Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to achieve, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Ji Lin Date

Likelihood Methods for Logistic Regression with Missing Data

By

Ji Lin

Doctor of Philosophy
Biostatistics

Robert H. Lyles, Ph.D.
Advisor

Eugene Huang, Ph.D.
Committee Member

Kyle Steenland, Ph.D.
Committee Member

Lance Waller, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Likelihood Methods for Logistic Regression with Missing Data

By

Ji Lin

B.S., Peking University, 2003
M.S., University of Texas at Dallas, 2005

Advisor: Robert H. Lyles, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2012

Abstract

Likelihood Methods for Logistic Regression with Missing Data

By Ji Lin

In biometric research, missing data are often encountered due to many reasons. This dissertation explores methods to deal with missing data in statistical analysis of logistic regression. The disease status and risk exposure could be subject to missing data separately or together. The interest is on identifying the covariate-adjusted association between the disease status and the risk exposure, with consideration of the potential impact of the missing data.

The first research topic was focused on providing an intuitive and computationally accessible approach when the assumption of missing at random (MAR) was imposed. We proposed a weighting method, utilizing an expanded dataset with two approaches to estimation in different situations. The first one makes use of a "flipped-around" logistic model and behaves similarly to multiple imputation, and the second one iterates like the expectation-maximization algorithm but in a simplified fashion. Simulation studies were performed to demonstrate the performance of the methods under different scenarios.

The assumption of MAR is usually imposed in practice but often not testable. It is then important to assess how sensitive the results are to the violation of this assumption. In the second research topic, a framework of sensitivity analysis was proposed by specification of alternative missing data mechanisms. The result from each specified scenario is compared to that from MAR so that to assess the magnitude of change of parameter estimates relative to deviation from MAR. Examples and simulation results suggest that the proposed method succeeds in detecting the direction and magnitude of bias in parameter estimates even if the specification of the alternative missing data mechanism is not completely correct.

In the third research topic, we explore the reassessment design, where a second wave of sampling is made in an attempt to recover some portion of the missing data in the original data collection. We construct a joint likelihood based on the original model of interest and a model for the missing data mechanism, with emphasis upon "non-ignorable" missingness. The estimation is carried out by numerical maximization of the joint likelihood and standard errors are estimated via a close approximation of the Hessian matrix. We show how likelihood ratio tests can be used for model selection and how they facilitate hypothesis testing for whether missingness is at random, which is an assumption that can be suspect in many practical applications. Examples and simulations are presented to demonstrate the performance of the proposed method.

Likelihood Methods for Logistic Regression with Missing Data

By

Ji Lin

B.S., Peking University, 2003
M.S., University of Texas at Dallas, 2005

Advisor: Robert H. Lyles, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2012

ACKNOWLEDGEMENT

At the completion of this dissertation, I realized that I have not only completed an academic work that I have been enthusiastic about, but also finished an extraordinary voyage in my life leading to who I am. The doctorate degree in Philosophy means literally what it exactly means, representing not only an in-depth understanding of the nature and science, but also an extended understanding to a world within an individual through the reflection of the world outside.

Looking back, I am very grateful for all I have received from everyone around me. I want to first thank my advisor Prof. Robert H. Lyles for his continuous guidance and support. He has enlightened me through his wide knowledge and his deep intuitions about the scientific research; guided me throughout the mist, yet offered me the freedom to explore the unknown.

My gratuity also goes to Profs. Eugene Huang, Kyle Steenland, and Lance Waller, for motivating and inspiring me throughout my research with constructive feedback and suggestions. I am also appreciative for the valuable contribution and support from Dr. Candice Y. Johnson and Prof. Dana W. Flanders for their insights into the motivating study. I owe my thanks to Prof. Amita K. Manatunga for her support during my study and the opportunity she offered to me of valuable experience working with her in statistical research.

I could not have finished my dissertation without the unconditional love and support from my parents. Their voice through the phone helped me through the years taking on the challenges and holding on straight to the end. I especially owe it to my special other half, who has always been there with me, supporting me, cheering me up and celebrating with me through the ups and downs.

Table of Contents

List of Figures

List of Tables

Chapter 1. INTRODUCTION AND BACKGROUND

1.1. Introduction

In biometric research, missing data are often encountered due to many reasons, including the unavailability of measurements, survey non-response, study subjects failing to report for evaluations, respondents refusing to answer certain items on a questionnaire, and loss of data.

Here I will propose a likelihood based weighting approach that applies appropriate weights to the records of a properly constructed expanded data set, which is designed to represent the underlying unobserved complete data set. Under the common assumption of missing at random (MAR), this approach can produce results that are comparable to existing likelihood based methods (e.g. expectation-maximization method), semi-parametric methods (e.g. weighted estimating equations), and simulation based methods (e.g. multiple imputation). When the assumption of missing at random is questionable based on prior knowledge, this weighting approach can be used to construct sensitivity analysis on the violation of the MAR assumption. The assumption on the missingness mechanism can be formulated in terms of conditional probability of missingness, risk ratio of missingness, and odds ratio of missingness. When re-assessment data are available, this approach can be used to incorporate these data and produce maximum likelihood estimates.

1.2. Background

1.2.1. Missing-Data Mechanisms

When dealing with missing data, a critical issue is to define the mechanism leading to missing data. The missing-data mechanism describes the relationship between the occurrence that values of certain variables are missing and the values of all the variables, including the underlying values that would have been observed. It determines what information is unreachable by the researcher, and what information can be potentially recovered via due diligence. The missing-data

mechanism was first formalized by Rubin (1976) by introducing random indicators of missingness and describing the missing-data mechanism via probability distribution. Little and Rubin (1987, 2002) proposed a nice framework for categorizing missing-data mechanisms, namely missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR), the later sometimes referred to as missing not at random (MNAR). Using the notation and terminology in Little and Rubin's classic textbook (2002), define the complete data as $Y=(y_{ij})$, which denotes an $(n\times K)$ rectangular data set without missing values, with ith row $y_i=(y_{i1},\ldots,y_{iK})$ for the subject i. Define the missing-data indicators as $M=(m_{ij})$, such that $m_{ij}=1$ if y_{ij} is missing and $m_{ij}=0$ if y_{ij} is present. The missing-data mechanism is described by the conditional distribution of M given Y, $f(M\mid Y,\phi)$, where ϕ denotes unknown parameters. Let Y_{obs} denote the observed components and Y_{mis} the missing components. Then the missing-data mechanism is called missing completely at random (MCAR) if missingness does not depend on the values of the data Y, that is,

$$f(M\mid Y,\phi)=f(M\mid\phi) \text{ for all } Y,\phi.$$

If the missing-data mechanism is MCAR, the observed data set is a random subset of the complete data set, and in turn, a random subset of the population of interest. Therefore the observed data set can be treated as a standalone random sample drawn from the population of interest, and inference can be carried out as usual without any problem.

A less restrictive assumption is missing at random (MAR), when the missingness depends only on the components Y_{obs} of Y that are observed, and not on the components that are missing, after controlling for the components Y_{obs} that are observed. That is,

$$f(M\mid Y,\phi)=f(M\mid Y_{\text{obs}},\phi) \text{ for all } Y_{\text{mis}},\phi.$$

The mechanism is called not missing at random (NMAR) if the distribution of M depends on the missing values Y_{mis}.

With data missing at random but not missing completely at random, inference on the

marginal distribution of the variable affected by missing values is not valid. However, regression analysis based on complete cases can still produce valid parameter estimates of the regression coefficients of the variables under MAR. Multivariate analysis might not be appropriate and efficiency might be lost.

With data NMAR, it is generally not appropriate to conduct any analysis based only on the complete cases. However, at least with binary data in logistic regression, analyzing only the observed data does not produce biased estimates and tests if missingness is related to treatment assignment but not outcome (e.g., because of side effects unrelated to outcome) or outcome but not treatment assignment (e.g., those with events are more likely to be missing, but to the same degree in both arms). Biased estimates and tests can arise in the more likely scenario that missingness is related to both treatment assignment and outcome (Jones 1996, Proschan, McMahon, Shih 2001).

1.2.2. Complete-Case Analysis

As we know, standard statistical methods for regression analysis are designed for rectangular data sets, where all variables in the regression model are observed for all of the subjects. Therefore, when some variables are not observed for some subjects, the straightforward option is to analyze only those subjects that are completely observed for all of the variables. This method, known as complete case analysis (CC), which simply excludes records with missing data from statistical analysis, is the technique most commonly used when missing values are present. When the data are MCAR, CC analysis leads to consistent estimates. Therefore, it generally requires the MCAR assumption to apply CC analysis on a data set with missing values, because the CC analysis could be inconsistent when the data are not MCAR. However, as has been discussed in detail by Little and Rubin (1987, 2002), in some cases, CC analysis can be applied with less restrictive conditions, e.g. MAR, and the parameter estimates are still consistent. In regression analysis, when the response is subject to missing data, it has been shown that CC analysis produces consistent estimates if the missingness in the response is unrelated to the response itself

(Glynn 1993; Little 1992). When covariates are subject to missing data, Jones (Jones, MP 1996) proved theoretically that complete case analysis produces consistent estimates provided that the missingness in the covariates is unrelated to the outcome.

Table 1.1 Validity of complete case analysis in logistic regression

Variable missing	Missingness Mechanism	Odds Ratio on X	Odds Ratio on $\mathbf{C}$
Outcome	$m_y \mid x,c$ (MAR)	Valid	Valid
	$m_y \mid y$ (NMAR)	Valid	Valid
	$m_y \mid y,x,c$ (NMAR)	Not Valid	Not Valid*
Exposure	$m_x \mid y,c$ (MAR)	Valid	Not Valid**
	$m_x \mid x$ (NMAR)	Valid	Valid
	$m_x \mid y,x,c$ (NMAR)	Not Valid	Not Valid*

* Assuming the covariates $\mathbf{C}$ and the exposure are not independent
** As pointed out by Robins, Rotnitzky and Zhao (1994)

On the other hand, NMAR does not necessarily lead to inconsistent estimates. One example is the case-control study in retrospective epidemiology studies. The subjects are included in the study with unbalanced screening rate where the group with smaller prevalence will be included with higher rate, so that to reach a relative balanced number of subjects in case and control. The selection is an analogous process of missingness, and it depends on the underlying value, which constitute a NMAR regarding the outcome. However, the estimates of the logistic regression parameters obtained from the unbalance-selection data set are still valid in such studies.

1.2.3. Maximum Likelihood Method

The maximum likelihood (ML) approach was proposed to deal with missing data problems under the generalized linear model setup by Little and Schluchter (1985), Ibrahim (1990), Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996) and Ibrahim, Chen, Lipsitz et al. (2005) by constructing the likelihood function of the outcome, the covariates, together with the missingness indicator. The joint distribution could be written as the product of a series of conditional probabilities, as proposed by Lipsitz and Ibrahim (1996) and Ibrahim, Lipsitz and Chen (1999) and summarized by Little and Rubin (2002). Chen, Ibrahim, Shao (2004) put the ML approach into a Bayesian framework. The joint distribution of the covariates was modeled, in the monotone missing case, as a product of one-dimensional parametric conditional distributions for the covariates that have missing values. On the other hand, nonparametric and semi-parametric approaches for specifying the covariate distribution have been considered by Chen and Little (1999) and Chen (2002, 2004). In application of ML, numerical maximization is usually needed to obtain the maximum likelihood estimate (MLE).

Dempster, Laird and Rubin (1977) proposed a general algorithm to obtain MLEs when incomplete data present. A general method for estimation in the presence of missing covariates has been proposed by Ibrahim (1990), who used EM via a method of weights to find the MLEs. Ibrahim and Lipsitz (1996) applied Ibrahim's weighting approach with the EM algorithm in binomial regression with non-ignorable non-response. Lipsitz, Ibrahim, Chen and Peterson (1999) and Lipsitz, Ibrahim, Chen (1999) proposed a likelihood method for estimating parameters in generalized linear models with NMAR missing covariates. The EM algorithm was used and a closed form of the E step was given in case of categorical covariates and Monte Carlo EM algorithm was used in the case of continuous covariates. Horton and Laird (1999) discussed the application of EM via the method of weights in detail, with illustrative examples. Although commonly cited as a different estimation method, the EM algorithm can be considered as an alternative approach that applies maximum likelihood estimation to a general model with additional assumptions on the distribution of the variables subject to missingness. Examples were

given to show that the distributional form assumed for the unobserved variables may actually determine the missing mechanism (Little and Rubin 1987, Kenward 1998). Likelihood methods by the EM algorithm usually require complex self-written programs, which prevent it from wide application by epidemiologists.

1.2.4. Inverse Propensity Weighting

The treatment of missing data via a weighting adjustment arises in the survey sampling literature. The non-response weight is a factor that multiplies the sampling weight to account for the differential sampling probability, as in the Horvitz-Thompson estimator (1952). Rosenbaum and Rubin's (1983) theory of propensity scores (so called "estimated propensity scores" in that context) shows that response propensity weighting effectively removes non-response bias when non-response is random within subpopulations (David et al. 1983). Little (1988) proposed to model the binary non-response indicator on the other variables using either logistic or probit regression when the number of variables is relatively large for respondents and non-respondents. Little (1986, 1991) compared response propensity weighting with mean imputation within subclasses. Flanders and Greenland (1991) and Zhao and Lipsitz (1992) suggested a weighted estimator. Wang, Wang, Zhao et al. (1997) proposed semi-parametric modeling of the missing probability via kernel smoothing. Zhao and Lipsitz (1992) and Little and David (1983) showed how the weighting method can be extended to handle monotone patterns of non-response, such as those occurring with attrition from a panel study.

