p_x)\pi_0.\end{aligned}$$

Similarly

$$\begin{aligned}& E_{XC}[cxh_x] - E_{X,C}[cxh_x]\\ =\ & E_X[xh_xE_{C|X}(c)] - E_X[xh_x]E_C(c)\\ =\ & \rho\frac{\sigma_c}{\sigma_x}\{E_X(x^2h_x) - E(x)E_X(xh_x)\}\\ =\ & \rho\frac{\sigma_c}{\sigma_x}(p_x\pi_1 - p_xp_x\pi_1)\\ =\ & \mathrm{cov}(x,c)\pi_1.\end{aligned}$$

Then incomplete data bias under MAR is

$$\boldsymbol{b} = \frac{\beta\mathrm{cov}(x,c)}{\sigma_Y^2}\boldsymbol{I_Y^{-1}}\begin{pmatrix}\pi_1-\pi_0\\ \pi_1\\ 0\end{pmatrix}. \tag{3.9}$$

If $\pi_1 = \pi_0 = \pi$, it reduces as MCAR, and the incomplete data is expressed as:

$$\boldsymbol{b_{XC}} = \frac{\beta}{\sigma_Y^2}\boldsymbol{I_Y^{-1}}E(cu_{XC}\boldsymbol{v_0}\pi) = \frac{\beta\mathrm{cov}(x,c)}{\sigma_Y^2}\boldsymbol{I_Y^{-1}}\begin{pmatrix}0\\ \pi\\ 0\end{pmatrix}. \tag{3.10}$$

But this creats the MDM bias and it can be estimated using Theorem 3.1:

$$\begin{aligned}
\boldsymbol{b_R} &= \epsilon_R \boldsymbol{I_Y^{-1}} E(u_R \boldsymbol{s_Y}) - \epsilon_R \boldsymbol{I_Z^{-1}} E(u_R \boldsymbol{s_Z}) \\
&= \boldsymbol{I_Y^{-1}} E(\pi \log(\frac{h_x}{\pi}) \boldsymbol{s_{Y|r=0}}) + E((1-\pi)\log(\frac{1-h_x}{1-\pi}) \boldsymbol{s_{Y|r=1}}) \\
&\approx \frac{\beta}{\sigma_Y^2} \boldsymbol{I_Y^{-1}} E[(h_x - \pi) c \boldsymbol{v_0}] \\
&= \frac{\beta \text{cov}(x,c)}{\sigma_Y^2} \boldsymbol{I_Y^{-1}} \begin{pmatrix} \pi_1 - \pi_0 \\ \pi_1 - \pi \\ 0 \end{pmatrix}
\end{aligned}$$

which is approximately equal to the departure between the incomplete data bias under MAR (3.9) and MCAR (3.10). The differences $(\pi_1 - \pi_0)$ and $(\pi_1 - \pi)$ measure how much the missingness depends on the covariate X, which also index the cost of treating MAR as MCAR in this specific issue.

Below we perform a simulation study to compare the size of both bias sources and identify the importance of correctly specifying the MDM model. We let $(\alpha, \theta, \beta) = (0.2, 0.6, 1)$, $p_x = 0.3$ and $\sigma_c^2 = 1$ such that the size of bias is relatively large. The sample size is chosen as n=(50, 200, 1000) and simulation study is conducted with 100 replications. The simulation results are shown in Table 3.5 where corr(x, c) is fixed at a medium level: corr(x, c)= 0.5. We can see that both the covariate bias and MDM bias increase with smaller ψ_1 (moving further from MCAR). MDM bias is almost in the same scale as covariate bias when corr(x, c) is large or MDM is more far away from MCAR.

Confidence interval of effect size is calculated and adjusted: $\hat{\theta} = \hat{\theta}_{gY} - \boldsymbol{b_{XC}} - \boldsymbol{b_R}$, and coverage rates (CR) of 100 replications are shown in the tables. CR1 is the coverage rate with covariate bias adjustment only $\hat{\theta} = \hat{\theta}_{gY} - \boldsymbol{b_{XC}}$, as seen from the table, CR is usually better than CR1. The difference shows the cost of MDM misspecification, which is apparent to be related with the sample size (where the estimators are more accurate and bias adjustment seems more necessary).

3.5.2 Misspecified MDM Model under MAR

Much of literature discussing the missing data problem is based on MAR assumption. The logistic linear model is popularly used for MDM specification. In practice,

Table 3.5: Covariate bias and MDM bias

		α					θ					β				
n	ψ_1	EB	b_{XC}	b_R	CR	CR1	EB	b_{XC}	b_R	CR	CR1	EB	b_{XC}	b_R	CR	CR1
n=50	0	-0.052	-0.038	-0.001	95	95	0.171	0.135	0.004	83	82	-0.066	-0.019	-0.001	93	87
	-0.5	-0.052	-0.053	0.002	97	97	0.182	0.186	0.012	90	90	-0.060	-0.041	-0.001	93	93
	-1	-0.044	-0.062	0.006	95	95	0.252	0.229	0.065	89	92	-0.084	-0.048	-0.009	94	90
	-2	-0.050	-0.060	0.018	99	99	0.535	0.303	0.198	88	76	-0.118	-0.028	-0.014	95	91
	-3	-0.045	-0.064	0.025	98	98	0.635	0.355	0.263	95	82	-0.125	-0.018	-0.011	97	87
	-5	-0.049	-0.068	0.029	96	96	0.763	0.433	0.348	87	77	-0.114	-0.002	-0.002	98	90
n=200	0	-0.045	-0.049	0.001	90	90	0.147	0.165	0.001	86	89	-0.058	-0.039	0.001	95	92
	-0.5	-0.044	-0.049	0.001	91	93	0.181	0.173	0.011	89	90	-0.066	-0.039	-0.002	93	92
	-1	-0.057	-0.055	0.007	95	95	0.277	0.210	0.062	85	81	-0.075	-0.037	-0.010	92	87
	-2	-0.055	-0.059	0.017	98	98	0.434	0.281	0.161	90	75	-0.103	-0.027	-0.015	97	77
	-3	-0.072	-0.062	0.024	90	89	0.639	0.347	0.252	92	46	-0.117	-0.017	-0.012	96	69
	-5	-0.078	-0.067	0.029	92	91	0.780	0.429	0.346	92	44	-0.116	-0.003	-0.002	93	66
n=1000	0	-0.042	-0.047	-0.001	88	91	0.156	0.158	-0.001	90	91	-0.058	-0.036	0.001	99	86
	-0.5	-0.048	-0.048	0.001	91	97	0.181	0.166	0.009	88	85	-0.059	-0.037	-0.002	96	86
	-1	-0.056	-0.054	0.007	96	94	0.279	0.207	0.062	92	66	-0.076	-0.036	-0.011	97	64
	-2	-0.059	-0.059	0.017	96	94	0.46	0.278	0.156	90	16	-0.102	-0.028	-0.015	95	35
	-3	-0.061	-0.063	0.024	95	93	0.624	0.349	0.246	90	5	-0.110	-0.018	-0.012	90	10
	-5	-0.061	-0.065	0.029	92	89	0.785	0.422	0.344	92	3	-0.099	-0.003	-0.003	79	13

MDM model $h(r=1|x) = \text{expit}(1-\psi_1 x)$. EB: empirical bias $(\hat{\theta}_{gY} - \theta)$.

however, the true MDM is often complicated. For example, we can have the missing data mechanism models in the form of

M1 (Logistic Quadratic)

$$h(r=1|x) = \text{expit}(\psi_0 + \psi_1 x + \psi_2 x^2);$$

or M2 (Log-Log Quadratic)

$$h(r=1|x) = 1 - \exp\{-\exp(\psi_0 + \psi_1 x + \psi_2 x^2)\}.$$

We now use Theorem 3.1 to analyse the bias caused by a misspecified MDM model.

Suppose the true missing data mechanism depends on X, denoted as $h(r|x)$, while the actually working model is denoted as $h_1(r|x)$ (e.g. logistic linear model). Then the misspecification for MDM component is

$$\exp(\epsilon_R u_R) = \frac{h(r|x)}{h_1(r|x)}. \tag{3.11}$$

As in equation (2.24) we use Taylor series approximation:

$$h_x = h(r=0|x) + h'(r=0|x)x + O(h''(r=0|x)).$$

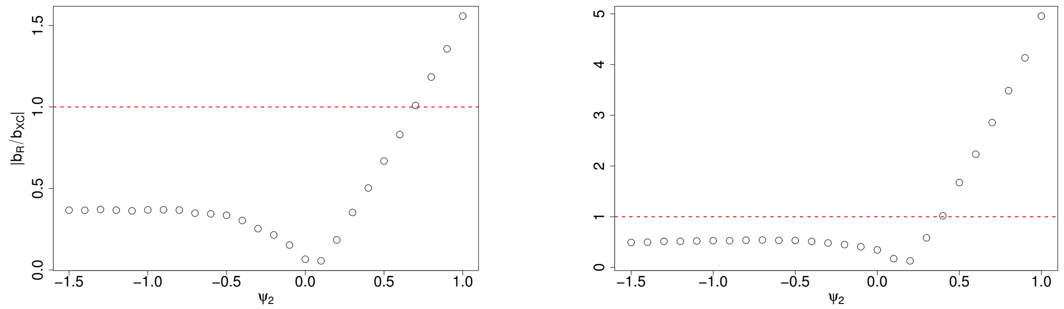
In Theorem 2.1, we calculated the incomplete data bias for the linear regression model, which depends on the expectation of the censoring probability h_x and includes up to its first derivative. But for complex models the high order items such as $E(h''_x)$ can no longer be abandoned. We should add this item to the bias expression, for example, the incomplete data bias is

$$b_{XC} = \frac{\beta \boldsymbol{I_Y^{-1}}}{\sigma_Y^2} \text{cov}(x,c) \begin{pmatrix} E_X(h'_x) \\ E_X(h_x) + E_X(h''_x) \\ 0 \end{pmatrix},$$

under linear regression model $t|(x,c) \sim N(\alpha+\theta x+\beta c, \sigma^2)$ with $X \sim N(0, \sigma_x^2)$. Details were given in Appendix 2.7.4.

A simulation study is conducted below. The true value is $(\alpha, \theta, \beta) = (0.2, 0.6, 1)$ with both X and C assumed standardised normally distributed with $\text{corr}(x,c) = 0.5$. Moderate variation for $t|(x,c)$ is taken from $U(0.16, 1)$. 1000 observations are generated with C missing through logistic quadratic (M1) or log-log quadratic model (M2). Here we fix $\psi_0 = 1$ and $\psi_1 = 0.5$, but vary ψ_2 between(-1.5, 1). Then, we calculate the covariate bias b_{XC} and missing data mechanism bias b_R, and obtain the average ratio $|b_R|/|b_{XC}|$ for 100 replications. The results are shown in Figure 3.4. The horizontal line corresponds to $|b_R|/|b_{XC}| = 1$, indicating the same size for both biases. As seen from the figures, although the missing data mechanism bias is often smaller than the covariate bias (i.e. ratio around 0.5), this is not uniformly true when ψ_2 getting larger, both bias sizes actually decrease but covariate bias apparently decrease much faster than MDM bias. More attentions should be paid for these cases.

More simulations under various missing data mechanisms are given in Appendix 3.7.3.



(a) logistic Quadratic

(b) log-log Quadratic

Figure 3.4: Simulation study: ratio of missing data bias and covariate bias. $\psi_2 \in (-1.5, 1)$.

3.6 Discussion

In this chapter we addressed another uncertainty problem in missing data mechanism specification. We realized that the usual parametric model fitting may cause additional bias, which is termed as missing data mechanism bias. It calculates the departure of the conditional working model from the true model. And this bias is compared with covariate bias which was discussed in Chapter 2.

The model uncertainties involved in the missing data were explained in Figure 3.1, where we first consider the misspecification of MDM and then misspecification of covariate density. These two uncertainties may coexist in many missing data problems.

A simulation study based on fuel consumption data was given in Section 3.4 . Incomplete data was generated artifically, then model uncertainties were assessed and the incomplete data bias analysis was calculated. When we consider the MDM misspecification only, three different models were used to fit the MDM, and the simulation results showed that nonparametric fitting works better than the other two. When we further consider the covariate density misspecification, serious bias exists. This example draws our attention to the model misspecification issue for ignorable missing data, which requests a careful concern about the model assumptions.

Then we discussed two simulation studies to compare the size of both biases (covariate bias and MDM bias). The first example had a MCAR assumption but the true model was MAR. Ignoring differences between these two assumptions resulted in MDM bias. The second example assumed a logistic linear model while the true missing data mechanism was more complicated. The MDM bias calculated in the simulation studies was considerably large, suggesting MDM model selection should be made properly.

There are some difficulties in estimating the bias parameters in the incomplete data bias analysis based on observed knowledge, and these parameters are treated as uncertainty parameters in sensitivity analysis. The detailed discussions will be given in Chapter 4. Local bias analysis is a general tool to analyse the uncertainty problems, and we will further consider the non-ignorable missing data in Chapter 5.

3.7 Appendix

3.7.1 Proof of Theorem 3.1

The true distribution g_Z for $\boldsymbol{Z} = (T, X, C, R)$ is

$$g_Z = f_{T|XC} f_{XC} h(r|t, x, c)$$

with c partially missing and the model is assumed as $h(r|t, x, c)$. The fitting model is double misspecified from g_Z:

$$g_Z = f_Z \exp(\epsilon_R u_R) = f_Z^* \exp(\epsilon_R u_R) \exp(\epsilon_{XC} u_{XC})$$

where

$$f_Z = f_{T|XC} f_{XC} h_1(r|t, x)$$

$$f_Z^* = f_{T|XC} f_X f_C h_1(r|t, x).$$

From step g_Z to f_Z, it is a MDM misspecification problem and from step f_Z to f_Z^*, it is a missing covariate problem. We fit f_Z^* to a random sample of n observations from g_Z, the limiting value of the MLE $\hat{\theta}_Z$ as $n \to \infty$ is

$$\theta_{gZ} = arg_\theta[E_{gZ}\{s_Z^*(z; \theta)\} = 0] \approx \theta + \epsilon_{XC} E_{f_Z^*}(u_{XC} {I_Z^*}^{-1} s_Z^*) + \epsilon_R E_{f_Z}(u_R {I_Z^*}^{-1} s_Z^*).$$

where I_Z^*, s_Z^* are Fisher information matrix and score function under model f_Z^*.

Under incomplete data Y, the marginal distribution of g_Z is g_Y

$$\begin{aligned} g_Y &= \int_{(y)} g_Z dy \\ &= \int_{(y)} f_Z \exp(\epsilon_R u_R) dy \approx f_Y \exp(\epsilon u_{R|Y}) \\ &= \int_{(y)} f_Z^* \exp(\epsilon_{XC} u_{XC|Z}) \exp(\epsilon_R u_R) dy \approx f_Y^* \exp(\epsilon_{XC} u_{XC|Y}) \exp(\epsilon_R u_{R|Y}) \end{aligned}$$

where

$$u_{R|Y} = E_{f_Z}(u_R(z;\theta)|Y),$$

$$u_{XC|Y} = E_{f_Z^*}(u_{XC}(z;\theta)|Y).$$

The actually working model for incomplete data Y is $f_Y^*(y;\theta)$, and

$$f_Y^*(y;\theta) = f_{T|XC}^r f_X f_c^r h_1(r|t,x).$$

We assume the difference between f_Y^* and f_Y is small, so that for a small value of ϵ_R, we can write

$$f_Y = f_Y^* \exp(\epsilon_{XC} u_{XC|Y}(y;\theta)).$$

The working model f_Y^* can be regarded as misspecified with misspecification quantities $\epsilon u_Y = \epsilon_{XC} u_{XC|Y}(y;\theta) + \epsilon_R u_{R|Y}(y;\theta)$. Then we fit f_Y^* to a random sample of n observations from g_Y, the limiting value of MLE $\hat{\theta}_Y$ as $n \to \infty$ is

$$\begin{aligned} \theta_{gY} &= arg_\theta[E_{gY}\{s_Y^*(y;\theta)\} = 0] \approx \theta + \epsilon E_{f_Y^*}(u_y {I_Y^*}^{-1} s_Y^*) \\ &= \theta + \epsilon_{XC} E_{f_Y^*}(u_{XC|Y} {I_Y^*}^{-1} s_Y^*) + \epsilon_R E_{f_Y^*}\{u_{R|Y} {I_Y^*}^{-1} s_Y^*\}. \end{aligned}$$

Following Theorem 1, we can have the incomplete-data bias as

$$\begin{aligned} b_\theta &\approx \theta_{g_Y} - \theta_{g_Z} \\ &= \epsilon_{XC} E_{f_Y^*}(u_{XC|Y} {I_Y^*}^{-1} s_Y^*) + \epsilon_R E_{f_Y^*}\{u_{R|Y} {I_Y^*}^{-1} s_Y^*\} \\ &\quad -(\epsilon_{XC} E_{f_Z^*}(u_{XC} {I_Z^*}^{-1} s_Z^*) + \epsilon_R E_{f_Z}(u_R {I_Z^*}^{-1} s_Z^*)). \end{aligned}$$

Because

$$
\begin{aligned}
& \epsilon_{XC} E_{f_Y^*}(I_Y^{*\,-1} u_{XC|Y} s_Y^*) \\
= \quad & \epsilon_{XC} I_Y^{*\,-1} \int u_{XC|Y} s_Y^* f_Y^* dy \\
\overset{f_Y = f_Y^* \exp(\epsilon_{XC} u_{XC|Y})}{=} \quad & \epsilon_{XC} I_Y^{*\,-1} \int u_{XC|Y} s_Y^* f_Y \exp(-\epsilon_{XC} u_{XC|Y}) dy \\
\approx \quad & \epsilon_{XC} I_Y^{*\,-1} \int u_{XC|Y} s_Y^* f_Y (1 - \epsilon_{XC} u_{XC|Y}) dy \\
= \quad & \epsilon_{XC} I_Y^{*\,-1} \int u_{XC|Y} s_Y^* f_Y dy - \epsilon_{XC} I_Y^{*\,-1} \int u_{XC|Y} s_Y^* f_Y \epsilon_{XC} u_{XC|Y} dy \\
\overset{u_{XC|Y} = E_{f_Z^*}(u_{XC})}{=} \quad & \epsilon_{XC} I_Y^{*\,-1} \int (\int_{(y)} u_{XC} f_Z^* dy) s_Y^* f_Y dy - \epsilon_{XC}^2 I_Y^{*\,-1} E[u_{XC|Y} s_Y^* u_{XC|Y}] \\
= \quad & \epsilon_{XC} I_Y^{*\,-1} \int_{(y)} u_{XC} s_Y^* f_Z^* dz - O(\epsilon_{XC} \epsilon_{XC}) \\
\approx \quad & \epsilon_{XC} E_{f_Z^*}(I_Y^{*\,-1} u_{XC} s_Y^*)
\end{aligned}
$$

and

$$
\begin{aligned}
& \epsilon_R E_{f_Y^*}(I_Y^{*\,-1} u_{R|Y} s_Y^*) \\
= \quad & \epsilon_R I_Y^{*\,-1} \int u_{R|Y} s_Y^* f_Y^* dy \\
\overset{u_{R|Y} = E_{f_Z}(u_R)}{=} \quad & \epsilon_R I_Y^{*\,-1} \int (\int_{(y)} u_{R|Z} f_Z d(y)) s_Y^* f_Y^* dy \\
= \quad & \epsilon_R I_Y^{*\,-1} \int_{(y)} u_{R|Z} s_Y^* f_Z dz \\
\approx \quad & \epsilon_R E_{f_Z}(I_Y^{*\,-1} u_R s_Y^*).
\end{aligned}
$$

So we have

$$\epsilon_{XC} E_{f_Y^*}(I_Y^{*\,-1} u_{XC|Y} s_Y^*) \approx \epsilon_{XC} E_{f_Z^*}(I_Y^{*\,-1} u_{XC} s_Y^*),$$

$$\epsilon_R E_{f_Y^*}\{u_{R|Y} I_Y^{*\,-1} s_Y^*\} \approx \epsilon_R E_{f_Z}\{u_R I_Y^{*\,-1} s_Y^*\}.$$

Thus

$$
\begin{aligned}
b \quad \approx \quad & \theta_{gY} - \theta_{gZ} \\
= \quad & \epsilon_{XC} I_Y^{*\,-1} E_{f_Z^*}[u_{XC} s_Y^*] - \epsilon_{XC} I_Z^{*\,-1} E_{f_Z^*}[u_{XC} s_Z^*] \\
& \quad + \epsilon_R E_{f_Z}(u_R I_Y^{*\,-1} s_Y^*) - \epsilon_R E_{f_Z^*}(u_R I_Z^{*\,-1} s_Z^*).
\end{aligned}
$$

One thing to notice is that the information matrix (I_Z^* under f_Z^* or I_Z under f_Z) and score function (s_Z^* under f_Z^* or s_Z under f_Z) are not changing during the bias models (g_Z, f_Z, f_Z^*), which means $I_Z^* = I_Z$ for example. Incomplete data is in the same case with complete data.

3.7.2 Fuel Consumption Data

```
       Fuel      Dlic   Income Logmiles   Tax
AL 690.2644 1031.3801 23.471    16.52711 18.00
AK 514.2792 1031.6411 30.064 * 13.73429  8.00
AZ 621.4751  908.5972 25.578    15.75356 18.00
AR 655.2927  946.5706 22.257 * 16.58244 21.70
CA 573.9129  844.7033 32.275 * 17.36471 18.00
CO 616.6115  989.6062 32.949 * 16.38960 22.00
CT 549.9926  999.5934  40.64 * 14.35191 25.00
DE 626.0239  924.3448 31.255 * 12.50532 23.00
DC 317.4924  700.1953 37.383 * 10.58308 20.00
FL 586.3461 1000.1242 28.145    16.83983 13.60
GA 750.9074  933.3026  27.94 * 16.81796  7.50
HI 426.3494  829.9971 28.221 * 12.06272 16.00
ID 628.4279  925.1934  24.18    15.49904 25.00
IL 526.2377  819.4367 32.259    17.07806 19.00
IN 666.5365  879.2352 27.011 * 16.52096 15.00
IA 647.0016  867.4907 26.723    16.79153 20.00
KS 600.9024  909.0653 27.816    17.03966 21.00
KY 659.7413  871.9985 24.294    16.26799 16.40
LA 633.7348  800.6851 23.334 * 15.89247 20.00
ME 584.0926  932.9716 25.623 * 14.46862 22.00
MD 602.2862  844.9638 33.872 * 14.90228 23.50
MA 543.2321  920.6589 37.992 * 15.11179 21.00
MI 642.9706  914.6338 29.612    16.89404 19.00
MN 672.9191  782.8124 32.101    17.01324 20.00
MS 683.5020  860.8079 20.993    16.16940 18.40
MO 689.3661  899.8468 27.445    16.92375 17.00
MT 666.5978  974.2352 22.569    16.08479 27.00
```

```
NE 617.6905  963.7331 27.829   16.50131 24.50
NV 614.8940  923.8037 30.529   15.23848 24.75
NH 689.6521  980.4662 33.332 * 13.92073 19.50
NJ 597.6403  873.1364 36.983   15.14271 10.50
NM 646.5273  898.9639 22.203   15.86986 18.50
NY 374.1641  744.3802 34.547 * 16.78547 22.00
NC 645.4418  935.3808 27.194 * 16.62678 24.10
ND 666.1887  907.8909 25.068 * 16.40193 21.00
OH 572.0756  880.1512   28.4   16.83944 22.00
OK 657.0605  814.8619 23.517   16.78205 17.00
OR 556.3455  948.0717  28.35   16.02721 24.00
PA 518.3286  848.5881 29.539   16.87249 26.00
RI 482.3269  798.1338 29.685 * 12.56343 29.00
SC 711.7331  914.8527 24.321   16.01382 16.00
SD 697.0528  943.8959 26.115   16.35052 22.00
TN 638.2311  942.0444 26.239   16.42236 20.00
TX 681.1001  835.2956 27.871 * 18.19829 20.00
UT 591.4999  935.7885 23.907 * 15.36523 24.50
VT 691.0227 1075.2882 26.901 * 13.80282 20.00
VA 681.0311  889.9195 31.162 * 16.10985 17.50
WA 576.0697  930.8562 31.528   16.30537 23.00
WV 562.4109  904.8936 21.915   15.17512 25.65
WI 581.7937  882.3291 28.232   16.78165 27.30
WY 842.7918  970.7527  27.23   14.73619 14.00

* index the data (Income) is designed to be missing
```

3.7.3 Simulation Studies for Complex Misspecified Models

For a linear regression model

$$t|(x, c) \sim N(\alpha + \theta x + \beta c, \sigma^2)$$

with covariates as multivariate normal distributed. Data is generated with true value $(\alpha, \theta, \beta) = (0.2, 0.6, 1)$ and σ^2 takes value from $U(0.16, 1)$. Mean and variance is 0 and 1 for both x and c respectively. The correlation ρ=corr(x, c) is selected to vary between (0,0.1,0.3,0.5) for different studies.

M1 (Logit Linear)

$$h(r = 1|x) = \text{expit}(\psi_1 + \psi_2 x)$$

M2 (Logit Quadratic)

$$h(r = 1|x) = \text{expit}(\psi_0 + \psi_1 x + \psi_2 x^2)$$

M3 (Log-Log Linear)

$$h(r = 1|x) = 1 - \exp\{-\exp(\psi_0 + \psi_1 x)\}$$

M4 (Log-Log Quadratic)

$$h(r = 1|x) = 1 - \exp\{-\exp(\psi_0 + \psi_1 x + \psi_2 x^2)\}$$

M5 (Jump)

$$h(r = 1|x) = \begin{cases} \psi_1, & if\, x \leq 0; \\ \psi_2, & if\, x \geq 0. \end{cases}$$

M6 (Fragment)

$$h(r = 1|x) = \begin{cases} 0, & if\, \psi_1 \leq x \leq 0; \\ \psi_2, & \text{others}. \end{cases}$$

Confounder C is partly missing by a complex missing data mechanism as listed (M1–M6), while three working models are assumed: 1) MCAR; 2) logistic linear model (MAR); 3) non-parametric model (Nonp). The average estimation of $\hat{\theta}$ for 100 replications is shown in the tables below. For each study we generate 1000 samples to have precise parameter estimation and small standard error (se). As seen from tables, fitting from non-parametric methods is often better than the other two models, especially for continuous complex models (in M2, M3 and M4). MCAR assumption does not work well in many cases, and logistic linear model fitting does not work well in M4 and some cases of M6. Model selection or sensitivity analysis is necessary to obtain more accurate results.

Table 3.6: MAR simulation 1

MDM			θ			Incomplete data bias					Coverage Rate			MDM Bias			MP
Model	ψ	corr(x,c)	θ_{gZ}	θ_{gY}	se	EB	IB	MCAR	MAR	Nonp	MCAR	MAR	Nonp	MCAR	MAR	Nonp	(%)
M1	2	ρ= 0	0.598	0.598	0.027	-0.001	0	0	0	0	0.98	0.98	0.98	0	0	0	35.4
		ρ= 0.1	0.600	0.624	0.027	0.024	0.026	0.017	0.021	0.025	0.94	0.96	0.97	-0.008	-0.005	-0.001	35.1
		ρ= 0.3	0.601	0.675	0.026	0.073	0.083	0.058	0.069	0.082	0.86	0.94	0.97	-0.025	-0.014	-0.001	35.1
		ρ= 0.5	0.597	0.736	0.027	0.139	0.158	0.111	0.132	0.156	0.68	0.93	0.83	-0.048	-0.026	-0.002	35.0
	0.5	ρ= 0	0.596	0.601	0.025	0.004	0	0	0	0	0.99	0.99	0.99	0	0	0	27.8
		ρ= 0.1	0.603	0.610	0.025	0.007	0.013	0.012	0.012	0.013	0.96	0.96	0.97	-0.001	-0.001	0.001	28.0
		ρ= 0.3	0.600	0.635	0.025	0.034	0.043	0.039	0.040	0.042	0.96	0.98	0.98	-0.003	-0.002	0.001	27.9
		ρ= 0.5	0.599	0.668	0.025	0.068	0.084	0.076	0.078	0.083	0.95	0.97	0.95	-0.007	-0.005	-0.001	27.8
	-0.5	ρ= 0	0.601	0.601	0.025	-0.001	0	0	0	0	0.99	0.99	0.99	0	0	0	28.0
		ρ= 0.1	0.603	0.612	0.025	0.008	0.013	0.012	0.012	0.013	0.97	0.98	0.97	-0.001	-0.001	0.001	28.1
		ρ= 0.3	0.601	0.634	0.025	0.033	0.042	0.039	0.040	0.042	0.97	0.95	0.95	-0.003	-0.002	-0.001	28
		ρ= 0.5	0.601	0.672	0.025	0.071	0.084	0.077	0.079	0.084	0.91	0.93	0.91	-0.006	-0.004	0.001	28.1
	-2	ρ= 0	0.598	0.599	0.027	0.001	0	0	0	0	0.97	0.97	0.97	0	0	0	35.1
		ρ= 0.1	0.598	0.621	0.027	0.023	0.026	0.018	0.022	0.025	0.96	0.99	0.97	-0.007	-0.004	-0.001	35.1
		ρ= 0.3	0.599	0.672	0.027	0.073	0.083	0.058	0.069	0.082	0.89	0.95	0.94	-0.025	-0.014	-0.001	35.2
		ρ= 0.5	0.595	0.731	0.027	0.136	0.159	0.111	0.133	0.157	0.68	0.94	0.84	-0.047	-0.026	-0.001	35.4
M2	2	ρ= 0	0.600	0.601	0.023	0.001	0	0	0	0	0.99	0.99	0.99	0	0	0	13.2
		ρ= 0.1	0.601	0.603	0.023	0.001	0.001	0.004	0.004	0.001	0.96	0.97	0.96	0.003	0.004	0.001	13.2
		ρ= 0.3	0.596	0.602	0.023	0.006	0.003	0.015	0.015	0.005	0.97	0.94	0.99	0.012	0.012	0.001	13.2
		ρ= 0.5	0.604	0.61	0.023	0.006	0.007	0.032	0.032	0.009	0.86	0.84	0.94	0.024	0.025	0.002	13.3
	0.5	ρ= 0	0.597	0.597	0.024	-0.001	0	0	0	0	1	1	1	0	0	0	20.9
		ρ= 0.1	0.604	0.606	0.024	0.002	0.004	0.007	0.008	0.004	0.97	0.96	0.97	0.003	0.003	0.001	20.8
		ρ= 0.3	0.601	0.615	0.024	0.014	0.015	0.026	0.026	0.016	0.97	0.96	0.97	0.011	0.011	0.001	21.0
		ρ= 0.5	0.598	0.627	0.024	0.029	0.032	0.053	0.054	0.033	0.91	0.78	0.96	0.021	0.022	0.001	20.9
	-0.5	ρ= 0	0.599	0.602	0.028	0.003	0	0	0	0	0.99	1	1	0	0	0	37.7
		ρ= 0.1	0.602	0.627	0.028	0.026	0.032	0.021	0.021	0.031	0.98	0.98	0.96	-0.011	-0.010	-0.001	37.8
		ρ= 0.3	0.600	0.682	0.028	0.082	0.098	0.064	0.066	0.096	0.84	0.87	0.93	-0.034	-0.032	-0.001	37.4
		ρ= 0.5	0.600	0.756	0.029	0.156	0.180	0.119	0.122	0.177	0.50	0.78	0.88	-0.061	-0.058	-0.002	37.8
	-2	ρ= 0	0.599	0.599	0.033	0.001	0	0	0	0	0.93	0.96	0.96	0	0	0	56.5
		ρ= 0.1	0.599	0.657	0.034	0.058	0.069	0.045	0.045	0.066	0.95	0.96	0.96	-0.024	-0.024	-0.003	56.1
		ρ= 0.3	0.603	0.800	0.034	0.197	0.209	0.136	0.136	0.202	0.44	0.56	0.92	-0.073	-0.073	-0.007	56.1
		ρ= 0.5	0.598	0.944	0.035	0.346	0.354	0.232	0.233	0.343	0.03	0.11	0.96	-0.122	-0.121	-0.011	56.4

M1: Log Linear $h(r=1|x) = \text{expit}(1+\psi x)$; M2 : Log Quadrant $h(r=1|x) = \text{expit}(1+0.5x+\psi x^2)$. EB: empirical bias $(\hat{\theta}_{gY} - \theta)$; IB: incomplete data bias $(\hat{\theta}_{gY} - \hat{\theta}gZ)$; MCAR, MAR and Nonp indicate the results based the relevant MDM assumption. corr(x,c) is assumed given.