Although its appropriateness can and should be assessed for a given data set, a second logistic regression model for missingness is a convenient and intuitive choice in practice, as has been done by Rosenbaum and Rubin (1983, 1984, and 1985), Rosenbaum (1984), and Little (1988). Then the predicted response propensities can be derived for respondents and non-respondents, and the weights applied to respondents are proportional to the inverse of the response rates. Although the logistic regression for missingness on the other variables is univariate, analysis this regression is still a nontrivial task if the set of variables is very large. In

practice, judicious selection from the available variables, based on a priori knowledge and preliminary analysis, may be necessary (Little 1988). Propensity score weighting can lead to estimates with large variance, as discussed in Little (1986), where empirical Bayesian methods were proposed to smooth the weights. D'Agostino (1998) made a comparative review of several ways of using propensity scores to estimate the treatment effect. More reference for IPW can be found in Rosenbaum (1987) and Xie and Liu (2005). Successful examples of application of propensity weighting can be found in Czajka Hirabayashi, Little, and Rubin (1987), Little and Rubin (2002) and Hogan and Lancaster (2004).

1.2.5. Weighted Estimating Equations

Robins, Rotnitzky and Zhao (1994) proposed a semi-parametric approach using weighted estimation equations (WEE), which was then extended by Robins and Ritov (1997) to obtain consistent estimates of the regression parameters when either the missing data mechanism or the score vector for the missing data given the observed data (or both) can be correctly specified. This attractive property against model misspecification is often cited as "double robustness". More discussions on this method can be found in Robins and Rotnitzky (1995, 2001), Rotnitzky, Robins, and Scharfstein (1998), Robins, Rotnitzky, and Scharfstein (2000). Lipsitz, Ibrahim and Zhao (1999) proposed a WEE method for missing covariate data with close underlying connections to the maximum likelihood approach. The proposed WEE has an almost identical form to the ML estimating equations. To be more specific, suppose the problem of interest is to estimate the regression parameter β associating a binary outcome Y and a binary exposure X, where X is subject to missing data for some observations. There are additional covariates $\mathbf{C}$ that are completely observed. Then consider a logistic model:

$$p(Y=1 \mid x, \mathbf{c}; \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2' \mathbf{c})}{1+\exp(\beta_0 + \beta_1 x + \beta_2' \mathbf{c})}, \qquad (x = 0,1)$$

An additional logistic model

$$p(X=1 \mid \mathbf{C}; \Theta) = \frac{\exp(\theta_0 + \theta_1' \mathbf{C})}{1+\exp(\theta_0 + \theta_1' \mathbf{C})}$$

and a logistic model for the probability of being observed

$$\pi_i = \pi_i(\mathrm{T}) = \frac{exp(\tau_0 + \tau_1 y + \tau_2' \mathbf{c})}{1 + exp(\tau_0 + \tau_1 y + \tau_2' \mathbf{c})}$$

The density of $(y_i, x_i, r_i \mid \mathbf{c}_i)$ for subject i is given by

$$\begin{aligned} & p(y_i, x_i, r_i \mid \mathbf{c}_i; \boldsymbol{\beta}, \Theta, \mathrm{T}) \\ & \quad = p(y_i \mid x_i, \mathbf{c}_i; \boldsymbol{\beta})\, p(x_i \mid \mathbf{c}_i; \Theta)\, p(r_i \mid y_i, \mathbf{c}_i, x_i; \mathrm{T}) \\ & \quad = p(y_i \mid x_i, \mathbf{c}_i; \boldsymbol{\beta})\, p(x_i \mid \mathbf{c}_i; \Theta)\, p(r_i \mid y_i, \mathbf{c}_i; \mathrm{T}) \end{aligned}$$

The main interest is in estimation of $\boldsymbol{\beta}$. Then the score functions are given by

$$\begin{aligned} u_1(\boldsymbol{\beta}) &= \sum_{i=1}^{n} u_{1i}(\boldsymbol{\beta}; y_i, x_i, \mathbf{c}_i) = \sum_{i=1}^{n} \frac{\partial \log \Pr(y_i \mid x_i, \mathbf{c}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ u_2(\Theta) &= \sum_{i=1}^{n} u_{2i}(\Theta; x_i, \mathbf{c}_i) = \sum_{i=1}^{n} \frac{\partial \log \Pr(x_i \mid \mathbf{c}_i; \Theta)}{\partial \Theta} \\ u_3(\mathrm{T}) &= \sum_{i=1}^{n} u_{3i}(\mathrm{T}; r_i, y_i, \mathbf{c}_i) = \sum_{i=1}^{n} \frac{\partial \log \Pr(r_i \mid y_i, \mathbf{c}_i; \mathrm{T})}{\partial \mathrm{T}} \end{aligned}$$

Define a set of weighted score functions as follows:

$$\begin{aligned} S(\Gamma) &= \begin{bmatrix} S_1(\boldsymbol{\beta}, \Theta, \mathrm{T}) \\ S_2(\boldsymbol{\beta}, \Theta, \mathrm{T}) \\ S_3(\mathrm{T}) \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} S_{1i}(\boldsymbol{\beta}, \Theta, \mathrm{T}) \\ S_{2i}(\boldsymbol{\beta}, \Theta, \mathrm{T}) \\ S_{3i}(\mathrm{T}) \end{bmatrix} \\ &= \sum_{i=1}^{n} \begin{bmatrix} \frac{r_i}{\pi_i} u_{1i}(\boldsymbol{\beta}; y_i, x_i, \mathbf{c}_i) + \left(1 - \frac{r_i}{\pi_i}\right) E_{x_i \mid y_i, \mathbf{c}_i}[u_{1i}(\boldsymbol{\beta}; y_i, x_i, \mathbf{c}_i)] \\ \frac{r_i}{\pi_i} u_{2i}(\Theta; x_i, \mathbf{c}_i) + \left(1 - \frac{r_i}{\pi_i}\right) E_{x_i \mid y_i, \mathbf{c}_i}[u_{2i}(\Theta; x_i, \mathbf{c}_i)] \\ u_{3i}(\mathrm{T}) \end{bmatrix} \end{aligned} \tag{1.1}$$

where $\Gamma' = (\boldsymbol{\beta}', \Theta', \mathrm{T}')$. Lipsitz et al. (1999) show that $E[S(\Gamma)] = 0$, as long as either the missing data mechanism π_i or the score vector $p(x_i \mid \mathbf{c}_i)$ is correctly specified, but not necessarily both. Therefore we obtain a set of WEE by setting (1.1) equal to zero. This approach can be implemented via an EM-type algorithm, which provides relatively easy access to investigators.

1.2.6. Multiple Imputation

Multiple imputation was developed as an improvement upon single imputation. In single

imputation, one attempts to construct a dataset that would have been observed if missingness did not occur. The missing values are filled in artificially by researchers using a variety of approaches, and then standard statistical analysis is applied to the filled-in data set as if it was observed without any missing values. Treating the imputed dataset as a truly observed one, single imputation is not recommended due to underestimation of variability. Multiple imputation was formulated by Rubin to correct upon single imputation (Rubin 1987). Rubin (1996) and Schafer (1997) gave comprehensive reviews of multiple imputation in both its fundamental theoretical results and practical objectives. The imputation process is repeated multiple times and standard statistical analysis is performed on each imputed dataset. The results from each single imputation are combined using Rubin's method to account for the uncertainty that should be built in to correct for treating each imputed dataset as if truly observed. The basic idea was first proposed by Rubin (1987) and elaborated in his book. Additional discussions of multiple imputation in regression analysis can be found in Little and Rubin (1989) and Landerman, Land and Pieper (1997).

A typical multiple imputation inference is performed in three steps:

1. The missing values are artificially generated m times to reconstruct m fully observed datasets;
2. The m reconstructed datasets are analyzed as complete data separately with any standard statistical methods that are designated for complete data analysis;
3. The results of each of the above analyses are combined using Rubin's method

Let β be a $k \times 1$ vector of parameters. Suppose that the vector-valued point estimate and the covariance matrix for the parameter vector β from the i th imputed data set are $\widehat{\beta}_i$ and $\widehat{\mathbf{W}}_i$, $i = 1, 2, \ldots, m$. Then the Rubin's method concludes that the combined point estimate is

$$\overline{\beta} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\beta}_i \tag{1.2}$$

and the associated covariance matrix is

$$\mathbf{T} = \overline{\mathbf{W}} + (1 + \frac{1}{m})\mathbf{B} \tag{1.3}$$

where

$$\overline{\mathbf{W}} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\mathbf{W}}_i$$

represents the within-imputation covariance matrix, and

$$\mathbf{B} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\beta}_i - \overline{\beta})(\hat{\beta}_i - \overline{\beta})'$$

represents the between-imputation covariance matrix.

Of course, certain requirements must be met for MI to have these desirable properties. First, the data must be missing at random (MAR), meaning that the probability of missing data on a particular variable can depend on other observed variables, but not on itself after controlling for the other observed variables. Second, the model used to generate the imputed values must be "correct" in some sense. Third, the model used for the analysis must match up, in some sense, with the model used in the imputation. All these conditions have been rigorously described by Rubin (1987, 1996).

1.2.7. Predictive Probability Weighting

Lyles and Lin (2010) proposed a weighting method to conduct sensitivity analysis for misclassification in logistic regression. Observed data together with investigator-supplied values for sensitivity and specificity parameters were used to produce corresponding positive and negative predictive values. These values were used to reconstruct an appropriately defined expanded data set with appropriate weights to be used in fitting the model of interest using standard statistical software. A close form of the weights was derived to facilitate convenient utilization of standard software packages with a weighting option to avoid numerical algorithms. The Jackknife method was proposed to incorporate uncertainty in the estimated weights into valid standard errors (see next section). A similar idea was also discussed by Fleiss, Levin, and Paik (2004), where weighted-type methods were applied via expanded datasets, within the framework of random 2×2 table and then extended to incorporate covariate.

1.2.8. Jackknife Resampling Method

The idea of the Jackknife method was first proposed by Quenoulle (1949) to reduce bias of point estimates, and then proposed by Tukey (1958) to provide estimates of standard errors. The Jackknife and bootstrap are now widely used in survey sampling and other applications (Efron and Tibshirani, 1993). There are underlying relationships between the methods. For the purposes of this dissertation, the Jackknife is found to be more stable.

As a resampling method, the Jackknife estimator is based on dropping a single or a set of observations from the sample. Following the same symbols as in Wu (1986), the process of Jackknife method is as follows. Suppose β is a $k \times 1$ vector of parameters, and let $\hat{\beta}$ be its estimate obtained from the original sample. Let $\hat{\beta}_{(i)}$ be the estimate of β obtained from the resampled data with the i-th observation dropped out. Then, for estimating a function of β, say $\theta = g(\beta)$, define $\hat{\theta} = g(\hat{\beta})$, $(\hat{\theta})_{(i)} = g(\hat{\beta}_{(i)})$, then the pseudo-values are defined as:

$$p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$$

Then the jackknife point estimator of θ is given by

$$\tilde{\theta} = \frac{1}{n}\sum_{1}^{n} p_i ,$$

and the jackknife variance estimator for $\hat{\theta}$ is given by

$$v_J = \frac{1}{n(n-1)}\sum_{1}^{n}(p_i - \tilde{\theta})(p_i - \tilde{\theta})' .$$

The asymptotic properties of the jackknife method were studied by Miller (1974).

1.2.9. Sensitivity Analysis with Data Missing Not-At-Random

The concept of sensitivity analysis has a long history. As Little (1982) noted, if the response mechanism is non-ignorable, one can eliminate bias only by constructing "a model that correctly represents the response mechanism". Nordhein (1984) studied the prevalence of a genetic abnormality with sensitivity analysis by assuming the missing mechanism through the relative risk of missing rates. A closed form of the MLE was provided. The method is derived in the case

of a 2×2 table, but extension to higher dimensional contingency tables was provided. In Vach and Blettner (1995), the importance of sensitivity analysis is addressed. The work was based on an ML framework with a self-written program, and sensitivity analysis is based on the framework of specifying relative risks and odds ratios. Non-differential violation of MAR was assumed to simplify the specification of the missing mechanism. "Non-differential' violation here means that the dependence of a missing rate on the true value of a covariate does not depend on the outcome variable. On the other hand, Molenberghs, Goetghebeur, Lipsitz, and Kenward (1999) pointed out that in the contingency table setting, different models on the missing mechanism might give different prediction of the unobserved values, even though they all produce the same fit to the observed data. Therefore, Molenberghs, Kenward, Goetghebeur (2001) argued that the role of such sensitivity analysis is to supplement information obtained from the MAR model.

In traditional sensitivity analysis, a series of alternatives are specified and treated as known and fixed. This aids detection of potential bias in point estimate, but underestimate the variability of the estimate by ignoring the uncertainty in the specified alternatives. Monte Carlo sensitivity was proposed to corporate this additional variability from a simulation perspective (Lash and Fink 2003; Fox, Lash and Greenland 2005). An informative prior is specified to describe the probability distribution of the underlying alternatives based on previous knowledge. The estimate of the parameter of interest is summarized by the posterior by pooling the estimates from a collection of alternatives randomly generated from the prior.

1.2.10. Reassessment Data in Missing Data Problems

The idea of two stage sampling has been discussed to deal with misclassification problems by Breslow and Cain (1988), Flanders and Greenland (1991) and Zhao and Lipsitz (1992). Lyles and Allen (2002) explored a similar idea in case-control studies with non-ignorably missing exposure status by incorporating supplemental data from a second stage of sampling, termed a 'reassessment' study. This approach involves selecting a subset only among those with missing data and applying an intensive effort to obtain the information. Analytic expressions for the odds

ratio and relative risk were given by Lyles and Allen (2003) for cross-sectional studies with a binary outcome and/or a simple binary exposure subject to missingness. They estimated these parameters via a likelihood approach under five different missing data patterns.

Chapter 2. A WEIGHTING METHOD FOR LOGISTIC REGRESSION WITH DATA MISSING-AT-RANDOM

2.1. Introduction

In biometric research, missing data are often encountered due to many reasons, including the unavailability of measurements, survey non-response, study subjects failing to report for evaluations, respondents refusing to answer certain items on a questionnaire, and loss of data. Because standard techniques for regression models are designed for complete data sets, the straightforward option is to analyze only those subjects that are completely observed. For instance, suppose one is interested in the logistic regression model of a binary outcome variable Y (e.g., lung cancer or not) on the risk exposure X, a binary predictor, controlling for some other factors $\mathbf{C}$. Suppose that the predictor X is subject to missing values and the probability of X missing could depend on Y and/or $\mathbf{C}$, but not on X itself after conditioning on Y and $\mathbf{C}$ (X is MAR). In this case, a common practice is to perform the complete-case (CC) analysis on records $(y, x, \mathbf{c})$, for which values of all variables are observed. The usual parameter estimates are then still consistent, as discussed by Glynn (1985), Little (1992), and Jones (1996).