Table 3.7: MAR simulation 2

MDM			θ			Incomplete data bias					Coverage Rate			MDM Bias			MP
Model	ψ	$\text{corr}(x, c)$	θ_{gZ}	θ_{gY}	se	EB	IB	MCAR	MAR	Nonp	MCAR	MAR	Nonp	MCAR	MAR	Nonp	(%)
M3	2	ρ= 0	0.599	0.600	0.026	0.001	0	0	0	0	0.99	0.98	0.98	0	0	0	25.5
		ρ= 0.1	0.600	0.621	0.026	0.021	0.023	0.012	0.016	0.023	0.95	0.95	0.96	-0.010	-0.007	-0.001	25.4
		ρ= 0.3	0.601	0.662	0.026	0.061	0.075	0.040	0.0508	0.073	0.83	0.91	0.91	-0.034	-0.024	-0.002	25.6
		ρ= 0.5	0.602	0.719	0.026	0.118	0.142	0.077	0.0963	0.139	0.51	0.82	0.86	-0.064	-0.045	-0.003	25.6
	0.5	ρ= 0	0.601	0.599	0.024	-0.001	0	0	0	0	0.98	0.98	0.98	0	0	0	9.78
		ρ= 0.1	0.595	0.602	0.024	0.007	0.006	0.004	0.004	0.006	0.98	0.98	0.98	-0.002	-0.002	-0.001	9.85
		ρ= 0.3	0.602	0.615	0.024	0.013	0.019	0.012	0.0129	0.019	0.97	0.97	0.95	-0.007	-0.006	-0.001	9.71
		ρ= 0.5	0.598	0.627	0.024	0.029	0.038	0.024	0.0262	0.039	0.9	0.91	0.90	-0.014	-0.012	0.001	9.93
	-0.5	ρ= 0	0.600	0.600	0.024	-0.001	0	0	0	0	0.98	0.98	0.98	0	0	0	9.88
		ρ= 0.1	0.596	0.602	0.024	0.005	0.006	0.004	0.004	0.006	0.98	0.98	0.98	-0.002	-0.002	-0.001	9.79
		ρ= 0.3	0.601	0.617	0.024	0.016	0.019	0.012	0.013	0.0193	0.94	0.94	0.96	-0.007	-0.006	-0.001	9.65
		ρ= 0.5	0.602	0.632	0.024	0.030	0.039	0.0242	0.026	0.039	0.92	0.96	0.96	-0.015	-0.014	-0.001	9.90
	-2	ρ= 0	0.603	0.606	0.026	0.002	0	0	0	0	0.96	0.96	0.96	0	0	0	25.5
		ρ= 0.1	0.601	0.622	0.027	0.021	0.024	0.013	0.016	0.023	0.96	0.98	0.99	-0.011	-0.007	-0.001	25.6
		ρ= 0.3	0.600	0.664	0.028	0.064	0.075	0.041	0.051	0.074	0.80	0.97	0.94	-0.035	-0.025	-0.002	25.3
		ρ= 0.5	0.601	0.718	0.027	0.119	0.142	0.077	0.095	0.139	0.48	0.86	0.87	-0.066	-0.047	-0.003	25.5
M4	2	ρ= 0	0.601	0.600	0.023	-0.001	0	0	0	0	0.99	0.99	0.99	0	0	0	1.93
		ρ= 0.1	0.598	0.598	0.022	0.001	0.001	0.001	0.001	0.001	0.94	0.94	0.94	0.001	0.001	0.001	1.95
		ρ= 0.3	0.597	0.595	0.022	-0.002	0.001	0.002	0.002	0.001	0.99	0.99	0.99	0.002	0.002	0.001	1.90
		ρ= 0.5	0.597	0.599	0.022	0.002	0.001	0.004	0.004	0.001	0.95	0.93	0.96	0.004	0.004	0.001	1.91
	0.5	ρ= 0	0.601	0.603	0.022	0.002	0	0	0	0	0.96	0.96	0.96	0	0	0	4.17
		ρ= 0.1	0.603	0.602	0.022	-0.001	0.001	0.001	0.001	0.001	0.97	0.98	0.97	0.001	0.001	0.001	4.19
		ρ= 0.3	0.595	0.600	0.022	0.005	0.002	0.004	0.005	0.002	0.95	0.95	0.95	0.003	0.003	0.001	4.20
		ρ= 0.5	0.601	0.604	0.022	0.003	0.004	0.009	0.009	0.004	0.94	0.94	0.94	0.006	0.006	0.001	4.19
	-0.5	ρ= 0	0.600	0.600	0.027	0	0	0	0	0	0.98	0.98	0.98	0	0	0	20.7
		ρ= 0.1	0.598	0.620	0.027	0.022	0.023	0.010	0.011	0.023	0.98	0.98	0.96	-0.013	-0.012	-0.001	20.6
		ρ= 0.3	0.602	0.663	0.027	0.061	0.072	0.032	0.034	0.071	0.78	0.84	0.95	-0.040	-0.038	-0.002	20.6
		ρ= 0.5	0.598	0.708	0.027	0.110	0.134	0.059	0.063	0.131	0.45	0.56	0.86	-0.075	-0.071	-0.003	20.6
	-2	ρ= 0	0.600	0.602	0.034	0.001	0	0	0	0	0.95	0.95	0.95	0	0	0	44.9
		ρ= 0.1	0.595	0.660	0.034	0.065	0.064	0.034	0.035	0.062	0.90	0.90	0.96	-0.030	-0.029	-0.003	44.9
		ρ= 0.3	0.598	0.773	0.034	0.174	0.195	0.104	0.104	0.186	0.41	0.51	0.93	-0.091	-0.090	-0.009	44.7
		ρ= 0.5	0.595	0.913	0.034	0.319	0.331	0.176	0.176	0.315	0	0	0.99	-0.154	-0.154	-0.014	44.5

M3: Log Log Linear $h(r = 1|x) = 1 - \exp\{-\exp(1 + \psi x)\}$; M4 : Log Log Quadrant $h(r = 1|x) = 1 - \exp\{-\exp(1 + 0.5x + \psi x^2)\}$.

Table 3.8: MAR simulation 3

MDM			θ			Incomplete data bias					Coverage Rate			MDM Bias			MP
Model	ψ	corr(x, c)	θ_{gZ}	θ_{gY}	se	EB	IB	MCAR	MAR	Nonp	MCAR	MAR	Nonp	MCAR	MAR	Nonp	(%)
M5	0	ρ= 0	0.600	0.599	0.029	-0.001	0	0	0	0	0.96	0.96	0.96	0	0	0	60.2
		ρ= 0.1	0.598	0.642	0.029	0.043	0.046	0.039	0.047	0.046	0.96	0.94	0.95	-0.006	0.001	0.001	60.2
		ρ= 0.3	0.601	0.734	0.029	0.134	0.146	0.124	0.148	0.148	0.81	0.88	0.88	-0.022	0.001	0.002	59.8
		ρ= 0.5	0.597	0.838	0.029	0.241	0.283	0.239	0.283	0.288	0.76	0.66	0.59	-0.044	-0.000	0.004	60.0
	0.2	ρ= 0	0.597	0.599	0.027	0.002	0	0	0	0	0.97	0.96	0.96	0	0	0	50.0
		ρ= 0.1	0.598	0.627	0.027	0.029	0.030	0.027	0.030	0.030	0.94	0.95	0.95	-0.003	0.001	0.001	49.9
		ρ= 0.3	0.601	0.686	0.027	0.084	0.097	0.087	0.097	0.098	0.95	0.91	0.91	-0.009	0.001	0.001	50.0
		ρ= 0.5	0.593	0.759	0.027	0.166	0.187	0.169	0.187	0.188	0.88	0.8	0.85	-0.018	-0.001	0.001	50.0
	0.4	ρ= 0	0.599	0.599	0.026	0.001	0	0	0	0	0.95	0.95	0.95	0	0	0	40.0
		ρ= 0.1	0.602	0.616	0.026	0.013	0.020	0.019	0.020	0.020	0.96	0.96	0.97	-0.001	0.001	0.001	40.0
		ρ= 0.3	0.598	0.652	0.026	0.053	0.065	0.062	0.065	0.066	0.96	0.94	0.93	-0.003	-0.001	0.001	40.1
		ρ= 0.5	0.599	0.709	0.026	0.112	0.127	0.121	0.128	0.128	0.87	0.9	0.87	-0.006	0.001	0.001	39.9
	0.6	ρ= 0	0.607	0.604	0.025	-0.003	0	0	0	0	0.98	0.98	0.98	0	0	0	29.9
		ρ= 0.1	0.600	0.613	0.025	0.013	0.013	0.012	0.013	0.013	0.98	0.98	0.99	-0.001	-0.001	0.001	30.1
		ρ= 0.3	0.603	0.639	0.025	0.035	0.042	0.041	0.042	0.041	0.94	0.97	0.97	-0.001	0.001	-0.005	29.9
		ρ= 0.5	0.599	0.671	0.025	0.072	0.083	0.082	0.083	0.083	0.93	0.94	0.91	-0.001	0.001	0.001	30.0
M6	-3	ρ= 0	0.599	0.603	0.028	0.004	0	0	0	0	0.97	0.97	0.97	0	0	0	55.0
		ρ= 0.1	0.595	0.638	0.028	0.042	0.041	0.033	0.041	0.0417	0.92	0.95	0.95	-0.007	0.001	0.001	55.0
		ρ= 0.3	0.599	0.718	0.028	0.119	0.132	0.108	0.134	0.133	0.91	0.88	0.87	-0.023	0.002	0.001	54.8
		ρ= 0.5	0.604	0.817	0.028	0.212	0.262	0.215	0.266	0.265	0.82	0.54	0.60	-0.047	0.004	0.003	54.9
	-2.4	ρ= 0	0.602	0.597	0.027	-0.004	0	0	0	0	0.95	0.95	0.95	0	0	0	54.4
		ρ= 0.1	0.596	0.633	0.027	0.035	0.035	0.031	0.038	0.035	0.93	0.96	0.95	-0.003	0.003	0.001	54.2
		ρ= 0.3	0.600	0.703	0.027	0.103	0.114	0.102	0.124	0.115	0.87	0.87	0.90	-0.012	0.009	0.001	54.3
		ρ= 0.5	0.599	0.788	0.027	0.188	0.234	0.207	0.252	0.234	0.94	0.43	0.58	-0.026	0.018	0.001	54.3
	-1.8	ρ= 0	0.601	0.602	0.026	0.001	0	0	0	0	0.96	0.96	0.96	0	0	0	51.8
		ρ= 0.1	0.597	0.621	0.026	0.023	0.023	0.026	0.030	0.024	0.97	0.96	0.98	0.002	0.006	0.001	51.9
		ρ= 0.3	0.599	0.668	0.026	0.069	0.078	0.086	0.101	0.079	0.97	0.76	0.95	0.008	0.022	0.001	51.7
		ρ= 0.5	0.598	0.730	0.026	0.131	0.163	0.18	0.209	0.166	0.88	0.22	0.72	0.017	0.046	0.003	51.8
	-1.2	ρ= 0	0.596	0.598	0.024	0.002	0	0	0	0	1	1	1	0	0	0	44.5
		ρ= 0.1	0.602	0.608	0.024	0.008	0.011	0.019	0.020	0.012	0.97	0.95	0.99	0.007	0.009	0.001	44.8
		ρ= 0.3	0.598	0.632	0.024	0.034	0.037	0.062	0.066	0.041	0.91	0.72	0.95	0.024	0.029	0.004	44.2
		ρ= 0.5	0.599	0.671	0.024	0.072	0.083	0.138	0.147	0.092	0.58	0.13	0.86	0.054	0.064	0.009	44.7

M5: Jump function $h(r = 1|x) = \psi(x \leq 0) + 0.8(x > 0)$; M6 : Fragment model $h(r = 1|x) = 0.9(x \leq \psi) + 0.9(x > 0)$.

Chapter 4

Monte Carlo Sensitivity Analysis and Selection of Bias Models

4.1 Introduction

Local sensitivity analysis for model uncertainty problem with missing data was addressed by incomplete data bias analysis. Unfortunately, it is often difficult to calculate the bias parameters in experimental trials due to lack of knowledge. The examples we previously discussed show that the evaluation of bias parameters such as $\operatorname{corr}(x, c)$ can be a realistic problem for ignorable missing data. Also if the missingness is under so-called non-ignorable missing data mechanism (Little and Rubin, 2002), bias model may suffer from identification problem and valid inference is restricted at the stage of limited observed information; see Chapter 5 and Chapter 6 for details. In some special cases we may utilize a follow-up study to estimate those bias parameters (see e.g. Kim and Yu, 2011). However, it may be difficult to conduct further investigation in most of the cases, for example epidemiology designs, and also extra bias may result from lack of randomization and independence between former observations and follow up samples.

Sensitivity analysis is one of commonly used approaches in assessing uncertainty via bias parameter η or the related bias model. Conventional sensitivity analysis considers the range of all the plausible results, while a Monte Carlo sensitivity analysis sample the bias parameters from a prior distribution and then inverts the bias model to provide a distribution of bias-corrected estimates (Greenland, 2005, p.269). We follow the idea, but instead of obtaining a 'posterior' average for the parameter of interest based on a prior density which is usually difficult to justify, we attempt to select one bias model based on the nearest neighbour distance between the observed data and the data simulated from bias models. This model can be treated as the most plausible model in all the considered bias models.

The rest of the chapter is organized as follows. Section 4.2 will first describe briefly the idea of Monte Carlo sensitivity analysis, followed by a detailed discussion of our proposed method for bias model selection. The method will be applied to several missing data problems in Section 4.3 and Chapter 6 and will be illustrated by numerical results of simulation studies and real data problems.

4.2 Monte Carlo Sensitivity Analysis and Bias Model Selection

4.2.1 Sensitivity Analysis and Bias Parameter

Let $D = (D_{obs}, D_{mis})$ be a set of complete data including observed and unobserved data and R be missingness indicator vector which takes the value 1 if data observed or 0 otherwise. The complete data model can be factorized into an extrapolation model and an observed data model as follows.

$$f(D, R|\theta) = f(D_{mis}|D_{obs}, R, \theta_{mis})f(D_{obs}, R|\theta_{obs}). \tag{4.1}$$

Here, θ_{mis} and θ_{obs} denote parameters indexing the extrapolation and observed data models respectively. The observed data distribution $f(D_{obs}, R|\theta_{obs})$ is identifiable and can be fitted by a parametric or nonparametric model. However, the extrapolation distribution $f(D_{mis}|D_{obs}, R, \theta_{mis})$ cannot be identified unless extra assumptions are made. Those parameters are described as *sensitivity parameters* or *bias parameters* (Daniels and Hogan, 2008; Greenland, 2005), denoted by η. The following are some features: (i) η is a function of the parameter θ_{mis}; (ii) fit of the model to the observed data $f(D_{obs}, R|\theta_{obs})$ is independent from the bias parameter; and (iii) when the bias parameter is fixed, the full data model $f(D, R|\theta)$ is identified. One way to identify the bias parameter under selection model frame is to use the following equation:

$$f(D_{mis}|D_{obs}, R, \theta_{mis}) \propto f(R|D_{mis}, D_{obs}, \psi) f(D_{mis}|D_{obs}, \theta_*),$$

where ψ is the parameter describing missing data mechanism and θ_* is the parameter of conditional distribution of missing variables given observed variables. Thus the bias parameter η is a function of (ψ, θ_*). We need to bear in mind that D_{mis} component is unobservable and therefore at least part of (ψ, θ_*) are inestimable under non-ignorable missingness. In some cases, for example the models discussed in Section 6.3, we need to consider ψ only.

Local sensitivity analysis is based on derivatives of θ, the parameters of interest, evaluated at some belief $\eta = \eta_0$ where the model with η_0 is usually the practical model used in inference. This method indicates how the estimate of θ changes corresponding to the input values of η which are allowed to be perturbed in a neighborhood of η_0. This helps to understand the robustness of the practical model in a local area but has limited value in understanding the consequences of global uncertainty about η. In contrast, the global sensitivity analysis considers those more substantial changes individually without limitation based on its sample space (see e.g. Oakley and O'Hagan, 2004) although an unrealistically wide range is frequently a problem. Bayesian techniques in some sense partly overcome the difficulty (see e.g. McCandless et al., 2007, 2008; Gustafson et al., 2010), offering a route to sample smoothly via a prior distribution which weights possible scenarios rather than the traditional method which only reflects the investigator's plausible beliefs.

Monte Carlo Sensitivity Analysis (MCSA) is a type of Bayesian sensitivity analysis with modifications. Assuming that $f(\theta|\eta)$ is uniformly distributed and posterior distribution $f(\eta|D_{obs}, R)$ is close to the prior distribution $f(\eta)$, the MCSA procedure is to sample from

$$f(\theta|D_{obs}, R) = \int f(\theta|D_{obs}, R, \eta) f(\eta|D_{obs}, R) d\eta \approx \int f(\theta|D_{obs}, R, \eta) f(\eta) d\eta.$$

The details can be found in Greenland (2005), and the problems caused by replacing the posterior of the bias parameters by its prior are discussed by Daniels and Hogan (2008). Bayesian sensitivity analysis however relies on the prior distribution of the bias parameters and the hierarchical bias model. The posterior average may be asymptotically biased and credible intervals may not have expected coverage probability due to possible wrong prior choice, according to Gustafson (2005). Monte Carlo sensitivity analysis may cause extra bias due to an incorrect sampling distribution. We propose a novel method in the next subsection by combining the idea of traditional and Bayesian approaches and by focusing on the influence of each individual η and then select the most plausible value from all possible values.

4.2.2 Bias Model Selection

Let F be a population of the complete data, and we wish to infer the parameter of interest θ using model $L(F;\theta)$. An experimental design sample D is drawn randomly from F, and $\hat{\theta}$ calculated from the model $L(D;\theta)$ is usually unbiased without missing data and model misspecification. However, observed data, denoted by D_{obs}, often conceal some values under a certain missing data mechanism. Conventional inference employs a model $L(D_{obs};\theta)$ under assumptions such as identification of the model or missing at random (MAR). Those assumptions are often invalid under some 'imperfect' situations such as missing confounders or measurement errors with non-ignorable missingness. This results in bias. The effect of biased sources on L may be modelled by a bias model via a bias parameter η. For missing data problem, η is a function of (ψ, θ_*) as discussed in the previous subsection. Once η is given, the evaluation is available by the model $L(D_{obs};\theta,\eta)$. For example, we may use multiple imputation by resampling the missing values by their conditional distributions given observed variables and η; the generated data is denoted by $D_{mis,\eta}$. Thus the imputed dataset $D_\eta = (D_{obs}, D_{mis,\eta})$ is complete and inference can be carried out in the usual way for complete data.

Assume that, given η_{true}, we can get unbiased estimation from the corresponding model $L(D_{obs};\theta,\eta_{true})$. Unfortunately, η_{true} is usually unknown and it cannot be estimated from D_{obs} under some conditions, e.g. non-ignorable missingness. In MCSA (Monte Carlo sensitivity analysis) a prior distribution is assumed, $\eta \sim f(\eta),\ \eta \in \mathit{\Gamma}$. Inference is based on the average of the marginal posterior of bias model $L(D_{obs};\theta,\eta)$ on its prior density:

$$\begin{aligned} f(\theta|D_{obs}) &= \int f(\theta|D_{obs},\eta) f(\eta|D_{obs}) d\eta \\ &\propto \int L(D_{obs};\theta,\eta) f(\theta,\eta) d\eta. \end{aligned}$$

An incorrect choice of this prior distribution may, however, lead to extra bias. The method we proposed below will avoid this problem. Instead of using posterior average, we attempt to find one η or a small set of η's which may be close to η_{true}. We call this or these values as 'most plausible' value(s).

$$\begin{tikzcd}
F \arrow[r, "\text{sample}"] \arrow[d, "L"] & D \arrow[r, "\text{conceal}"] & D_{obs} \arrow[r, "\eta", "\text{conditional resample}"'] \arrow[d] & D_{\eta} \arrow[r, "\text{conceal}"] \arrow[d, "L"] & D_{\eta,\, obs} \arrow[r, "\text{measure}"] & s(D_{\eta,\, obs}, D_{obs}) \arrow[dll] \\
\theta & & \hat{\theta} & \hat{\theta}_{\eta} & &
\end{tikzcd}$$

Figure 4.1: Diagram of bias model

For any given η, let $\hat{\theta}_\eta$ be the estimate of θ obtained from $L(D_{obs}; \theta, \eta)$, for example the maximum likelihood estimate by maximizing its marginal likelihood or by using the imputed method as discussed before. Consider a series of possible $\eta \in \mathit{\Gamma}$ to return a set of estimates:

$$\{\hat{\theta}_\eta : L(D_{obs}; \theta, \eta),\ \eta \in \mathit{\Gamma}\}, \tag{4.2}$$

where η can be for example generated from its prior distribution. For each of given η's, $\hat{\theta}_\eta$ is calculated and the bias resource is described by $L(D_{obs}; \theta, \eta)$. We can, therefore, generate a 'complete' data set of D_η. Using the MDM specified given η, an 'incomplete' data set $D_{\eta,obs}$ can be simulated from the 'complete' data set D_η. If the value of η is close to η_{true}, D_η and D would come from the same population distribution; so do the simulated 'incomplete' data set $D_{\eta,obs}$ and the raw 'incomplete' data D_{obs}. We then define a distance $s(D_{\eta,obs}, D_{obs})$ to measure the 'closeness' or 'similarity' between them. The model with the smallest distance can be selected as the most plausible model. We call this method as Monte Carlo bias model selection (MC-BMS or BMS). Its procedure is described as follows (also see Figure 4.1).

(i) Select one η in $\mathit{\Gamma}$, or generate it from a prior distribution $f(\eta)$ if we have prior knowledge about η;

(ii) Estimate $\hat{\theta}_\eta$ using bias model $L(D_{obs}; \theta, \eta)$ given η;

(iii) Simulate a complete dataset D_η from $L(D_{obs}; \hat{\theta}_\eta, \eta)$ given $\hat{\theta}_\eta$ and η and censor the simulated sample D_η into an incomplete dataset $D_{\eta,obs}$ using the MDM model specified by η ($D_{\eta,obs}$ is comparable with D_{obs});

(iv) Calculate distance $s(D_{\eta,obs}, D_{obs})$;

(v) Repeat Steps (i) to (iv) for a set of η and select the one with the smallest distance or select a small set of η if the distance is very close to the smallest one for each of them.

In Steps (i) and (ii) we use methods of conventional sensitivity analysis and calculate a series of estimation $\hat{\theta}_\eta$ for a set of η. This can be used to investigate how the estimation changes along η where η is usually associated with an interpretable quantity for example partial correlation between an observed covariate and a missing confounder (see e.g. Lin et al., 2012). Conclusions can be made based on prior knowledge or historic data for the interpretable quantity. The MCSA method needs to select a prior distribution and generate random numbers from the selected prior distribution $f(\eta)$. An overall estimate of θ is calculated via Bayesian average.

In Step (iii) we first sample D_η from its distribution conditional on the observed data, given bias parameter η and the corresponding estimation of $\hat{\theta}_\eta$. We further censor D_η into an incomplete dataset $D_{\eta,obs}$, which is comparable with D_{obs}. This requires a missing data mechanism (MDM) model which may depend on bias parameter η. We may use a parametric model such as a logistic linear model or a semiparametric model as we will use in chapter 6. There is no unified method on how to simulate D_η or $D_{\eta,obs}$. Specific technique is upon individual problem; see more discussion for specific examples given in Section 4.3.

The last two steps are to calculate the distance between simulated data set and the observed data set and to select the most plausible bias model or a small set of the most plausible models. The key here is which distance should be used to measure the 'closeness' or 'similarity' between datasets $D_{\eta,obs}$ and D_{obs}. This is particularly important for large-dimensional cases. To measure similarity or dissimilarity of two clusters, various statistical distances are available to be considered. We may calculate the distance for each pair of data points in $D_{\eta,obs}$ and D_{obs}, and then use the minimum distance (single linkage by Sneath (1957)), maximum distance (complete linkage by Sorensen (1948)) or the average distance (Sokal and Michener, 1958). An alternative method is to use the K-nearest neighbour (KNN) method, which was first introduced by Fix and Hodges (1951) as a nonparametric density measure. This measure works well in most of the examples. The detailed description is given in Appendix 4.6.1.

Remark 1. Bias model $L(D_{obs}; \theta, \eta)$ depends on the bias parameter η and it depends on the hierarchical structure as well, as discussed around equation (4.1); so do $D_{\eta,obs}$.

When we compare the models by using the distance between $D_{\eta,obs}$ and D_{obs}, we actually consider both the bias model structure and the value of bias parameter. This will be further illustrated in the next sections.

Remark 2. In Steps (iv) and (v), it may be numerically unstable if we compare bias models based on the distance between D_{obs} and one set of $D_{\eta,obs}$. One way to solve this problem is to use an average distance by sampling more than one set of $D_{\eta,obs}$ for the same η.

4.2.3 Hypothesis Test for η

D We considered some dissimilarity measures $s(D_{obs}, D_{\eta,obs})$ in Section 4.2.2, and they can also be used as test statistics. We expect that if the H_0^* is not true, the value of distance $s(D_{obs}, D_{\eta,obs})$ will be larger than if H_0^* is true. The achieved significance level (ASL) in the permutation test (Fisher, 1971) is defined as the probability of observing a larger value s^* when the null hypothesis is true

$$ASL = \Pr_{H_0^*}\{s^* \geq s\}.$$

We can also calculate the critical value (denote as s_α) at certain significance level α based on the estimators s^* from permutation samples.

Before we show the examples, we introduce some rules for the bias selection procedure.

Plausible Set Rule. For any $\eta_i \in \mathit{\Gamma}$, if

$$\Pr_{H_0^*}\{s(D_{obs}, D_{\eta_i,obs}) < s_\alpha\} < \alpha$$

then η_i will be rejected. In practice, we may choose $\alpha = 0.05$.

Bias Model Selection Option 1 (BMS-1). Plausible set of bias parameters is given as:

$$\mathit{\Gamma}_\alpha = \{\eta_i : s(D_{obs}, D_{\eta_i,obs}) < s_\alpha, \eta_i \in \mathit{\Gamma}\}.$$

All the values of η_i excluding in $\mathit{\Gamma}_\alpha$ will not be chosen according to CI Rule.

Bias Model Selection Option 2 (BMS-2). If we are interested in obtaining one selection of η, it is reasonable to choose the value with largest ASL (smallest distance) since the smaller the ASL, the stronger the evidence against H_0^*.

$$\tilde{\eta} = \arg_\eta \min\{s(D_{obs}, D_{\eta,obs})\},$$

$$\tilde{\theta} = \arg_\theta \max\{L(D_{obs}; \theta, \tilde{\eta})\}.$$

Remark 3. The plausible subset $\mathit{\Gamma}_\alpha$ may be used in Bayesian sensitivity analysis, where the posterior distribution $f(\eta|D_{obs})$ can be approximately calculated as

$$f(\eta|D_{obs}) = \frac{\Pr_{H_0}(\eta = \eta_i|D_{obs})}{\sum_{\eta_i \in \mathit{\Gamma}} \Pr_{H_0}(\eta = \eta_i|D_{obs})}$$

where $\mathit{\Gamma}$ should be replaced by $\mathit{\Gamma}_\alpha$, and a prior $p(\eta)$ is defined on $\mathit{\Gamma}_\alpha$. And for each η_i, the probability $\Pr_{H_0}(\eta = \eta_i|D_{obs})$ is proportional to $p(\eta_i)f(D_{obs}|\eta_i)$, denoted as w_i. And the parameter θ is estimated as

$$\tilde{\theta} = \sum_i w_i \hat{\theta}_{\eta_i} / \sum_i w_i. \tag{4.3}$$

This method actually improves the MCSA by Greenland (2005), and we call it Monte Carlo sensitivity analysis with Bayesian model average (MC-BMA or BMA).

BMS Method-1 evaluates a plausible set $\mathit{\Gamma}_\alpha \subset \mathit{\Gamma}$ at a certain significance level (e.g. 5%). BMS Method-2 concerns the ‘maximum likely’ one from all plausible values and this method performs efficiently since the hypothesis test is not required.

4.3 Numerical Result

4.3.1 An Example

We first apply the Monte Carlo bias model selection method (MC-BMS) to ignorable missing data problems. We used the local bias analysis method to address the model uncertainty in missing covariate problems in previous chapters and calculated the incomplete data bias for the effect size estimation. Under a linear regression model

$t|(x,c) \sim N(\alpha + \theta x + \beta c, \sigma^2)$, the incomplete data bias is given as

$$\boldsymbol{b_{XC}} \propto \text{corr}(x,c)\boldsymbol{I_Y}^{-1}(E(h_x'), E(h_x), 0)^T. \tag{4.4}$$

according to Section 2.4. The covariate correlation corr(x,c) generates and governs the incomplete data bias, and it is usually difficult to measure in practice.

Now we treat it as a bias parameter and we use MC-BMS to estimate ρ=corr(x,c). Covariate x is assumed as normal distributed $x \sim N(5,1)$ and conditional distribution $(c|x)$ follow uniform distribution $U(1+\rho\frac{\sigma_c}{\sigma_x}x, 3+\rho\frac{\sigma_c}{\sigma_x}x)$. The variance of c is σ_c^2 which is assumed as 0.444 and the covariate correlation is corr(x,c)=0.5. Here c is partly missing with probability $h_x = 1 - \text{expit}\{1 + (x - \bar{x}) - 2(x - \bar{x})^2\}$ with $\bar{x}$=E(x). True value is (α, θ, β)=(0.5,1,1). The sample size is 100.

In this study, we perform a Monte Carlo sensitivity analysis first to discover the uncertainty of output from the input: ρ. We sample under a series of scenarios, specifically, ρ = (-1,-0.99,-0.98,. . . , 1). Given each ρ, we use local bias analysis to adjust the parameter estimations.

Next we conduct a bias model selection process, to select the best value of ρ from all the plausible scenarios by comparing the observed data and simulated data. To do this, we borrow the idea of bootstrapping residuals method to simulate the incomplete data. We first sample the missing values of c by a conditional model $f(c|x,\rho)$ to obtain the imputed c^*. Then the response variable t is bootstrapped by adding the residuals $\epsilon^* \sim N(0, \hat{\sigma}^2)$:

$$t^* = \hat{\alpha} + \hat{\theta}x + \hat{\beta}c^* + \epsilon^*.$$

We then calculate the distance between cluster $D_\rho^* = (t^*, x)$ and cluster $D^* = (t, x)$, and the distance measures introduced in Section 4.2.2 are used. As pointed out in Remark 2, we repeat the procedure several times and use the average distance to reduce sampling errors. We call the repeated times as Monte Carlo sample size, or MC size. Here MC size is 10.