Although the complete case analysis works well in certain situations, this strategy obviously causes loss of information. As the fraction of missing data increases, the deletion of all subjects with missing values is unnecessarily wasteful and quite inefficient. In addition, complete-case analysis violates the intention-to-treat principles currently widespread in biometric research (Nich and Carroll 2002; Liu and Gould 2002; Hollis 2002). Finally, by excluding incomplete records under MAR, it might result in biased estimates of the regression coefficients for the covariates $\mathbf{C}$ that are not subject to missing values (Robins, Rotnitzky and Zhao 1994). If the probability of X being missing depends on $\mathbf{C}$, by omitting the unit all together, complete-case analysis results in NMAR for covariates $\mathbf{C}$. Therefore it is very important to point out a common misunderstanding, and emphasize it that if one is interested in multivariable analysis where the parameter estimates

of the controlling variables are also of interest, then the CC analysis is generally not recommended even if the missing-data mechanism is MAR. This is one of the motivations of the proposed method. By augmentation of the missing values and making use of all the available data, the parameters of the controlling variables could be estimated consistently.

It has been an active area of research to develop methods for regression analysis with missing data. Methods have been proposed in the past decades, i.e. maximum likelihood approaches (ML), the inverse propensity weighting approach (IPW), multiple imputation (MI), and so on. Little (1992) gave an exclusive review focused on multivariate normal models, whereas Horton and Laird (1999) focused exclusively on ML methods for GLMs with MAR categorical covariates. Ibrahim, Chen, Lipsitz et al. (2005) made a comprehensive review for these methods for cases with categorical or continuous covariates. As the development of these methods is active and becomes fruitful, we do have to note that they are built on certain assumptions on the distribution of the variables with missing values and/or the mechanism that generates the missing values. Thus they retain information at the cost of sensitivity to model specification of their own kind.

In this chapter, I will review the theoretical background of several proposed methods, summarize important issues of their application, and finally propose a novel method in the sense of its straightforward implementation in application with standard statistical software. In this chapter, I will specifically focus on logistic regression analysis of a binary outcome Y on the exposure X, and a vector of covariates $\mathbf{C}=(C_1,\ldots,C_p)'$. The outcome Y and the exposure X can be missing for some records, whilst the vector of covariate $\mathbf{C}$ is observed for all records. The mechanism inducing missing values is assumed to be independent of the missing values itself after conditioning on the other variables (MAR).

2.2. Methods

Missing data is a common problem in statistical analysis. In this chapter, model building and inference are built on a parametric framework throughout. We focus on binary exposure and

outcome, but the extension of the proposed method to handle missing values in categorical exposure with multiple levels will be discussed. Suppose one is interested in a logistic regression model of the outcome on the predictor variable and other covariates, i.e.,

$$\begin{aligned} & \text{logit}[\Pr(Y=1 \mid X=x, \mathbf{C}=\mathbf{c})] = \alpha + \beta x + \gamma' \mathbf{c} \\ \text{or} \quad & \Pr(Y=1 \mid X=x, \mathbf{C}=\mathbf{c}) = \frac{\exp(\alpha + \beta x + \gamma' \mathbf{c})}{1+\exp(\alpha + \beta x + \gamma' \mathbf{c})} \end{aligned} \tag{2.1}$$

where Y is a binary outcome, such as an indicator for disease, and X is a binary predictor variable, representing exposure to a certain risk factor. $\mathbf{C}$ is a vector of adjusting covariates that takes any form (continuous or categorical or a mixture of both). In the following discussion, we focus on cases where only one variable is subject to missing data, but not both. The method to handle multiple missing variables can be extended easily with some additional assumptions as discussed later.

2.2.1. Outcome Missing

There have been many discussions on the outcome missingness problem in a logistic regression analysis. The related problem of missing values in the outcome Y was prominent in the early history of missing data methods, but is less interesting in the following sense: if the $X's$ are complete and the missing values of Y are missing at random, then the incomplete cases contribute no information to the regression relationship of Y on the $X's$. In a logistic regression model as in (2.1), the contribution of each record is $\Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i)$. Define m_i as the missingness indicator for subject i such that $m_i = 1$ if the outcome Y_i is missing and $m_i = 0$ otherwise. If a record is complete, then its contribution to the likelihood funcion is

$$\begin{aligned} & \Pr(Y_i = y_i, m_i = 0 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \\ = \; & \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \Pr(m_i = 0 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \end{aligned} \tag{2.2}$$

If the outcome Y of a record is missing, then the contribution of the record becomes

$$
\begin{aligned}
& \Pr(m_i = 1 \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \\
= \; & \Pr(Y_i = 1, m_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) + \Pr(Y_i = 0, m_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \\
= \; & \Pr(Y_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \Pr(m_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \\
& \quad + \Pr(Y_i = 0 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \Pr(m_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \\
= \; & [\Pr(Y_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) + \Pr(Y_i = 0 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i)] \\
& \quad \times \Pr(m_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i)
\end{aligned}
\tag{2.3}
$$

The above equations hold under the assumption of "ignorable" missingness as defined by Rubin (1976) and Little and Rubin (2002). Obviously, when Y is missing, the terms containing the parameters of the model of interest are factorized out and take summation to 1, leaving only the term containing the parameters of the missingness model. Therefore, the likelihood contribution of the records with $(m_i = 1 \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i)$ actually does not involve the parameters of the model of interest. Thus in a logistic regression model, if the outcome is missing, there is no information gained regarding the regression relationship. Therefore, the likelihood function with consideration of the missing mechanism is identical to the likelihood function using only complete cases with respect to estimating the primary parameters. In conclusion, in logistic regression analysis, the estimates of the regression parameters by CC analysis are consistent and efficient when the outcome is missing.

2.2.2. Predictor Variable Missing

The problem of covariates missing at random has been studied extensively. Complete case analysis (CC) is commonly used, with which one would again simply ignore the incomplete cases and conduct statistical inference as usual with the complete cases. This approach is used blindly in many statistical analyses, although it is not fully efficient, and sometimes induces bias. There are many approaches proposed in previous literature targeted at improvements upon the CC analysis. In this chapter, the following widely used methods are reviewed: a maximum likelihood approach (ML), the inverse propensity weighting (IPW) approach, multiple imputation (MI) and weighted estimating equations (WEE). Furthermore, a novel approach is proposed, which can be easily implemented with standard statistical software, namely the predictive probability weighting

approach (PPW). The performances of these methods will be compared to the CC analysis by simulation studies under different settings.

Consider a simple 2×2 table where Y indicates a binary outcome (e.g. hypertensive versus normal) and X indicates a binary predictor indicating a risk exposure (e.g. smoker versus nonsmoker). Suppose data are missing on the predictor variable for some observations. Such a dataset can be expressed in the following table.

Table 2.1 Data missing at random in a 2×2 table

$n_{m,y,x}$	Smoker not missing $(X=1,m=0)$	Nonsmoker not missing $(X=0,m=0)$	Smoking status missing $(m=1)$
Hypertensive $(Y=1)$	n_{011}	n_{010}	$n_{11\cdot}$
Normal $(Y=0)$	n_{001}	n_{000}	$n_{10\cdot}$

Define m as the missing indicator that takes value 1 if the value of X is missing and 0 if it is not missing (note that in some literature, R is used to represent the missing/observed status, with $R=1$ as observed (response) and $R=0$ as missing (non-response)). We are interested in statistical inference on the odds ratio or the logarithm transformation of the odds ratio.

In this case, the assumption of MAR defined by Rubin (1976) can be expressed as

$$\Pr(m=1\mid Y=y,X=1)=\Pr(m=1\mid Y=y,X=0)=\Pr(m=1\mid Y=y) \tag{2.4}$$

$$\text{or}\quad \Pr(X=1\mid Y=y,m=1)=\Pr(X=1\mid Y=y,m=0)=\Pr(X=1\mid Y=y) \tag{2.5}$$

The two equations (2.4) and (2.5) are equivalent. With additional covariates $\mathbf{C}$, the MAR assumption can be stated as

$$\begin{aligned}&\Pr(m=1\mid Y=y,X=x,\mathbf{C}=\mathbf{c})=\Pr(m=1\mid Y=y,\mathbf{C}=\mathbf{c})=Pm_{y,\mathbf{c}}\\ \text{or}\quad &\Pr(X=1\mid Y=y,\mathbf{C}=\mathbf{c},m=1)=\Pr(X=1\mid Y=y,\mathbf{C}=\mathbf{c},m=0)=Pr(X=1\mid Y=y,\mathbf{C}=\mathbf{c})\end{aligned} \tag{2.6}$$

In other words, the conditional probability that the variable of interest X missing, defined as $Pm_{y,\mathbf{c}}$, does not depend on the value of X itself after conditioning on Y and $\mathbf{C}$.

2.2.2.1. Complete-Case Analysis

In complete-case analysis, the subjects with missing data are totally ignored. In case of a 2×2 table, as in Table 2.1, the last column with $m=1$ is omitted from analysis. The odds ratio is directly estimated by

$$\widehat{OR}_{\text{CC}} = \frac{n_{011}\times n_{000}}{n_{010}\times n_{001}} \tag{2.7}$$

Assuming MAR, the above odds ratio estimator is unbiased, and its standard error can be estimated via the Delta method in the usual fashion. With or without covariates, if one is only interested in inference about the effect of X on Y, CC analysis can be used as a valid approach, although not efficient. When there are covariate(s) $\mathbf{C}$ where $\mathbf{C}$ is a vector of categorical covariates, an odds ratio estimate can be easily achieved in the same way as above, stratified for each set of values of $\mathbf{C}$. However, we are often interested in estimating a common odds ratio across $\mathbf{C}$, especially when there are continuous components in $\mathbf{C}$. To be more specific, with binary outcome Y, a binary predictor variable X and covariate(s) $\mathbf{C}$, one is usually interested in fitting a logistic regression model

$$\text{logit}[\Pr(Y=1\mid X=x,\mathbf{C}=\mathbf{c})] = \alpha+\beta x+\gamma'\mathbf{c}, \qquad (x=0,1) \tag{2.8}$$

where $\mathbf{C}$ can contain either categorical or continuous variables, or a mixture of both. In such cases, if one is interested in making multivariable inference on the data, CC analysis could lead to biased estimates for the regression coefficients γ. In fact, if $Pm_{y,\mathbf{c}} = \Pr(m_x=1\mid y,\mathbf{c})$ truly depends on the values of $\mathbf{C}$, i.e.

$$Pm_{y,\mathbf{c}} \neq Pm_y,$$

then the missingness is informative regarding $\mathbf{C}$ (NMAR for $\mathbf{C}$). And as will be discussed in the next chapter, if the missingness depends on both the outcome Y and the covariates $\mathbf{C}$, i.e. if $Pm_{y,\mathbf{c}} \neq Pm_y$ and $Pm_{y,\mathbf{c}} \neq Pm_{\mathbf{c}}$ both hold at the same time, then the CC estimates of the regression coefficients of $\mathbf{C}$ would be biased. Thus if multivariable estimation and inference is of interest for both X and $\mathbf{C}$, complete-case analysis may not be appropriate even if the less restrictive assumption MAR holds.

2.2.2.2. Maximum Likelihood Approach

In cases of a 2×2 table like in Table 2.1 where estimating a crude odds ratio is of interest, one can construct a logistic regression model as in equation (1), but without any covariates.

$$\begin{aligned} &\text{logit}[\Pr(Y=1\mid X=x)]=\alpha+\beta x \\ \text{or}\quad &\Pr(Y=1\mid X=x)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}, \quad x=0,1 \end{aligned} \tag{2.9}$$

Thus we can make statistical inference on the odds ratio through $\log(\widehat{OR})=\hat{\beta}$. To construct the likelihood function, one could consider the likelihood contribution of each subject. One can specify the joint distribution of (Y,X) by the conditional probability of Y given X and a marginal distribution of X. This approach for modeling the joint distribution of Y and X has been considered for GLMs by Rubin (1976), Little and Schluchter (1985), Ibrahim (1990), Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Lipsitz, Ibrahim and Zhao (1999), and Ibrahim, Chen, Lipsitz et al. (2005). Specifically with no additional covariates $\mathbf{C}$, a subject with X observed contributes the term

$$\begin{aligned} \Pr(Y_i=y_i,X_i=x_i,m_i=0) &= \Pr(m_i=0\mid Y_i=y_i,X_i=x_i)\Pr(Y_i=y_i\mid X_i=x_i)\Pr(X_i=x_i) \\ &= \Pr(m_i=0\mid Y_i=y_i)\Pr(Y_i=y_i\mid X_i=x_i)\Pr(X_i=x_i) \end{aligned} \tag{2.10}$$

and a subject with X missing contributes the term

$$\begin{aligned} \Pr(Y_i=y_i,m_i=1) &= \sum_{x=0}^{1}\Pr(m_i=1\mid Y_i=y_i,X_i=x)\Pr(Y_i=y_i\mid X_i=x)\Pr(X_i=x) \\ &= \sum_{x=0}^{1}\Pr(m_i=1\mid Y_i=y_i)\Pr(Y_i=y_i\mid X_i=x)\Pr(X_i=x) \end{aligned} \tag{2.11}$$

where the second equality of (2.10) and (2.11) holds under the MAR assumption. For ease of exposition, suppose that X is completely observed for the first n_{cc} cases and missing for the remaining $n-n_{cc}$ cases. Assuming ordering of the index i to reflect observations of the respective types, the likelihood function is proportional to

$$
\begin{aligned}
L \quad &= \prod_{i=1}^{n_{cc}} [\Pr(m_i = 0 \mid Y_i = y_i) \Pr(Y_i = y_i \mid X_i = x_i) \Pr(X_i = x_i)] \times \\
&\qquad \prod_{i=n_{cc}+1}^{n} [\sum_{x=0}^{1} \Pr(m_i = 1 \mid Y_i = y_i) \Pr(Y_i = y_i \mid X_i = x) \Pr(X_i = x)] \\
&= \prod_{i=1}^{n_{cc}} \Pr(m_i = 0 \mid Y_i = y_i) \times \prod_{i=n_{cc}+1}^{n} \Pr(m_i = 1 \mid Y_i = y_i) \times \\
&\qquad \prod_{i=1}^{n_{cc}} [\Pr(Y_i = y_i \mid X_i = x_i) \Pr(X_i = x_i)] \times \prod_{i=n_{cc}+1}^{n} [\sum_{x=0}^{1} \Pr(Y_i = y_i \mid X_i = x) \Pr(X_i = x)]
\end{aligned}
\tag{2.12}
$$

where n represents the sample size and n_{cc} represents the number of complete cases. The first two terms define the missingness mechanism, namely the missingness model, whilst the remaining terms are determined by the logistic regression model of interest, and a sub-model for the marginal distribution of X. The latter follows a Bernoulli distribution, with nuisance parameter $p = \Pr(X = 1)$. The justification of the likelihood function can be found in (Rubin 1976). With the setup in this chapter, the likelihood function (2.12) does have a closed form and can be factorized and numerically maximized directly. However, if the observed data likelihood in (2.10) does not have a closed form and cannot be factorized, approaches such as the EM algorithm are generally needed to obtain MLEs from (2.10) (Ibrahim, Chen, Lipsitz et al. 2005). A general method for estimation in the presence of missing covariates has been proposed by Ibrahim (1990), who used EM via a method of weights to find the MLEs. As a convenient closed form of the likelihood function exists under the set-up of this chapter, it is simple and clear to fixate on the direct numerical maximization of the closed-form log-likelihood function.