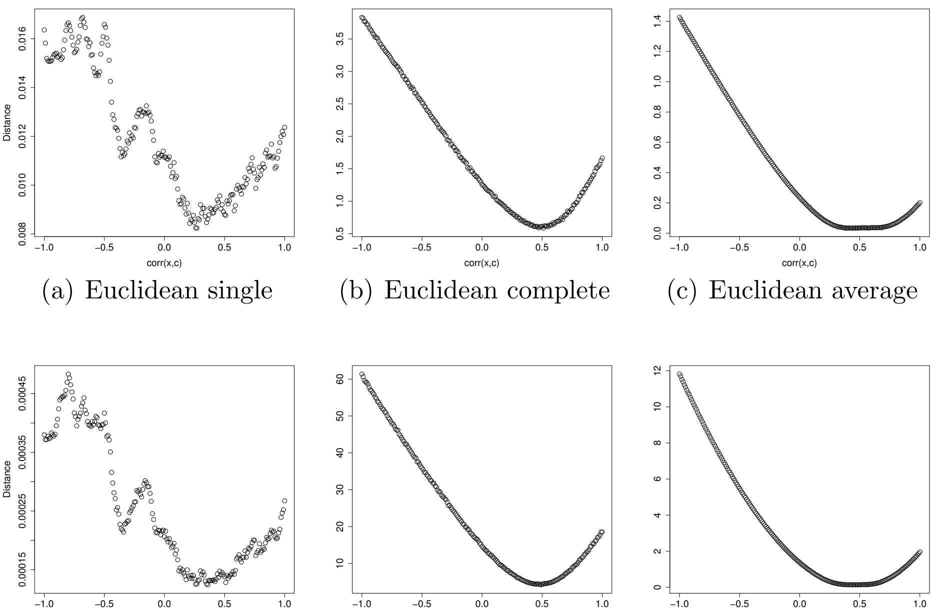
We present the bias parameter ρ against the corresponding average measurements for 100 repeated studies $s(D^*, D_\rho^*)$ at Figures 4.2 and 4.3 and show the evaluations of $\tilde{\rho}$ in Table 4.1.

- Single linkage distance is not a suitable measure for bias model selection. The shortest distance of the clusters seems too sensitive to edge effect.

- Complete linkage distance has a better trend than single distance, but we may find relatively large validation during the replications.

- Selections under average distance give relatively robust results for both Euclidean and Mahalanobis metric as show in figure 4.2. There is clearly a concave shaped trend with a bottom around true value 0.5, and the confidence intervals are relatively narrow.

- Results of KNN measure under Euclidean and Mahalanobis metric (presented in figure 4.3) are robust for all cases with parameter K chosen as 2,3,5 respectively. KNN method is described as a nonparametric method, and thus not restricted by a certain statistical density.



(a) Euclidean single (b) Euclidean complete (c) Euclidean average

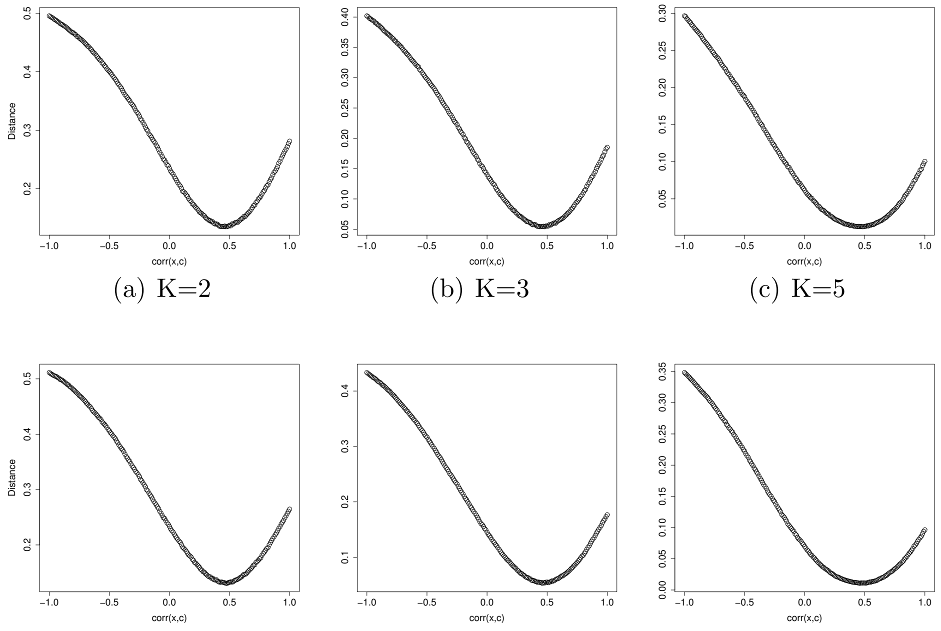
(d) Mahalanobis single (e) Mahalanobis complete (f) Mahalanobis average

Figure 4.2: Selection of corr(x, c) under hierarchical measure.

Now we test the null hypothesis, H_0^*, of no difference between observed cluster D_{obs} and test cluster $D_{\rho,obs}$.

$$H_0^* : D_{obs} = D_{\rho,obs}$$

As was shown in the above figures, the average dissimilarity and KNN measure are during the best criteria, they are then used as the test statistics in the Fisher permutation test, which is introduced in the first chapter. The achieved significance level



(a) K=2 (b) K=3 (c) K=5

(d) K=2 (e) K=3 (f) K=5

Figure 4.3: Selection of corr(x, c) under KNN measure. Upper panel use Euclidean metric and lower panel use Mahalanobis metric.

(ASL) in the permutation test is calculated as the probability of observing a larger value s^* when the null hypothesis is true

$$ASL = \Pr_{H_0}\{s^* \geq s\}.$$

To use the permutation test, we combine the samples first, denote $D_A = (D_{obs}, D_{\rho,obs})$. We resample $n = 100$ data from D_A as the first cluster $D_1^{(i)}$ and the rest as second cluster $D_2^{(i)}$, for $i = 1, \dots B$ and B is the permutation time. We choose $B = 1000$. Then we calculate the values of $\{s^* : s_i^* = s(D_1^{(i)}, D_2^{(i)})\}$ under the proposed distance measure (i.e. Euclidean average distance). The ASL is estimated by

$$\widehat{ASL}_{perm} = \#\{s^* \geq s\}/B.$$

where $\#\{.\}$ denotes the cardinality of the set, and s is the plausible distances $s(D_{obs}, D_{\rho,obs})$ given a fixed ρ in this case. If we choose the significant level at 5%, then the critical value (denoted as s_α) can be calculated from the resampled data, i.e. $s_\alpha = \text{Quantile}_{0.95}(s^*)$. So the interval of accepted ρ at 5% significant level (denoted as $\Gamma_\alpha \subset \Gamma$) is considered as collection of possible values for bias parameter. For example, if we use the Euclidean average distance as the test statistics, the average acceptable corr(x, c) through 100 replications is (0.207,0.831). And the plausible set for θ estimation is given as (0.78, 1.13). Prior is defined in Γ_α, i.e. uniform distribution. BMS and BMA results are shown in Table 4.1, and these two methods work quite similarly.

Table 4.1: Sensitivity analysis results of selecting corr(x, c)

	E-S	E-C	E-A	E-KNN
BMS-1	(-0.458, 0.944)	(-0.144, 0.856)	(0.207, 0.831)	(0.170, 0.733)
BMS-2	-0.004	0.444	0.541	0.457
BMA	0.065	0.421	0.528	0.453
	M-S	M-C	M-A	M-KNN
BMS-1	(-0.424, 0.905)	(0.150, 0.752)	(0.256, 0.685)	(0.208 0.726)
BMS-2	-0.001	0.454	0.468	0.452
BMA	0.0479	0.465	0.469	0.464

E: Euclidean metric; M: Mahalanobis metric. S: Single distance; C: Complete distance; A: Average distance.

Except ASL, the power of the test at 5% significant level can be further calculated by Monte Carlo methods, which is shown in Figure 4.4, with the grey dashed lined indicating 80% power ratio.

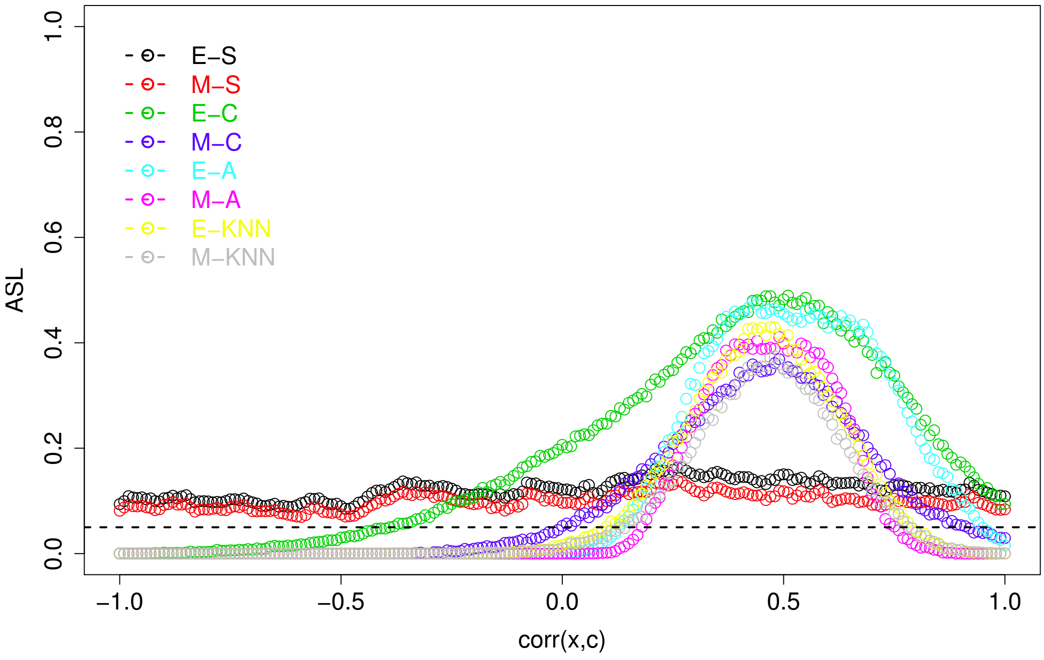
- Average distance and KNN distance work well to have relatively narrow confidence interval under both Euclidean and Mahalanobis metric. Mahalanobis average (MA) distance is the best in this example.

- KNN distance under Euclidean metric (E-KNN) seems to be the most sensitive measure to distinguish different ρ, as seen from ASL plot (Figure 4.4) that it has the apparent peak.

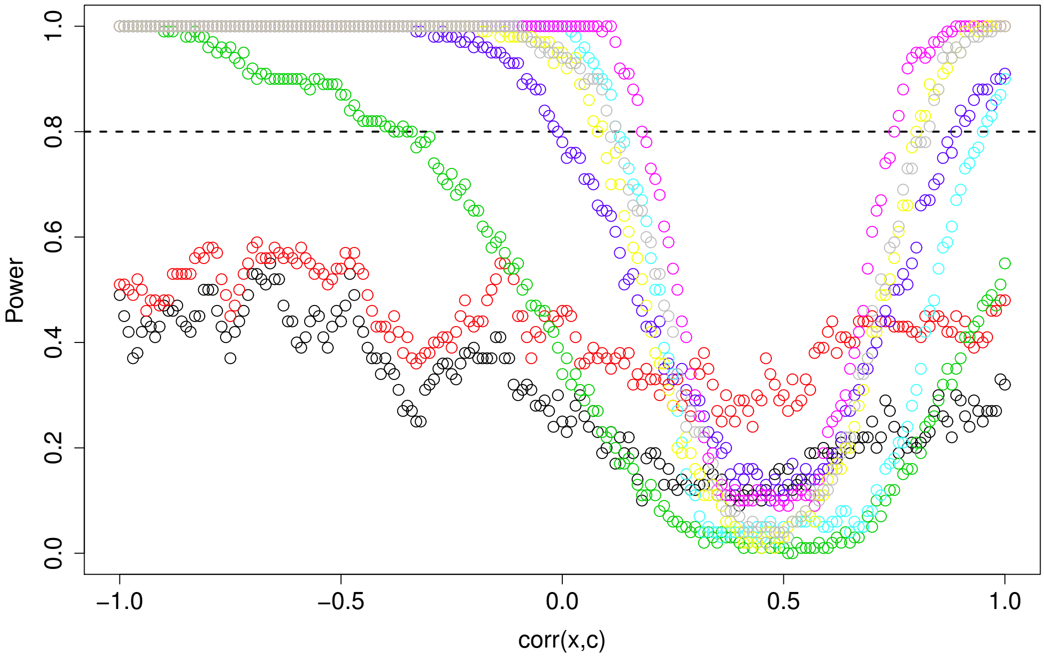
- The power of test at points excluded from Γ is relative large (over 80%), and the smallest power ratio is located around true value.

4.3.2 More Simulation Studies

More simulation studies are conducted aiming to examine the performance of MC-BMS under the average distances and KNN methods. The settings are the same



(a) ASL



(b) Power

Figure 4.4: Achieved significance level and power of test

with Section 4.3.1, but variable c is assumed to censor under different missing data mechanisms:

- MCAR: $h(r=1)$=expit(1)
- Logit Linear (LL): $h(r=1|x) = \text{expit}(1-(x-\bar{x}))$
- Logit Quadratic (LQ): $h(r=1|x) = \text{expit}(1+(x-\bar{x})-(x-\bar{x})^2)$.

and covariate correlation ρ=corr(x,c) is allowed to vary between (0.1, 0.2, …, 0.7). We use three working models to fit the missing data mechanism: 1) MCAR 2) parametric model (logistic linear) and 3) nonparametric modelling; and Bayesian Information Criteria (BIC) is used to make MDM models selection. We repeat 100 times with each sample size equaling 100. Simply, ρ is considered as the single bias parameter which is then measured by MC-BMS method, and we sample a number of values from its sample space: $\rho = (-0.5, -0.4, \ldots, 1)$. The estimator ($\tilde{\rho}$) is calculated with MC-BMS and MC-BMA methods, and the average of the 100 replications and its root mean square error (RMSE) are shown in Table 4.2. It is found that both BMS and BMA method perform very well for most cases and it is robust under all the four distance measures. Inference is assumed to be efficient locally, but we found the results are also robust for large values of ρ.

4.4 Monte Carlo Sensitivity Analysis for Fuel Consumption Data

We apply the MC-BMS method into the Fuel consumption data example to address the uncertainty analysis with missing data issue discussed in Section 3.4. As we pointed previously, evaluation of the incomplete data bias

$$\text{Bias}(\hat{\theta}_{gY}) = \theta_2 \frac{\boldsymbol{I_Y}^{-1}}{\sigma_Y^2} E\{x_2 \boldsymbol{v_0} h(r=0|x_1)\}$$

requires the specification of the distribution $f(x_2|x_1,x_3,x_4)$. We perform a Monte Carlo sensitivity analysis on two dimensional bias parameters (corr$(x1,x2)$,corr$(x2,x3)$). Since gasoline Tax (x_4) has little correlation with Income (x_2), we ignore corr(x_2,x_4).

Table 4.2: Sensitivity analysis and selection of corr(x, c) by MC-BMS

		$\tilde{\rho}$ [BMS]				$\tilde{\rho}$ [BMA]			
	ρ	E-A(RMSE)	M-A	E-KNN *	M-KNN	E-A(RMSE)	M-A	E-KNN	M-KNN
MCAR	0.1	0.091 (0.009)	0.087 (0.013)	0.120 (0.020)	0.104 (0.004)	0.105 (0.005)	0.102 (0.002)	0.154 (0.054)	0.147 (0.047)
	0.2	0.179 (0.021)	0.186 (0.014)	0.191 (0.009)	0.202 (0.001)	0.197 (0.003)	0.195 (0.005)	0.217 (0.017)	0.204 (0.004)
	0.3	0.259 (0.041)	0.305 (0.005)	0.275 (0.025)	0.310 (0.010)	0.28 (0.019)	0.294 (0.006)	0.287 (0.013)	0.300 (0.001)
	0.4	0.375 (0.025)	0.405 (0.005)	0.391 (0.011)	0.370 (0.030)	0.376 (0.024)	0.387 (0.013)	0.371 (0.029)	0.363 (0.037)
	0.5	0.462 (0.038)	0.511 (0.011)	0.483 (0.017)	0.491 (0.009)	0.474 (0.026)	0.491 (0.009)	0.467 (0.033)	0.475 (0.025)
	0.6	0.512 (0.088)	0.608 (0.008)	0.582 (0.018)	0.566 (0.034)	0.545 (0.055)	0.589 (0.011)	0.535 (0.065)	0.535 (0.064)
	0.7	0.593 (0.107)	0.653 (0.047)	0.680 (0.020)	0.676 (0.024)	0.616 (0.084)	0.672 (0.028)	0.616 (0.084)	0.618 (0.082)
LL	0.1	-0.011 (0.111)	0.099 (0.001)	0.095 (0.005)	0.097 (0.003)	-0.022 (0.122)	0.087 (0.013)	0.119 (0.019)	0.110 (0.009)
	0.2	0.036 (0.164)	0.186 (0.014)	0.220 (0.020)	0.181 (0.019)	0.041 (0.159)	0.180 (0.021)	0.230 (0.030)	0.194 (0.005)
	0.3	0.106 (0.194)	0.287 (0.013)	0.320 (0.020)	0.271 (0.029)	0.115 (0.185)	0.279 (0.021)	0.299 (0.001)	0.282 (0.017)
	0.4	0.160 (0.240)	0.409 (0.009)	0.395 (0.005)	0.379 (0.021)	0.200 (0.200)	0.388 (0.012)	0.402 (0.002)	0.379 (0.021)
	0.5	0.286 (0.214)	0.468 (0.032)	0.527 (0.027)	0.471 (0.030)	0.271 (0.229)	0.477 (0.023)	0.492 (0.007)	0.455 (0.045)
	0.6	0.363 (0.237)	0.590 (0.010)	0.595 (0.005)	0.554 (0.046)	0.373 (0.227)	0.580 (0.020)	0.563 (0.036)	0.538 (0.062)
	0.7	0.441 (0.259)	0.680 (0.020)	0.717 (0.017)	0.708 (0.008)	0.457 (0.243)	0.676 (0.024)	0.669 (0.032)	0.654 (0.046)
LQ	0.1	0.194 (0.094)	0.092 (0.008)	0.069 (0.031)	0.101 (0.001)	0.192 (0.093)	0.089 (0.011)	0.085 (0.015)	0.097 (0.002)
	0.2	0.244 (0.044)	0.179 (0.021)	0.148 (0.052)	0.183 (0.017)	0.269 (0.068)	0.175 (0.025)	0.159 (0.041)	0.181 (0.018)
	0.3	0.365 (0.065)	0.277 (0.023)	0.243 (0.057)	0.273 (0.027)	0.376 (0.075)	0.273 (0.027)	0.237 (0.063)	0.262 (0.038)
	0.4	0.455 (0.055)	0.350 (0.050)	0.366 (0.034)	0.355 (0.045)	0.461 (0.061)	0.366 (0.034)	0.339 (0.061)	0.347 (0.053)
	0.5	0.509 (0.009)	0.443 (0.057)	0.404 (0.096)	0.401 (0.099)	0.534 (0.034)	0.470 (0.033)	0.416 (0.084)	0.428 (0.072)
	0.6	0.582 (0.018)	0.532 (0.068)	0.535 (0.065)	0.525 (0.075)	0.617 (0.016)	0.564 (0.041)	0.533 (0.067)	0.541 (0.059)
	0.7	0.585 (0.115)	0.612 (0.088)	0.557 (0.143)	0.561 (0.142)	0.618 (0.081)	0.628 (0.072)	0.567 (0.133)	0.590 (0.110)

* K=2 in KNN.

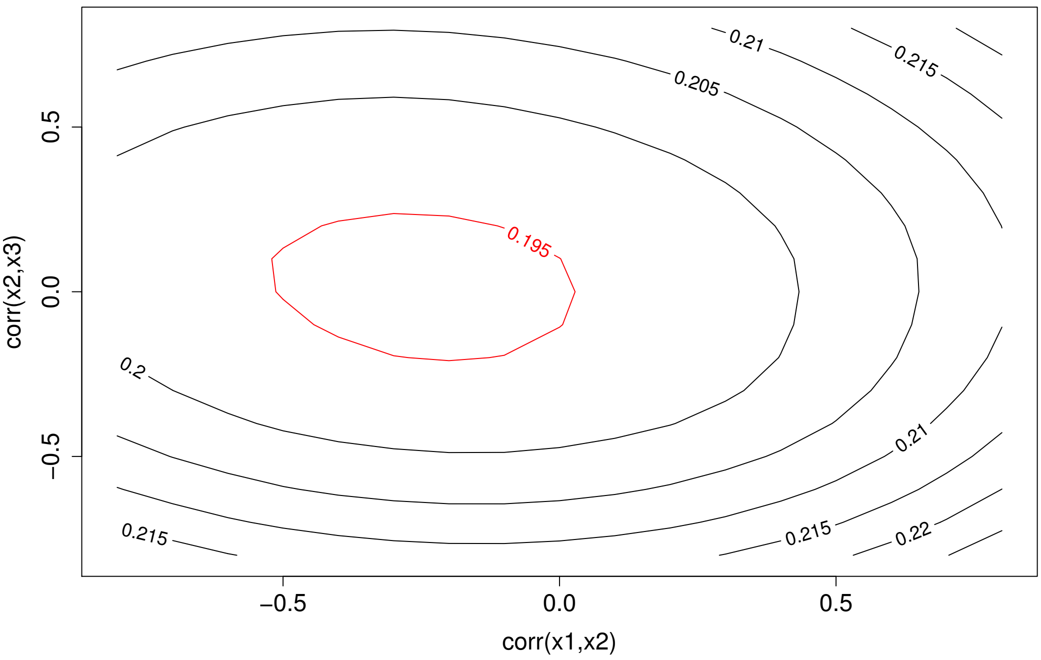
First of all, we choose the correlation ρ_1= corr(x_1, x_2) and ρ_2=corr(x_2, x_3) between -1 to 1 with interval 0.1. Given any pair of the values $\rho = (\rho_1, \rho_2)$, $\hat{\theta}$ can be evaluated through local bias analysis, and it is denoted as $\hat{\theta}_\rho$. Then the missing value of income x_2 is imputed according to conditional distribution $f(x_2|x_1, x_3, x_4; \rho_1, \rho_2)$ while the fuel consumption t is bootstrapped through linear regression model $f(t|x_1, x_2, x_3, x_4; \hat{\theta}_\rho)$. To reduce sampling error we choose the MC size as 100. The bias model selection process is conducted by calculating the distance between the simulated data $D^*_\rho = (t^*, x_1, x_3, x_4)$ and corresponding observed data $D^* = (t, x_1, x_3, x_4)$. And we use four distance measures: 1) Euclidean average distance; 2) Mahalanobis average distance; 3) KNN measure under Euclidean metric; 4) KNN measure under Mahalanobis metric.

The contour plot in Figure 4.5 (a) shows the selection results of bias parameters by average distance under Mahalanobis metric and Figure 4.5 (b) shows the results of averaged ASL during the replications. We noticed that no pair of bias parameters is rejected if the significance level is selected at 5%, however, we can still find some pairs may perform better than the others. As shown in Figure 4.5 (b), the area inside the red color line is considered as the most possible location of bias parameters.

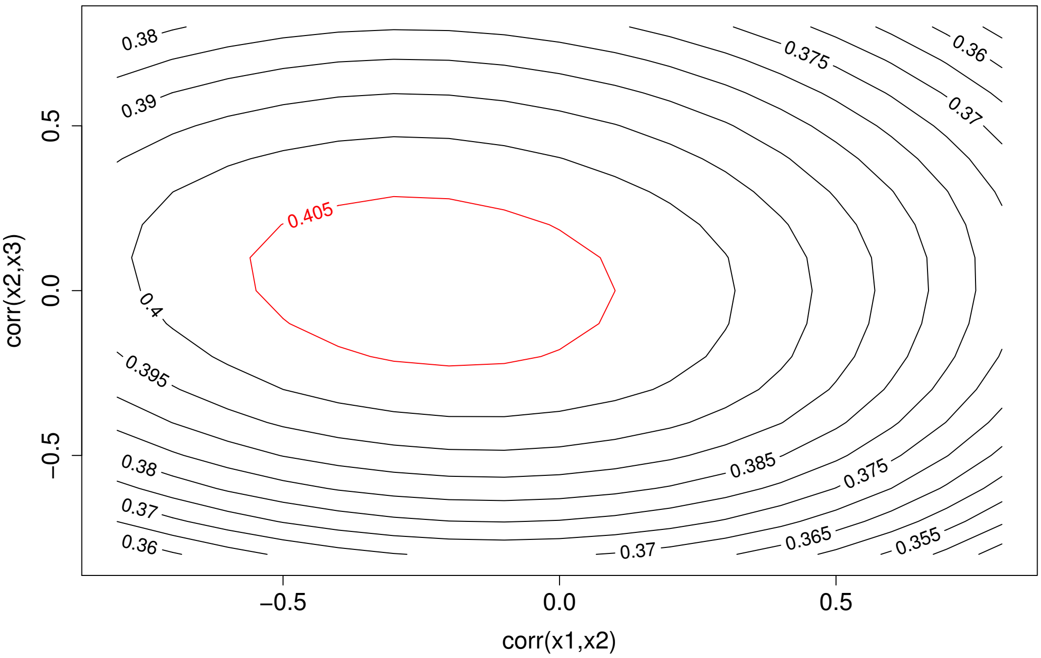
One simplest way to calculate the BMA estimator is to choose the achieved significance level as the weight w_i as in formula (4.3), and results are presented at Table 4.3. The bias model selection methods perform robustly for all the four measures, and it works better than inference without consideration of the model uncertainty (results shown as complete case (CC) analysis or Multiple imputation (MI)). BMA method works quite similarly as BMS, but how to identify the posterior weight (or specify prior) is always difficult.

4.5 Discussion

In this Chapter, we were concerned with the sensitivity analysis for the missing data problems and developed a new Monte Carlo bias model selection method. Conventional inference based on observed data D_{obs} only is usually short of knowledge or lack of randomization thus it is always difficult to approach the true value of θ with missing data problems. Sensitivity analysis treats the tilting parameter η as the input of uncertainty analysis, and we calculate the plausible values for the outcome given inputs with Monte Carlo sampling procedure. We then aim to find the best value (or



(a) Distance



(b) ASL

Figure 4.5: Contour plots for selection of $\mathrm{corr}(x_1, x_2)$ and $\mathrm{corr}(x_2, x_3)$ for fuel consumption data. Fig (a): distance; Fig (b): achieved significance level. Mahalanobis average distance is used as in MC-BMS method.

Table 4.3: Simulation results with covariate density uncertainty

		θ_0	θ_1	θ_2	θ_3	θ_4
MLE	$\hat{\theta}_{gZ}$	154.193	18.545	-6.135	0.472	-4.228
BMS	E-A	154.597	17.553	-6.073	0.482	-4.252
	M-A	154.249	18.389	-6.137	0.472	-4.257
	E-KNN	154.249	18.466	-6.324	0.394	-4.357
	M-KNN	153.558	20.128	-6.114	0.471	-4.237
BMA	E-A	153.862	19.359	-6.075	0.494	-4.215
	M-A	154.597	18.554	-6.073	0.482	-4.252
	E-KNN	152.188	18.649	-6.146	0.468	-4.261
	M-KNN	154.144	18.653	-6.133	0.473	-4.253
Others	MI	155.176	11.808	-8.300	0.393	-5.474
	CC	164.041	-1.247	-6.998	0.331	-7.689

a small set of the best values) among all the considered values for η. The selection process is taken as measuring the distance between observed data and simulated data given each η, and several clustering distances are used.

We conducted some simulation studies to examine this method. The advantages of KNN and average distance have been discovered clearly. In particular, the average distance under Mahalanobis metric perform robustly in the hypothesis test, but non-parametric measure KNN may work better to describe small pattern differences.

We further applied the MC-BMS method into the missing covariate problems for Fuel consumption data, and found that the MC-BMS method works robustly and the advantage is found compared with conventional missing data analysis methods. This novel approach is very flexible and useful, and we will apply it into the non-ignorable missing data problems in the following chapters.

4.6 Appendix

4.6.1 K Nearest Neighbour

Two observations x_i and x_j are defined as k-neighbours if (see Definition 1, page 364 Wong and Lane, 1983):

$$d(x_i, x_j) \leq d_k(x_i) \,\text{or}\, d_k(x_j),$$

where d is the Euclidean metric and $d_k(x_i)$ is the kth nearest-neighbour distance to point x_i. To define a distance between two clusters by KNN method, we need a definition of KNN between an individual point and a cluster first.

Definition 4. *Given a cluster $D = \{x_i, i = 1, \ldots, n\}$, an individual observation x_j is said to be neighbour of cluster D if there exists at least one point x_i in cluster D that*

$$d(x_i, x_j) \leq d_k(x_i),$$

where d is the Euclidean metric and $d_k(x_i)$ is the kth nearest-neighbour within cluster distance to point x_i.

If a test sample $D^* = \{x_j, j = 1, \ldots, m\}$ is distributed similarly with cluster D, then we expect most of observations in D^* to be the nearest neighbour of D. But if not, that means the difference between the two clusters is apparent and the distance should be large. Mathematically, we write I as indicator function, which takes value 1 if the condition is satisfied or 0 if the condition is failed. Then the percentage of observations in test sample D^* with the nearest neighbour relationship of cluster D can be calculated by the average $E(I_1)$, with each element defined as:

$$I_1^{(j)} = \begin{cases} 1, & \sum_{x_i \in D}\{I(d(x_i, x_j) < d_k(x_i)\} > 0; \\ 0, & \qquad\qquad \text{otherwise.} \end{cases} \tag{4.5}$$

Here $\sum_{x_i \in D}\{I(d(x_i, x_j) < d_k(x_i)\}$ takes integer in $\{0, 1, 2, ..n\}$. Only when it equals 0, the observation x_j is not the nearest neighbour of cluster D.

When we compare two clusters D and D^*, this measure is symmetric. Similarly, the percentage of points in cluster D which are the nearest neighbours of test sample D^*

is calculated by $E(I_2)$:

$$I_2^{(i)} = \begin{cases} 1, & \sum_{x_j \in D^*} \{I(d(x_i, x_j) < d_k(x_j)\} > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{4.6}$$

with $d_k(x_j)$ is kth nearest-neighbour distance to point x_j within cluster D^*. The average of $E(I_1)$ and $E(I_2)$

$$s(D, D^*) = \frac{1}{2}(E(I_1) + E(I_2))$$

is considered an similarity measure and $1 - s(D, D^*)$ taken as the 'KNN distance' measure in this paper.

Other measures such as Mahalanobis metric may also be used.

Chapter 5

Local Bias Analysis for Non-Ignorable Missing Data

5.1 Introduction

Beyond ignorable missingness, non-ignorable missing data mechanism is also very common. What this means is: even accounting for all the observed variables, the reason for the missingness of observations still depends on the values of the unseen observations. We consider the multivariate regression analysis of a n-dimensional vector of an incomplete data set $D = (D_1, \ldots, D_n)$ where each D_i is independent from the other and includes the response variable t_i and covariate variables x_i, c_i. We consider the problem when confounder C is not always observed, and denote R as an indicator vector $R = (r_1, \ldots, r_n)^T$ such that r_i=1 when c_i is observed or 0 when c_i is missing; $i = 1, \ldots, n$. In regression analysis the focus is on inferring the conditional distribution of response variable T given these covariates X, C: $[T|XC]$ with a given joint distribution of $[XC]$. The fully observed information will comprise $Z = (T, X, C, R)$, and the incomplete data is $Y = (T, X, C_{obs}, R)$ correspondingly. The difference in modelling $[R|D]$ between non-ignorable and ignorable missing data assumptions will be our interest in this chapter.

There is an extensive literature regarding regression models with non-ignorable missing responses and covariates. For example, Ibrahim et al. (1999) discussed non-ignorable missing covariates in generalized linear models, and Paik (2004) considered matched case control analysis with non-ignorable missingness and Saha and Jones (2005) estimate asymptotic bias of the linear mixed effects with non-ignorable assumption. A standard approach is to assume a parametric model for $[R, D]$ and parameters are estimated by the maximum likelihood method. Selection models (Ch.12 Little and Rubin, 2002) factorize the distribution $[R, D]$ into a model for $[D]$ and a model for $[R|D]$. The distribution $[R|D]$ needs to be identified when the missing data mechanism is non-ignorable (Rubin, 1976), which in our context means the MDM depends on C.

Unfortunately, little information is known about the form of $[R|D]$ in practice, and bias in the estimation of parameter θ can be resulted from misspecification of the models; see for example Diggle and Kenward (1994). Besides the ML method, the inverse probability weighted estimating equations approach (Robins et al., 1994) and the multiple imputation method (Ibrahim et al., 2005) also require correct specification of the form of MDM to ensure unbiased analysis. Unlike ignorable missing data, non-identifiability in modelling $[R|D]$ is the real problem of non-ignorable missing data. Characterizing model identifiability is a very difficult task requiring deep technical machinery. Chen et al. (2004) presented necessary and sufficient conditions for model identifiability in generalized linear models for missing covariates problem. Some extensions have been given in Huang et al. (2005).