In addition to the MAR assumption, if we further assume that the parameters of the missing model are "distinct" from the parameters of the logistic regression model of interest, we can isolate the first two terms and focus on the remaining terms. These two assumptions together are referred to as an "ignorable" missing scenario by Rubin (1976). Thus the likelihood function of interest becomes

$$
L = \prod_{i=1}^{n_{cc}} \left[\Pr(Y_i = y_i \mid X_i = x_i) \Pr(X_i = x_i)\right] \prod_{i=n_{cc}+1}^{n} \left[\sum_{x_i=0}^{1} \Pr(Y_i = y_i \mid X_i = x_i) \Pr(X_i = x_i) \right] \tag{2.13}
$$

The logarithm transformation of the above likelihood function can be easily maximized through available numerical routines. We find quasi-Newton routines in SAS/IML program (SAS Institute, Cary, NC) quite straightforward and computationally stable.

In cases where there is a covariate vector $\mathbf{C}$, one can construct a logistic regression model as in (2.1):

$$\text{logit}[\Pr(Y=1 \mid X=x, \mathbf{C}=\mathbf{c})] = \alpha + \beta x + \gamma' \mathbf{c}$$

$$\text{or} \quad \Pr(Y=1 \mid X=x, \mathbf{C}=\mathbf{c}) = \frac{\exp(\alpha + \beta x + \gamma' \mathbf{c})}{1+\exp(\alpha + \beta x + \gamma' \mathbf{c})}, \qquad (x=0,1)$$

Thus we can make statistical inference on the common odds ratio through $\log(\widehat{OR}) = \hat{\beta}$.

To construct the likelihood function, one would again consider a full likelihood contribution for each subject. One useful strategy to model the joint distribution of $(X, \mathbf{C}')'$ was proposed by Lipsitz and Ibrahim (1996) and Ibrahim, Lipsitz and Chen (1999) and summarized by Little and Rubin (2002), where the joint distribution of the covariates was modeled, in the monotone missing case, as a product of one-dimensional parametric conditional distributions for the covariates that have missing values. Nonparametric and semi-parametric approaches for specifying the covariate distribution have been considered by Chen and Little (1999) and Chen (2002, 2004). In any case, a subject with X observed contributes the term

$$\begin{aligned} \Pr(Y_i = y_i, X_i = x_i, m_i = 0 \mid \mathbf{C_i} = \mathbf{c_i}) &= \Pr(m_i = 0 \mid Y_i = y_i, X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \times \\ &\quad \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \\ &= \Pr(m_i = 0 \mid Y_i = y_i, \mathbf{C_i} = \mathbf{c_i}) \times \\ &\quad \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \end{aligned}, \quad (2.14)$$

and a subject with X missing contributes the term

$$\begin{aligned} \Pr(Y_i = y_i, m_i = 1 \mid \mathbf{C_i} = \mathbf{c_i}) &= \sum_{x_i=0}^{1} \big[\Pr(m_i = 1 \mid Y_i = y_i, X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \times \\ &\quad \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \big] \\ &= \sum_{x_i=0}^{1} \big[\Pr(m_i = 1 \mid Y_i = y_i, \mathbf{C_i} = \mathbf{c_i}) \times \\ &\quad \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \big] \end{aligned}, \quad (2.15)$$

where the second equality of (2.14) and (2.15) holds under the MAR assumption. Assuming

ordering of the index i to reflect observations of the respective types, the likelihood function becomes

$$\begin{aligned} L &= \prod_{i=1}^{n_{cc}} \left[\Pr(m_i = 0 \mid Y_i = y_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \right] \\ &\quad \times \prod_{i=n_{cc}+1}^{n} \sum_{x_i=0}^{1} \left[\Pr(m_i = 1 \mid Y_i = y_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \right] \\ &= \prod_{i=1}^{n_{cc}} \Pr(m_i = 0 \mid Y_i = y_i, \mathbf{C_i} = \mathbf{c_i}) \prod_{i=n_{cc}+1}^{n} \Pr(m_i = 1 \mid Y_i = y_i, \mathbf{C_i} = \mathbf{c_i}) \\ &\quad \times \prod_{i=1}^{n_{cc}} \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \\ &\quad \times \prod_{i=n_{cc}+1}^{n} \sum_{x_i=0}^{1} \left[\Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C_i} = \mathbf{c_i}) \Pr(X_i = x_i \mid \mathbf{C_i} = \mathbf{c_i}) \right] \end{aligned} \quad ,(2.16)$$

where n represents the sample size and n_{cc} represents the number of complete cases. As before, the first two terms define the missing mechanism, whilst the following two terms are determined by the logistic regression model of interest, and a sub-model for the marginal distribution of X conditional on $\mathbf{C}$. A logistic regression may be a natural choice:

$$\text{logit}[\Pr(X = 1 \mid \mathbf{C} = \mathbf{c})] = \theta_0 + \boldsymbol{\theta}_1' \mathbf{c}, \quad (2.17)$$

The modeling of the conditional distribution of X on $\mathbf{C}$ may need other covariates and/or higher order terms of the original $\mathbf{C}$. One can use standard procedures of model selection to identify a proper form. Again, we can isolate the first two terms. Then we end up with the likelihood under ignorability, i.e.,

$$\begin{aligned} L &\propto \prod_{i=1}^{n_{cc}} \Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \Pr(X_i = x_i \mid \mathbf{C}_i = \mathbf{c}_i) \\ &\quad \times \prod_{i=n_{cc}+1}^{n} \sum_{x_i=0}^{1} [\Pr(Y_i = y_i \mid X_i = x_i, \mathbf{C}_i = \mathbf{c}_i) \Pr(X_i = x_i \mid \mathbf{C}_i = \mathbf{c}_i)] \end{aligned} \quad , \quad (2.18)$$

The parameters in the first probability term $(\alpha, \beta, \boldsymbol{\gamma})$ are of main interest, whilst the parameters in the second probability term $(\theta_0, \boldsymbol{\theta}_1)$ are nuisance parameters, as similarly proposed in Lipsitz, Ibrahim and Zhao (1999) and Ibrahim, Chen and Lipsitz (1999). Again, the logarithm transformation of the likelihood function can again be maximized through standard

statistical software.

The likelihood function in (2.16) is a joint distribution of the outcome Y and the exposure X. Thus, it is most clearly suitable in a cross-sectional study. It is very common in epidemiologic research, however, that data are collected for a case-control study where the cases are subject to oversampling to achieve relative balance between cases and controls for small prevalence incidences. In such a case-control study, the likelihood function in (2.16) does not describe the true joint distribution with cases over sampled. However, as we are only interested in the odds ratio between the outcome and the exposure, with similar argument by Prentice and Pyke (1979), one can still fit a logistic regression model to the case-control data via the same likelihood function with a prospective (random sampling) formulation as in (2.16). Prentice and Pyke showed that the odds ratio estimates together with the covariance matrix remained valid in a case-control study. Carroll, Wang and Wang (1995) extended this result to applications for dealing with measurement error and missing data. They proved theoretically that the estimators of the non-intercept parameters are consistent and asymptotically normally distributed in many cases. Assuming appropriate model specifications in (2.18), the estimates of the parameters of interest for multivariate analysis, β and γ, are still valid, although the parameter estimates for the conditional marginal distribution of X, $\Pr(X \mid \mathbf{C})$, would be invalid.

2.2.2.3. Inverse Propensity Weighting

The treatment of missing data via a weighting adjustment arises in the survey sampling literature. The non-response weight is used as a factor that multiplies the sampling weight to account for the differential sampling probability, as in the Horvitz-Thompson estimator (1952). Basically, this method reconstructs the information that would have been observed via weighting the complete records with proper weights, the inverse of the propensity of the complete record. Little (1988) proposed to model the binary non-response indicator R on the other variables X using either logistic or probit regression when the number of variables is relatively large for respondents and non-respondents. Then the predicted response propensities can be derived for

respondents and non-respondents, and the weights applied to respondents are proportional to the inverse of the response rates.

To be more specific, under the 2×2 table setting, each observed subject is weighted by the inverse of the probability that it is observed

$$w_{y,x} = \frac{1}{\Pr(m=0 \mid Y=y, X=x)} = \frac{1}{\Pr(m=0 \mid Y=y)} \tag{2.19}$$

The second equality holds under the MAR assumption. Thus the weights can be estimated via the marginal rates of Table 2.1:

$$\hat{w}_{y,x} = \frac{1}{\widehat{\Pr}(m=0 \mid Y=y)} = \left(\frac{n_{0,y,1}+n_{0,y,0}}{n_{0,y,1}+n_{0,y,0}+n_{1,y}}\right)^{-1} = \frac{n_{0,y,1}+n_{0,y,0}+n_{1,y}}{n_{0,y,1}+n_{0,y,0}} \tag{2.20}$$

After applying the weights, the odds ratio can be estimated by

$$\widehat{OR}_{\text{IPW}} = \frac{n^*_{011}\times n^*_{000}}{n^*_{010}\times n^*_{001}} \tag{2.21}$$

where $n^*_{0,y,x} = n_{0,y,x}\times \hat{w}_{y,x} = n_{0,y,x}\times \frac{n_{0,y,1}+n_{0,y,0}+n_{1,y}}{n_{0,y,1}+n_{0,y,0}}$. This is essentially the same as estimating the odds ratio from a reconstructed 2×2 table as in Table 2.2. The standard error estimate can be obtained via the Delta method.

Table 2.2 Reconstructed data set by IPW

$n_{0,y,x}^{\ *}$	Smoker not missing (X=1, m=0)	Nonsmoker not missing (X=0, m=0)
Hypertensive (Y=1)	$n_{011}\times\frac{n_{011}+n_{010}+n_{11}}{n_{011}+n_{010}}$	$n_{010}\times\frac{n_{011}+n_{010}+n_{11}}{n_{011}+n_{010}}$
Normal (Y=0)	$n_{001}\times\frac{n_{001}+n_{000}+n_{10}}{n_{001}+n_{000}}$	$n_{000}\times\frac{n_{001}+n_{000}+n_{10}}{n_{001}+n_{000}}$

Complete cases are weighted by the inverse propensity to reflect the underlying sample that would have been observed if there were no missing values.

In cases where there are covariates, if a subject has observation $(y, x, \mathbf{c})$, then the probability of it being completely observed can be defined as $\Pr(m = 0 \mid Y = y, X = x, \mathbf{C} = \mathbf{c})$. Since we assume MAR, the above probability does not depend on the value of X, and thus it reduces to $\Pr(m = 0 \mid Y = y, \mathbf{C} = \mathbf{c}) = 1 - \Pr(m = 1 \mid Y = y, \mathbf{C} = \mathbf{c}) = 1 - Pm_{y,\mathbf{c}}$. Following the same rationale of the inverse propensity weighting approach as above, the inverse of the above probability $\frac{1}{1 - Pm_{y,\mathbf{c}}}$ is used to weight each of the complete cases. To estimate $Pm_{y,\mathbf{c}}$, we notice that the sets $(m, Y, \mathbf{C})$ are completely observed for all subjects. We can assume a logistic regression model for the binary indicator of missingness m as the outcome, with the disease status Y and covariates $\mathbf{C}$ from that model as predictors (Cox 1970). Then the statistical analysis of this model could be conducted with the entire data set where all observations contain complete values. The only assumption needed here is the validity of such a logistic model for missingness fitted to the set $(m, Y, \mathbf{C})$. The logistic regression model for missingness assumed here is a convenient and intuitive choice in practice, as has been done by Rosenbaum and Rubin (1983, 1984, and 1985), Rosenbaum (1984), and Little (1988):

$$\text{logit}[\Pr(m = 1 \mid Y = y, \mathbf{C} = \mathbf{c})] = \text{logit}(Pm_{y,\mathbf{c}}) = \psi_0 + \psi_1 y + \psi_2' \mathbf{c} \tag{2.22}$$

The model assessment is indeed very important for this logistic model for the potential polynomial and interaction terms in Y and $\mathbf{C}$ (Rosenbaum and Rubin 1983, 1984). Incorrect model specification can lead to bias. Our simulation results also reveal the impressive sensitivity to model specification.

With the predicted probability $\widehat{Pm}_{y,\mathbf{c}}$ from the above model, one can make inference by maximizing a weighted likelihood function corresponding to the primary logistic model. Therefore, assuming ordering of the index i to reflect observations of the respective types, we are maximizing a weighted log-likelihood

$$l(\alpha, \beta, \gamma) = \sum_{i=1}^{n_{cc}} w_{y\mathbf{c}} l_{yx\mathbf{c}}(\alpha, \beta, \boldsymbol{\gamma}) \tag{2.23}.$$

Here, $l_{yx\mathbf{c}}(\alpha,\beta,\boldsymbol{\gamma})$ is the usual log-likelihood contribution for a $(y,x,\mathbf{c})$ set. For simplicity, the (i) subscripts on the l and w terms are suppressed. This practically means that we fit a logistic regression model to the complete part of the data set, applying weights $w_{y\mathbf{c}}=\dfrac{1}{1-\widehat{Pm}_{y\mathbf{c}}}$ to each complete subject. This can be accomplished through standard statistical software in one single step with a weighting option (e.g. using SAS PROC LOGISTIC (SAS Institute, Cary, NC)).

Pros and cons of IPW versus the other approaches will be discussed at the end of this chapter.

Direct estimate of variance is not appropriate because it does not account for the uncertainty in propensity scores estimated by the logistic regression (Jones and Chromy 1982 and Little 1988), which is generally difficult to quantify by a closed-form expression. However, the point estimator for the vector of logistic regression parameters remains valid. The variance in the point estimate could be evaluated by jackknife or bootstrap methods (see section 1.2.8).

2.2.2.4. Weighted Estimating Equations

Robins et al. (1994) proposed a class of estimators via weighted estimating equations (WEE). With WEE, the contribution to the estimating equation from a complete observation is weighted by the inverse of the probability that the observation is complete (propensity). As a method via estimating equations, it does not require full specification of the likelihood, therefore, provides attractive double robustness against model misspecification. The estimators are consistent if the missing data mechanism model, $(m\mid y,\mathbf{c})$, is correctly specified or the score vector of the missing data given the observed data, $(x\mid y,\mathbf{c})$, is correctly specified, but not necessarily both. Lipsitz et al. (1999) proposed a WEE method for missing data in covariates which can be implemented via an EM-type algorithm. To be more specific, under the setting of this chapter, the conditional expectation terms of the score functions (1.1) can be written as a summation of weighted score functions:

$$E_{x_i|y_i,\mathbf{c}_i}[u_{1i}(\boldsymbol{\beta};y_i,x_i,\mathbf{c}_i)]=\sum_{x=0,1}\Pr[x_i=x\mid y_i,\mathbf{c}_i]u_{1i}(\boldsymbol{\beta};y_i,x,\mathbf{c}_i)$$
$$E_{x_i|y_i,\mathbf{c}_i}[u_{2i}(\Theta;x_i,\mathbf{c}_i)]=\sum_{x=0,1}\Pr[x_i=x\mid y_i,\mathbf{c}_i]u_{2i}(\Theta;x,\mathbf{c}_i)$$

Thus, the score vector in (1.1) becomes

$$S(\Gamma)=\sum_{i=1}^{n}\begin{bmatrix}\dfrac{r_i}{\pi_i}u_{1i}(\boldsymbol{\beta};y_i,x_i,\mathbf{c}_i)+\left(1-\dfrac{r_i}{\pi_i}\right)\left[w_iu_{1i}(\boldsymbol{\beta};y_i,x=1,\mathbf{c}_i)+(1-w_i)u_{1i}(\boldsymbol{\beta};y_i,x=0,\mathbf{c}_i)\right]\\ \dfrac{r_i}{\pi_i}u_{2i}(\Theta;x_i,\mathbf{c}_i)+\left(1-\dfrac{r_i}{\pi_i}\right)\left[w_iu_{2i}(\Theta;x=1,\mathbf{c}_i)+(1-w_i)u_{2i}(\Theta;x=0,\mathbf{c}_i)\right]\\ u_{3i}(\mathrm{T})\end{bmatrix} \quad (2.24)$$

with weights $w_i=\Pr(x_i=1\mid y_i,\mathbf{c}_i)$. The WEEs are then found by setting $S(\hat{\Gamma}_{WEE})=0$. As shown by Lipsitz et al. (1999), the estimating equation for the missing mechanism $\pi_i(\mathrm{T})$ does not involve $\boldsymbol{\beta}$ or Θ. As a matter of fact, $\hat{\mathrm{T}}_{WEE}=\hat{\mathrm{T}}$ is the MLE obtained via ordinary logistic regression. Therefore, given the MLE $\hat{\pi}_i$ as known, an EM-type algorithm can be applied to obtain the estimate of $\hat{\beta}_{WEE}$ and $\hat{\Theta}_{WEE}$. In passing we note here that this implies that the MLE $\hat{\pi}_i$ is treated as known therefore the variability in it is not taken into account. If we define the function

$$S(\Gamma\mid\Gamma^{(t)})=\sum_{i=1}^{n}\begin{bmatrix}\dfrac{r_i}{\pi_i}u_{1i}(\boldsymbol{\beta};y_i,x_i,\mathbf{c}_i)+\left(1-\dfrac{r_i}{\pi_i}\right)\left[w_i^{(t)}u_{1i}(\boldsymbol{\beta};y_i,x=1,\mathbf{c}_i)+(1-w_i^{(t)})u_{1i}(\boldsymbol{\beta};y_i,x=0,\mathbf{c}_i)\right]\\ \dfrac{r_i}{\pi_i}u_{2i}(\Theta;x_i,\mathbf{c}_i)+\left(1-\dfrac{r_i}{\pi_i}\right)\left[w_i^{(t)}u_{2i}(\Theta;x=1,\mathbf{c}_i)+(1-w_i^{(t)})u_{2i}(\Theta;x=0,\mathbf{c}_i)\right]\\ u_{3i}(\mathrm{T})\end{bmatrix}$$

where $w_i^{(t)}$ is the conditional probability in (2.24) evaluated at $\Gamma^{(t)}$. Then the EM-type algorithm is as follows:

1. Obtain initial estimate $\Gamma=\Gamma^{(1)}$. At the t th step, we have $\Gamma^{(t)}$
2. Calculate $w_i^{(t)}=w_i(\Gamma^{(t)})$ using the current estimate

3. Treating $w_i^{(t)}$ as given and fixed, solve $S(\Gamma \mid \Gamma^{(t)}) = 0$ to get updated estimate $\Gamma^{(t+1)}$

4. Repeat step 2 and 3 until convergence. $\dot{S}$

The asymptotic variance has a "sandwich" form (White 1982),

$$Var(\hat{\Gamma}_{WEE}) = \left\{ \sum_{i=1}^{n} E[\dot{S}_i(\Gamma)] \right\}^{-1} \sum_{i=1}^{n} E[S_i(\Gamma) S_i(\Gamma)'] \left\{ \sum_{i=1}^{n} E[\dot{S}_i(\Gamma)]' \right\}^{-1}$$

where $\dot{S}_i(\Gamma) = \dfrac{\partial S_i(\Gamma)'}{\partial \Gamma}$ (Lipsitz et al. 1999).

We notice that the double robustness of the WEE is an asymptotic property. Ibrahim, Chen, Lipsitz et. al. (2005) noted that in small sample simulations, there were cases where negative weights arise, which often lead to no unique solution. Furthermore, there is underlying similarity between the WEE and the ML score estimating equations as discussed by Lipsitz, Ibrahim, and Zhao (1999). The ML requires correct specification of $(Y \mid X, \mathbf{C})$ and $(X \mid \mathbf{C})$, but does not rely on the model assumption of the missing mechanism $(m \mid Y, \mathbf{C})$. The WEE requires either a correct specification of $(X \mid \mathbf{C})$ or $(m \mid Y, \mathbf{C})$, but not both. In that sense it is arguably more robust than ML in terms of more flexible model assumptions. Therefore, the WEE is more promising in reducing bias in certain cases. However, again this double robustness is an asymptotic property. It requires large sample size and a relatively high missingness rate. The latter is needed to achieve precision in estimating the missing data mechanism. The simulation studies by Lipsitz, Ibrahim, and Zhao (1999) suggest a trend of relatively smaller bias using WEE compared to ML, but the conclusion is not definitive.

2.2.2.5. Multiple Imputation

As discussed in Chapter 1, in multiple imputation, the imputation process is repeated multiple times and standard statistical analysis is performed on each imputed dataset. The results from each single imputation are combined using Rubin's method to account for the uncertainty that should be built in to correct for treating each imputed dataset as if truly observed.

There are many ways that MI imputes artificial datasets. In the example with a 2×2 table, one can easily impute new datasets using marginal rates of the complete cases. Each imputed

dataset is a 2×2 table, for which many standard statistical methods can be applied. When there are covariates $\mathbf{C}$, especially when there are continuous components in $\mathbf{C}$, a natural approach is to assume a logistic regression model of X, the predictor variable with missing data, on the other covariates $\mathbf{C}$, and the outcome of the model of interest, Y (Rubin 1976), namely the "imputation model":

$$\text{logit}[\Pr(X=1|Y=y,\mathbf{C}=\mathbf{c})]=\tau_0+\tau_1 y+\tau_2'\mathbf{c} \tag{2.25}$$

Polynomial and interaction terms may apply here, and standard methods can be applied for model selection on those terms. When the missingness is monotonic, and the variable with missing values is binary, the standard SAS procedure MI would conduct multiple imputation based on the above imputation model. However, we note here in passing that the appropriateness of model (2.25) may be called into question in some extreme circumstances. We return to this issue in section 2.2.2.6.

For the logistic model in (2.25), where the outcome X has missing values, CC analysis is equivalent to the ML approach, as has been discussed in section 2.1. Thus one can achieve consistent estimates $\hat{\tau}=\begin{pmatrix}\hat{\tau}_0\\ \hat{\tau}_1\\ \hat{\tau}_2\end{pmatrix}$, with covariance matrix Σ_τ, by fitting the above model to the complete cases. Rubin suggested that a "proper" imputation (Rubin 1987) should first draw random values of the imputation parameters from the proper posterior distribution to construct an imputation model, and then use it to generate imputed values. Thus an imputed value of X will be generated as follows.

1. The predicted probability of X being 1 is generated as

$$\Pr^{\dagger}(X=1|Y=y,\mathbf{C}=\mathbf{c})=\frac{\exp(\tau_0^{\dagger}+\tau_1^{\dagger}y+\tau_2^{\dagger\prime}\mathbf{c})}{1+\exp(\tau_0^{\dagger}+\tau_1^{\dagger}y+\tau_2^{\dagger\prime}\mathbf{c})} \tag{2.26}$$

where $\begin{pmatrix}\tau_0^{\dagger}\\ \tau_1^{\dagger}\\ \tau_2^{\dagger}\end{pmatrix}\sim N(\hat{\tau},\Sigma_\tau)$ is randomly generated from the multivariate normal distribution,

where $(\hat{\tau},\Sigma_\tau)$ is the point estimate and variance-covariance matrix by fitting the model

(2.25) to the complete cases;

2. The imputed value of X, $x^{\dagger}$, is generated by a Bernoulli random value generator with the probability of $x^{\dagger}=1$ being $\Pr^{\dagger}(X=1|Y=y,\mathbf{C}=\mathbf{c})$ from step 1.

Then the imputed value $x^{\dagger}$ is used as an observed value, and the record $(Y=y,X=x^{\dagger},\mathbf{C}=\mathbf{c})$ is filled into the dataset as a complete case and an imputed rectangular data set is obtained. The imputation process above is repeated m times (usually $m=5$ is sufficient as suggested by Rubin), and m rectangular data sets are obtained. From here one can use standard statistical methods to analyze each of the imputed data sets. Suppose that the point and covariance matrix for the parameter vector β' from the ith imputed data set are $\hat{\beta}_i$ and $\widehat{\mathbf{W}}_i$, $i=1,2,\ldots,m$. Then the point estimate $\overline{\beta}$ and covariance matrix $\mathbf{T}$ can be obtained via Rubin's method as in equations (1.2) and (1.3).

In practice, one can use software packages that provide built-in multiple imputation procedures. We found that the procedures PROC MI and PROC MIANALYZE in SAS (SAS Institute, Cary, NC) are very convenient. The procedure MI would generate m imputed data sets, which are then supplied to standard statistical analytical procedures, such as PROC LOGISTIC, stratified by imputation groups $i=1,2,\ldots,m$. Results are obtained for each of the m stratifications. Finally the results from all stratifications are supplied to the procedure MIANALYZE for the combined result via Rubin's method.

2.2.2.6. Predictive Probability Weighting

In this section, I propose a "novel" weighting approach in the sense of its straightforward implementation in standard statistical software packages. Most statistical methods and software are designed for dealing with complete cases. Although working methods, such as the EM algorithm and weighted estimating equations, are developed as general guiding approaches that can be customized to deal with specific problems, researchers are generally not prepared to develop such applications for each individual problem. However, as many statistical software

packages are capable of carrying out weighting on subjects, it may be easier for researchers to implement via those widely available software packages by using our proposed weighting approach.

The idea of the proposed predictive probability weighting approach is from a previous work by Lyles and Lin (2010) dealing with misclassification problems, as discussed in Chapter 1. An expanded, or augmented, data set can be constructed by listing out all possible realizations of the missing values and assigning proper weights to each of them. The idea is similar to the inverse propensity weighting in the sense that both are trying to restore, via weights based on the probability of missingness, the original data set that would have been observed and then feed it to an appropriate statistical program that works on rectangular data sets. It is also similar to the multiple imputation method, whilst MI intends to impute a simulated value to replace the missing value, but the method proposed here intends to enumerate all possible values of the missing value and assign each of them a proper weight. When the assumed missingness models line up, the weight assigned to each possible value could be viewed as the mean of the conditional probability that is used by MI to generate the imputed values.