The problems with these concerns as we may encounter in the missing data problem comprise:

(i) *Model Uncertainty*: the respondent and non-respondent variables have exactly the same values conditional on observed variables for MAR missing data (Rubin, 1987), but model uncertainty exists in the sense that trials suffer ‘lack of randomization’. And these problems continue to happen in non-ignorable missing data. The uncertainty lies in the regression models, covariate density modelling and missing data mechanism specification. It was also discovered in missing confounder and publication bias problems.

(ii) *Non-identifiability*: a specific model for missing data mechanism is necessary for non-ignorable missing data, and it may be fitted by a parametric model (logistic, probit, e.g.) or a semiparametric model (see Kim and Yu, 2011). However, it is always difficult to judge whether the model is proper or not.

The inference in this chapter will continue the discussion with the bias adjustment from local sensitivity analysis, but missing data mechanism bias will no longer be easy to calculate because of the identifiability issue. Here we consider a transformation between a selection model and pattern mixture model to allow the MAR counterpart to fit the MNAR part.

5.2 Bias Analysis for MNAR

5.2.1 Double Misspecified Models

Theorem 3.1 provides a general tool to analyse bias model uncertainty for any missing data problems. This method is valid for MNAR assumption as well. For complete data $Z = (T, X, C, R)$, the joint distribution is:

$$g_Z = f_{T|XC}(t|x,c) f_{XC}(x,c) h(r|t,x,c).$$

Under the MNAR assumption, MDM model $h(r|t,x,c)$ depends on the missing covariate C. In practice, we usually consider an identifiable working model (i.e. MAR) $h_1(r|t,x)$:

$$f_Z = f_{T|XC}(t|x,c) f_{XC}(x,c) h_1(r|t,x).$$

The misspecification of f_Z from true model g_Z is $\exp(\epsilon_R u_R) = \frac{h(r|t,x,c)}{h_1(r|t,x)}$. Also with the assumption that X and C are independent, the working model will be

$$f_Z^* = f_{T|XC}(t|x,c) f_X(x) f_C(c) h_1(r|t,x),$$

with covariate distribution misspecification $\exp(\epsilon_{XC} u_{XC}) = \frac{f_{XC}(x,c)}{f_X(x) f_C(c)}$ induced.

The corresponding model with incomplete data will have the form

$$\begin{aligned} g_Y &= \int_{(y)} g_Z dz = \int_{(y)} f_Z \exp(\epsilon_R u_R) dz \\ &= f_Y \exp(\epsilon_R u_{R|Y}) \\ &\approx f_Y^* \exp(\epsilon_{XC} u_{XC|Y}) \exp(\epsilon_R u_{R|Y}). \end{aligned} \tag{5.1}$$

The models g_Y, f_Y, f_Y^* are the corresponding marginal models under complete data

g_Z, f_Z, f_Z^*. We are assuming the misspecification sources are in local analysis assumptions and thus we can have the approximation in equation (5.1) to the first order. The working model used to calculate the estimation of parameters $\hat{\theta}_{gY}$ is

$$f_Y^* = f_{T|XC}(t|x, c^{(r)}) f_X(x) f_C^r(c) h_1(r|t, x).$$

To simplify notations, we denote $h_Z = h(r = 0|t, x, c)$ and $h_Y = h_1(r = 0|t, x)$. Using Theorem 3.1, we have the covariate bias b_{XC} as follows:

$$\begin{aligned}
\boldsymbol{b_{XC}} &= \epsilon_{XC} E_{f_Z^*}[u_{XC}\{\boldsymbol{I_Y^*}^{-1}\boldsymbol{s_Y}^* - \boldsymbol{I_Z^*}^{-1}\boldsymbol{s_Z}^*\}] \\
&= \epsilon_{XC}\boldsymbol{I_Y^*}^{-1} E_{f_Z^*}[u_{XC}\boldsymbol{s^*_{Y|r=0}}] \\
&\approx \boldsymbol{I_Y^*}^{-1}\{E_{XC}[\boldsymbol{s^*_{Y|r=0}} h_Y] - E_{X,C}[\boldsymbol{s^*_{Y|r=0}} h_Y]\}
\end{aligned}$$

where MDM working model h_Y is assumed from MAR assumption, and covariate distribution $f(x, c)$ is required in order to evaluate the incomplete data bias. The MDM bias is given by:

$$\begin{aligned}
\boldsymbol{b_R} &= \epsilon_R E_{f_Z}[u_R \boldsymbol{I_Y^*}^{-1}\boldsymbol{s_Y^*}] \\
&= \epsilon_R \boldsymbol{I_Y^*}^{-1} E_{f_Z}[u_R h_Y \boldsymbol{s^*_{Y|r=0}}] \\
&= \boldsymbol{I_Y^*}^{-1} E_{f_Z}[\log(\frac{h_Z}{h_Y})\boldsymbol{s^*_{Y|r=0}} h_Y] \\
&\approx \boldsymbol{I_Y^*}^{-1} E_{f_Z}[(h_Z - h_Y)\boldsymbol{s^*_{Y|r=0}}]. \qquad (5.2)
\end{aligned}$$

5.2.2 Triple Misspecified Models

Following discussion in Section 2.5, when we consider the non-ignorable missingness in generalized linear models following discussion, a triple misspecification will ensue. The data sampling distribution under incomplete data Y is

$$\begin{aligned}
g_Y &\approx f_Y^* \exp(\epsilon_R u_{R|Y}) \exp(\epsilon_{XC} u_{XC|Y}) \\
&= f_Y^{**} \exp(\epsilon_M u_M) \exp(\epsilon_R u_{R|Y}) \exp(\epsilon_{XC} u_{XC|Y}) \qquad (5.3)
\end{aligned}$$

where the misspecification $\exp(\epsilon_M u_M)$ is the ratio of the marginal model and the working model

$$\exp(\epsilon_M u_M) = \frac{f_Y^*}{f_Y^{**}}.$$

The working model f_Y^{**} can be written as

$$f_Y^{**} = f_{T|XC}^r f_X f_C^r h_1(r|t,x),$$

where

$$f_{T|XC}^r = \begin{cases} f_{T|XC}, & r=1; \\ f_{T|X}^*, & r=0. \end{cases}$$

with $f_{T|X}^*$ given at equation (2.29) and

$$f_C^r = \begin{cases} f_C(c), & r=1; \\ 1, & r=0. \end{cases}$$

The distribution f_Y^{**} is triple misspecified from g_Y, and the three misspecification quantities are:

$$\epsilon u_Y = \epsilon_{XC} u_{XC|Y} + \epsilon_R u_{R|Y} + \epsilon_M u_M.$$

Correspondingly, the incomplete data bias are separated into three components according the bias sources:

$$\text{bias} = b_M + b_{XC} + b_R,$$

with marginal model bias

$$\begin{aligned} \boldsymbol{b_M} &= \epsilon_M E_{f_Y^{**}}(u_M \boldsymbol{I_Y}^{**-1} \boldsymbol{s_Y}^{**}) \\ &= \frac{\beta^2 \sigma_c^2}{2a(\phi)} \boldsymbol{I_Y}^{**-1} E_X\{\xi''(\pi_x) \boldsymbol{v_0} h_Y\}, \end{aligned}$$

covariate bias

$$\begin{aligned} \boldsymbol{b_{XC}} &= \epsilon_{XC} E_{f_Y^{**}}(u_{XC|Y} \boldsymbol{I_Y}^{**-1} \boldsymbol{s_Y}^{**}) - \epsilon_{XC} E_{f_Z^*}(u_{XC} \boldsymbol{I_Z}^{**-1} \boldsymbol{s_Z}^{**}) \\ &= \frac{\boldsymbol{I_Y}^{**-1}}{a(\phi)} E_{X,C}\{\log(\frac{f(c|x)}{f(c)}) \boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)] h_Y\} \\ &= \frac{\boldsymbol{I_Y}^{**-1}}{a(\phi)} [E_{XC}\{\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)] h_Y\} - E_{X,C}\{\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)] h_Y\}], \end{aligned}$$

and MDM bias

$$
\begin{aligned}
\boldsymbol{b_R} &= \epsilon_R E_{f_Z}\{u_R[\boldsymbol{I_Y}^{**-1}\boldsymbol{s_Y}^{**} - \boldsymbol{I_Z}^{**-1}\boldsymbol{s_Z}^{**}]\} \\
&= \frac{\boldsymbol{I_Y}^{**-1}}{a(\phi)} E\{log(\frac{h_Z}{h_Y})\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)h_Y]\} \\
&\approx \frac{\boldsymbol{I_Y}^{**-1}}{a(\phi)} E\{[h_Z - h_Y]\boldsymbol{v_0}[\xi(\pi_{xc}) - \xi(\pi_x)]\}.
\end{aligned}
$$

As we see in the equations, marginal model bias can be calculated based on identifiable model f_Y^{**} while covariate bias and missing data mechanism bias are obtained requiring two models: covariate distribution $f(x, c)$ and missing data mechanism. In practice the true MDM model $h(r|t, x, c)$ is always difficult to fit, especially because of the non-identifiablility issue for non-ignorable missing data, thus further consideration (for example, sensitivity analysis) is necessary.

5.3 Inference about MDM Bias

Under the ignorable missing data assumption, the MDM model can be fitted by a parametric model such as logistic linear model or non-parametric method such as generalized additive model (Hastie and Tibshirani, 1990). In non-ignorable missing data problems, we recognized that both the covariate distribution and the missing data mechanism modelling are unknown and this leads to a non-identifiability problem. Sensitivity analysis (see e.g. Cook, 1986; Oakley and O'Hagan, 2004; Greenland, 2005; McCandless et al., 2007; Daniels and Hogan, 2008; Gustafson et al., 2010) is a valid method to handle the uncertainty analysis, but how to reduce the dimension of bias parameters is a key step and will be discussed below. Specifically if we can deduce the evaluation of non-ignorable missing data mechanism through its marginal model, then we may make similar inferences with ignorable missingness.

Let D be the whole dataset, which can be divided into an observed variable component D^{obs} which contains variables which are always observed and a missing variable component D^{mis}, which contains variables with non-response values. [1] Missing data

[1] The use of superscript is different with subscript as used before, see Section 1.1 for their differences. For example, if complete data $D = (T, X, C)$ and variable C is partially missing, then $D^{obs} = (T, X)$ and $D^{mis} = (C)$; while D_{obs} includes all observed values $D_{obs} = (T, X, C_{obs})$ and $D_{mis} = (C_{mis})$.

indicator R is defined as before. According to Bayes' Theorem, the true MDM model $[R|D]$ can be written as

$$\begin{aligned} h(R|D) &= \frac{f(D|R)h(R)}{f(D)} \\ &= \frac{f(D^{mis}|D^{obs},R)f(D^{obs}|R)h(R)}{f(D^{mis}|D^{obs})f(D^{obs}).} \end{aligned} \tag{5.4}$$

A working model for MDM in incomplete data distribution f_Y^* (or f_Y^{**} in GLMs) is usually under an identifiable MAR model:

$$h(R|D^{obs}) = \frac{f(D^{obs}|R)h(R)}{f(D^{obs})}. \tag{5.5}$$

The ratio of equation (5.4) and (5.5) is

$$\frac{h(R|D)}{h(R|D^{obs})} = \frac{f(D^{mis}|D^{obs},R)}{f(D^{mis}|D^{obs})},$$

where $f(D^{mis}|D^{obs})$ is the conditional distribution of missing variables given observed variables D^{obs} and $f(D^{mis}|D^{obs},R)$ is the distribution on specific patterns only. This ratio equals 1 under ignorable assumption, where $h(R|D) = h(R|D^{obs})$. But $f(D^{mis}|D^{obs},R)$ is apparently different from $f(D^{mis}|D^{obs})$ under non-ignorable missingness, and it is also one of the significant differences between non-ignorable and ignorable missingness (Rubin, 1987). How to model the distributions for each pattern, especially for the incomplete cases, is our concern.

As we know, R is Bernoulli distributed, with probability $h(R = 1|D)$ and $h(R = 0|D)$. In this case, the distribution $[D^{mis}|D^{obs}]$ is the weighted average on both parts: $f(D^{mis}|D^{obs}, R = 1)$ and $f(D^{mis}|D^{obs}, R = 0)$, with the weights equal to the marginal density of missing data mechanism $h(R|D)$ for $R = 1$ and $R = 0$ respectively:

$$f(D^{mis}|D^{obs}) = f(D^{mis}|D^{obs}, R = 1)h(R = 1|D^{obs}) + f(D^{mis}|D^{obs}, R = 0)h(R = 0|D^{obs}). \tag{5.6}$$

We must emphasize that, the probability $h(R|D^{obs})$ is the marginal model of the true missing data mechanism, $h(R|D^{obs}) = E_{D^{mis}|D^{obs}}[h(R|D^{obs}, D^{mis})]$, which can be calculated by Bayesian inference through equation (5.5). It may also be fitted as a proper conditional model, such as a generalized additive model with non-parametric method (see Section 3.4), but using parametric models may lead to further misspecification. Below we will discuss one example.

For a linear regression model

$$t|(x, c) \sim N(\alpha + \theta x + \beta c, \sigma^2) \tag{5.7}$$

with c partially missing and suppose the true missing data mechanism depends on (x, c) by a logistic linear model:

$$R \sim \text{Bernoulli}(1, \pi_{xc})$$

$$\text{logit}(\pi_{xc}) = \psi_0 + \psi_1 x + \psi_2 c. \tag{5.8}$$

In this case the probability of data censoring is $h_Z = h(r = 0|x, c)$ and the corresponding model under MAR h_Y is the model without c: $h(r = 0|x)$. The ratio $\frac{h_Z}{h_Y}$ in bias expression is given as

$$\frac{h_Z}{h_Y} = \frac{h(r = 0|x, c)}{h(r = 0|x)} = \frac{f(c|x, r = 0)}{f(c|x)},$$

which is the ratio of conditional covariate distribution of $f(c|x)$ between incomplete cases and all cases. If covariate distribution $[XC]$ is assumed as multivariate normal distributed, then the marginal model is

$$\begin{aligned} h(r = 1|x) &= E_{c|x}(\pi_{xc}) \\ &\approx \pi_x + \frac{1}{2}\psi_2^2\sigma_c^2\xi''(\psi_0 + \psi_1 x) + O(\psi_2^2\sigma_c^4) \end{aligned} \tag{5.9}$$

with $\pi_x = \xi(\psi_0 + \psi_1 x)$ and link function $\xi(.) = \text{expit}(.)$. The symbol ξ'' indicates the second derivative of $\xi(.)$. Apparently, fit of the marginal model of missing data mechanism through a logistic linear model will be misspecified, because it ignores the high order items. But we may consider to fit nonparametrically. Below we conduct a simulation study to test the inference.

The true value of parameter is selected as $(\alpha, \theta, \beta) = (0.5, 0.5, 1)$, and σ^2 takes value from uniform distribution in (0.16,1). Covariate variables (X, C) are assumed to have a multivariate normal distribution, both have standard $N(0, 1)$ marginal distribution with corr(x, c) taken as (0,0.1,0.2,0.3,0.4,0.5). Sample size is 100. The confounder C

is designed to be missing with probability $h(r = 0|x, c) = 1 - \text{expit}(1 - x - 2c)$, which generates about 36% missing values. As we discussed, the marginal model of MDM is required to evaluate the bias quantities, and we use Bayes' approach (formula (5.5)), a logistic linear model and non-parametric method to fit it. Results are presented in Table 5.1, with estimations taken as an average of 100 replications.

Table 5.1: Simulation results for $\hat{\theta}$

		Bayes			Nonparametric			Logistic Linear		
corr(x, c)	EB	b_{XC}	b_R	CR	b_{XC}	b_R	CR	b_{XC}	b_R	CR
ρ= 0	0.039	0	0.034	0.95	0	0.035	0.95	0	0.109	0.88
ρ= 0.1	0.056	0.022	0.042	0.90	0.022	0.043	0.90	0.023	0.086	0.90
ρ= 0.2	0.081	0.049	0.039	0.97	0.047	0.042	0.94	0.048	0.098	0.92
ρ= 0.3	0.121	0.079	0.049	0.89	0.078	0.052	0.89	0.081	0.122	0.87
ρ= 0.4	0.145	0.111	0.050	0.92	0.108	0.053	0.92	0.113	0.137	0.77
ρ= 0.5	0.174	0.156	0.048	0.91	0.157	0.049	0.89	0.164	0.152	0.72

EB: empirical bias $(\theta_{gY} - \theta)$. CR: coverage rate for adjusted estimation.

Remarks:

- The marginal model is free of missing covariate C, which can be fitted by a parametric model (5.9) under MNAR form (5.8) assumed, however as we discussed, unless we have strong belief in the model form it is difficult to approve or oppose it.

- As we see in this example, an apparent parametric model such as the logistic linear regression model (π_x) is not a good fit of the marginal distribution $h(r|x)$, even if the true MDM is in the logistic linear regression form (equation 5.8).

- The proper fitting should follow equation (5.5) or be approached with a non-parametric method (such as general additive modelling) to take into account the high order terms as in equation (5.9). Without further information on the true MDM form, we suggest using nonparametric modelling.

- The covariate distribution $f(x, c)$ can not be fitted given observed data only, and more information may be collected from other literature sources. Sensitivity analysis is our preferred method and one example is give in Appendix 5.6.1.

More simulation studies can be found in Appendix 5.6.2.

5.4 Numeric Result for GLMs

5.4.1 Equine Data Example

It is worth noting that the local bias analysis can accommodate both continuous and discrete variables. In this section we present the evaluation of triple misspecification bias for a real data example under non-ignorable missing assumption. The equine epidemiology example is considered as a matched case-control study by Sinha et al. (2005). The aim is to investigate how a disease indicator T (a case of colic versus a control received for any condition other than colic) depends on age (C) measured on a continuous scale and a binary covariate (X) indicating whether the horse experienced a recent diet change or not. In total 998 cases are observed. The logistic regression model estimated by the maximum likelihood method based on the complete data set is

$$\text{logit}(Pr(t=1|x,c)) = -0.611 + 2.097x + 0.0474c.$$

The effect of recent diet change towards disease is of interest. Its estimate is $\hat{\theta} = 2.097$. Suppose that the individual is selected with probability function

$$h(r=1|t,x,c) = \text{expit}(\omega(t,x) + \psi_1 c + \psi_2 xc) \tag{5.10}$$

which induces about 43% missingness in exposure variable (C) given $(\psi_1, \psi_2) = (0.5, -1)$ and $\omega(t,x) = 1$. Incomplete data bias analysis is conducted and the result is listed in Table 5.2. In the study, the MAR counterpart of the missing data mechanism: $h(r=1|x)$ is evaluated by a generalized additive model with nonparametric method (Hastie and Tibshirani, 1990). As seen in Table 5.2, we find that covariate bias and marginal bias for MLEs under incomplete data are relatively small

Table 5.2: Bias analysis result for Equine data under MNAR

	θ_Z	SE	θ_Y	$\hat{\theta}_Y$	b_{XC}	b_M	b_R	$\hat{b}_R$
θ_0	-0.611	0.068	-0.712	-0.593	-0.006	0.002	-0.122	-0.115
θ_x	2.097	0.029	2.298	2.093	0.025	-0.016	0.237	0.196
θ_c	0.047	0.011	0.046	0.037	0.001	-0.001	0.010	0.009

$\hat{\theta}_Y$ is estimation with adjustment from incomplete data bias $b_{XC} + b_R + b_M$. MDM bias b_R is calculated given true mechanism (5.10), while $\hat{b}_R$ is evaluated based on inference in Section 5.3.

in this example, while the missing data mechanism bias predominates over all the incomplete data bias sources. The approximation $\hat{b}_R$ works very well when compared with b_R, and it supports our inference by transposing the MNAR problem into its MAR counterpart.

5.4.2 Simulation Study

A simulation study is conducted with the response variable distributed as a logistic linear model $T \sim B(1, \pi_{xc})$:

$$f_{T|XC} = \pi_{xc}^t (1 - \pi_{xc})^{1-t},\ \pi_{xc} = \frac{\exp\{\alpha + \theta x + \beta c\}}{1 + \exp\{\alpha + \theta x + \beta c\}}.$$

Covariate variables are generated as multivariate normal distribution as before, with corr(x, c) selected from (0,0.3,0.5). True values are $(\alpha, \theta, \beta) = (1, 1, 1)$. Variable C is designed to be partially missing in this simulation, with the missing data mechanism under non-ignorable assumption:

$$h(r = 1|x, c) = \text{expit}(\psi_0 + \psi_1 c + \psi_2 xc).$$

Incomplete data bias is calculated, and the average estimators of 100 replications are presented in Table 5.3. The marginal bias seems not vary much for different corr(x, c) and MDM models. The covariate bias doesn't exist when corr$(x, c) = 0$, but it increases with corr(x, c). The MDM bias exists in each study and has a relative large bias size compared with marginal bias and covariate bias, especially for α and β components. The evaluation of MDM bias from MAR counterpart modelling works well. This is clearly seen by comparing the two coverage probabilities (CR and CR1).

5.5 Discussion

The incomplete data bias under non-ignorable assumption is analysed for both linear and generalized linear regression models in this chapter. The specification of missing data mechanism is difficult to achieve under non-ignorable missing data because of identifiability issue. Consequently, we transposed the MNAR problem into its corresponding MAR counterpart. The marginal model of MDM is required, which was

Table 5.3: Incomplete data biases for GLM under MNAR

		Empirical Bias			Marginal Bias			Covariate Bias			MDM Bias			MDM Bias 1			CR			CR1		
		b^α	b^θ	b^β	b^α_M	b^θ_M	b^β_M	b^α_{XC}	b^θ_{XC}	b^β_{XC}	b^α_R	b^θ_R	b^β_R	$\hat{b}^\alpha_R$	$\hat{b}^\theta_R$	$\hat{b}^\beta_R$	α	θ	β	α	θ	β
$\rho = 0$	ψ_1= -3	0.312	0.011	0.211	-0.097	-0.076	-0.074	0.001	-0.002	-0.001	0.362	0.065	0.242	0.373	0.069	0.253	84	93	43	88	90	40
	ψ_1= -2	0.221	0.052	0.133	-0.087	-0.073	-0.063	0.001	0.001	-0.001	0.284	0.102	0.165	0.291	0.105	0.172	87	91	76	82	88	74
	ψ_1= -1	0.113	0.092	0.071	-0.073	-0.067	-0.046	0.002	0.001	0.001	0.163	0.144	0.099	0.163	0.151	0.115	88	90	97	89	93	97
	ψ_1= 0	-0.048	0.123	0.029	-0.063	-0.063	-0.036	0.001	0.002	0.001	0.006	0.155	0.056	0.002	0.169	0.059	94	94	93	96	91	89
	ψ_1= 1	-0.203	0.072	0.017	-0.062	-0.063	-0.031	-0.001	0.002	0.001	-0.154	0.125	0.047	-0.155	0.124	0.052	92	92	92	93	88	96
	ψ_1= 2	-0.330	0.024	0.120	-0.072	-0.072	-0.017	-0.002	0.002	0.001	-0.293	0.057	0.100	-0.293	0.053	0.108	92	91	82	89	91	83
	ψ_1= 3	-0.428	-0.017	0.207	-0.087	-0.079	0.005	-0.002	0.002	0.002	-0.394	0.013	0.212	-0.394	0.015	0.214	90	94	50	87	89	51
$\rho = 0.3$	ψ_1= -3	0.302	0.085	0.205	-0.110	-0.083	-0.074	0.028	0.105	0.008	0.332	0.035	0.234	0.362	0.031	0.253	78	92	67	85	88	60
	ψ_1= -2	0.224	0.132	0.132	-0.097	-0.083	-0.056	0.027	0.109	0.012	0.225	0.075	0.153	0.275	0.084	0.161	84	93	89	83	92	90
	ψ_1= -1	0.098	0.186	0.051	-0.083	-0.082	-0.042	0.024	0.110	0.015	0.146	0.123	0.071	0.149	0.141	0.074	91	88	85	91	89	83
	ψ_1= 0	-0.052	0.201	0.010	-0.071	-0.085	-0.033	0.010	0.132	0.021	-0.001	0.154	0.019	-0.003	0.152	0.021	84	90	67	83	91	67
	ψ_1= 1	-0.202	0.182	0.044	-0.067	-0.076	-0.022	-0.002	0.135	0.026	-0.152	0.096	0.012	-0.151	0.094	0.013	93	87	86	91	87	85
	ψ_1= 2	-0.324	0.133	0.132	-0.073	-0.078	-0.002	-0.009	0.133	0.036	-0.266	0.042	0.075	-0.275	0.033	0.077	92	88	96	91	87	94
	ψ_1= 3	-0.421	0.094	0.251	-0.086	-0.085	0.025	-0.009	0.142	0.043	-0.352	-0.006	0.190	-0.362	-0.014	0.195	89	91	80	88	91	78
$\rho = 0.5$	ψ_1= -3	0.262	0.141	0.164	-0.113	-0.093	-0.062	0.061	0.190	-0.011	0.282	0.009	0.215	0.312	0.005	0.231	88	90	92	86	89	90
	ψ_1= -2	0.204	0.194	0.100	-0.101	-0.094	-0.046	0.053	0.201	-0.015	0.216	0.047	0.132	0.246	0.056	0.152	87	87	83	87	88	81
	ψ_1= -1	0.092	0.254	0.019	-0.089	-0.096	-0.034	0.036	0.222	-0.017	0.123	0.092	0.049	0.133	0.104	0.056	84	91	52	83	92	50
	ψ_1= 0	-0.045	0.297	-0.023	-0.076	-0.093	-0.022	0.018	0.245	-0.009	-0.006	0.113	-0.009	-0.009	0.123	-0.009	94	88	36	91	91	36
	ψ_1= 1	-0.181	0.282	0.017	-0.071	-0.088	-0.005	-0.001	0.251	-0.001	-0.132	0.072	-0.009	-0.144	0.063	-0.009	93	87	63	90	86	66
	ψ_1= 2	-0.301	0.233	0.073	-0.074	-0.086	0.016	-0.019	0.254	0.008	-0.223	0.021	0.053	-0.231	0.006	0.056	95	94	95	91	89	95
	ψ_1= 3	-0.378	0.201	0.268	-0.084	-0.091	0.043	-0.021	0.253	0.017	-0.306	-0.022	0.164	-0.313	-0.042	0.172	96	76	83	92	72	81

The missing data mechanism is $h(r = 1|x, c) = \text{expit}(1+\psi_1 c - xc)$. CR is the coverage rate of $\hat{\theta}$ with bias adjustment $b_{XC}+b_R+b_M$; CR1 is $\hat{\theta}$ with bias adjustment $b_{XC} + \hat{b}_R + b_M$.

fitted nonparametrically by a generalized additive model. Simulation results show that these techniques work very well.

Under the pattern mixture model frame, the covariate distribution is the key but it is difficulty to identify in real terms. Further investigation, such as follow up study or sensitivity analysis, should be considered.

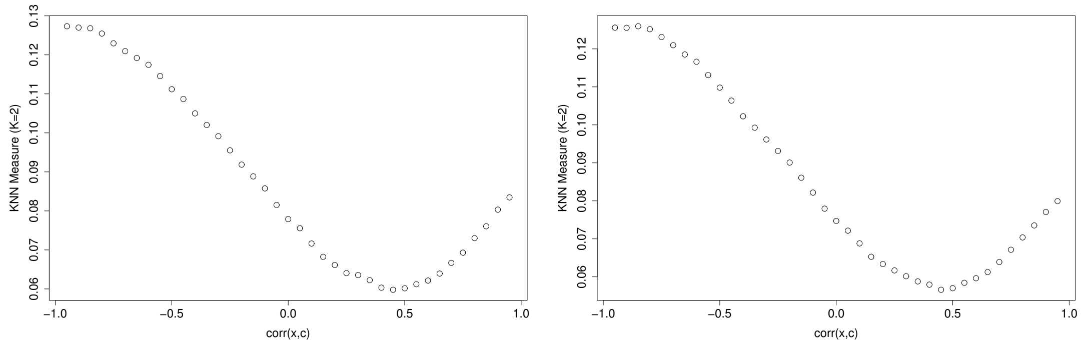
5.6 Appendix

5.6.1 Model Selection of Covariate Distribution

As noticed, the evaluation of incomplete data biases requires information on covariate distribution, which can be approximately calculated from complete cases under ignorable missingness. However this approach can be seriously biased under non-ignorable missing data, because of significant difference between complete case pattern and incomplete case pattern. In practice, we may use historic data or a follow up study (see e.g. Kim and Yu, 2011) to obtain this information, but it is not always simple to conduct a further investigation, and also extra bias may exist due to lack of randomization. One possible solution is to apply the MC-BMS method proposed in previous chapter.

Simply, we take one simulation study conduced from Table 5.1 with $\text{corr}(x, c) = 0.5$. The correlation $\text{corr}(x, c)$ is treated as the single bias parameter, with ρ sampled between (-1, 1) with an interval 0.05. We use a nonparametric model to fit the marginal distribution of missing data mechanism $h(r|x)$. MC-BMS approach is conducted similarly as ignorable missing data, but the bias adjustment may be slightly complex as with MDM bias evaluation.

Figure 5.1 shows the bias model selection result with nearest neighbour method measure under Euclidean and Mahalanobis metric. The averaged achieved significant level (ASL) is also calculated and shown in Figure 5.2. The power of testing pattern differences is much lower (below 0.2) in this case, then the method of calculating confidence interval (see Chapter 4) is no longer valid for this example. However, the smallest distance selection (BMS-2) by KNN have a clear selection result, as shown in Figure 5.1 (a) and (b). The bottom of the curve is very close to the true value 0.5. The KNN distance is clearly capable to discover small pattern differences. And it is robust for both Euclidean and Mahalanobis norm and also for different ‘K’ (parameter in KNN).



(a) KNN Euclidean(K=2)

(b) KNN Mahalanobis(K=2)

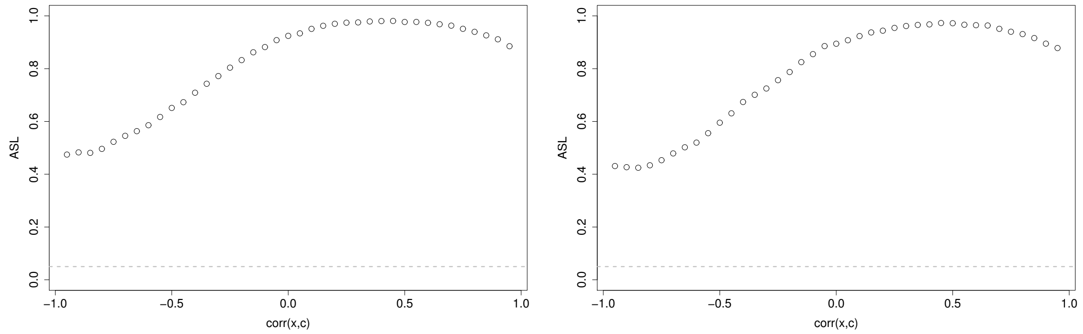
Figure 5.1: Selection of corr(x, c).

Monte Carlo sensitivity analysis (MCSA, Greenland, 2005) and multiple imputation method (under MAR) are also used to calculate $\hat{\theta}$, and we find MC-BMS gives the best result ($\hat{\theta}$=0.478). MCSA ($\hat{\theta}$=0.739) works even worse than the multiple imputation method under the ignorable assumption ($\hat{\theta}$=0.614), which indicates the uniform prior U(-1,1) is not a good choice. Bayesian model average (BMA) method improves the MCSA but the result is not as good as MC-BMS, as seen in Table 5.4.

Chapter 6 will extend the Monte Carlo bias model selection method to ‘selection model frame’ for non-ignorable missing data problem.

Table 5.4: Sensitivity analysis results

	α	θ	β
True value	0.500	0.500	1.000
BMS	0.477	0.478	1.052
MCSA	0.536	0.739	0.952
BMA	0.521	0.669	0.977
MAR	0.713	0.614	1.062



(a) KNN Euclidean

(b) KNN Mahalanobis

Figure 5.2: Achieved significance level

5.6.2 Simulation Studies for Complex MNAR models

Below we show more simulation studies under non-ignorable assumption. We generate two covariate variables by multivariate normal distribution:

$$\begin{pmatrix} x \\ c \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{5.11}$$

with $\rho = (0, 0.1, 0.3, 0.5)$. And response variable

$$t|(x, c) \sim N(\alpha + \theta x + \beta c, \sigma^2) \tag{5.12}$$

with setting the true value of parameter $(\alpha, \theta, \beta) = (0.5, 0.5, 1)$ and σ^2 takes value from $U(0.16, 1)$. We consider several MDM models for the procedure of dropping C, let $\pi = h(r = 1|t, x, c)$ and the designs are:

M1 (Expit Linear)

$$\pi = \frac{\exp(\psi_0 + \psi_1 c)}{1 + \exp(\psi_0 + \psi_1 c)}$$

where $(\psi_0, \psi_1) = (1, \psi_1)$.

M2 (Expit Linear: with x, c)

$$\pi = \frac{\exp(\psi_0 + \psi_1 c + \psi_2 x)}{1 + \exp(\psi_0 + \psi_1 c + \psi_2 x)}$$

where $(\psi_0, \psi_1, \psi_2) = (1, \psi_1, -1)$.

M3 (Expit Nonlinear: quadratic in c)

$$\pi = \frac{\exp(\psi_0 + \psi_1 c^2)}{1 + \exp(\psi_0 + \psi_1 c^2)}$$

where $(\psi_0, \psi_1) = (1, \psi_1)$.

M4 (Expit Nonlinear: interaction)

$$\pi = \frac{\exp(\psi_0 + \psi_2 c + \psi_3 x + \psi_1 cx)}{1 + \exp(\psi_0 + \psi_2 c + \psi_3 x + \psi_1 cx)}$$

where $(\psi_0, \psi_1, \psi_2, \psi_3) = (1, \psi_1, 1, 1)$.

M5 (Complementary log-log Linear: with x)

$$\pi = 1 - \exp\{-\exp(\psi_0 + \psi_1 c + \psi_2 x)\}$$

where $(\psi_0, \psi_1, \psi_2) = (1, \psi_1, 1)$.

M6 (Sin Linear)

$$\pi = \|\sin(\psi_0 + \psi_1 c + \psi_2 x)\|$$

where $(\psi_0, \psi_1, \psi_2) = (1, \psi_1, 1)$.

In all models, let $\psi_1 = c(2, 1, 0, -1, -2)$ in different instances. The incomplete data biases (b_{XC} and b_R) are estimated given the covariate distribution $[XC]$ and given the missing data mechanism $[R|T, X, C]$, while the estimation of MDM bias $\hat{b}_R$ is calculated with $h(r|t, x, c)$ assumed to be unknown, and we fit its marginal model $h_Y = h(r = 0|t, x)$ by a generalized additive model with nonparametric method and bias is then evaluated according to the inference in Section 5.3. In the simulations, sample size is chosen as $M = 1000$ and the studies are replicated 100 times. Estimation of θ is given as the average of all the replications.

The simulation results are listed in Tables 5.5 to 5.10. We show some significant findings below:

- Approximation about the missing data mechanism bias works well, when comparing the estimation of $\hat{b}_R$ with b_R or the coverage rates CR with CR1.

Table 5.5: MNAR simulation 1

MDM	corr(x,c)	EB	IB	b_{XC}	b_R	$\hat{b}_R$	CR	CR1
expit$(1+2c)$	$\rho=0$	-0.04	0.06	0	0.06	-0.02	98	96
	$\rho=0.1$	1.50	2.38	1.60	0.77	0.71	98	92
	$\rho=0.3$	4.34	7.25	5.26	1.99	1.99	84	84
	$\rho=0.5$	9.52	13.9	10.5	3.47	3.24	67	70
expit$(1+c)$	$\rho=0$	-0.19	0.06	0	0.07	-0.03	98	94
	$\rho=0.1$	1.46	1.70	1.30	0.40	0.35	99	97
	$\rho=0.3$	3.54	5.28	4.31	0.97	0.96	90	86
	$\rho=0.5$	7.55	10.2	8.45	1.78	1.57	76	75
expit$(1-c)$	$\rho=0$	-0.05	0	0	0	0.02	96	95
	$\rho=0.1$	1.21	1.65	1.31	0.33	0.24	95	92
	$\rho=0.3$	3.78	5.31	4.33	0.97	1.12	92	89
	$\rho=0.5$	7.11	10.1	8.33	1.77	1.57	80	78
expit$(1-2c)$	$\rho=0$	-0.17	0.01	0	0.01	0.03	96	97
	$\rho=0.1$	1.24	2.27	1.58	0.69	0.68	95	93
	$\rho=0.3$	4.61	7.32	5.21	2.11	1.89	82	83
	$\rho=0.5$	9.44	13.8	10.4	3.33	3.30	65	64

Note: EB is the empirical bias $(\hat{\theta}_Y - \theta)$. IB is the estimated incomplete data bias $(\hat{\theta}_Y - \hat{\theta}_Z)$, which is the sum of covariate bias b_{XC} and MDM bias b_R. $\hat{b}_R$ is the estimation based on inference section 5.3. CR is coverage rate of adjusted $\hat{\theta}_Y$, with incomplete data bias adjustment $b_{XC} + b_R$. CR1 is with adjustment of $b_{XC} + \hat{b}_R$. All outputs listed $\times 10^{-2}$.

Table 5.6: MNAR simulation 2

MDM	corr(x,c)	EB	IB	b_{XC}	b_R	$\hat{b}_R$	CR	CR1
expit$(1-x+2c)$	$\rho=0$	-2.89	-3.00	0	-3.01	-3.19	95	93
	$\rho=0.1$	-1.62	-0.69	1.75	-2.44	-2.53	93	92
	$\rho=0.3$	1.67	3.72	5.23	-1.51	-1.18	93	88
	$\rho=0.5$	6.17	9.59	9.61	-0.02	-0.12	73	74
expit$(1-x+c)$	$\rho=0$	-2.48	-2.71	0	-2.71	-2.66	100	98
	$\rho=0.1$	-0.73	-0.82	1.67	-2.50	-2.53	99	95
	$\rho=0.3$	1.85	2.87	4.97	-2.10	-2.17	96	93
	$\rho=0.5$	6.12	7.51	8.90	-1.40	-1.74	91	88
expit$(1-x-c)$	$\rho=0$	2.20	2.48	0	2.48	2.85	99	92
	$\rho=0.1$	3.66	4.49	1.77	2.72	2.87	94	93
	$\rho=0.3$	7.20	9.03	6.07	2.96	3.23	89	86
	$\rho=0.5$	12.2	15.6	12.4	3.20	3.11	71	70
expit$(1-x-2c)$	$\rho=0$	3.05	2.95	0	2.95	3.12	96	89
	$\rho=0.1$	4.82	5.48	1.87	3.61	3.47	95	90
	$\rho=0.3$	7.82	10.7	6.57	4.18	4.06	78	77
	$\rho=0.5$	13.3	18.2	14.0	4.25	4.36	54	46

Table 5.7: MNAR simulation 3

MDM	corr(x, c)	EB	IB	b_{XC}	b_R	$\hat{b}_R$	CR	CR1
$\text{expit}(1 + 2c^2)$	$\rho = 0$	-0.18	0	0	0	-0.17	99	95
	$\rho = 0.1$	-0.08	0.11	0.49	-0.38	-0.24	98	99
	$\rho = 0.3$	0.95	0.36	1.58	-1.22	-1.28	94	92
	$\rho = 0.5$	2.20	0.74	3.04	-2.29	-1.97	92	83
$\text{expit}(1 + c^2)$	$\rho = 0$	0.02	-0.01	0	-0.01	0.12	97	95
	$\rho = 0.1$	0.37	0.26	0.64	-0.38	-0.51	96	89
	$\rho = 0.3$	1.45	0.77	2.12	-1.35	-1.23	97	93
	$\rho = 0.5$	3.09	1.63	4.09	-2.46	-2.05	91	91
$\text{expit}(1 - c^2)$	$\rho = 0$	-0.15	-0.07	0	-0.07	-0.08	97	97
	$\rho = 0.1$	1.93	3.66	2.32	1.34	1.38	93	94
	$\rho = 0.3$	7.34	11.5	7.43	4.10	3.95	75	76
	$\rho = 0.5$	15.2	20.5	14.3	6.23	6.03	56	56
$\text{expit}(1 - 2c^2)$	$\rho = 0$	-0.19	0.12	0	0.12	0.03	100	94
	$\rho = 0.1$	3.45	4.74	3.11	1.62	1.64	97	92
	$\rho = 0.3$	9.90	14.8	9.78	5.03	4.76	66	69
	$\rho = 0.5$	20.8	26.1	18.9	7.18	7.17	58	57

Table 5.8: MNAR simulation 4

MDM	corr(x, c)	EB	IB	b_{XC}	b_R	$\hat{b}_R$	CR	CR1
$\text{expit}(1 + x + c + 2x \times c)$	$\rho = 0$	-9.02	-11.4	0	-11.4	-11.7	91	80
	$\rho = 0.1$	-6.94	-9.15	1.66	-10.8	-11.2	94	85
	$\rho = 0.3$	-3.74	-5.55	3.93	-9.47	-8.88	86	83
	$\rho = 0.5$	-0.470	-2.83	5.08	-7.90	-7.34	85	81
$\text{expit}(1 + x + c + x \times c)$	$\rho = 0$	-5.11	-7.03	0	-7.03	-7.48	95	89
	$\rho = 0.1$	-3.96	-5.4	1.33	-6.74	-6.99	96	88
	$\rho = 0.3$	-1.65	-2.62	3.46	-6.08	-5.85	95	93
	$\rho = 0.5$	1.24	-0.18	5.01	-5.19	-4.66	87	86
$\text{expit}(1 + x + c - x \times c)$	$\rho = 0$	5.39	7.16	0	7.16	7.43	96	93
	$\rho = 0.1$	7.54	9.16	1.69	7.48	7.51	97	95
	$\rho = 0.3$	11.8	14.4	6.62	7.77	7.68	83	82
	$\rho = 0.5$	18.8	22.1	14.8	7.33	7.18	77	79
$\text{expit}(1 + x + c - 2x \times c)$	$\rho = 0$	8.69	11.3	0	11.3	11.8	94	77
	$\rho = 0.1$	11.2	14.0	2.23	11.7	11.9	85	75
	$\rho = 0.3$	16.9	20.5	8.82	11.7	12.0	80	73
	$\rho = 0.5$	25.7	30.2	19.5	10.5	10.5	70	63

Table 5.9: MNAR simulation 5

MDM	corr(x, c)	EB	IB	b_{XC}	b_R	$\hat{b}_R$	CR	CR1
1-exp$\{-\exp(1 + x + 2c)\}$	$\rho = 0$	3.80	3.91	0	3.91	3.93	98	98
	$\rho = 0.1$	5.06	5.97	1.47	4.49	4.51	96	95
	$\rho = 0.3$	7.77	10.5	5.44	5.07	5.23	87	85
	$\rho = 0.5$	11.5	17.1	11.4	5.61	5.37	49	55
1-exp$\{-\exp(1 + x + c)\}$	$\rho = 0$	3.51	3.92	0	3.91	4.08	97	96
	$\rho = 0.1$	4.55	5.41	1.34	4.06	4.07	93	97
	$\rho = 0.3$	6.82	9.04	4.82	4.22	4.32	87	87
	$\rho = 0.5$	10.2	14.3	10.1	4.20	4.09	68	72
1-exp$\{-\exp(1 + x - c)\}$	$\rho = 0$	-3.47	-3.97	0	-3.97	-3.89	99	93
	$\rho = 0.1$	-2.86	-2.73	1.20	-3.92	-3.81	99	94
	$\rho = 0.3$	-0.63	-0.24	3.16	-3.40	-3.53	99	97
	$\rho = 0.5$	1.56	2.32	5.11	-2.79	-2.95	90	87
1-exp$\{-\exp(1 + x - 2c)\}$	$\rho = 0$	-3.98	-3.98	0	-3.98	-3.96	97	94
	$\rho = 0.1$	-2.99	-2.03	1.27	-3.30	-3.61	99	98
	$\rho = 0.3$	-0.82	1.74	3.56	-1.82	-2.02	85	85
	$\rho = 0.5$	2.24	6.04	6.14	-0.11	0.02	58	55

Table 5.10: MNAR simulation 6

MDM	corr(x, c)	EB	IB	b_{XC}	b_R	$\hat{b}_R$	CR	CR1
$\|sin(1 + x + 2c)\|$	$\rho = 0$	-0.05	-0.07	0	-0.07	0.024	99	96
	$\rho = 0.1$	1.46	1.74	1.66	0.08	-0.22	96	94
	$\rho = 0.3$	4.43	5.34	5.34	0.01	0.04	97	89
	$\rho = 0.5$	9.15	10.3	10.4	-0.03	-0.16	90	90
$\|sin(1 + x + c)\|$	$\rho = 0$	0.38	0.60	0	0.60	0.68	98	98
	$\rho = 0.1$	1.71	2.11	1.69	0.41	0.30	96	97
	$\rho = 0.3$	4.52	5.59	5.35	0.24	0.14	95	96
	$\rho = 0.5$	9.48	10.6	10.5	0.13	0.07	89	87
$\|sin(1 + x - c)\|$	$\rho = 0$	-0.24	-0.57	0	-0.57	-0.58	97	94
	$\rho = 0.1$	0.49	0.84	1.71	-0.87	-0.93	99	96
	$\rho = 0.3$	3.34	4.25	5.63	-1.38	-1.33	94	91
	$\rho = 0.5$	6.56	8.56	10.8	-2.22	-2.49	92	91
$\|sin(1 + x - 2c)\|$	$\rho = 0$	0.02	-0.01	0	-0.01	0.06	97	97
	$\rho = 0.1$	1.08	1.62	1.65	-0.02	-0.05	99	96
	$\rho = 0.3$	4.59	5.32	5.29	0.03	0.01	95	91
	$\rho = 0.5$	8.92	10.3	10.4	-0.08	-0.03	89	88

- When $\text{corr}(x, c) = 0$, the covariate bias is zero as expected, but the MDM bias exists. This is one of the significant differences from the MAR assumption.

- Comparing with b_{XC}, the MDM bias b_R is not large in all simulations. But when there is an interaction influence (M4) or quadratic form (M3), the MDM bias gets considerable large size. In this case, fitting MDM from ignorable missing data assumption or a logistic linear model may not work so well.

- In the simulation, we choose sample size to be large enough to have precise evaluations for two bias components. If the study sample size is too small, the variance of $\hat{\theta}$ will be too large to discover the significant differences. But we should always be aware of the uncertainty issue for the missing data mechanism specification, especially when there is uncertainty in covariates distribution modelling.

Chapter 6

Bias Model Selection for Non-Ignorable Missing Data

6.1 Introduction

The joint density of data D and the missingness R: $f(R, D)$ can be factorized as pattern mixture models $f(D|R)f(R)$ or selection models $f(R|D)f(D)$. Both of them have useful features, and the comparation of the two modelling approachs can be found in e.g. Glynn et al. (1986), Kenward and Molenberghs (1999), Little (1995) and Little and Rubin (2002). As we discussed, inference on pattern mixture models can avoid the non-ignorable missing data mechanism selection process. However, it requires priori knowledge on the distribution of all variables $f(D)$, or rather a model structure assumption on incomplete cases $f(D|R = 0)$. Since these information is very rare, we consider missing data problem in selection models frame in this chapter by specifying the missing data mechanism $f(R|D)$, following the work of Ibrahim et al. (2001), Oakley and O'Hagan (2004) , Molenberghs et al. (2001) and Tang et al. (2003). An explicit parametric model may be built as (Ibrahim et al., 2001)

$$\text{logit}\{h(R = 1|D^{obs}, D^{mis})\} = \psi_0 + \psi_1 D^{obs} + \psi_2 D^{mis}$$

where full specification is necessary under missing not at random (MNAR), and sensitivity analysis is advocated (Horowitz and Manski, 2000) because of lack of identifiability. This can be replaced by a semiparametric selection model (Kim and Yu, 2011)

$$\text{logit}\{h(R = 1|D^{obs}, D^{mis})\} = \omega(D^{obs}) + \psi D^{mis}, \tag{6.1}$$

where $\omega(\cdot)$ is a nonparametric function $\omega(.)$ and ψ is an unknown parameter. This model takes a nonparametric model on the observed part D^{obs} and a simplified form on the missing variable D^{mis}. The discussion in this chapter will use this semiparametric model to specify the missing data mechanism, and analysis will be conducted with the multiple imputation method to fill in missing values in the incomplete cases based conditional distribution $f(D^{mis}|D^{obs}, R = 0)$. Monte Carlo bias model selection method (BMS) will be applied to select a proper value of the bias parameter. We will discuss several specific missing data problems in this section, including mean estimation with non-ignorable missing data, model misspecification for missing data mechanism and regression analysis with missing covariates.

6.2 Mean Estimation with Non-Ignorable Missing Data

Assume that X is continuously distributed with its mean μ as the parameter of interest. The complete data set is $D = (X_{obs}, X_{mis})$ and MDM depends on the missing value itself. Let R be missing data indicator which is equal to 1 if the data is observed or 0 otherwise. Assume that MDM is modelled by a logistic model

$$h(R = 1|X = x) = \text{expit}\{\psi(x + \lambda)\} \tag{6.2}$$

where $\text{expit}(x) = \exp(x)/(1 + \exp(x))$ and λ is assumed to be known as in Tang et al. (2003) (it can be estimated if the proportion of the missing data is known). The parameter of interest is $\mu = E(X)$, the mean of the complete data. It can be expressed by

$$\begin{aligned} \mu &= E_R\left(E_X(X|R)\right) \\ &= E(X|R = 1)h(r = 1) + E(X|R = 0)h(r = 0) \\ &= \pi\mu_1 + (1 - \pi)\mu_2, \end{aligned}$$

where $\pi = h(r = 1)$ is the marginal probability which can be estimated by the observed proportion; μ_1 and μ_2 are the means of the observed data and the missing data respectively. So evaluation of μ_2 is the main task. Using Bayes theorem, we have

$$\begin{aligned} f(x|r=0) &= \frac{h(r=0|x)f(x)}{h(r=0)} \\ &= f(x|r=1)\frac{h(r=1)}{h(r=0)}\frac{h(r=0|x)}{h(r=1|x)}. \end{aligned} \tag{6.3}$$

Denote that $\pi_x = h(r = 0|x)$, then

$$\frac{h(r=0|x)}{h(r=1|x)} = \frac{1-\pi_x}{\pi_x} = \frac{1}{\exp(\psi(x+\lambda))}$$

is the odds of missing when $X = x$. The second equation comes from MDM model (6.2). The mean of the missing data can therefore be expressed by

$$\begin{aligned} \mu_2 &= \int xf(x|r=0)dx \\ &= \int xf(x|r=1)\frac{\pi}{1-\pi}\frac{1}{\exp(\psi(x+\lambda))}dx \\ &= \frac{\pi}{1-\pi}E_{X|R=1}\left[\frac{X}{\exp(\psi(X+\lambda))}\right]. \end{aligned}$$

In this example, ψ is the bias parameter. From the observed data, we are unable to estimate ψ since it depends on the missing data as well.

6.2.1 Simulation Study

We now conduct a simulation study to demonstrate how to use the proposed MC-BMS approach and how it performs. The true values are selected as $\mu = 28.4, \sigma^2 = 19.82$ and $\psi = -0.5, \lambda = -28$ in model (6.2), indicating the average missing proportion is about 51.3%. Sample size of the complete data is 51. In this example λ is assumed to be fixed and ψ is treated as an unknown bias parameter. A MC-BMS approach is designed as follows. We first select a series of ψ, and in this example we simply choose its value from the interval of (-1,1). For each selected ψ, we evaluate the density $f(x|r = 0)$ by (6.3) and then use the density function to sample the missing data, and denote the imputed values as $x_{mis,\psi}$. Thus, $D_\psi = (x_{obs}, x_{mis,\psi})$ forms a simulated complete data set. If the selected bias parameter is close to the 'true value' ψ_{true}, D_ψ should be close to the true complete data set $D = (x_{obs}, x_{mis})$. Since D_ψ and D cannot be compared directly since D involves unobserved data x_{mis}, we further generate a set of $x_{\psi,obs}$ from D_ψ using MDM (6.2) with the given value of ψ. The simulated set of $x_{\psi,obs}$ is comparable with the observed dataset x_{obs}. The closeness of $x_{\psi,obs}$ and x_{obs} is measured by K-nearest neighbour distance. We choose the ψ with the smallest distance.

Usually the sample size of missing data may be not very large (which is about 26 in this simulation study), and one run of the procedure may suffer from sampling error and result in unstable conclusion. Figure 6.1 shows the results with the MC size of 1000. KNN distance takes the minimum at $\psi = -0.53$ when $K = 2$. The corresponding estimate is $\hat{\mu} = 27.5$. We also consider the other values of K. As shown in the same figure, all of them give the similar results although the values of KNN distance is less sensitive to ψ for larger values of K. Discussion on how to choose K can be found in for example Hall et al. (2008) and Nigsch et al. (2006).

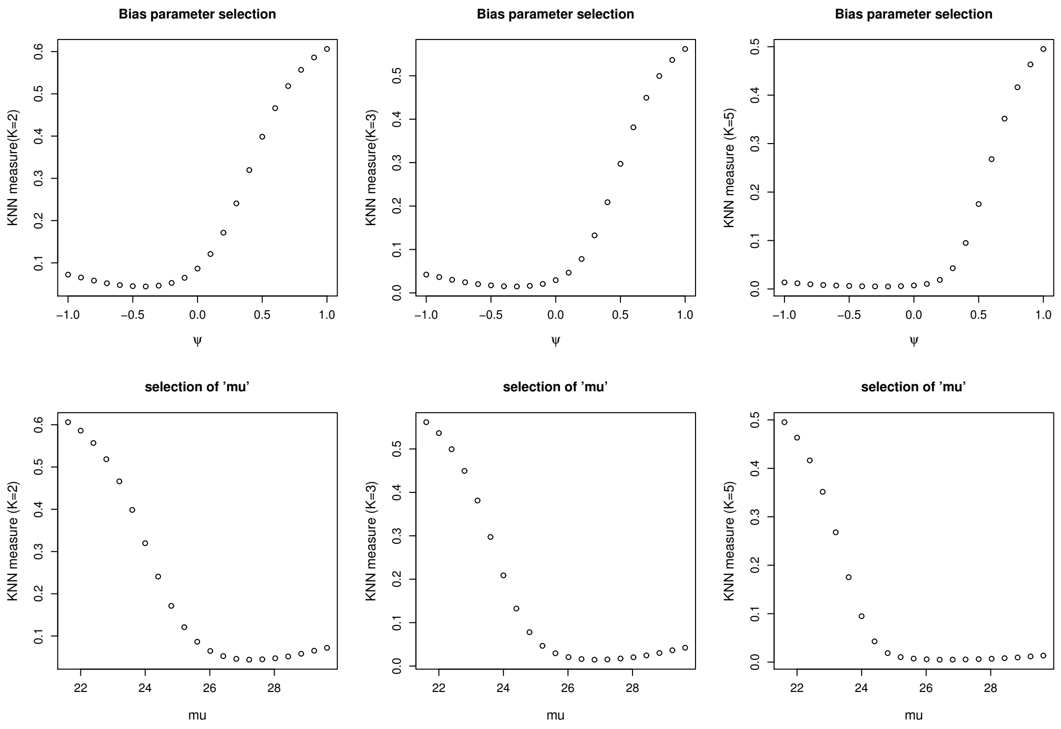


Figure 6.1: Bias parameter selection. Upper panel: KNN distances versus different values of ψ; Lower panel: KNN distance versus the corresponding estimate of μ for the given value of ψ.

We should point out that the approach with smaller value of K is often quite sensitive to the simulated data set D_ψ, we should use a relative large value of MC size in this case.

Table 6.1 presents the simulation study results of 100 replications with different MC sizes and different K's of KNN distance. The average of $\hat{\mu}$ calculated by using observed data only is 25.64. Table 6.1 shows that the MC-BMS approach gives much better results comparing with the true value of $\mu = 28.4$. It also shows that the estimates are quite consistent for different values of K even for small number of MC sizes. Figure 6.2 gives the histograms of the selected ψ with different MC sizes. It suggests that the method with MC size 100 or over usually gives quite robust result.

Table 6.1: Bias model selection: simulation study

	average of $\hat{\mu}$				average of selected ψ			
MC size	10	20	100	1000	10	20	100	1000
K=2	27.84	27.69	27.70	27.67	-0.572	-0.531	-0.541	-0.531
K=3	27.94	27.52	27.34	27.41	-0.572	-0.484	-0.390	-0.391
K=4	28.42	28.05	27.37	27.35	-0.690	-0.596	-0.435	-0.425
K=5	29.23	28.50	27.74	27.85	-0.883	-0.673	-0.512	-0.463

6.3 Regression Models with Non-Ignorable Missing Data

We now consider a regression problem with missing confounders. Let $D = (D^{obs}, D^{mis})$, where D^{obs} denotes the variables that are always observed; while D^{mis} denotes the variables that are totally or partly missing. We still use R as the missing indicator and assume that the MDM (missing data mechanism) depends on both D^{obs} and D^{mis}.

$$\text{logit}\{h(R = 1|D^{obs}, D^{mis})\} = \omega(D^{obs}) + \psi D^{mis}, \tag{6.4}$$

where D^{obs} may include all observed covariates and the observed response variable as well and $\omega(\cdot)$ is a nonparametric function $\omega(.)$ and ψ is an unknown parameter. This model takes a nonparametric model on the observed part D^{obs} and a simplified form on the missing variable D^{mis}.

Based on discussions similar to those around (6.3), we get the following result.

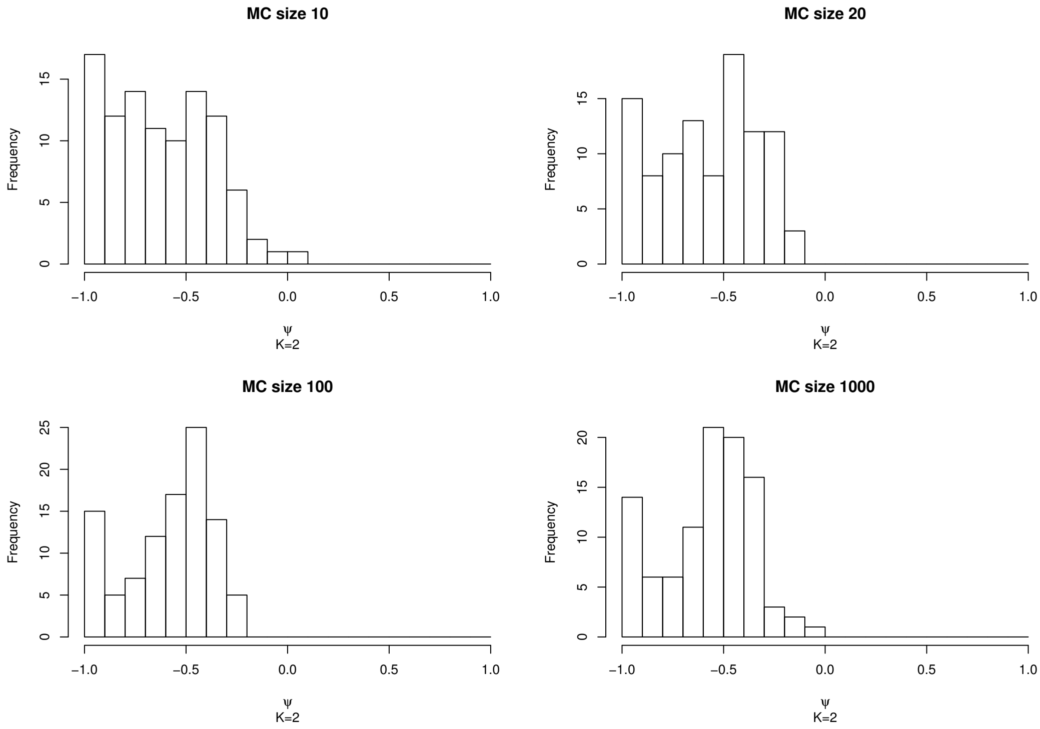


Figure 6.2: Bias parameter selection: histograms of the selected ψ with different MC sizes.

Lemma 6.1. *Let $D = (D^{obs}, D^{mis})$ and D^{obs} is a set of observed variables and D^{mis} is a set of variables with missing data. Let R be the missing data indicator. The conditional distribution of the missing data D^{mis} (i.e. when $R = 0$) given observed data D^{obs} is given by*

$$f(D^{mis}|D^{obs}, R = 0) = f(D^{mis}|D^{obs}, R = 1)\frac{Q(D)}{E(Q(D)|D^{obs}, R = 1)}, \qquad (6.5)$$

where

$$Q(D) = \frac{h(R = 0|D)}{h(R = 1|D)} \qquad (6.6)$$

is the conditional odds of nonresponse with $h(r|D)$ as the missing data mechanism.

The proof is given in Appendix 6.5.1.

The expression (6.5) gives the distribution of missing data given the observed data. It is the key to estimate parameters of interest with non-ignorable missing data. When this model can be determined, the parameters of interest can be estimated. We now apply the lemma to the semiparametric logistic regression model to fit MDM. Rewrite

(6.4) as

$$\pi_D = h(R=1|D) = \text{expit}(\omega(D^{obs}) + \psi D^{mis}). \tag{6.7}$$

Note that the component on observed part $\omega(D^{obs})$ will disappear in the fraction at equation (6.5). The formula is simplified as

$$f(D^{mis}|D^{obs}, R=0) = f(D^{mis}|D^{obs}, R=1)\frac{\exp(-\psi D^{mis})}{E(\exp(-\psi D^{mis})|D^{obs}, R=1)}. \tag{6.8}$$

The parameter ψ is considered as tilting parameter that determines the amount of departure from the ignorability of the MDM.

In formula (6.5) or in (6.8), we need two models to compute the conditional distribution of missing data: $f(D^{mis}|D^{obs}, R=1)$ and $h(R=1|D^{obs}, D^{mis})$. A consistent estimate of $f(D^{mis}|D^{obs}, R=1)$ can be parametrically fitted such as a conditional logistic model by Sinha et al. (2005) or nonparametrically fitted with a kernel estimator as discussed by Kim and Yu (2011). Thus the only uncertainty in formula (6.8) is the parameter ψ. This is the bias parameter which cannot be estimated from the observed data. We usually use a sensitivity analysis method to study how estimation of the parameter of interest depends on ψ or the associated interpretable quantities (see e.g. Kim and Yu, 2011).

Here we use the MC-BMS method discussed in Chapter 4 to select the most plausible value of ψ. We first choose a value of ψ, and then simulate missing data from (6.5) or (6.8). The simulated data are imputed to form a simulated complete data and then a subset $D^{\psi,obs}$ is resampled based on MDM with the given ψ and the simulated complete data. $D^{\psi,obs}$ is compared with the true observed data D^{obs} using the nearest neighbour distance. To eliminate sampling error, we used average distance calculated from repeated $D^{\psi,obs}$. As we suggested in the previous section, we usually use the MC size of 100. The details will be illustrated by two examples discussed in the following subsections.

6.3.1 Fuel Consumption Data Example

We now consider a missing not at random problem based on Fuel consumption data and let income (X_2) be partly missing with probability $h(R=0|D) = 1 - \text{expit}(1 + (x_1 - \bar{x}_1) - 0.5(x_2 - \bar{x}_2))$, where r is the missing data indicator and $\bar{x} = E(x)$. This model is used to simulate data in this example.