We have seen that the IPW approach would put weights on the subjects without missing data with appropriate weights. In predictive probability weighting, instead, each subject with missing X would be replaced by two artificial subjects, representing the two possible values of $X=1$ and $X=0$. A similar idea of reconstructing the data set and applying appropriate weights via the EM algorithm has been proposed for handling missing values in generalized linear models with missing covariates from discrete distributions with a finite range by Ibrahim (1990), Lipsitz and Ibrahim (1996) and Horton and Laird (1999), and extended to cooperate with continuous covariates by Ibrahim, Chen and Lipsitz (1999). A similar weighting approach has also been proposed via estimating equations (weighted estimating equations, or WEE) (Lipsitz, Ibrahim and Zhao 1999). The EM algorithm approach and the WEE method require both a model for the covariates and a model for the missing data mechanism, and thus a specific algorithm (e.g., EM algorithm) is required to achieve simultaneous model fitting. In contrast, we propose to weight

each likelihood contribution appropriately in order to take advantage of standard statistical software.

To be more specific, in the 2×2 case, the two artificial subjects are weighted by their own observable probability

$$\begin{cases} \Pr(X=1 \mid Y=y, m=1) \\ \Pr(X=0 \mid Y=y, m=1) \end{cases} \tag{2.27}$$

Under the assumption of MAR, we know that

$$\Pr(X=x \mid Y=y, m=1) = \Pr(X=x \mid Y=y, m=0), \tag{2.28}$$

so that the probability can be estimated based on the complete cases of the dataset without inducing any bias. Then each of the two "observations" for subjects with missing X is weighted by the estimated probability

$$\begin{cases} \hat{w}_y = \widehat{Pr}(X=1 \mid Y=y, m=0) = \dfrac{n_{0,y,1}}{n_{0,y,1}+n_{0,y,0}} \\ 1-\hat{w}_y = \widehat{Pr}(X=0 \mid Y=y, m=0) = \dfrac{n_{0,y,0}}{n_{0,y,1}+n_{0,y,0}} \end{cases} \tag{2.29}$$

The reconstructed data set has the following structure:

Table 2.3 Reconstructed data set by PPW

$n_{m,y,x}$	Smoker $(X=1, m=0)$	Nonsmoker $(X=0, m=0)$	Smoking status missing $(m=1)$	
			Reconstructed with $X=1$	Reconstructed with $X=0$
Hypertensive (Y=1)	n_{011}	n_{010}	$n_{11\cdot}\times\dfrac{n_{011}}{n_{011}+n_{010}}$	$n_{11\cdot}\times\dfrac{n_{010}}{n_{011}+n_{010}}$
Normal (Y=0)	n_{001}	n_{000}	$n_{10\cdot}\times\dfrac{n_{001}}{n_{001}+n_{000}}$	$n_{10\cdot}\times\dfrac{n_{000}}{n_{001}+n_{000}}$

Note: Cases with missing values are stratified and weighted by the corresponding predictive probabilities to reflect the underlying sample that would have been observed if there were no missing values.

After combining the augmented cells and the originally observed cells that have common X values, the above table becomes identical to Table 2.2. For example, if we combine the 2 cells

with $(Y=1, X=1)$, the number of cell counts would be

$$n_{011} + n_{11} \times \frac{n_{011}}{n_{011} + n_{010}} = n_{011} \times \frac{n_{011} + n_{010} + n_{11}}{n_{011} + n_{010}} \tag{2.30}$$

Thus, by comparing the cell counts of Table 2.2 and Table 2.3, it is clear that in the case of a 2×2 table, the inverse propensity weighting and the predictive probability weighting approaches are identical. More generally, we can easily show that the weighted log-likelihood based on the predictive probability weighting method is identical to that of the inverse propensity weighting if there are categorical covariates $\mathbf{C}$ by constructing 2×2 tables for each set of values of $\mathbf{C}$. To be more specific, if the covariate $\mathbf{C}$ contains only categorical components, we can stratify the observed data set by $\mathbf{C}$ and obtain a 2×2 table such as that in Table 2.1 for each of the strata. Then IPW can be used to reconstruct a new data set as in Table 2.2 and PPW can be used to reconstruct a new data set as in Table 2.3. Then each pair of the new data sets reconstructed by IPW and PPW are identical for each of the strata; therefore, the overall odds ratio estimates, as a weighted average of the stratified odds ratio estimates, would be identical between the IPW and the PPW methods.

Now consider the case where the covariate $\mathbf{C}$ contains one or more continuous components. If we have a subject $(y, m=1, \mathbf{c})$, for which the value of X is missing, and X is a binary variable, then the above observation can only be made as a consequence of two possible instances, an instance of $(y, x=1, \mathbf{c})$ or an instance of $(y, x=0, \mathbf{c})$, with probabilities $w_{1,y\mathbf{c}} = \Pr(x=1 \mid y, \mathbf{c})$ and $w_{0,y\mathbf{c}} = 1 - w_{1,y\mathbf{c}}$ respectively, and we have $w_{0,y\mathbf{c}} + w_{1,y\mathbf{c}} = 1$. The probabilities of each possible instance can be used as weights, which sum to one for all possible instances. To be more specific, each complete record $(y, x, \mathbf{c})$ is kept as is, and assigned weight 1 $(y, x, \mathbf{c}, wt=1)$. Each incomplete record with the value of X missing $(y, m=1, \mathbf{c})$ is replaced by two records $(y, x=1, \mathbf{c}, wt=w_{1,y\mathbf{c}})$ and $(y, x=0, \mathbf{c}, wt=w_{0,y\mathbf{c}})$ in the expanded data set. Practically, we are maximizing a weighted log-likelihood

$$l(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \{ I_i l_{yx\mathbf{c}}(\alpha, \beta, \gamma) + (1 - I_i)[l_{y, X=1, \mathbf{c}}(\alpha, \beta, \gamma) w_{1,y\mathbf{c}} + l_{y, X=0, \mathbf{c}}(\alpha, \beta, \gamma)(1 - w_{1,y\mathbf{c}})] \} \tag{2.31}$$

where $I_i = 1$ if the case is completely observed; $l_{yx\mathbf{c}}$ is the log-likelihood contribution of a complete record $(y, x, \mathbf{c})$; and $l_{y,X=x,\mathbf{c}}$ is the log-likelihood contribution of a reconstructed record $(y, x, \mathbf{c})$, $x = 0,1$, and $w_{1,y\mathbf{c}}$ and $w_{0,y\mathbf{c}}$ are weights as defined above. The (i) subscripts on the $l_{yx\mathbf{c}}$, $w_{1,y\mathbf{c}}$ and $w_{0,y\mathbf{c}}$ terms are suppressed for simplicity.

With this idea in mind, one can use a variety of methods to estimate the proper weights to be used. Two methods are proposed here. The first one is to assume a "flipped-around" logistic regression model with X as the outcome, and all the other variables $(Y, \mathbf{C})$ as predictors. The idea is similar to multiple imputation (2.25), and thus it shares similar properties as MI. Another method is through an iterative procedure, which is similar to an EM algorithm, but differs in implementation due to its nature as a pseudo-likelihood type of approach. The first method is more intuitive and offers the advantage of simple accessibility, therefore, is suggested in general cases. However, as the simulation study shows, in some extreme cases, it might suffer from confliction between the flipped-around model and the original model; therefore, the second method is recommended in these cases.

The proposed methods provide advantages over the IPW when $\mathbf{C}$ has continuous components. By applying weights only to complete cases and removing all incomplete cases, IPW might lose information regarding $\mathbf{C}$, whereas the proposed predictive probability weighting can preserve this part of the information and achieve better results. A simulation study will demonstrate this advantage of the proposed approach. The proposed methods provide advantages over the MI in the sense that MI is a simulation based method, which produces slight different result each time. The proposed iterative method also provides an alternative to MI when the validity of the imputation model (2.25) is questionable.

Due to the fact that the weights in the proposed method are constructed with predicted probabilities, the variability in the predictions should be considered in a proper estimation of standard error. Resampling based methods like bootstrap (Efron and Tibshirani 1993) or jackknife (Hinkley 1983) are recommended to properly account for such variability, and we recommend

jackknife over bootstrapping due to less numerical problems. For each leave-one-out sample from the original observed data, the point estimate is obtained by either of the following two methods. The final jackknife point estimate and standard error are calculated as discussed in Section 1.2.8.

2.2.2.6.1. "Flipped-Around" Regression Modeling

The first approach to predict the appropriate weight is simple and very intuitive. Assume a logistic regression model as in (2.25), and restated here

$$\text{logit}[\Pr(X=1 \mid Y=y, \mathbf{C}=\mathbf{c})] = \tau_0 + \tau_1 y + \tau_2' \mathbf{c},$$

with the incomplete predictor variable X as the dependent variable and the outcome Y and the rest of the covariates as independent variables. One can fit this logistic regression model using only the complete cases. That is, for the logistic model above where the outcome X has MAR missing values, the CC analysis is equivalent to the ML approach, as has been discussed in section 2.1. Thus one can achieve ML estimates $\hat{\tau} = \begin{pmatrix} \hat{\tau}_0 \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{pmatrix}$ together with covariance matrix Σ_τ by fitting the above model (2.25) to the complete cases. Nevertheless, under the MAR assumption, it is known that

$$\Pr(X=1 \mid Y=y, \mathbf{C}=\mathbf{c}, m=1) = \Pr(X=1 \mid Y=y, \mathbf{C}=\mathbf{c}, m=0) = \Pr(X=1 \mid Y=y, \mathbf{C}=\mathbf{c}) \quad (2.32)$$

Thus the parameter estimates from the above model can be directly used to predict the probability $\widehat{\Pr}(X=1 \mid Y=y, \mathbf{C}=\mathbf{c}, m=1)$. Here, I redefine this logistic regression model as

$$\text{logit}[\Pr(X=1 \mid y, \mathbf{c})] = \alpha^* + \beta^* y + \gamma^{*\prime} \mathbf{c} \quad (2.33)$$

where $(\alpha^*, \beta^*, \gamma^*)$ are distinct from the original model parameters (α, β, γ). Possible interaction and polynomial terms of Y and $\mathbf{C}$ can be included if necessary to approximate a saturated model (Breslow and Powers, 1978).

Then with each incomplete case $(y, m=1, \mathbf{c})$, one can replace it by the two possible records $(y, X=1, \mathbf{c})$ and $(y, X=0, \mathbf{c})$, and assign weights w_1 and w_0 respectively, where

$w_1 = \widehat{Pr}(X = 1 | y, \mathbf{c})$ and $w_0 = \widehat{Pr}(X = 0 | y, \mathbf{c})$ can be easily estimated by the above model.

This approach makes a similar assumption of a "flipped-around" logistic regression model of X on Y and $\mathbf{C}$ as MI does, suggesting that the two would share similar properties. The flipped-around logistic regression model is intuitive and simple. However in extreme situations, this assumption may conflict with the logistic regression model of Y on X and $\mathbf{C}$ in (2.1).

Intuitively, the odds ratio relating Y to X can be estimated in both ways, i.e.,

$$\log(\widehat{OR}) = \hat{\beta} = \hat{\beta}^* . \tag{2.34}$$

This will be true or nearly so under moderate circumstances, but could fall apart in extreme situations when there are continuous components in $\mathbf{C}$, and the regression parameters and the variances of $\mathbf{C}$ are large in magnitude. A logistic model of a binary outcome Y on predictors X and $\mathbf{C}$, where X is a binary predictor, does not directly imply a valid flipped-around logistic model of X on Y and $\mathbf{C}$, unless polynomial terms are added into model (2.33) properly to approximate a saturated model. With continuous covariate $\mathbf{C}$, the flipped-around logistic regression model fitting is not trivial. As it was discussed by Breslow and Powers (1978) and Prentice and Pyke (1979), the equivalence of the association relationship between the outcome and the exposure estimated by the retrospective and prospective models can only be approximated by further covariate adjustment. The polynomial terms of the continuous covariates need to be added into the linear term to approximate a saturated model so that the equivalence could be reached. However, when the covariates are "sufficiently continuous", it is neither feasible nor desirable to reach such saturated model. Therefore, the incompatibility of the "flipped-around" model and the original model with sufficiently continuous covariates induces the drawback of those methods based upon these two models. As a matter of fact, empirical studies show that if $\mathbf{C}$ has relatively large variance and/or a large regression coefficient γ, then the estimates $\hat{\beta}$ and $\hat{\beta}^*$ can be quite different (Masalovich, 2010). In such a case, the above equation (2.34) does not hold. As a result, the predicted probability estimated using model (2.33) deviates away from the true probability, which results in inaccurate predicted weights, and in turn

results in biased estimates in $(\hat{\beta},\hat{\gamma})$. Simulation studies were conducted to demonstrate this finding.

Furthermore, we also notice the tactic of relying on model (2.33) in this way would only allow estimation of β via β^* in an ideal case, while the estimate of α and γ is not available. This makes a multivariate analysis on (α,β,γ) not directly available if that is of interest.

In passing we notice that one may consider the alternative to estimating the odds ratio relating Y to X via the model as in (2.33). In this way one can translate the exposure MAR problem into an outcome MAR problem, and make use of the result found in Section 2.2.1. However, the tactic of relying on this "flipped-around" model would only allow one to estimate the odds ratio relating Y to X, but not the other covariates.

As pointed out above, the MI method makes similar assumptions as in the flipped-around modeling for PPW based on model (2.25). Thus, the MI could result in biased estimates as well, as we will show in the simulation study. Therefore, the following method is proposed as an alternative over the PPW and MI approaches in certain circumstances when they might fall apart due to inadequacy in model (2.33).