In our MC-BMS method, we use the following semiparametric MDM model:

$$h(r = 1|t, x_1, x_2, x_3, x_4) = \text{expit}(\omega(t, x_1, x_3, x_4) + \psi x_2). \tag{6.9}$$

Applying equation (6.8) to this example, we impute the missing values from:

$$f(x_2|t, x_1, x_3, x_4, r = 0) = f(x_2|t, x_1, x_3, x_4, r = 1)\frac{\exp(-\psi x_2)}{E(\exp(-\psi x_2)|t, x_1, x_3, x_4, r = 1)}.$$

We use a normal distribution to fit the conditional distribution:

$$(x_2|t, x_1, x_3, x_4, r = 1) \sim N(\gamma_0 + \gamma_1 t + \gamma_2 x_1 + \gamma_3 x_3 + \gamma_4 x_4, \tau^2).$$

Considering that X_2 is the personal income, normal distribution seems a reasonable assumption.

Now we can simulate a 'complete' dataset $D_\psi = (T, X_1, X_2^*, X_3, X_4)$ for each given ψ, and we can estimate parameter $(\hat{\theta}_\psi, \hat{\sigma}_\psi^2)$ from a linear regression model with dataset D_ψ. To conduct a stable selection step by using KNN distance, we use 'bootstrapping residuals' method to obtain $D_\psi^* = (T^*, X_1, X_2^*, X_3, X_4)$, where T^* is re-sampled conditional on the following linear regression model with the estimates $(\hat{\theta}_\psi, \hat{\sigma}^2)$ and imputed covariates $X^* = (X_1, X_2^*, X_3, X_4)$:

$$t^*|x_1, x_2^*, x_3, x_4 \sim N(\hat{\theta}_\psi^T x^*, \hat{\sigma}_\psi^2).$$

So T^* is simulated by adding residuals on the predicted values, where the residuals are sampled from a normal distribution $N(0, \hat{\sigma}_\psi^2)$. We then calculate the distance between $D_\psi^{o*} = (T^*, X_1, X_3, X_4)$ and its associated observed data set $D^{obs} = (T, X_1, X_3, X_4)$. We still use the average distance with MC size of 100 to eliminate sampling errors.

In this example, we choose ψ in $(-5, 5)$ with interval of 0.2. Figure 6.3 shows the KNN distances with $K = 2$ against the values of ψ. It achieves minimum at $\psi = -0.6$ and we consider it as the 'most plausible' value of ψ. The corresponding estimates are very close to the ones obtained from the complete data; see the results in Table 6.2.

As comparison, we also considered the MCSA by Greenland (2005) and BSA by McCandless et al. (2007) (see the detailed procedure in Appendix 6.5.2). Table 6.2 shows the simulation results with 100 replications. Estimation based on MAR assumption is also listed. MCSA method works well except the estimation of θ_1 and θ_2, this may be because the missingness depends on x_1 and x_2 but it is usually not easy to give a good prior distribution for bias parameter ψ. BSA gives an even worse result indicating the uniform prior U(-5, 5) is not a good choice. Overall the MC-BMS method gives a much better result than the others. All the estimates are very close to the estimates calculated from the complete data.

Table 6.2: Simulation study for fuel consumption data

	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$
Complete data	154.19	18.55	-6.14	0.472	-4.228
MAR	128.62	33.65	-10.90	0.440	-4.843
MCSA	153.60	22.87	-3.44	0.494	-4.292
BSA	131.51	31.77	-10.17	0.445	-4.753
MC-BMS	152.33	18.24	-3.99	0.493	-4.394

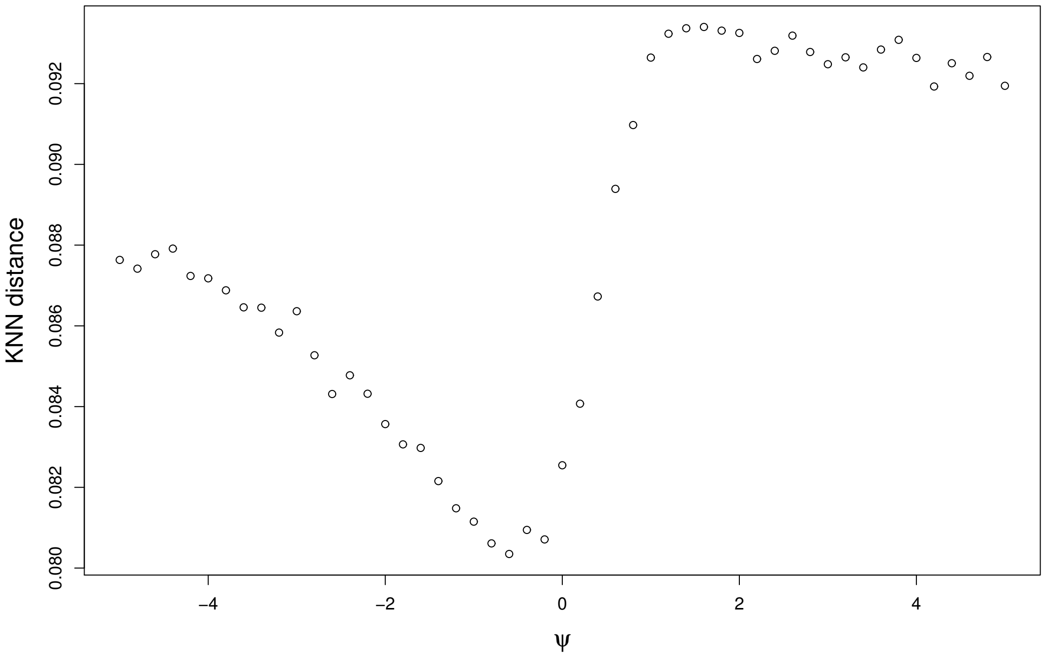


Figure 6.3: Selection of bias parameter ψ for fuel consumption data: KNN (K=2) distance versus values of ψ.

6.3.2 Simulation Studies for Misspecified Models

Now we conduct a simulation study to further examine the proposed MC-BMS method for non-ignorable missing covariates. We consider a complete data example D=(T, X, C) from the following linear regression model:

$$(t|x, c) \sim N(\alpha + \theta x + \beta c, \sigma^2), \tag{6.10}$$

where C is partly missing. The observed data is $D_{obs} = (T, X, C_{obs})$, meaning that all of (T, X, C) are observed for the complete cases $(R = 1)$ and only (T, X) are observed for the missing cases $(R = 0)$. We used a semiparametric MDM model in this example:

$$h(r = 1|t, x, c) = \text{expit}(\omega(t, x) + \psi c). \tag{6.11}$$

Here ψ is the bias parameter and it is inestimable since it depends on the missing values of c. The conditional distribution of the missing variable C given (T, X) can be derived by using Lemma 6.1, which is

$$f(c|t, x, r = 0) = f(c|t, x, r = 1)\frac{\exp(-\psi c)}{E[\exp(-\psi c)|t, x, r = 1]}. \tag{6.12}$$

The conditional distribution of c for the complete cases $f(c|t, x, r = 1)$ is fitted by a normal distribution:

$$c_{obs}|(t_{obs}, x_{obs}) \sim N(\gamma_0 + \gamma_1 x_{obs} + \gamma_2 t_{obs}, \tau_c^2).$$

The unknown parameters $(\gamma_0, \gamma_1, \gamma_2, \tau_c^2)$ are estimated from D_{obs}.

In the simulation study, the true values of the parameters are $(\alpha, \theta, \beta) = (1, 1, 1)$, and σ^2 takes value from a uniform distribution in $(0.16, 1)$. Covariates variables (X, C) are assumed to be continous distributed, with $X \sim U(0, 2)$ and $(C|X) \sim U(\rho\frac{\sigma_c}{\sigma_x}x + 1, \rho\frac{\sigma_c}{\sigma_x}x + 4)$ with corr(x, c) taken as either 0.5 or -0.5. We considered the following MDM models in four different scenarios:

S1. Logit Linear: $h(r = 1|x, c) = \text{expit}(2 - 0.6c + 0.2x^2)$, $\text{corr}(x, c) = 0.5$;

S2. Logit Interaction: $h(r = 1|x, c) = \text{expit}(-1 + c + x - xc)$, $\text{corr}(x, c) = -0.5$;

S3. Logit Quadratic: $h(r = 1|x, c) = \text{expit}(3 - 0.3c^2)$, $\text{corr}(x, c) = 0.5$;

S4. Log Log Linear: $h(r = 1|x, c) = 1 - \exp\{-\exp(0.5 - c + x)\}$, $\text{corr}(x, c) = -0.5$.

Note that the true mechanism may have the interaction or quadratic component, but the fitting model (6.11) has no consideration on it and can be biased. This study also aims to show how MC-BMS method performs when the MDM is misspecified.

For each given ψ, we first generate $c^{\psi,mis}$ from (6.12), and then estimate (α, θ, β) and σ^2 using the simulated 'complete' data set with c_{mis} imputed by $c^{\psi,mis}$. We then use the estimates to generate a new set of complete data $D^*_\psi = \{(t^*_i, x_i, c^*_i), i = 1, \ldots, n\}$ where c^*_i takes either the observed data $c_{obs,i}$ or the imputed data $c_{\psi,mis,i}$, and t^*_i is generated using model (6.10) with x_i and c^*_i by adding a 'bootstrapping residual' (i.e. the one generated from $N(0, \sigma^2_\psi)$). The MC-BMS approach is to compare $D^*_{\psi 1} = \{(t^*_i, x_i), i = 1, \ldots, n\}$ with $D_{obs1} = \{(t_i, x_i), i = 1, \ldots, n\}$ using a KNN distance.

To consider the comparison we also used the model with MAR assumption, i.e. using (6.11) without the item of ψc. Table 6.3 presents the average values of the estimates via 100 replications. The values of RMSE (root mean squared error) are listed in brackets. It shows that the MC-BMS performs very well even for S2 to S4. The MDM (6.11) we used in MC-BMS is actually misspecified in S2 to S4. In S2 it ignores the interaction; S3 (6.11) uses a linear predictor for c with logistic link function but the actual one is nonlinear; while in S4 different link function is used. However the selected bias model using MC-BMS still give quite good results. In all the scenarios, MC-BMS performs better than the model with MAR assumption.

Table 6.3: Simulation study: average estimates and RMSE (in brackets)

	MC-BMS			MAR		
	α	θ	β	α	θ	β
True	1	1	1	1	1	1
S1	0.993 (0.130)	0.993 (0.075)	1.010 (0.046)	1.003 (0.521)	0.988 (0.493)	1.013 (0.047)
S2	0.884 (0.494)	1.129 (0.275)	0.977 (0.165)	0.949 (0.620)	1.191 (0.732)	0.981 (0.143)
S3	0.739 (0.386)	0.981 (0.116)	1.113 (0.162)	0.732 (0.368)	0.992 (0.504)	1.129 (0.172)
S4	0.919 (0.503)	1.038 (0.246)	1.007 (0.199)	1.378 (1.010)	0.834 (0.408)	1.038 (0.201)

6.4 Discussion

In this Chapter, we were concerned with the sensitivity analysis for non-ignorable missing data problems under the selection models framework. The missing data mechanism is specified as a semiparametric model, with bias parameter ($\eta = \psi$) evaluated through Monte Carlo bias model selection (MC-BMS) method. Given a value of η, we make imputation on D^{mis} (when $R = 0$) to obtain a complete dataset D_η. How to generate D^{mis} from a bias model is the key step of MC-BMS approach. Several examples are demonstrated in Section 6.2 and 6.3, in which Lemma 6.1 plays a key role. The detailed technique has been reported for those examples, and they can be extended to other missing data problems.

The 'closeness' of the simulated data and the observed data is measured by the distance between the two sets of samples. We have tried a variety of distances and found that KNN distance is proper for the approach. The advantage has also been discovered in previour chapters.

Mean function example has been discussed in sensitivity analysis for many years (see e.g. Rubin, 1987; Daniels and Hogan, 2008), and it is always difficult to calculate the sample mean since the concealed observations are unknown. But MC-BMS method performs very well for this non-ignorable missing data problem. And we noticed although the results can be slightly different on different values of K (parameter in KNN) and Monte Carlo size, the method actually performs quite robustly.

We further applied the MC-BMS method in regression models under non-ignorable missing covariates. We used a semiparametric model and keep the dimension of bias parameters low. As we discussed before the key of success is to find how we can simulate D^{mis} from $f(D^{mis}|D^{obs}, R = 0)$. We used the formula in Lemma 6.1 and used a linear regression model to fit $f(D^{mis}|D^{obs}, R = 1)$ in our examples. This can certainly be improved. Since the fit for $f(D^{mis}|D^{obs}, R = 1)$ involves no missing data, many parametric or nonparametric methods can be used. The MC-BMS method works robustly and it always make a proper vote on selection of the 'best' from plausible values. This method is indeed very flexible and useful, it can be extended into many other missing data problems and uncertainty analysis.

6.5 Appendix

6.5.1 Proof of Lemma 6.1

Using Bayes Theorem, we have

$$f(D^{mis}|D^{obs}, R = 0) = \frac{h(R = 0|D)f(D^{mis}|D^{obs})}{h(R = 0|D^{obs})}.$$

Similarly, we have

$$f(D^{mis}|D^{obs}, R = 1) = \frac{h(R = 1|D)f(D^{mis}|D^{obs})}{h(R = 1|D^{obs})}.$$

This leads to the following equation

$$f(D^{mis}|D^{obs}, R = 0) = f(D^{mis}|D^{obs}, R = 1)\frac{h(R = 0|D)}{h(R = 1|D)}\frac{h(R = 1|D^{obs})}{h(R = 0|D^{obs})}$$

Let $Q(D)$ be the one defined in (6.6), then we have

$$\begin{aligned}
E(Q(D)|D^{obs}, R = 1) &= \int \frac{h(R = 0|D^{obs}, D^{mis})}{h(R = 1|D^{obs}, D^{mis})} f(D^{mis}|D_0, R = 1)dD^{mis} \\
&= \int \frac{f(D^{mis}|D_0, R = 1)}{h(R = 1|D^{obs}, D^{mis})} h(R = 0|D^{obs}, D^{mis})dD^{mis} \\
&= \int \frac{f(D^{mis}|D^{obs})}{h(R = 1|D^{obs})} h(R = 0|D^{obs}, D^{mis})dD^{mis} \\
&= \int f(D^{mis}|D^{obs}, R = 0)\frac{h(R = 0|D^{obs})}{h(R = 1|D^{obs})} dD^{mis} \\
&= \frac{h(R = 0|D^{obs})}{h(R = 1|D^{obs})}.
\end{aligned}$$

This proves the Lemma.

6.5.2 BSA Details used in Section 6.3.1

Assume that the prior of ψ is a uniform distribution, then Bayesian Sensitivity Analysis (BSA) can be conducted by the following Gibbs sampler (McCandless et al., 2007):

1. Obtain a reasonably starting value for (θ, ψ);
2. For j=1, 2 . . . , sample $D_{mis}^{(j)}$ from its conditional distribution in (6.5) given $\theta^{(j-1)}$ and $\psi^{(j-1)}$;
3. Sample $\theta^{(j)}$ using a Metropolis Hastings step with target density $f(\theta|D_{obs}, D_{mis}^{(j)})$ and proposal distribution obtained by regression model of response variable on covariates;
4. Sample $\psi^{(j)}$ using a Metropolis Hastings step with target density $f(\psi|D_{obs}, D_{mis}^{(j)})$ and proposal distribution obtained by semiparametric MDM model.

Discard a suitable number of initial iterations, and the sequence $(\theta^{(j)}, \psi^{(j)})$ comprise a sample from the required posterior distribution.

Chapter 7

Robust Confidence Interval with Missing Data in Meta Analysis

7.1 Introduction

Meta-analysis is frequently used in medical research to estimate the overall effect of an experience or exposure towards the risk of diseases. For example, Longnecker et al. (1988) reviewed the multiple studies on the association between alcohol consumption and risk of breast cancer, and further discussion of allowing the correlation between estimated log-odds ratios was considered by Greenland and Longnecker (1992) and publication bias problem was considered by Shi and Copas (2004). Fixed-effects model and random-effects model are two widely used procedures. Various type of confidence intervals (CI) for treatment effect have been proposed for those two models. The discussion of identifying a proper CI for meta analysis averaged effect size has continued for decades, good literature includes DerSimonian and Laird (1986) which used random-effects models with a normal distribution assumed for between study effects; Hardy and Thompson (1996) which used a likelihood method; and Sidik and Jonkman (2002) which used the odds of two chi-square distributed statistics as a t-test ratio being expected to work well specially for small sample size trials. More recently, Henmi and Copas (2010) centred the confidence interval on a fixed-effects estimator, *but allow for heterogeneity by including an assessment of the extra uncertainty induced by the random-effects setting.* They found that this method, namely

HC method, is more robust than the others particularly when there is publication bias. We use the idea of the HC confidence interval but extended into meta-regression analysis with trend estimation.

The rest of the chapter is arranged as follows. In Section 7.2, we first introduce fixed-effects model and random-effects model in meta regression analysis and we will review several commonly used confidence intervals including DerSimonian-Laird method(DL), Likelihood ratio method (LR), Restricted maximum likelihood method (RM) and Sidik and Jonkman's method (SJ). Section 7.3 will give the detailed discussion on how to extend Henmi and Copas's method (HC) to meta regression model. The variance of effect size is evaluated by an approximated gamma distribution. A bootstrap method is also presented in this section. In section 7.4, we conduct simulation studies on cases without and with publication bias. The comparison between all discussed methods is presented. We further consider missing confounder problems in Section 7.5. Conclusion will be made in Section 7.6.

7.2 Meta Regression Model and Confidence Intervals

We consider a meta-analysis model for trend estimation with heterogeneity in this section, but the results can be easily extended to a general multi-level regression model.

For a meta-analysis with m studies, a model for trend estimation is defined as follows (see the details in Shi and Copas, 2004). For the i-th study,

$$\boldsymbol{t}_i = \theta_i \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, m \tag{7.1}$$

where $\boldsymbol{t}_i = (t_{i1}, \dots, t_{in_i})^T$ and the notations $\boldsymbol{x}_i$ and $\boldsymbol{\epsilon}_i$ are defined accordingly. The response variable t_{ij} in dose-analysis is usually a log-odds ratio for a group with dosage x_{ij} against a control group. When sample sizes are not very small, t_{ij} has an approximate normal distribution. However, since t_{ij}'s are calculated from groups with different dose-levels against the same control group, they are not independent. The error items $\boldsymbol{\epsilon}_i$ have zero means and the covariance matrix $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Omega}_i$. The covariance matrix can be calculated from the original data (see Shi and Copas, 2004).

If we consider a fixed-effects model, i.e. $\theta_i = \theta$ in (7.1), we can easily get an estimate of θ by using either least square or maximum likelihood method, which is

$$\hat{\theta}_F = \frac{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i}{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i}. \tag{7.2}$$

We call it a fixed-effects estimate.

If we consider a random-effects model, we can further assume that

$$\theta_i \sim N(\theta,\ \tau^2). \tag{7.3}$$

If both $\boldsymbol{\Omega}_i$ and τ^2 are given, the estimate of θ is given by (using either least square or maximum likelihood method, although the latter needs normal assumptions)

$$\hat{\theta}_R = \frac{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{t}_i}{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{x}_i}, \tag{7.4}$$

where

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Omega}_i + \tau^2 \boldsymbol{x}_i \boldsymbol{x}_i^T.$$

To estimate τ^2, we may consider the following Q-statistics

$$Q = \sum_{i=1}^m (\boldsymbol{t}_i - \hat{\theta}_F \boldsymbol{x}_i)^T \boldsymbol{\Omega}_i^{-1} (\boldsymbol{t}_i - \hat{\theta}_F \boldsymbol{x}_i), \tag{7.5}$$

where $\hat{\theta}_F$ is the estimate from a fixed-effects model given by (7.2). The DerSimonian-Laird estimate is given by

$$\hat{\tau}^2 = \max\left\{0,\ \frac{Q-(N-1)}{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i - \frac{\sum_{i=1}^m (\boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i)^2}{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i}}\right\} \tag{7.6}$$

where $N = \sum_{i=1}^m n_i$.

We now use some conventional approaches to construct confidence intervals for θ, the parameter of interest in meta-analysis and dose-analysis model (7.1), and will propose a new one in the next section.

From (7.2) it is easy to know that $\mathrm{Var}(\hat{\theta}_F) = 1/\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i$, thus approximately

$$Z = \frac{\hat{\theta}_F - \theta}{(\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i)^{-1/2}} \sim N(0,1).$$

The confidence interval with confidence level of $1-\alpha$ constructed from the fixed-effects model is

$$\left(\hat{\theta}_F - z_{\alpha/2}(\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i)^{-1/2},\ \hat{\theta}_F + z_{\alpha/2}(\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i)^{-1/2}\right) \tag{7.7}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution. We call this a fixed-effects (FE) confidence interval.

7.1 DerSimonian-Laird Method (DL)

From (7.4) we get that $\mathrm{Var}(\hat{\theta}_R) = 1/\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{x}_i$, thus the confidence interval based on the random-effects model can be constructed similarly to the one for the fixed-effects model. This leads to the following result

$$\left(\hat{\theta}_R - z_{\alpha/2}(\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{x}_i)^{-1/2},\ \hat{\theta}_R + z_{\alpha/2}(\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{x}_i)^{-1/2}\right) \tag{7.8}$$

for level $1-\alpha$, where τ^2 is evaluated by the DerSimonian-Laird estimate given in (7.6). We therefore call it the DerSimonian-Laird method.

7.2 Likelihood Ratio Method (LR)

Given m studies, the log-likelihood for (θ, τ^2) is expressed by

$$l(\theta, \tau^2) = \frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{m} \log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m} (\boldsymbol{t}_i - \theta\boldsymbol{x}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{t}_i - \theta\boldsymbol{x}_i), \tag{7.9}$$

where $N = \sum n_i$. The profile log-likelihood for θ is therefore

$$l_p(\theta) = l(\theta, \hat{\tau}^2_{ML}(\theta)), \tag{7.10}$$

where $\hat{\tau}^2_{ML}(\theta)$ is the maximizer of (7.9) given θ.

Let $\hat{\theta}_{ML}$ and $\hat{\tau}^2_{ML}$ be the maximum likelihood calculated from (7.9), then we have the following approximation result

$$-2\left(l_p(\theta) - l(\hat{\theta}_{ML}, \hat{\tau}^2_{ML})\right) \sim \chi^2_1.$$

This results in the following LR confidence interval

$$\left\{\theta : l_p(\theta) > l(\hat{\theta}_{ML}, \hat{\tau}^2_{ML}) - \frac{1}{2} z^2_{\alpha/2}\right\}. \tag{7.11}$$

There is no analytical form so a numerical method should be used. The right-hand side of the inequality is a constant. So the confidence interval can be constructed based on the profile log-likelihood given in (7.10).

7.3 Restricted Maximum Likelihood Method (RM)

The profile log-likelihood for τ^2 is given by

$$l_p(\tau^2) = l(\hat{\theta}_{ML}(\tau^2), \tau^2),$$

where $\hat{\theta}_{ML}(\tau^2)$ is the maximizer of (7.9) given τ. It has an analytical form given by (7.4). The restricted maximum likelihood estimate of $\hat{\tau}^2_{RE}$ is the one calculated by maximizing the above profile likelihood. We then construct a confidence interval by (7.8) but evaluated at $\tau^2 = \hat{\tau}^2_{RE}$. This is called the restricted maximum likelihood method.

7.4 Sidik and Jonkman's method (SJ)

Using the fact that

$$\sum_{i=1}^{m} (\boldsymbol{t}_i - \hat{\theta}_R \boldsymbol{x}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{t}_i - \hat{\theta}_R \boldsymbol{x}_i) \sim \chi^2_{N-1} \ \text{ approximately}$$

we know that

$$\frac{(\hat{\theta}_R-\theta)/(\sum_{i=1}^m \boldsymbol{x}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}_i)^{-1/2}}{\sqrt{\sum_{i=1}^m(\boldsymbol{t}_i-\hat{\theta}_R\boldsymbol{x}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{t}_i-\hat{\theta}_R\boldsymbol{x}_i)/(N-1)}}$$

has a t-distribution with $N-1$ degrees of freedom. This leads to the following approximate confidence interval

$$\hat{\theta}_R \pm t_{N-1,\alpha/2}\sqrt{\frac{\sum_{i=1}^m(\boldsymbol{t}_i-\hat{\theta}_R\boldsymbol{x}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{t}_i-\hat{\theta}_R\boldsymbol{x}_i)}{(N-1)\sum_{i=1}^m \boldsymbol{x}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}_i}}$$

where $t_{N-1,\alpha/2}$ is the upper $\alpha/2$ quantile of the related t-distribution.

7.3 Extension of HC methods

The basic idea of Henmi and Copas (2010) is to construct a confidence interval centred on a fixed-effects estimate although the model we are using is a random-effects model. They argued that the method is more robust than the conventional one particularly when there is publication bias. We extend the method to the trend estimation model (7.1) in this section and will investigate if this will also provide a robust result in meta-regression analysis. The results can be applied directly to a general multi-level model. A new bootstrap method is also proposed.

7.3.1 Trend Estimation in Meta-Analysis

We start our derivation from the fixed-effects estimate (7.2), i.e.

$$\hat{\theta}_F=\frac{\sum_{i=1}^m \boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{t}_i}{\sum_{i=1}^m \boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i}=\frac{\sum_{i=1}^m \boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{t}_i}{W_1}.$$

Here we define the following notations

$$W_1=\sum_{i=1}^m \boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i, \quad \text{and} \quad W_j=\frac{\sum_{i=1}^m(\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i)^j}{W_1},\ j=2,3,4. \tag{7.12}$$

Bear in mind that $\boldsymbol{t}_i$ follows a random-effects model (7.1) and (7.3), the variance of $\hat{\theta}_F$ is calculated by

$$\begin{aligned} s_{HC}^2 = \mathrm{Var}(\hat{\theta}_F) &= \frac{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1}\{\boldsymbol{\Omega}_i + \tau^2 \boldsymbol{x}_i \boldsymbol{x}_i^T\}\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i}{W_1^2} \\ &= \frac{1+\tau^2 W_2}{W_1}. \end{aligned} \tag{7.13}$$

We are interested in the following quantity:

$$Z_{HC} = \frac{\hat{\theta}_F - \theta}{\hat{s}_{HC}}, \tag{7.14}$$

where $\hat{s}_{HC}$ is given by (7.13) with τ^2 replaced by $\hat{\tau}^2$ in (7.6), the DerSimonian-Laird estimate, which can be rewritten as

$$\hat{\tau}^2 = \max\left\{0,\ \frac{Q-(N-1)}{W_1 - W_2}\right\}.$$

Thus, the quantity Z_{HC} is expressed by

$$Z_{HC} = \begin{cases} \frac{V}{f(Q)} & \text{if } Q \geq N-1; \\ \\ V & \text{if } Q < N-1, \end{cases} \tag{7.15}$$

where Q is given by (7.5),

$$V = \frac{\sum_{i=1}^m \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1}(\boldsymbol{t}_i - \theta\boldsymbol{x}_i)}{\sqrt{W_1}}. \tag{7.16}$$

and

$$f(Q) = \sqrt{1 + \frac{W_2(Q-(N-1))}{W_1 - W_2}}.$$

To construct confidence interval of θ from (7.14) or (7.15), we need to derive the distribution of Z_{HC} or the quantile of the distribution. It can be calculated by the following formula through a conditional distribution given V.

$$P(Z_{HC} \leq z) = \begin{cases} 1 - \int_z^\infty P(Q \leq f^{-1}(\frac{v}{z})|V = v)f(v)dv & \text{if } z \geq 0; \\ \\ \int_{-\infty}^z P(Q \leq f^{-1}(\frac{v}{z})|V = v)f(v)dv & \text{if } z < 0, \end{cases} \tag{7.17}$$

where $f(v)$ is the density function of V, which is the normal distribution with zero mean and variance $(1 + \tau^2 W_2)$.

The conditional distribution of Q given V can be approximated by a Gamma distribution following the argument given in Henmi and Copas (2010). The conditional mean and variance are given respectively by

$$\begin{aligned} \mu_v &= \mathrm{E}(Q|V = v) \\ &= (N-1) + \tau^2(W_1 - W_2) + \tau^4 d(W_3 - W_2^2); \qquad (7.18) \\ \sigma_v^2 &= \mathrm{Var}(Q|V = v) \\ &= 2(N-1) + 4\tau^2(W_1 - W_2) + 2\tau^4(W_1W_2 - 2W_3 + W_2^2) + 4\tau^4 d(W_3 - W_2^2) \\ &\quad +4\tau^6 d(W_4 - 2W_2W_3 + W_2^3) + 2\tau^8(d^2 - d_1^2)(W_3 - W_2^2)^2, \qquad (7.19) \end{aligned}$$

where

$$d = d_1 - (1 + \tau^2 W_2)^{-1} \text{ and } d_1 = (1 + \tau^2 W_2)^{-2} v^2. \tag{7.20}$$

The proof of (7.18) is given in Appendix 7.7.1 and (7.19) in Appendix 7.7.2. Thus,

$$P\left(Q \leq f^{-1}(\frac{v}{z})|V = v\right) \approx g(z, v) = \mathit{\Gamma}\left(f^{-1}(\frac{v}{z});\ \frac{\sigma_v^2}{\mu_v}, \frac{\mu_v^2}{\sigma_v^2}\right),$$

where $\mathit{\Gamma}(x; a, b)$ is the cumulative distribution function of the Gamma distribution $\mathit{\Gamma}(a, b)$. From (7.17), the $1-\alpha/2$ quantile, $z_{1-\alpha/2}$, of Z_F is the solution of the following equation

$$\int_z^\infty g(z, v)f(v)dv = \frac{\alpha}{2}. \tag{7.21}$$

Similarly, the $\alpha/2$ quantile, $z_{\alpha/2}$, is the solution of the following equation

$$\int_{-\infty}^z g(z, v)f(v)dv = \frac{\alpha}{2}. \tag{7.22}$$

The solutions can be calculated numerically.

The confidence interval of θ with level $1-\alpha$ is given by

$$(\hat{\theta}_F + z_{\alpha/2}\hat{s}_{HC},\ \hat{\theta}_F + z_{1-\alpha/2}\hat{s}_{HC}). \tag{7.23}$$

7.3.2 Bootstrap Methods (BS)

An alternative way to calculate quantiles of Z_{HC} in (7.14) or (7.15) is bootstrap (Efron and Tibshirani, 1993), which is a popular nonparametric way in mean estimation, variance evaluation and confidence interval construction in regression analysis. The resampling technique is considered either parametrically or nonparametrically with sampling lots of replicated new sample from a replacement allowed bootstrapping pairs or bootstrapping residuals (see Chapter 9 Efron and Tibshirani, 1993).

We use the idea introduced in Noortgate and Onghena (2005), bootstrap samples are obtained by resampling residuals and random effects from the related parametric models. Suppose that the unknown parameters θ and τ have been estimated, the procedure used to calculate quantiles of Z_{HC} is described as follows.

1. Draw a set of random errors $\boldsymbol{\epsilon}_i^*$ from the n_i-dimensional normal distribution $N(\mathbf{0}, \hat{\boldsymbol{\Omega}}_i)$ for $i = 1, \ldots, m$;

2. Draw a random effect $\hat{\theta}_i^*$ from the normal distribution $N(\hat{\theta}, \hat{\tau}^2)$ for $i = 1, \ldots, m$;

3. Use the samples generated in Steps 1 and 2 to obtain samples of response variable $\boldsymbol{t}_i^*$ for $i = 1, \ldots, m$ using (7.1) and the original covariates.

4. Use the generated data set to estimate unknown parameters θ and τ and calculate the value of z_{HC} using formula (7.14) or (7.15).

The procedure is usually repeated a large number of times and a set of samples of z_{HC} are calculated based on the bootstrap samples. The numerical 2.5% quantiles and 97.5% quantiles can be used to replace the theoretical results given in (7.21) and (7.22) to construct 95% bootstrap confidence interval. Based on the large sample theory for bootstrap, the numerical bootstrap quantile converges to the true quantile when the bootstrap sample size is sufficiently large.

7.4 Simulation Study with Publication Bias

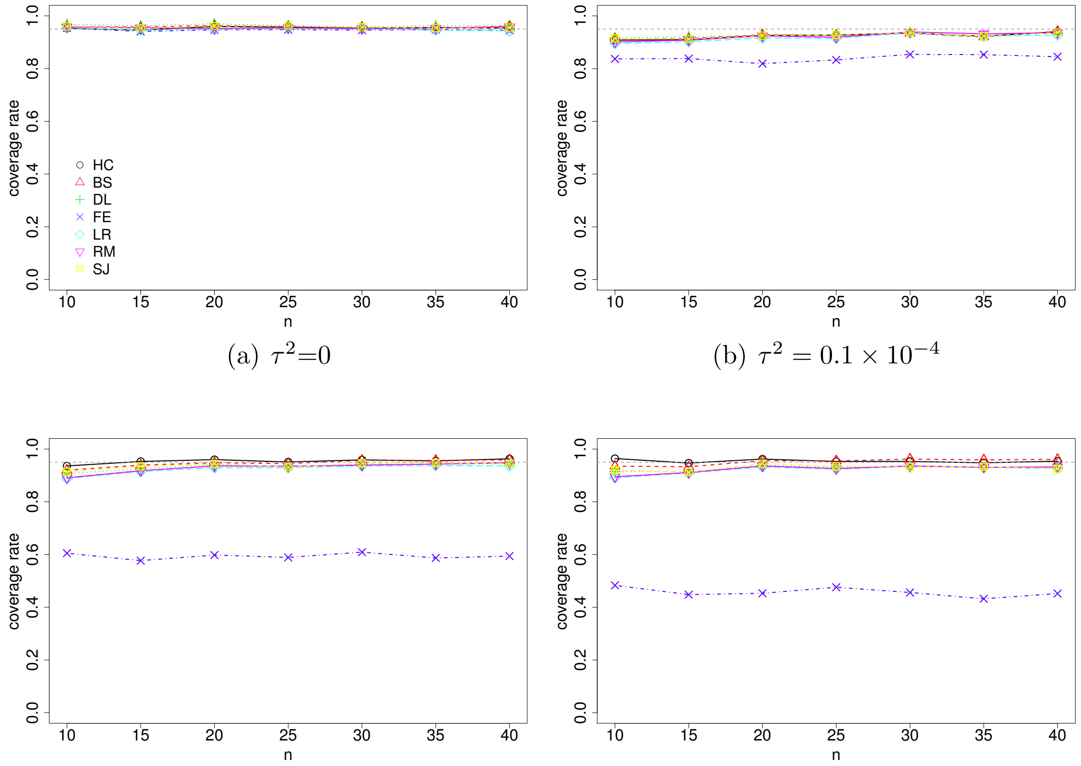
We conduct simulation studies based on the alcohol consumption and breast cancer example discussed in Shi and Copas (2004). It includes 14 studies (two studies are removed from the original sixteen studies due to difficulty of extracting raw data). We choose n studies randomly with replacement, where values for exposure variable $\boldsymbol{x}_i$ are taken from the original data, but the response variable $\boldsymbol{t_i}$ (the log-odds ratio between cases and controls) is generated by (7.1) with random effects generated from $\theta_i \sim N(\theta, \tau^2)$ with $\theta = 0.01$ and $\tau^2 \in \{(0, 0.1, 0.5, 1) \times 10^{-4}\}$. Residuals $\boldsymbol{e_i}$ is generated from the normal distribution with $N(\boldsymbol{0}, \hat{\boldsymbol{\Omega}}_{\boldsymbol{i}})$, and $\hat{\boldsymbol{\Omega}}_{\boldsymbol{i}}$ is estimated from the original data. Then for each study, we obtain the pairs of $(\boldsymbol{t}_i, \boldsymbol{x}_i)$, denoting the generated data by $D = \{(\boldsymbol{t}_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. The effect size θ_F and θ_R are respectively calculated from the fixed-effects and the random-effects models based on the simulated data set D. The whole procedure is repeated 1000 times and the coverage rates (probabilities) are calculated for different types of confidence intervals.

Henmi and Copas (2010) observed that the coverage rate of confidence intervals is affected by the sample size, and HC confidence interval is more robust than others especially when $n > 10$. To investigate the association of the performance with the sample size, we take $n = (10, 15, 20, 25, 30, 35, 40)$.

7.4.1 Confidence Intervals without Publication Bias

We first conduct a simulation study for the meta regression model without assuming publication bias. The true value of θ is 0.01. We calculate the estimation from the random-effects model and fixed-effects model according to the discussion given in Sections 2 and 3. And 95% level confidence intervals for the estimations under HC, BS (with 500 samples), DL, FE, LR, RM, SJ methods are calculated. We present the coverage rates for these CIs in Figure 7.1 and Table 7.1.

As we can see, all the methods perform very well expect the FE method, and those coverage probabilities approach to 95% as we expected. FE works well only at the first case when there is no heterogeneity ($\tau^2 = 0$) introduced in the meta analysis, but fails in other plots when there is heterogeneity. Overall HC method fits the 95% nominal probability very well. The bootstrap has a slightly narrower range for the CI which makes the coverage rate a little smaller than 95% when the sample size $n < 30$.



(a) τ^2=0

(b) $\tau^2 = 0.1 \times 10^{-4}$

(c) $\tau^2 = 0.5 \times 10^{-4}$

(d) $\tau^2 = 10^{-4}$

Figure 7.1: Coverage probabilities of the confidence intervals without assuming publication bias. The dotted line stands for the 95% nominal probability.

This may be improved by increasing bootstrap sample sizes. The coverage rates for other methods are all close but slightly below 95%.

For all the methods the coverage rates are improved with the increase of sample size, which matches the finding by Henmi and Copas (2010) for a simple mean model in meta-analysis.

Table 7.1: Coverage probabilities without publication bias ($\tau^2 = 0.5 \times 10^{-4}$)

n	Bootstrap	HC	DL	FE	LR	RM	SJ
10	92.0	93.6	90.7	60.5	88.8	89.1	91.4
15	93.9	95.3	92.5	57.7	91.5	91.8	93.4
20	94.8	96.0	94.1	59.8	92.9	93.6	94.4
25	94.5	95.1	93.3	58.9	92.9	93.5	93.2
30	95.7	95.9	94.2	60.9	93.6	93.9	94.6
35	95.7	95.5	94.3	58.7	93.7	94.3	94.7
40	95.9	96.3	94.4	59.4	93.6	94.8	94.5

7.4.2 Confidence Intervals with Publication Bias

We further conduct two simulation studies concerning the publication bias problem in meta analysis. We use the following selection model (Begg and Mazumda, 1994):

$$Pr(\text{selected}|\theta_i) = \exp\left(-b\{\varPhi(-\frac{\theta_i - \theta}{\sqrt{\text{Var}(\theta_i)}})\}^{\gamma}\right)$$

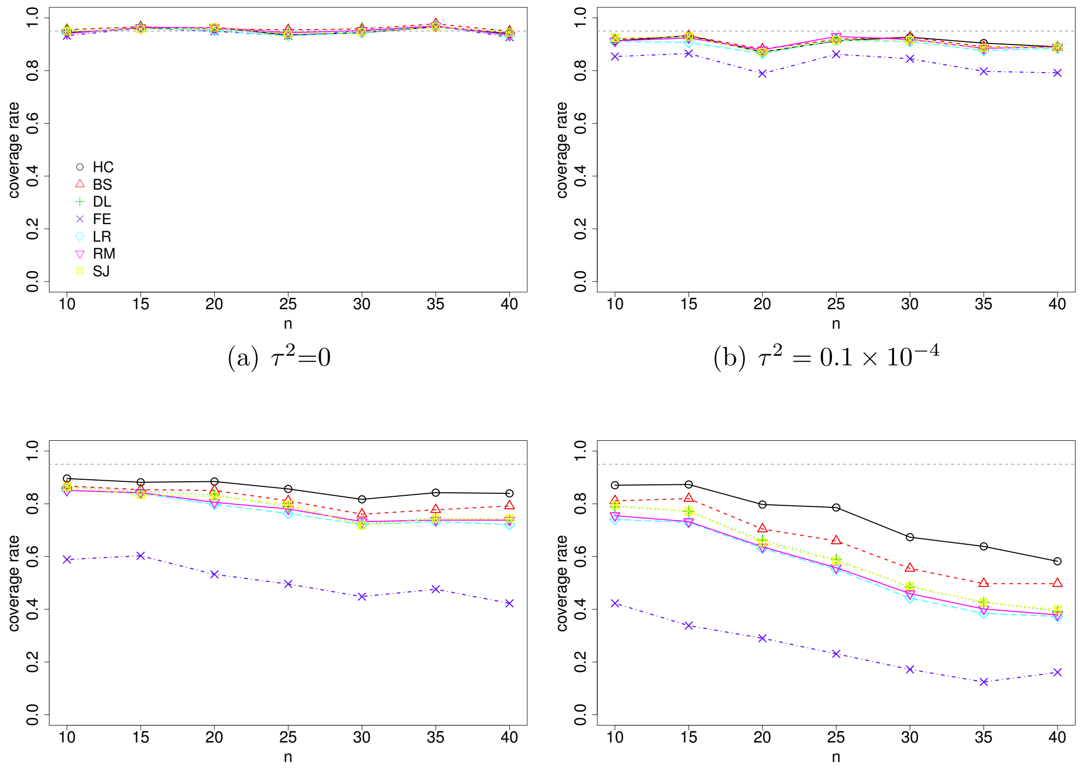
where the selection probability depends on the significance of effect size. Here $\text{Var}(\theta_i)$ is the variance of θ_i in the ith study and $\text{Var}(\theta_i) = (\boldsymbol{x_i}^T \boldsymbol{\Omega_i}^{-1} \boldsymbol{x_i})^{-1}$. We consider the parameter $b = 4$ and γ to be 1.5 and 3 corresponding to strong and moderate selection. These selection functions imply that studies with small effect size θ_i's are less likely to be published than studies with large effect size, which reflects the motivation of the design. Results are shown in Figures 7.2 and 7.3 under the two settings.

Due to the publication bias, the estimation is biased as shown in Figure 7.4. When τ^2 increases, the publication bias gets larger, and consequently the coverage probabilities decrease more as shown in Figures 7.2 and 7.3. Under both moderate and strong selection bias, we see that the bias of $\hat{\theta}_F$ obtained from the fixed-effects model is typically less than the bias of $\hat{\theta}_R$ from the random-effects model. HC and BS are centred on $\hat{\theta}_F$ while the others (except FE) are centred on $\hat{\theta}_R$. This explains why HC and BS perform better than the others when there is publication bias.

We also noticed that the coverage probabilities decrease with larger sample size and they are below 95%. Where there are publication bias particularly when the heterogeneity is serious, the HC and BS methods work less sensitive than others.

7.5 Missing Confounder Problems in Meta-Analysis

We continue the CIs comparison with missing data issue to further consider missing confounder problem in meta-analysis, which is not a new topic in clinical studies. Copas and Eguchi (2005) pointed that if a hidden variable c is independent of x, then the missing c is ignorable as the estimation of effect size θ is not influenced. But if c is associated with x as well as t, then it is a potential confounder and effect evaluation $\hat{\theta}$ is significantly biased due to ignoring the confounder. The discussions of missing confounder in linear regression model can be found in Copas and Eguchi (2005) and



(a) τ^2=0

(b) $\tau^2 = 0.1 \times 10^{-4}$

(c) $\tau^2 = 0.5 \times 10^{-4}$

(d) $\tau^2 = 10^{-4}$

Figure 7.2: Coverage probabilities of the confidence intervals under the moderate publication bias ($\gamma = 3$).

generalized linear model in Lin et al. (2012). Below we use a simulation study to illustrate how the coverage probabilities of confidence intervals for θ are affected by this problem.

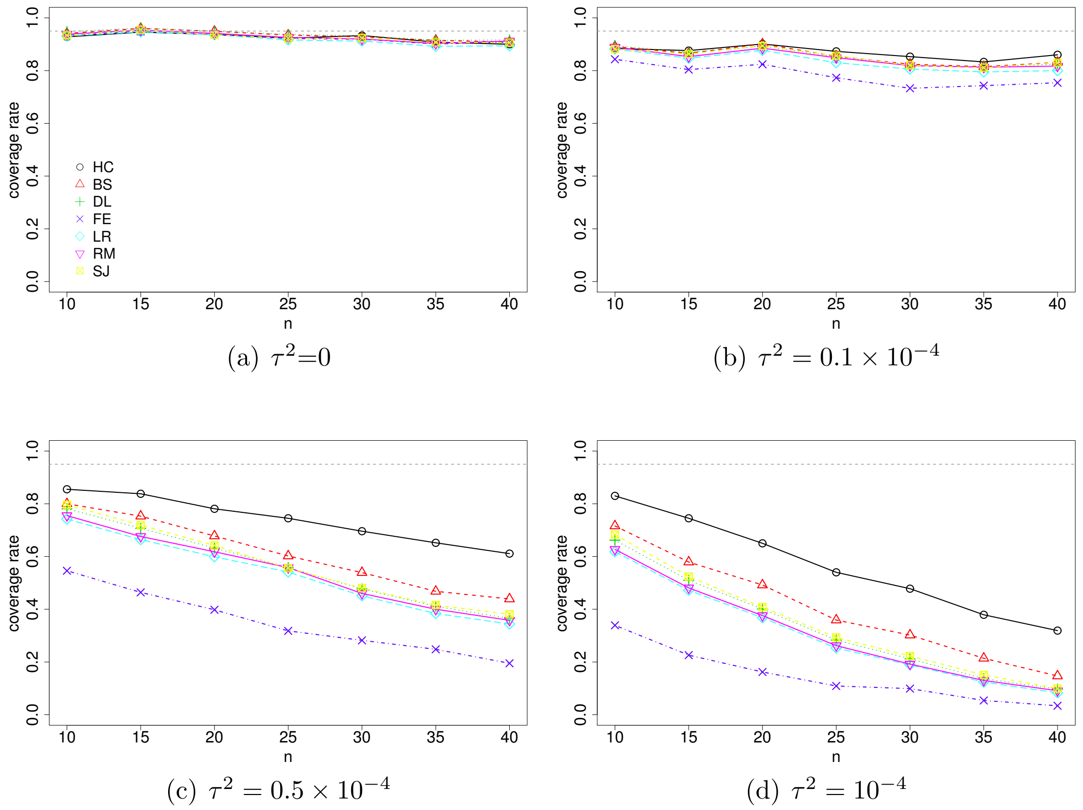
7.5.1 Simulation Study

The true model for the i-th study is assumed as

$$\boldsymbol{t_i} = \theta \boldsymbol{x_i} + \beta \boldsymbol{c_i} + \boldsymbol{\epsilon_i}.$$

We take true value of 0.5 for both θ and β. Confounder $\boldsymbol{c_i}$ is designed to be continuously distributed:

$$\boldsymbol{c_i} \sim N(0, \sigma_c^2).$$



(a) τ^2=0

(b) $\tau^2 = 0.1 \times 10^{-4}$

(c) $\tau^2 = 0.5 \times 10^{-4}$

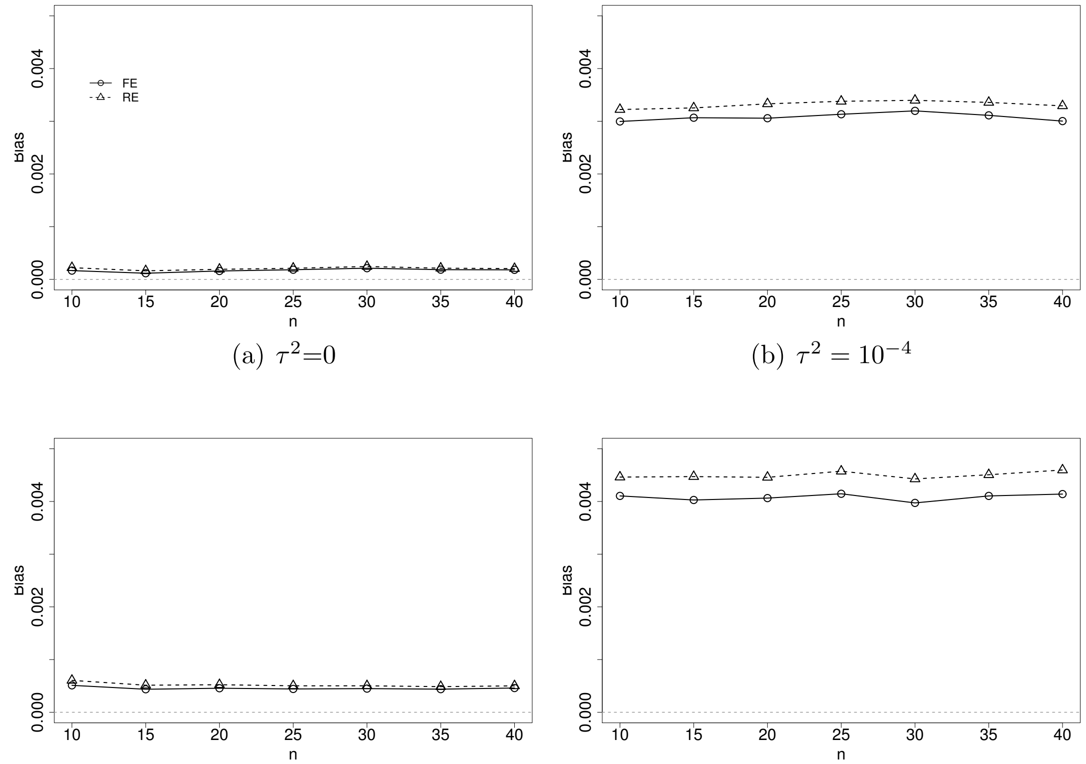
(d) $\tau^2 = 10^{-4}$

Figure 7.3: Coverage probabilities of the confidence intervals under the strong publication bias ($\gamma = 1.5$).

The correlation coefficient $\text{corr}(x, c)$ is selected as 0.3 (moderate correlation) and 0.5 (strong correlation), and the standard deviation of c is assumed as half (when $\text{corr}(x, c) = 0.3$) and one third (when $\text{corr}(x, c) = 0.5$) of x. In this case the biases of $\hat{\theta}$ will be approximately 0.004 for study with $\text{corr}(x, c) = 0.3$ and 0.008 for $\text{corr}(x, c) = 0.5$, and we should remember the standard deviation of $\hat{\theta}$ is about 0.002.

Meta regression analysis is carried out by formula (7.1), which is actually misspecified and confidence intervals are calculated under the discussed methods. Figures 7.5 and 7.6 show the coverage probabilities with 1000 replications.

The advantage of HC and BS are also clearly discovered in the simulation study although DL and SJ methods works also quite well. It is interesting to notice that the coverage probabilities improve when the heterogeneity becomes stronger. This is probably because the item $\theta_i x_i$ is more dominated in the model for the larger τ^2, and the influence due to the missing c becomes smaller. This results in smaller bias. The coverage probability also deteriorates for larger sample size since the standard devi-



(a) τ^2=0

(b) $\tau^2 = 10^{-4}$

(c) τ^2=0

(d) $\tau^2 = 10^{-4}$

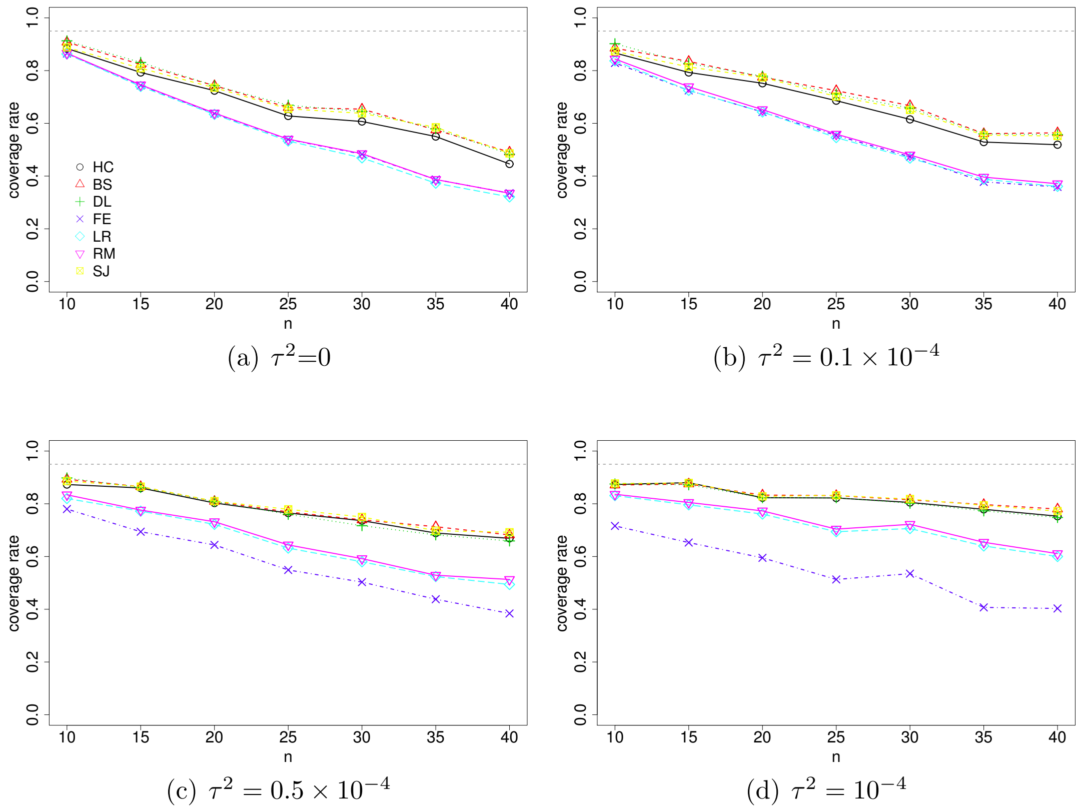
Figure 7.4: Biases of the fixed-effects and random-effects estimates under moderate and strong publication bias: (a)(b) with moderate publication bias ($\gamma = 3$); (c)(d) with strong publication bias ($\gamma = 1.5$).

ation of the parameter becomes smaller and the estimator without bias adjustment seems more serious. This is also discovered in the simulation study in Section 3.5.1.

7.6 Discussion

In this paper, we extended the HC method for calculating confidence interval to meta regression model. A bootstrap model is also presented. Simulation studies show that the HC and BS methods consistently perform better than other methods in almost all the cases particularly for the problems with publication bias.

However we should point out that although HC and BS methods perform quite well and robustly, it still give bias coverage when there is missing data particularly the non-ignorable missing data as discussed in this paper. For non-ignorable missing data problem some other methods should also be used, for example sensitivity analysis



(a) τ^2=0

(b) $\tau^2 = 0.1 \times 10^{-4}$

(c) $\tau^2 = 0.5 \times 10^{-4}$

(d) $\tau^2 = 10^{-4}$

Figure 7.5: Coverage probabilities of the confidence intervals with moderate correlated missing confounder.

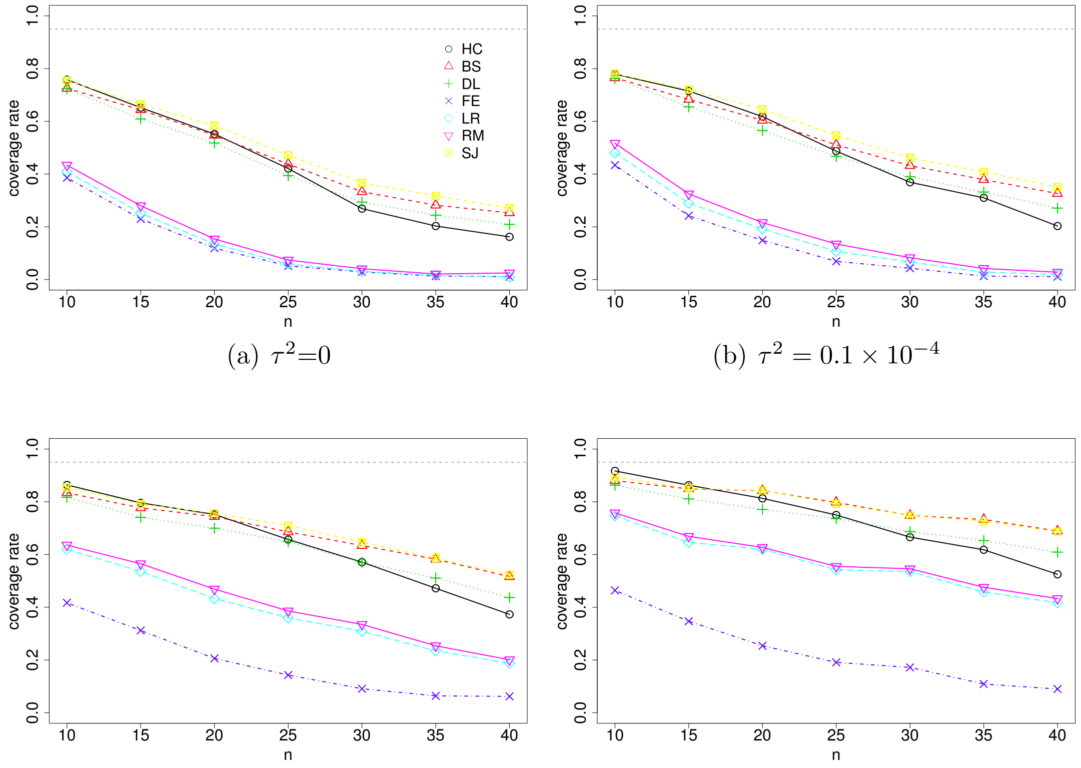
(Copas and Eguchi, 2005; Shi and Copas, 2004), Monte-Carlo sensitivity analysis (Greenland, 2005) or Bias model selection method.

7.7 Appendix

7.7.1 Proof of Equation (7.6)

Consider a fixed-effects model from (7.1) and the estimate of θ given in (7.2). Let

$$\boldsymbol{t}_i^* = \boldsymbol{t}_i - \theta \boldsymbol{x}_i, \quad \hat{\theta}^* = \hat{\theta}_F - \theta,$$



(a) τ^2=0

(b) $\tau^2 = 0.1 \times 10^{-4}$

(c) $\tau^2 = 0.5 \times 10^{-4}$

(d) $\tau^2 = 10^{-4}$

Figure 7.6: Coverage probabilities of the confidence intervals with strong correlated confounder missing.

the Q-statistic defined in (7.5) can be expressed as

$$
\begin{aligned}
Q &= \sum_{i=1}^{m} (\boldsymbol{t}_i^* - \hat{\theta}^* \boldsymbol{x}_i)^T \boldsymbol{\Omega}_i^{-1} (\boldsymbol{t}_i^* - \hat{\theta}^* \boldsymbol{x}_i) \\
&= \sum_{i=1}^{m} \left[\boldsymbol{t}_i^{*T} \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^* - 2\hat{\theta}^* \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^* + \hat{\theta}^{*2} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i \right].
\end{aligned}
$$

Note the fact that

$$
\hat{\theta}^* = \frac{\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^*}{\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i},
$$

we have

$$\begin{aligned}
& \sum_{i=1}^{m} \mathrm{E}\left[\hat{\theta}^{*2} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i\right] \\
= & \sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i \mathrm{E}\left(\hat{\theta}^{*2}\right) \\
= & \frac{\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \mathrm{E}\left(\boldsymbol{t}_i^* \boldsymbol{t}_i^{*T}\right) \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i}{\sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \sum_{i=1}^{m} \mathrm{E}\left[\hat{\theta}^{*} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^*\right] \\
= & \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \mathrm{E}\left(\boldsymbol{t}_i^* \boldsymbol{t}_i^{*T}\right) \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i}{\sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i}.
\end{aligned}$$

For the random-effects model (7.1),

$$\mathrm{Var}(\boldsymbol{t}_i^*) = \boldsymbol{\Sigma}_i = \boldsymbol{\Omega}_i + \tau^2 \boldsymbol{x}_i \boldsymbol{x}_i^T,$$

and thus

$$\mathrm{Var}(\boldsymbol{\Omega}_i^{-1/2} \boldsymbol{t}_i^*) = \boldsymbol{I}_{n_i} + \tau^2 \boldsymbol{\Omega}_i^{-1/2} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1/2}.$$

This leads to

$$\begin{aligned}
& \mathrm{E}\left[\boldsymbol{t}_i^{*T} \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^*\right] \\
= & \mathrm{E}\left[\left(\boldsymbol{\Omega}_i^{-1/2} \boldsymbol{t}_i^*\right)^T \left(\boldsymbol{\Omega}_i^{-1/2} \boldsymbol{t}_i^*\right)\right] \\
= & \mathrm{trace}\left[\mathrm{Var}\left(\boldsymbol{\Omega}_i^{-1/2} \boldsymbol{t}_i^*\right)\right] \\
= & n_i + \tau^2 \mathrm{trace}\left[\boldsymbol{\Omega}_i^{-1/2} \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1/2}\right] \\
= & n_i + \tau^2 \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i.
\end{aligned}$$

We therefore have the following result

$$\begin{aligned}\mathrm{E}(Q) &= \sum n_i + \tau^2 \sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i - \sum_{i=1}^{m} \frac{\boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \left(\boldsymbol{\Omega}_i + \tau^2 \boldsymbol{x}_i \boldsymbol{x}_i^T\right) \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i}{\sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i} \\ &= \sum n_i + \tau^2 \sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i - 1 - \frac{\sum (\boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i)^2}{\sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i} \tau^2 \\ &= N - 1 + \tau^2 \left(\sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i - \frac{\sum (\boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i)^2}{\sum \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{x}_i} \right).\end{aligned}$$

Equating the above expectation with the sample statistic Q yields the DerSimonian-Laird estimate in (7.6).