2.2.2.6.2. Iterative Predictive Probability Weighting Method

Following the same idea of assigning a proper weight to each possible instance that would have been observed if there was no missingness, there is another way to predict the values of the weights. In the above method, we assume a flipped-around logistic regression model with X as the dependent variable to make the prediction. Alternatively, one can use Bayes' rule in conjunction with the assumed model (2.1) and rewrite the conditional probability of X given the rest of the variables in the following way.

$$\Pr(X=1\mid Y=y,\mathbf{C}=\mathbf{c})=\frac{\Pr(Y=y\mid X=1,\mathbf{C}=\mathbf{c})\Pr(X=1\mid\mathbf{C}=\mathbf{c})}{\sum_{x=0}^{1}\Pr(Y=y\mid X=x,\mathbf{C}=\mathbf{c})\Pr(X=x\mid\mathbf{C}=\mathbf{c})} \tag{2.35}$$

Ibrahim (1990) proposed a similar weighting approach using the predictive probability when implementing the E-step of an EM algorithm. The difference is that, assuming a marginal distribution of the vector of predictors, denoted as $(X, \mathbf{C})$ here, Ibrahim dealt with the joint distribution of $(Y, X, \mathbf{C})$ directly, resulting in the necessity to correctly model the whole joint distribution. This approach was applied to an example (Ibrahim 1990) with only dichotomized covariates so that a saturated model could be applied in modeling the joint distribution. Note, however, that if there are continuous variables present, fitting a saturated model will become complicated. In addition to this difficulty, the numerical maximization needed in the application of Ibrahim's method presents an obstacle to general application. Lipsitz, Ibrahim, Chen and Peterson (1999) proposed a similar idea with the EM algorithm through weights for generalized linear models with NMAR missing covariates. This method can be easily transformed to deal with a MAR covariate, as in the setting of this chapter. However, additional computational complexity in the EM algorithm is caused by simultaneous maximization of the likelihood functions of more than one model, although only one model is of interest.

Because we are only interested with the model $(Y \mid X, \mathbf{C})$ rather than the model $(X \mid \mathbf{C})$, a pseudo-likelihood (Gong and Samaniego, 1981) type of approach is proposed here, where the nuisance parameters involved in modeling $(X \mid \mathbf{C})$ will be pre-estimated and the MLE's are plugged into the likelihood directly. Let us now assume a sub-logistic regression model of X on the rest of the covariates $\mathbf{C}$:

$$\text{logit}[\Pr(X = 1 \mid \mathbf{C} = \mathbf{c})] = \theta_0 + \theta_1' \mathbf{c} \tag{2.36}$$

If the missingness of X is MAR and depends on Y and $\mathbf{C}$ but not their interaction, then the estimate $(\hat{\theta}_0, \hat{\theta}_1')'$ by CC is unbiased. Therefore the probability $\Pr(X = 1 \mid \mathbf{C} = \mathbf{c})$ can be predicted by fitting model (2.36) directly to the complete cases. With the resulting predictive probability, together with the main model of interest (2.1), one can show with Bayes' rule that

$$\begin{aligned}\Pr(X=1\mid Y=y,\mathbf{C}=\mathbf{c}) &= \frac{\Pr(Y=y\mid X=1,\mathbf{c})\times\Pr(X=1\mid\mathbf{c})}{\Pr(Y=y\mid X=1,\mathbf{c})\Pr(X=1\mid\mathbf{c})+\Pr(Y=y\mid X=0,\mathbf{c})(1-\Pr(X=1\mid\mathbf{c}))} \\ &= \frac{1}{1+\frac{1-\Pr(X=1\mid\mathbf{c})}{\Pr(X=1\mid\mathbf{c})}\times\frac{e^{-\beta y}(1+e^{\alpha+\beta+\gamma'\mathbf{c}})}{1+e^{\alpha+\gamma'\mathbf{c}}}}\end{aligned} \tag{2.37}$$

This equation (Masalovich 2010) can be used to achieve prediction of the probability $\Pr(X=1\mid Y=y,\mathbf{C}=\mathbf{c})$ as long as $\Pr(X=1\mid\mathbf{C})$ is correctly predicted with model (2.36) as discussed above. Furthermore, the prediction of the probability $\Pr(X=1\mid Y=y,\mathbf{C}=\mathbf{c})$ can be made in more general cases. For example, as discussed by Lipsitz, Ibrahim and Zhao (1999) under the WEE setting, one can adopt a non-parametric model for $\Pr(X=1\mid\mathbf{C})$ as long as it can model the conditional probability correctly. Then the predicted conditional probability $\widehat{\Pr}(X=1\mid\mathbf{C})$ can be plugged in directly to (2.37) and the rest of the procedures follow the same approach previously proposed for the PPW method.

To estimate the above probability, we need estimates of the parameters $(\hat{\alpha},\hat{\beta},\hat{\gamma})$ and $(\hat{\theta}_0,\hat{\theta}_1')$. In the EM algorithm proposed by Ibrahim (1990), a numerical maximization procedure is needed to get simultaneous estimates for the above two sets of parameters. Thus, no standard statistical software package is available for ready implementation. Users need to write their own program for each specific problem and worry about iteration convergence, etc. As proposed here, the latter set of parameters can be estimated through fitting the sub- logistic regression model (2.36) on the complete cases, leaving only the parameters of interest for the main model (2.1) to be estimated. Thus, standard statistical software packages with a weighting option can be utilized to carry out the maximization step. The whole process still requires iteration, but compared to the EM approach, it eliminates the numerical algorithms needed both in the E-step and the M-step. Thus, it becomes much more user-friendly in real application. The difference between the proposed PPW method via iteration and the Ibrahim's ML method via EM algorithm is summarized in Table 2.4.

One can begin with the complete-case estimates of the parameters of interest by fitting

model (2.1) to the complete cases, obtaining, say, $(\hat{\alpha}^0, \hat{\beta}^0, \hat{\gamma}^0)$ as a reasonable starting point. Then we can calculate initial weights as follows:

$$w_1^0 = \widehat{\Pr}^0(X=1 \mid y, \mathbf{c}) = \frac{1}{1+\frac{1-\widehat{\Pr}(X=1\mid \mathbf{c})}{\widehat{\Pr}(X=1\mid \mathbf{c})}\frac{e^{-\hat{\beta}^0 y}(1+e^{\hat{\alpha}^0+\hat{\beta}^0+\hat{\gamma}^{0\prime}\mathbf{c}})}{1+e^{\hat{\alpha}^0+\hat{\gamma}^{0\prime}\mathbf{c}}}}$$
$$w_0^0 = 1 - w_1^0 \tag{2.38}$$

Using these weights, one can fit a weighted logistic regression model with the reconstructed dataset, and maximize the weighted log-likelihood in (2.31) via standard software to get a new set of parameter estimates $(\hat{\alpha}^1, \hat{\beta}^1, \hat{\gamma}^1)$. Then, one can update the predicted weights via equations

$$w_1^{(t)} = \widehat{\Pr}^{(t)}(X=1 \mid y, \mathbf{c}) = \frac{1}{1+\frac{1-\widehat{\Pr}(X=1\mid \mathbf{c})}{\widehat{\Pr}(X=1\mid \mathbf{c})}\frac{e^{-\hat{\beta}^{(t-1)} y}(1+e^{\hat{\alpha}^{(t-1)}+\hat{\beta}^{(t-1)}+\hat{\gamma}^{(t-1)\prime}\mathbf{c}})}{1+e^{\hat{\alpha}^{(t-1)}+\hat{\gamma}^{(t-1)\prime}\mathbf{c}}}}$$
$$w_0^{(t)} = 1 - w_1^{(t)} \tag{2.39}$$

and fit the weighted logistic regression model again to obtain an updated set of parameter estimates $(\hat{\alpha}^{(t)}, \hat{\beta}^{(t)}, \hat{\gamma}^{(t)})$. Based on a preset convergence criterion, the estimates $(\hat{\alpha}^{(t)}, \hat{\beta}^{(t)}, \hat{\gamma}^{(t)})$ stabilize to the final estimate as t increases. Formally, the estimate from this approach, namely $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$, maximizes the following weighted likelihood function,

$$l(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \sum_{i=1}^{n} \{ I_i l_{y,x\mathbf{c}}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) + (1-I_i)[l_{y,X=1,\mathbf{c}}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) w_1 + l_{y,X=0,\mathbf{c}}(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) w_0] \} \tag{2.40}$$

where $w_1 = \dfrac{1}{1+\frac{1-\widehat{\Pr}(X=1\mid \mathbf{c})}{\widehat{\Pr}(X=1\mid \mathbf{c})} \times \frac{e^{-\hat{\beta} y}(1+e^{\hat{\alpha}+\hat{\beta}+\hat{\gamma}'\mathbf{c}})}{1+e^{\hat{\alpha}+\hat{\gamma}'\mathbf{c}}}}$ and $w_0 = 1 - w_1$

To sum up, the proposed iterative process for predictive value weighting contains the following steps:

1. Fit a sub- logistic regression model of X on $\mathbf{C}$ to obtain $(\hat{\theta}_0, \hat{\theta}_1')$. Use $(\hat{\theta}_0, \hat{\theta}_1')$ to calculate the predicted probability $\widehat{\Pr}(X=1 \mid \mathbf{C}=\mathbf{c})$ for each of the incomplete cases.
2. Fit a logistic regression model as in (2.8) to the complete cases to get a set of starting

values $(\hat{\alpha}^0, \hat{\beta}^0, \hat{\gamma}^0)$, and calculate the starting value of the weights w_1^0 and w_0^0 via (2.38)

3. Maximize the weighted log-likelihood function (2.31) via standard software, and get a new set of parameter estimates $(\hat{\alpha}^{(t)}, \hat{\beta}^{(t)}, \hat{\gamma}^{(t)})$. Update the weights via (2.39)
4. Repeat step 3 until convergence criterion is met

Simulation results show that this iterative process converges very quickly, and no non-convergence cases have been observed so far under the setup described.

Table 2.4 Comparison of the iterative PPW and Ibrahim's ML approach via EM algorithm

Step	EM algorithm by Ibrahim (1990)	Proposed PPW via iteration
0		Get the MLE $(\hat{\theta}_0, \hat{\theta}_1)$ with CC
1	Supply initial values $(\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}, \theta_0^{(0)}, \theta_1^{(0)})$	Supply initial values $(\alpha^{(0)}, \beta^{(0)}, \theta^{(0)})$
2	Calculate the proper weight by (2.35) with $(\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)}, \theta_0^{(i)}, \theta_1^{(i)})$, and get the weighted log-likelihood function of the joint distribution function of $f(Y, X \mid C; \alpha, \beta, \gamma, \theta_0, \theta_1)$	Calculate the proper weight by (2.35) with $(\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)}, \hat{\theta}_0, \hat{\theta}_1)$, and get the weighted log-likelihood function of $f(Y \mid X, C; \alpha, \beta, \gamma)$
3	Use numerical algorithms to obtain the simultaneous MLE $(\alpha^{(i+1)}, \beta^{(i+1)}, \gamma^{(i+1)}, \theta_0^{(i+1)}, \theta_1^{(i+1)})$	Use standard statistical software to obtain the MLE $(\alpha^{(i+1)}, \beta^{(i+1)}, \gamma^{(i+1)})$
4	Repeat step 2 (E-step) and 3 (M-step) until convergence	Repeat step 2 (E-step) and 3 (M-step) until convergence

2.2.3. More Than One Variable Missing

Now suppose that more than one variable is subject to incomplete values. For simplicity, suppose at most two variables can be missing at the same time.

First, suppose the outcome Y and the binary exposure X can be missing at the same time. In this case, it can be readily shown that records with outcome Y missing contribute no information to the regression relationship between the outcome and the exposure. Therefore, if one is interested in estimation of the regression coefficients, one can simply omit the records with

the outcome missing, leaving only the complete records and the records with only X missing. Let m_y and m_x be missingness indicators for Y and X. This can be further illustrated by the following derivation of the likelihood contribution of such observations. Consider a record with Y missing and X observed. The likelihood contribution is as follows:

$$\begin{aligned} & \Pr(m_{y_i}=1, m_{x_i}=0, X_i=x_i \mid \mathbf{C}_i=\mathbf{c}_i) \\ = \; & \Pr(m_{y_i}=1 \mid m_{x_i}=0, X_i=x_i, \mathbf{C}_i=\mathbf{c}_i)\Pr(m_{x_i}=0 \mid X_i=x_i, \mathbf{C}_i=\mathbf{c}_i)\Pr(X_i=x_i \mid \mathbf{C}_i=\mathbf{c}_i) \\ = \; & \Pr(m_{y_i}=1 \mid X_i=x_i, \mathbf{C}_i=\mathbf{c}_i))\Pr(m_{x_i}=0 \mid \mathbf{C}_i=\mathbf{c}_i)\Pr(X_i=x_i \mid \mathbf{C}_i=\mathbf{c}_i) \end{aligned} \quad (2.41)$$

For a record with both Y and X missing, we have

$$\begin{aligned} & \Pr(m_{y_i}=1, m_{x_i}=1 \mid \mathbf{C}_i=\mathbf{c}_i) \\ = \; & \Pr(m_{y_i}=1 \mid \mathbf{C}_i=\mathbf{c}_i)\Pr(m_{x_i}=1 \mid \mathbf{C}_i=\mathbf{c}_i) \end{aligned} \quad (2.42)$$

The above equations hold because of the assumptions of MAR for both Y and X and independence between missingness indicators. They show that the contribution of these records cannot be factorized into stand-alone terms regarding the regression relationship between Y and $(X, \mathbf{C})$. Therefore, these records contribute no information regarding the regression relationship of Y on X and $\mathbf{C}$. Conclusively, if the outcome and the exposure can be missing at the same time, one can simply omit the records with the outcome missing. Then the question turns out to be the case where only the exposure X is missing at random. From here, one can adopt the proposed method (PPW), or any other methods that deal with MAR exposure problems.