7.7.2 Proof of Equation (7.19)

We use an idea similar to the one used in Henmi and Copas (2010) to derive the conditional mean and variance of Q given V. We first define the following statistical variables.

$$\boldsymbol{u} = (u_1, u_2, \ldots, u_N)^T = \boldsymbol{E}^T \boldsymbol{M} \boldsymbol{t}^*, \tag{7.24}$$

where $\boldsymbol{E} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N)$, $\boldsymbol{e}_1 = W_1^{-1/2} \boldsymbol{x}$,

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix}, \quad \boldsymbol{M} = \begin{pmatrix} \boldsymbol{\Omega}_1^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}_2^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\Omega}_m^{-1} \end{pmatrix}, \quad \boldsymbol{t}^* = \begin{pmatrix} \boldsymbol{t}_1 - \theta \boldsymbol{x}_1 \\ \boldsymbol{t}_2 - \theta \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{t}_m - \theta \boldsymbol{x}_m \end{pmatrix},$$

and thus $\boldsymbol{u}$, $\boldsymbol{t}^*$ are $N \times 1$ vectors while $\boldsymbol{E}$ and $\boldsymbol{M}$ are $N \times N$ matrices. Note that $N = \sum_i n_i$ and n_i is the dimension of $\boldsymbol{t}_i$ for the i-th study. We further assume that $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N\}$ is an orthonormal basis of $\mathcal{R}^N$ with respect to the following inner product

$$\langle \boldsymbol{t},\ \boldsymbol{t}^{'} \rangle = \boldsymbol{t}^T \boldsymbol{M} \boldsymbol{t}^{'} = \sum_{i=1}^{m} \boldsymbol{t}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^{'}, \tag{7.25}$$

where $\boldsymbol{t}$ is partitioned into $\boldsymbol{t}_i$'s with dimension n_i for $i = 1, \ldots, m$, so is $\boldsymbol{t}^{'}$. This leads to $\boldsymbol{E}^T \boldsymbol{M} \boldsymbol{E} = \boldsymbol{I}_N$ and $\boldsymbol{u} = \boldsymbol{E}^T \boldsymbol{M} \boldsymbol{E} \boldsymbol{E}^{-1} \boldsymbol{t}^*$. Consequently we have

$$\boldsymbol{t}^* = \boldsymbol{E} \boldsymbol{u} = u_1 \boldsymbol{e}_1 + u_2 \boldsymbol{e}_2 + \ldots + u_N \boldsymbol{e}_N. \tag{7.26}$$

From (7.24) and (7.16), we know that

$$u_1 = W_1^{-1/2} \sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^* = V.$$

In addition, we have

$$u_1 \boldsymbol{e}_1 = \frac{\sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{t}_i^*}{W_1} \boldsymbol{x}.$$

Substituting $\hat{\theta}_F$ by equation (7.2) in (7.5), Q is expressed as

$$\begin{aligned} Q &= \sum_{i=1}^{m} \left\{ (\boldsymbol{t}_i - \theta \boldsymbol{x}_i) - \frac{\sum_j \boldsymbol{x}_j^T \boldsymbol{\Omega}_j^{-1} (\boldsymbol{t}_j - \theta \boldsymbol{x}_j)}{W_1} \boldsymbol{x}_i \right\}^T \boldsymbol{\Omega}_i^{-1} \\ & \quad \left\{ (\boldsymbol{t}_i - \theta \boldsymbol{x}_i) - \frac{\sum_j \boldsymbol{x}_j^T \boldsymbol{\Omega}_j^{-1} (\boldsymbol{t}_j - \theta \boldsymbol{x}_j)}{W_1} \boldsymbol{x}_i \right\} \\ &= \langle \boldsymbol{t}^* - u_1 \boldsymbol{e}_1,\ \boldsymbol{t}^* - u_1 \boldsymbol{e}_1 \rangle \\ &= \langle u_2 \boldsymbol{e}_2 + \ldots + u_N \boldsymbol{e}_N,\ u_2 \boldsymbol{e}_2 + \ldots + u_N \boldsymbol{e}_N \rangle \\ &= u_2^2 + \ldots + u_N^2. \end{aligned}$$

The conditional distribution of Q given V can therefore be derived from the conditional distribution of $\boldsymbol{u}_{-1} = (\boldsymbol{u}_2, \ldots, \boldsymbol{u}_N)^T$ given $\boldsymbol{u}_1$. From the definition given in (7.24) we know that $\boldsymbol{u}$ has a multivariate normal distribution with zero mean and the following covariance matrix

$$\text{Var}(\boldsymbol{u}) = \boldsymbol{E}^T \boldsymbol{M} (\boldsymbol{M}^{-1} + \tau^2 \boldsymbol{A}) \boldsymbol{M} \boldsymbol{E} = \boldsymbol{I}_N + \tau^2 \boldsymbol{E}^T \boldsymbol{M}_A \boldsymbol{E},$$

where $\boldsymbol{A}$ is an $N \times N$ matrix defined as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{x}_1 \boldsymbol{x}_1^T & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_2 \boldsymbol{x}_2^T & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{x}_m \boldsymbol{x}_m^T \end{pmatrix},$$

and $\boldsymbol{M}_A$ is defined as

$$\boldsymbol{M}_A = \boldsymbol{MAM} = \begin{pmatrix} \boldsymbol{\Omega}_1^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^T\boldsymbol{\Omega}_1^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}_2^{-1}\boldsymbol{x}_2\boldsymbol{x}_2^T\boldsymbol{\Omega}_2^{-1} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\Omega}_m^{-1}\boldsymbol{x}_m\boldsymbol{x}_m^T\boldsymbol{\Omega}_m^{-1} \end{pmatrix}.$$

We partition $\mathrm{Var}(\boldsymbol{u})$ into the following form.

$$\begin{pmatrix} 1+\tau^2\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{e}_1, & \tau^2\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{E}_{-1} \\ \tau^2\boldsymbol{E}_{-1}^T\boldsymbol{M}_A\boldsymbol{e}_1, & \boldsymbol{I}_{N-1}+\tau^2\boldsymbol{E}_{-1}^T\boldsymbol{M}_A\boldsymbol{E}_{-1} \end{pmatrix} = \begin{pmatrix} 1+\tau^2 W_2, & \tau^2\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{E}_{-1} \\ \tau^2\boldsymbol{E}_{-1}^T\boldsymbol{M}_A\boldsymbol{e}_1, & \boldsymbol{I}_{N-1}+\tau^2\boldsymbol{E}_{-1}^T\boldsymbol{M}_A\boldsymbol{E}_{-1} \end{pmatrix},$$

where $\boldsymbol{E}_{-1} = (\boldsymbol{e}_2, \ldots, \boldsymbol{e}_N)$. In the above equation we used the fact that

$$\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{e}_1 = W_1^{-1}\sum_{i=1}^m \left(\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i\right) = W_1^{-1}\sum_{i=1}^m \left(\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i\right)^2 = W_2. \quad (7.27)$$

The conditional distribution of $\boldsymbol{u}_{-1}$ is therefore a normal distribution. The conditional mean and the conditional covariance matrix are respectively given by

$$\begin{aligned} \boldsymbol{\mu} &= \mathrm{E}(\boldsymbol{u}_{-1}|u_1) = \tau^2(1+\tau^2W_2)^{-1}u_1\boldsymbol{E}_{-1}^T\boldsymbol{M}_A\boldsymbol{e}_1, && (7.28) \\ \boldsymbol{\Gamma} &= \mathrm{Var}(\boldsymbol{u}_{-1}|u_1) \\ &= \boldsymbol{I}_{N-1} + \boldsymbol{E}_{-1}^T\left\{\tau^2\boldsymbol{M}_A - \tau^4(1+\tau^2W_2)^{-1}\boldsymbol{M}_A\boldsymbol{e}_1\boldsymbol{e}_1^T\boldsymbol{M}_A\right\}\boldsymbol{E}_{-1}. && (7.29) \end{aligned}$$

The conditional mean of Q given R is calculated from the above conditional mean and covariance matrix.

$$\begin{aligned} \mathrm{E}(Q|V) &= \sum_{i=2}^N \mathrm{E}(u_i^2|u_1) = \sum_{i=2}^N \mathrm{Var}(u_i|u_1) + \sum_{i=2}^N \{\mathrm{E}(u_i|u_1)\}^2 \\ &= \mathrm{tr}(\boldsymbol{\Gamma}) + \mathrm{tr}(\boldsymbol{\mu}\boldsymbol{\mu}^T) \\ &= (N-1) + \mathrm{tr}\{(\boldsymbol{M}^{1/2}\boldsymbol{E}_{-1})^T\boldsymbol{B}(\boldsymbol{M}^{1/2}\boldsymbol{E}_{-1})\}. \end{aligned}$$

From (7.28) and (7.29), $\boldsymbol{B}$ is expressed by

$$\boldsymbol{B} = \tau^2(\boldsymbol{M}^{-1/2}\boldsymbol{M}_A\boldsymbol{M}^{-1/2}) + \tau^4 d(\boldsymbol{M}^{-1/2}\boldsymbol{M}_A\boldsymbol{e}_1\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{M}^{-1/2}),$$

where d is given in (7.20). We note the fact that

$$\begin{aligned}
\mathrm{tr}\{(\boldsymbol{M}^{1/2}\boldsymbol{E}_{-1})^T\boldsymbol{B}(\boldsymbol{M}^{1/2}\boldsymbol{E}_{-1})\} &= \mathrm{tr}\{(\boldsymbol{M}^{1/2}\boldsymbol{E})^T\boldsymbol{B}(\boldsymbol{M}^{1/2}\boldsymbol{E})\} - \boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{B}\boldsymbol{M}^{1/2}\boldsymbol{e}_1 \\
&= \mathrm{tr}\{(\boldsymbol{M}^{1/2}\boldsymbol{E})(\boldsymbol{M}^{1/2}\boldsymbol{E})^T\boldsymbol{B}\} - \boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{B}\boldsymbol{M}^{1/2}\boldsymbol{e}_1 \\
&= \mathrm{tr}(\boldsymbol{B}) - \boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{B}\boldsymbol{M}^{1/2}\boldsymbol{e}_1. \qquad (7.30)
\end{aligned}$$

Note that the above equation is true for any matrix $\boldsymbol{B}$. Thus,

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{B}) &= \tau^2\mathrm{tr}(\boldsymbol{M}^{-1/2}\boldsymbol{M}_A\boldsymbol{M}^{-1/2}) + d\tau^4\mathrm{tr}(\boldsymbol{M}^{-1/2}\boldsymbol{M}_A\boldsymbol{e}^1\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{M}^{-1/2}) \\
&= \tau^2\mathrm{tr}\{\sum_i \boldsymbol{\Omega}_i^{-1/2}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1/2}\} + d\tau^4\mathrm{tr}(\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{M}^{-1}\boldsymbol{M}_A\boldsymbol{e}_1\} \\
&= \tau^2\mathrm{tr}\{\sum_i(\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i)\} + \tau^4 dW_1^{-1}\mathrm{tr}(\sum_i(\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{x}_i)\} \\
&= \tau^2 W_1 + \tau^4 dW_3
\end{aligned}$$

where W_3 is given in (7.12). In addition, we have the following formula by using (7.27).

$$\boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{B}\boldsymbol{M}^{1/2}\boldsymbol{e}_1 = \tau^2\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{e}_1 + \tau^4 d\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{e}_1\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{e}_1 = \tau^2 W_2 + \tau^4 dW_2^2.$$

Applying the above equations, we have obtained the conditional mean as

$$\mathrm{E}(Q|V) = (N-1) + \tau^2(W_1 - W_2) + \tau^4 d(W_3 - W_2^2). \qquad (7.31)$$

The conditional variance of Q given V is calculated by

$$\begin{aligned}
\mathrm{Var}(Q|V) = \sum_{i,j=2}^N \mathrm{Cov}(u_i^2, u_j^2|u_1) &= \sum_{ij=2}^N \mathrm{E}(u_i^2u_j^2|u_1) - \left(\sum_{i=2}^N \mathrm{E}(u_i^2|u_1)\right)^2 \\
&= 2\mathrm{tr}\{(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Gamma})^2\} - 2\{\mathrm{tr}(\boldsymbol{\mu}\boldsymbol{\mu}^T)\}^2.
\end{aligned}$$

The proof of the last equation above can be referred to equations (A7) and (A9) in Appendix A.1 in Henmi and Copas (2010). Define

$$\boldsymbol{F} = \boldsymbol{I}_N + \boldsymbol{B} \;\text{ and }\; \boldsymbol{G} = \tau^4 d_1\boldsymbol{M}^{-1/2}\boldsymbol{M}_A\boldsymbol{e}_1\boldsymbol{e}_1^T\boldsymbol{M}_A\boldsymbol{M}^{-1/2}$$

where d_1 is defined in (7.20), we have

$$\begin{aligned}
\mathrm{Var}(Q|V) &= 2\mathrm{tr}\{(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})\boldsymbol{F}(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})^T(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})\boldsymbol{F}(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})^T\} \\
&\quad -2[\mathrm{tr}\{(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})\boldsymbol{G}(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})^T\}]^2 \\
&= 2\mathrm{tr}(\boldsymbol{F}^2) - 4\boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{F}^2\boldsymbol{M}^{1/2}\boldsymbol{e}_1 + 2\{\boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{F}\boldsymbol{M}^{1/2}\boldsymbol{e}_1\}^2 \\
&\quad -2[\mathrm{tr}\{(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})\boldsymbol{G}(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})^T\}]^2.
\end{aligned}$$

To get the equations above, we used (7.30) repeatedly. It is not difficult (although it is tedious) to get the following results.

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{F}^2) &= N + 2\mathrm{tr}(\boldsymbol{B}) + 2\mathrm{tr}(\boldsymbol{B}^2) = N + 2(\tau^2 W_1 + \tau^4 d W_3) + 2\mathrm{tr}(\boldsymbol{B}^2), \\
\mathrm{tr}(\boldsymbol{B}^2) &= \tau^4 W_1 W_2 + 2\tau^6 d W_4 + \tau^8 d^2 W_3^2,
\end{aligned}$$

where W_4 is defined in (7.12). Similarly, we have

$$\begin{aligned}
\boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{F}\boldsymbol{M}^{1/2}\boldsymbol{e}_1 &= \tau^2 W_2 + \tau^4 d W_2^2 + 1, \\
\boldsymbol{e}_1^T\boldsymbol{M}^{1/2}\boldsymbol{F}^2\boldsymbol{M}^{1/2}\boldsymbol{e}_1 &= \tau^4 W_3 + 2\tau^6 d W_3 W_2 + \tau^8 d^2 W_2^2 W_3 + 2(\tau^2 W_2 + \tau^4 d W_2^2) + 1, \\
\mathrm{tr}(\boldsymbol{\mu}\boldsymbol{\mu}^T) &= \mathrm{tr}\{(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})\boldsymbol{G}(\boldsymbol{E}_{-1}\boldsymbol{M}^{1/2})^T = \tau^4 d_1(W_3 - W_2^2),
\end{aligned}$$

where d_1 is defined in (7.20). Finally, we have the following result.

$$\begin{aligned}
\mathrm{Var}(Q|V) &= 2(N-1) + 4\tau^2(W_1 - W_2) + 2\tau^4(W_1W_2 - 2W_3 + W_2^2) \\
&\quad +4\tau^4 d(W_3 - W_2^2) + 4\tau^6 d(W_4 - 2W_2W_3 + W_2^3) + 2\tau^8(d_2 - d_1^2)(W_3 - W_2^2)^2.
\end{aligned}$$

Chapter 8

Conclusions and Future Work

In this chapter we summarize the statistical problems discussed in this thesis and highlight the main findings and our contributions to the literature.

The first objective of the thesis was to assess model uncertainty, particularly with the missing data problems. We discovered the limitations of conventional analysis in exploring model uncertainty due to lack of knowledge based on observed data only, and we evaluated and interpreted those uncertainties through local sensitivity analysis. In our inference, we start from a working model we usually used for the observed data and then use the bias analysis by measuring the departure from the true model. In Chapters 2, 3, and 5, we applied the incomplete data bias analysis, which was first introduced by Copas and Eguchi (2005), to missing covariate problems for linear regression or GLM regression models. Analysis was carried out based on new terms such as 'bias models', which index the models involved in bias analysis and also sensitivity analysis, and 'bias parameters' that indicate those uncertainty factors which dominate the incomplete data bias and are difficult to measure in practice, and can be also described as 'sensitivity parameters' in the sensitivity analysis area.

The analysis for misspecified bias models can be different with different missing data mechanisms and regression models. Under ignorable missing data assumption, as discussed in Chapters 2, 3 and 4, the primary uncertainty comes from the misspecification of covariate distributions. The incomplete data bias (termed covariate bias according to the bias sources) towards the parameter of interest is mainly generated by the correlations between observed covariate variables and missing confounders. Examples under linear regression and GLMs are discussed separately, since for nonlinear models or GLMs, the identifiable bias model is not the marginal model but rather a misspecified conditional model based on observed variables. In this case, the marginal bias adds on to the total incomplete data biases.

And we also recognized that the full missing data mechanism is required for both ignorable and non-ignorable missing data, and model misspecification for the missing data mechanism can result in a substantial bias. We handled these problems generally through bias models identification and local bias analysis, with detailed discussion in Chapter 3 for ignorable missing data and Chapter 5 for non-ignorable missing data. The missing data mechanism bias is first introduced in Chapter 3, where the two types biases (covariate bias and MDM bias) are investigated by simulation studies under ignorable missing data. The MDM misspecification issue is a difficult problem but would not cause too many worries for ignorable missing data since we may consider proper model selection techniques, for example, we suggested nonparametrical models in complex cases. However, it is a serious problem for non-ignorable missing data, this is because of the non-identifiability problem of the missing data mechanism. The general idea of dealing with this problem was given in Chapter 3 and the details were provided in Chapter 5, where we identify an ignorable missing data mechanism and measure the difference from non-ignorability through sensitivity analysis. The covariate density misspecification and missing data mechanism misspecification can be complex in practice and efficiency and robustness of inferences are questioned using the usual working models. Thus we suggested a different method by transferring the non-ignorable missing data problem to the equivalent ignorable missing data counterpart. We are able to avoid the problem of identifying the non-ignorable missing data mechanism specification in this case, and only its marginal density is required. It is found that Bayes approach and nonparametric conditional model can fit the marginal model well.

Another approach to handling non-ignorable missing data is through the selection model frame, as discussed in Chapter 6, where a semiparametric model is assumed. The uncertainty was identified from the nonparametric component in the missing data mechanism.

The second objective of this thesis was to consider a proper sensitivity analysis and make a valid selection of bias models. Monte Carlo sensitivity analysis and Bayesian sensitivity analysis have been studied for many years (see e.g. Greenland, 2005; McCandless et al., 2007, 2008; Gustafson et al., 2010), and both methods average estimation over all competing models. But these methods can be sensitive to the prior selection, so we proposed a novel method named bias model selection with a Monte Carlo method. This approach actually contains two parts: 1) Monte Carlo sensitivity analysis and 2) Bias model selection. The first step carries out inferences for any given bias model and observed data; and the second step simulates an artificial set of 'observed data' using the fitted model obtained in the first step. The distances between the real observed data set and the simulated 'observed' data set are used to select bias model. Model selection versus model averaging is compared by simulation study. It shows that the former performs better than the latter in almost all the cases.

A test is developed to check how close the 'simulated' data set and the real observed data set; or check how close the selected bias model and the true bias model. This can help to remove some unreasonable bias models or remove some implausible values of bias parameters. We applied the idea to determine the range of bias model, and combine the method with Bayesian and Monte Carlo sensitivity analysis. Simulations given in Chapter 4 showed that the method improves the results obtained by using conventional MCSA.

This MC-BMS technique requires a replicated sampling procedure, and it is usually useful when combined with multiple imputation and bootstrapping methods for missing data problems. For example, we used it in Chapter 6 for the non-ignorable missing data selecting a missing data mechanism model. The method can be applied into a wide area. We found the K nearest neighbour distance works very well after comparing with some other types of distances.

The third objective of this thesis was to build a robust confidence interval when there is an uncertainty or bias. In Chapter 7, we considered robust confidence intervals for meta-regression models and found the confidence interval proposed by Henmi and Copas (2010) gives the most robust results when there is publication bias and missing covariates.

Missing data and model misspecification are difficult problems. Local sensitivity analysis provides a tool to assess the uncertainty and bias. We will carry on the research along this direction, particularly on studying the difference between ignorable and non-ignorable MDM model. This can be considered in a selection model framework as given in Chapter 6. The existence of nonparametric part to the first and second order terms is to be tested.

We will further apply the local sensitivity analysis to other regression models, such as survival analysis, longitudinal data analysis. The problems of non-response or missing multivariate variables may not be difficult to solve, although non-monotone missing data may be more challenging, and high dimensions of bias parameters will be involved.

MC-BMS method is a very flexible method, and it can be used combining with other techniques, such as prior selection, Bayesian model average, and the Expectation-Maximization algorithm. Furthermore, its efficiency depends on the choice of distance measure, and can surely be improved in the future. The related theory is yet to develop.

Bibliography

Amari, S. (1985). *Differential geometric methods in statistics.* Springer New York.

Arnold, B. and Beaver, R. (2000). Hidden trunction models. *The Indian Journal of Statistics*, 62:23–35.

Auffermann, W. F., Ngan, S. C., and Hu, X. (2002). Cluster significance testing using the bootstrap. *NeuroImage*, 17:583–591.

Begg, C. and Mazumda, M. (1994). Operation characteristics of a rank correlation test for publication bias. *Biometrics*, 50:1088–1101.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, 83:377–385.

Bock, H. H. (1979). Clustering by density estimation. *Analyse des Donnees et Informatique*, pages 173–186.

Bock, H. H. (1985). On some significant tests in cluster analysis. *Journal of Classfication*, 2:77–108.

Breslow, N. and Lin, X. (1995). Bias correction in linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91.

Cailliez, F. and Kuntz, P. (1996). A contribution to the study of metric and euclidean structures of dissimilarities. *Psychometrika*, 61:241–253.

Chen, B., Yi, G., and Cook, R. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association.*, 105(489):336–353.

Chen, B. and Zhou, X, H. (2011). Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. *biometrics*, 67:830–842.

Chen, M. and Ibrahim, J. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, 1:43–52.

Chen, M., Ibrahim, J., and Shao, Q. (2004). Propriety of the posterior distribution and existence of the MLE for regression models with covariates missing at random. *Journal of the American Statistical Association.*, 99:121–130.

Colledge, M., Johnson, J., Pare, R., and Sande, I. (1978). Large scale imputation of survey data. *Journal of the American Statistical Association.*, pages 431–436.

Cook, R. (1986). Assessment of local influence. *J.R.Statist.Soc.B.*, 48:133–169.

Copas, J. and Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *J.R.Statist.Soc.B.*, 63(4):871–895.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias. *J.R.Statist.Soc.B.*, 67:459–513.

Copas, J. and Li, H. (1997). Inference for non-random samples. *J.R.Statist.Soc.B.*, 59:55–95.

Copas, J. and Shi, J. (2000a). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*, 1:247–262.

Copas, J. and Shi, J. (2000b). A sensitivity analysis for publication bias in systematic reviews. *Statist. Meth. in Med. Res.*, 10:251–265.

Cornfield, J., Haenszel, W., Hammond, W., Lilienfeld, A. M., Shimkin, M., and Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natn. Cancer Inst.*, 22:173–203.

Cunningham, K. M. and Ogilvie, L. (1972). Evaluation of hierarchical grouping techniques: a preliminary study. *Computer Journal*, 15:209–213.

Daniels, M. and Hogan, J. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis.* Chapman & Hall/CRC.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trails*, 7:177–188.

Diggle, P. and Kenward, M. (1994). Informative dropout in longitudinal data analysis. *Appl. Statist*, 43:49–94.

Dox, D. (1972). *The Analysis of Binary Data.* London: Methuen.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J.R.Statist.Soc.B.*, 57(1):45–97.

Duflou, H. and Maenhaut, W. (1990). Application of principal component and cluster analysis to the study of the distribution of minor and trace elements in the normal human brain. *Chemometrics and Intelligent Laboratory Systems.*, 9:273–286.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap.* Chapman & Hall.

Estabrook, C. G. and Rodgers, D. J. (1966). A general method of taxonomic description for a computed similarity measure. *Bioscience*, 16:789–793.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis.* John Wiley & Sons, Ltd, 5th edition.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7:179–188.

Fisher, R. A. (1971). *The Design of Experiiments.* New York: Hafner Publishing, 8th edition.

Fix, E. and Hodges, J. (1951). Discriminatory analysis- nonparametric discrimination: Consistency properties. Technical report, University of California, Berkeley.

Florek, K., Lukaszewiez, L., and Perkal, L. (1951). Sur la liaison et la divison des points d'un ensemble fini. *Colloquium Mathematicum*, 2:282–285.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., and Cooper, T. (2013). *caret: Classification and Regression Training.* R package version 5.16-24.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, 77:270–278.

Gail, M., Wieand, S., and Piantadosi, S. (1984). Biased estimated of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431–444.

Garland, M., Hunter, D., Colditz, G., Spiegelman, D., Manson, J., Stampfer, M., and Willett, W. (1999). Alcohol consumption in relation to breast cancer risk in a cohort of United States women 25-42 years of age. *Cancer Epidemiology*, 8:10171021.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modelling versus mixture modelling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples,*, pages 115–142, New York: Springer. In H. Wainer (ed.).

Gower, J. C. (1971). A general coeffcient of similarity and some of its properties. *Biometrics*, 27:857–872.

Gower, J. C. (1985). Measures of similarity, dissimilarity and distance. *Encyclopaedia of Statistical Science*, 5:397–405.

Gower, J. C. and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 5:5–48.

Greenland, S. (2005). Multiple bias modelling for analysis of observational data. *J.R.Statist.Soc.A.*, 168:267–306.

Greenland, S. and Longnecker, M. (1992). Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology*, 135:1301–1309.

Gustafson, P. (2005). The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. *Statistics in Medicine*, 24:1203–1217.

Gustafson, P., McCandless, L., Levy, A., and Richardson, S. (2010). Simplified Bayesian sensivitity analysis for mismeasured and unobserved confounders. *Biometrics*, 66:1129–1137.

Hall, P., Park, B., and Samworth, R. J. (2008). Choice of neighbor order in nearest neighbor classification. *The Annals of Statistics*, 36(5):2135–2152.

Hardy, R. J. and Thompson, S. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, 15:619–629.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* Chapman and Hall, 1st edition.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrics*, 47:153–161.

Henmi, M. and Copas, J. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, 29:2969–2983.

Herring, A., Ibrahim, J., and Lipsitz, S. (2004). Non-ignorable missing covariate data in survival analysis: a case-study of an international breast cancer study group trial. *Journal of the American Statistical Association.*, 53:293–310.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95:77–84.

Huang, L., Chen, M., and Ibrahim, J. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics*, 61(3):767–780.

Hubert, L. (1974). Approximate evaluation techniques for the single link and complete link hierarchical clustering procedures. *Journal of the American Statistical Association.*, 69:698–704.

Ibrahim, J. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769.

Ibrahim, J., Chen, M., and Lipsitz, S. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88:551–564.

Ibrahim, J., Chen, M., Lipsitz, S., and Herring, A. (2005). Missing-data methods for generalized linear model: A comparative review. *Journal of the American Statistical Association.*, 100(469).

Ibrahim, J. and Lipsitz, S. (1999). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *J.R.Statist.Soc.B.*, 61(1):173–190.

Ibrahim, J., Lipsitz, S., and Chen, M. (1999). Missing covariates in generalized linear model when missing data mechanism is nonignorable. *J.R.Statist.Soc.B.*, 61(1):173–190.

Ichino, M. and Yaguchi, H. (1994). Generalized minkowski metrics for mixed featuretype data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 24:698–708.

Jain, A. and Dubes, R. C. (1988). *Algorithms for Clustering data.* Prentice Hall, Englewood Cliffs, NJ.

Jain, A. K., Murty, M., and Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

Jajuga, K., Walesiak, M., and Bak, A. (2003). On the general distance measure. *Exploratory data analysis in Empirical Research*, pages 104–109.

Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika*, 32:241–254.

Kenward, M. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statist. Sci*, 13:236–247.

Kenward, M. G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8:51–83.

Kim, J. and Yu, C. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association.*, 106(493).

Lawless, J. F., Kalbfleisch, J. D., and J., W. C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J.R.Statist.Soc.B.*, 61:413–438.

Lee, K. (1979). Multivariate tests for clusters. *Journal of the American Statistical Association.*, 74:708–714.

Legendre, P. and Chodorowski, A. (1977). A generalisation of jaccard's association coefficient for q-analysis of multi-state ecological data metrices. *Ekologia Polska*, 25:297–308.

Lerman, I. (1987). Construction d'un indice de similarite entre objets decrits par des variables d'un type quelconque. application au probleme du consensun en classification (1). *Revue de Statistique Appliquee*, 25:39–60.

Liang, H., WANG, S., ROBINS, J., and CARROLL, R. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association.*, 99(466):357–367.

Lin, N., Shi, J., and Henderson, R. (2012). Doubly misspecifed models. *Biometrika*, 99:285–298.

Ling, R. F. (1972). On the theory and construction of k-clusters. *Computer J.*, 15:326–332.

Little, R. (1993). Pattern mixture models for multivariate incomplete data. *Journal of the American Statistical Association.*, 88(421):125–134.

Little, R. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81:471–483.

Little, R. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association.*, 90(431):1112–1121.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data.* Wiley, New York, second edition.

Little, R. and Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *J.R.Statist.Soc.C.*, 60(4).

Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72:497–512.

Longnecker, M., Berlin, J., Orza, M., and Chalmers, T. (1988). A meta-analysis of alcohol consumption in relation to risk of breast cancer. *Journal of American Medical Association*, 260:652–656.

Lu, G. and Copas, J. (2004). Missing at random, likelihood ignorability and model completeness. *The Annals of Statistics*, 32(2):754–765.

Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems.*, 50:1–18.

Mao, J. and Jain, A. K. (1996). A self organizing network for hyperellipsoidal clustering (hec). *IEEE Trans. Neural Netw*, 7:16–29.

McCandless, L., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, 26:2331–2347.

McCandless, L., Gustafson, P., and Levy, A. (2008). A sensitivity analysis using information about measured confounders yielded improved assessments of uncertainty from unmeausred confounding. *Journal of Clinical Epidemiology*, 61:247–255.

McQuitty, L. (1966). Similarity analysis by reciprocal pairs for discrete and continous data. *Educational and Psychological Measurement*, 27:21–46.

Milligan, G. W. (1981). A review of monte carlo tests of cluster analysis. *Multivariate Behavioral Research*, 16:379–407.

Molenberghs, G., Benunckens, C., Sotto, C., and Kenward, M. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J.R.Statist.Soc.B.*, 70:371–388.

Molenberghs, G., Kenward, M., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the slovenian plebiscite case. *Appl. Statist*, 50(1):15–29.

Nigsch, F., Bender, A., Buuren, B., Tissen, J., Nigsch, E., and Mitchell, J. (2006). Melting point prediction employing k-nearest neighbor algorithm and genetic parameter optimization. *J. Chem. Inf. Model.*, 46:2142–2422.

Noortgate, W. and Onghena, P. (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods.*, 37(1):11–22.

Oakley, J. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J.R.Statist.Soc.B.*, 66(3):751–769.

Paik, M. (2004). Nonignorable missingness in matched case-control data analyses. *biometrics*, 60:306–314.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J.R.Statist.Soc.B.*, 69:101–122.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001). Sequential regression imputation for survey data. *Survey Methodology*, 27:85–96.

Rao, J. and Wang, Q. (2002). Empirical likelihood-based inference under imputation for missing response data. *The annals of Statistics*, 30(3):896–924.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82:299–314.

Robins, J. and Rotnitzky, A. (1995). Semiparametric effciency in multivariate regression models with missing data. *Journal of the American Statistical Association.*, 90:122–129.

Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association.*, 89(846):66.

Royall, R. and Tsou, T. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *J.R.Statist.Soc.B.*, 65:391–404.

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

Rubin, D. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Imputation and Editing of Faulty or Missing Survey Data*, pages 1–23.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Saha, C. and Jones, M. (2005). Asymptotic bias in the linear mixed effects model under non-ignorable missing data mechanisms. *J.R.Statist.Soc.B.*, 67(1):167–182.

Saltelli, A., Ratto, M., Andres, T., and Corporation, E. (2008). *Global sensitivity analysis : The Primer.* John Wiley & Sons, Ltd.

Saltelli, A., Tarantola, S., Campolongo, F., and Corporation, E. (2004). *Sensitivity analysis in practice : A guide to assessing scientific models.* John Wiley & Sons, Ltd.

Sander, I. (1983). Hot deck imputation procedures. *Incomplete data in sample surveys. Vol 3 Symposium on Incomplete Data, Proceedings(W.G. Madow and I.Olkin,Eds.). New York: Academic Press*, 3.

Schafer, J. (1997). *Analysis of incomplete multivariate data.* Chapman and Hall, London.

Scott, A. and Wild, C. (2002). On the robustness of weighted methods for fitting models to case-control data. *J.R.Statist.Soc.B.*, 64(2):207–219.

Shi, J. and Copas, J. (2004). Meta-analysis for trend estimation. *Statistics in Medicine*, 23:3–19.

Sidik, K. and Jonkman, J. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, 21:3153–3159.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B., and Carrol, R. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association.*, 100(470):591–601.

Sneath, P. (1957). The application of computers to taxonomy. *Journal of General Mocrobiology*, 17:201–226.

Sokal, R. and Michener, C. (1958). A statistical method for evaulating systematic relationships. *University of Kansas Science Bulletin.*, 38:1409–1438.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34.

Struthers, C. and Kalbfleisch, J. (1986). Misspecified proportional hazard models. *Biometrika*, 73:363–369.

Stubbendick, A. and Ibrahim, J. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometric.*, 59(4):1140–1150.

Tang, G., Little, R., and Raghunathan, T. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764.

Tanner, M. (1993). *Tools for statistical inference:methods for the exploration of posterior distributions and likelihood functions.* Springer, New York, second edition.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Vemuri, V., Dracup, J., Erdmann, R., and Vemuri, N. (1969). Sensitivity analysis method of system identification and its potential in hydrologic research. *Water Resources Research*, 5:341–349.

Weisberg, S. (2005). *Applied linear regression.* John Wiley & Sons, INC., 3rd edition.

Wishart, D. (1969). Mode analysis. *Numerical Taxonomy*, page New York: Academic Press.

Wong, M. and Lane, T. (1983). A kth nearest neighbour clustering procedure. *Journal of the American Statistical Association.*, 45:362–368.