Now, suppose there are two binary exposure variables, X_1 and X_2, and both are subject to missing values but that Y and $\mathbf{C}$ are fully observed. Assume that missingness indicators are independent, i.e.

$$p(m_{X_1}, m_{X_2} \mid Y, \mathbf{C}) = p(m_{X_1} \mid Y, \mathbf{C}) \times p(m_{X_2} \mid Y, \mathbf{C}).$$

Then the proposed PPW method via "flipped-around" model can be extended as follows. A record with both X_1 and X_2 observed is kept as is. A record with only X_1 missing, $(Y, \bullet, X_2, \mathbf{C})$, is replaced with two artificial records $(Y, X_1=1, X_2, \mathbf{C})$ and $(Y, X_1=0, X_2, \mathbf{C})$, with weights w_{1x_2} and w_{0x_2} respectively, where $w_{1x_2} = \Pr(X_1=1 \mid Y, X_2, \mathbf{C})$ and

$w_{0x_2} = 1 - w_{1x_2}$. A record with only X_2 missing, $(Y, X_1, \bullet, \mathbf{C})$, can be replaced similarly. A record with both X_1 and X_2 missing, $(Y, \bullet, \bullet, \mathbf{C})$, is replaced with four artificial records, representing the four possibilities of the combinations of values of X_1 and X_2, with weights w_{11}, w_{10}, w_{01} and w_{00}, where

$w_{11} = \Pr(X_1 = 1, X_2 = 1 \mid Y, \mathbf{C})$,
$w_{10} = \Pr(X_1 = 1, X_2 = 0 \mid Y, \mathbf{C})$,
$w_{01} = \Pr(X_1 = 0, X_2 = 1 \mid Y, \mathbf{C})$, and
$w_{00} = \Pr(X_1 = 0, X_2 = 0 \mid Y, \mathbf{C})$.

The weights w_{1x_2}, w_{0x_2} and $w_{x_1 1}$, $w_{x_1 0}$ could be readily predicted via logistic regression models on the corresponding "complete cases" as MAR outcome cases. For the weights w_{11}, w_{10}, w_{01} and w_{00}, assuming the independence between missingness indicators, there are two ways to factorize them.

$$w_{x_1x_2} = \Pr(X_1 = x_1, X_2 = x_2 \mid Y, \mathbf{C}) = \Pr(X_1 = x_1 \mid X_2 = x_2, Y, \mathbf{C}) \Pr(X_2 = x_2 \mid Y, \mathbf{C})$$

or $$w_{x_1x_2} = \Pr(X_1 = x_1, X_2 = x_2 \mid Y, \mathbf{C}) = \Pr(X_2 = x_2 \mid X_1 = x_1, Y, \mathbf{C}) \Pr(X_1 = x_1 \mid Y, \mathbf{C})$$

Depending on one's preference about the true underlying model, either of the two ways of factorization can be applied. Then each of the conditional probabilities can be predicted by fitting a logistic regression model on the corresponding "complete cases". The missing values in these model fittings occur in the outcomes, and are MAR. Therefore the parameter estimates from the CC analysis are consistent, as has been discussed in section 2.2.1.

2.3. Simulation Results

Simulation studies were performed to compare the performances of the discussed methods, namely the CC analysis, ML, IPW, WEE, PPW and MI. 500 simulations were conducted to evaluate the point estimates and the coverage rate of the 95% confidence intervals. In each simulation, 300 subjects were generated. First, covariate $\mathbf{C}$ was generated from a random Bernoulli distribution to compare the performance of the methods in the categorical covariate

case. Then covariate $\mathbf{C}$ was generated from a Normal distribution to compare the performance of the methods in the continuous covariate case. The exposure X was then generated from a logistic model as in (2.17) to induce correlation between X and $\mathbf{C}$. Then the outcome Y was generated from a logistic model as in (2.1). To induce missing data, a random binary indicator m was generated following the model (2.22), and the observation of X was set to be missing for a subject if $m=1$, so that the probability that X is missing for a subject depends on the values of the outcome Y and the covariate $\mathbf{C}$, but not on the value of X itself given Y and $\mathbf{C}$ (MAR).

Within the continuous covariate case, three sets of simulation studies were conducted to examine the three questions of interest discussed above.

1. The "flipped-around" logistic model might not be compatible with the original logistic model when there are continuous covariates (see discussion in Section 2.2.2.6.1). This might impact the performance of the classic multiple imputation method and the proposed predictive probability weighting method via "flipped-around" regression modeling.
2. The IPW method only retains, and applies appropriate weights to the complete cases. When there are continuous covariates, the information regarding these continuous covariates in the incomplete cases is lost, which might result in poor performance of IPW, especially when the missing rate is high (see discussion in 2.2.2.3).
3. The semi-parametric method WEE provides potential advantages in terms of robustness against model misspecification (see discussion in Section 2.2.2.4). The proposed PPW methods provide advantage in terms of simple implementation with standard software. Therefore it is of interest to assess how robust the PPW method is compared to the WEE.

2.3.1. With Categorical Covariate C

In cases where the covariate C contains only categorical components, the complete case analysis gives a valid estimate for the odds ratio of interest with respect to X. However, as

discussed in Section 2.2.2.1, the parameter estimates by CC analysis regarding covariate C are invalid. The other methods discussed in Section 2.2 provide improvements upon the result from CC analysis, especially in that they are capable to provide valid estimates to the parameters related to covariate C. In this simulation we only compare the point and standard error estimates, although the variance-covariance matrix can be achieved by standard multivariable Jackknife methods for IPW and PPW.

Assume C is a random binary variable, with probability 0.5 being one and 0.5 being zero. Suppose X is correlated with C through a logistic model with intercept zero and coefficient one. The true coefficients under model (2.1) were selected as follows: $\alpha = 1, \beta = 1.5, \text{and } \gamma = -2$. A missingness indicator m was generated from a logistic model (2.22) involving only Y and C, which resulted in a MAR case. 500 data sets were generated. Under the setup of the MAR case, it results in a 22.3% missing rate overall, with 36.4% with $Y = 0$ and 15.7% with $Y = 1$. Point estimates, standard deviations, mean estimated standard errors and 95% confidence interval coverage rates were compared in the following table.

In a case with only categorical covariates, Table 2.5 shows that CC analysis can provide an acceptable estimate regarding the effect of exposure X, which is MAR. However, the estimate of γ is quite biased due to the fact that the association between X and C produces a NMAR case regarding the covariate C. Therefore, even in a MAR case, the CC analysis is not appropriate if one is interested in a multivariable analysis. Compared to the CC analysis, all the other methods make satisfying improvement on the estimates of all the parameters. The PPW based on flipped-around logistic model produced almost identical result to that by the IPW, due to the fact that in case of all categorical variables, these two methods are essentially identical, as discussed in section 2.2.5.

Table 2.5 Comparison of the methods when the covariate is categorical

	α	β	γ
True Values	1	1.5	-2
CC	1.85 (0.41) [0.37] {36.0%}	1.55 (0.40) [0.40] {96.2%}	-2.70 (0.49) [0.47] {73.4%}
ML	1.02 (0.25) [0.25] {96.0%}	1.55 (0.40) [0.40] {96.0%}	-2.07 (0.37) [0.36] {94.6%}
PPW (Flipped Logistic Model Based)	1.00 (0.24) [0.25] {96.6%}	1.51 (0.38) [0.42] {97.4%}	-2.01 (0.35) [0.37] {97.0%}
PPW (Iteration Based)	1.00 (0.25) [0.26] {96.4%}	1.46 (0.38) [0.42] {97.0%}	-1.98 (0.35) [0.37] {96.4%}
IPW	1.00 (0.25) [0.26] {96.8%}	1.50 (0.42) [0.47] {95.8%}	-2.01 (0.37) [0.40] {96.4%}
MI	1.03 (0.25) [0.25] {96.2%}	1.55 (0.40) [0.41] {96.6%}	-2.07 (0.36) [0.36] {94.8%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates.

2.3.2. With Continuous Covariates C

Now let us consider the case when there are continuous components in covariate C. The true coefficients under model (2.1) were selected as follows: $\alpha = 1, \beta = 2, \text{and } \gamma = -1$. C is generated from a normal distribution $N(0,1)$, then X is generated from a logistic model as (2.36), with $\theta_0 = 0$ and $\theta_1 = 2$. A missingness indicator m was generated from a logistic model (2.22) involving only Y and C. Under the setup of the MAR case, it results in a 24.5%

missing rate overall, with 47.5% with $Y=0$ and 20.1% $Y=1$.

Under this moderate condition where the magnitude of variation and the regression coefficient of C were not very large, all the methods performed well except CC analysis (Table 2.6). The suggestion is that under moderate conditions, the proposed methods and the current widely used methods can all improve the inference upon the complete case analysis.

Table 2.6 Comparison of the methods when the covariate is continuous

	α	β	γ
Model Setup	1	2	-1
CC	1.90 (0.38) [0.37] {27.4%}	2.06 (0.79) [0.69] {92.7%}	-1.55 (0.40) [0.37] {71.5%}
ML	1.03 (0.26) [0.26] {95.6%}	2.06 (0.68) [0.63] {93.6%}	-1.03 (0.29) [0.28] {94.4%}
PPW (Flipped Logistic Model Based)	1.01 (0.25) [0.27] {96.9%}	1.98 (0.65) [0.66] {95.6%}	-1.00 (0.28) [0.28] {94.8%}
PPW (Iteration Based)	1.01 (0.26) [0.27] {96.8%}	1.94 (0.65) [0.66] {95.5%}	-0.98 (0.28) [0.29] {94.3%}
IPW	0.98 (0.61) [0.38] {97.2%}	2.04 (0.98) [0.86] {96.0%}	-1.01 (0.60) [0.45] {92.6}
MI	1.05 (0.27) [0.27] {96.1%}	2.01 (0.69) [0.65] {94.2%}	-1.01 (0.29) [0.28] {94.3%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates

In passing we also notice that the IPW lacks efficiency compared to the other methods, due

to restricting analysis to only those subjects with all values observed (see section 2.3.2.3 for details).

2.3.2.1. The Problem of the "Flipped-Around" Model Assumption

We are aware of the potential incompatibility of the "flipped-around" model and the original model (e.g., Breslow and Powers 1978). It is of interest to see how the flipped-around logistic model could fail in some circumstances. The true coefficients under model (2.1) were selected as follows: $\alpha = 1, \beta = 2, \text{and } \gamma = -2$, with relatively large coefficient for C. C is then generated from a normal distribution $N(0, 2^2)$, with relatively large variance. Missing data were introduced similarly as before, resulting in a 33.1% overall missing rate, with 38.7% with $Y = 0$ and 30.7% with $Y = 1$.

As discussed in section 2.2.2.5.1, the first version of the proposed method suffers from the assumption of a "flipped-around" logistic regression model of the exposure on the outcome and the other covariates (equation (2.25)). With continuous covariate C, fitting the flipped-around logistic regression model is not trivial. The polynomial terms of the continuous covariates need to be added into the linear term to approximate a saturated model. However, when the covariates are "sufficiently continuous", it is neither feasible nor desirable to reach such a saturated model. Under the setup of this simulation, the flipped-around logistic regression model without additional covariate adjustment could not provide a valid estimate to the conditional probability modeled by (2.25). The proposed PPW via a "flipped-around" model and the MI with "flipped-around" imputation model both result in biased estimates. Therefore, no matter if the probability is used to randomly generate an artificial instance for X as in MI, or to properly weight each possible instance of X as in PPW, bias is induced (Table 2.7). ML and PPW based on iteration are recommended in this case. They are based on an additional logistic regression model of the exposure on the rest of the covariates, but not on the outcome. Therefore, the paradox between the flipped-around model and the original model in logistic regression can be avoided.

Table 2.7 Comparison of the PPW and MI methods when the "flipped-around" logistic regression model is not appropriate

	α	β	γ
Model Setup	1	2	-2
CC	1.55 (0.55) [0.52] {86.6%}	2.25 (1.01) [0.85] {93.2%}	-2.71 (0.57) [0.49] {83.2%}
ML	1.00 (0.38) [0.36] {94.0%}	2.16 (0.84) [0.76] {94.0%}	-2.11 (0.35) [0.33] {96.0%}
PPW (Flipped Logistic Model Based)	1.04 (0.35) [0.37] {96.4%}	1.78 (0.67) [0.69] {93.6%}**	-1.92 (0.27) [0.29] {94.4%}
PPW (Iteration Based)	0.97 (0.36) [0.38] {95.6%}	2.04 (0.80) [0.80] {95.8%}	-2.01 (0.32) [0.35] {96.2%}
MI	1.08 (0.37) [0.37] {95.2%}	1.86 (0.71) [0.75] {95.2%}**	-1.99 (0.28) [0.41] {96.6%}

* Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates

** Though the overall coverage rate of the 95% confidence interval for parameter β looks reasonable, the confidence intervals actually lean to the left by a great deal. The un-coverage rate for the PPW (flipped-around logistic model) is 5.2% for the upper side and 1.2% for the lower side, due to the biased point estimate; and for the MI 3.8% for the upper side and 1% for the lower side.

2.3.2.2. The Problem of IPW When the Missing Rate Increases

By eliminating the whole records with missing values, the IPW method loses information regarding the variables without missing values. This can result in reduced precision, or even biased estimates. As the missing rate increases, this problem becomes more and more critical. A simulation study was conducted to demonstrate the drawback of the IPW. Here, the true coefficients under model (2.8) were selected as follows: $\alpha = 1, \beta = 2, \text{and } \gamma = -2$. C is

generated from a normal distribution $N(0,2^2)$. The missing indicator was generated from a logistic model (2.22) based on the outcome Y and covariate C. A set of values was chosen for the intercept to adjust the overall missing rates so that it increased from 10% to 50% by 10%, and one set of simulation was conducted under each level of missing rates. As summarized in the following tables (Table 2.8-Table 2.12), the estimate by IPW works fine when the missing rate is low, but becomes biased as the missing rate increases. When the overall missing rate reaches about 50%, the estimate by IPW becomes dramatically biased. Extreme estimates were obtained in some simulations because one subject in the complete case sub-data set could be assigned a weight as high as around 15, as observed in the simulation, while the complete case sub-data set contains only about 150 subjects. This subject then becomes an influential point and therefore disturbs the inference significantly. The result should no longer be trusted. Some modified versions of the IPW method have been proposed recently. One naïve approach is to simply truncate the weights when they are extremely large. Bodnar et al. (2004) and Cao et al. (2009) proposed alternative ways to construct the weights for better stability in extreme cases.

Table 2.8 Comparison of PPW and IPW as the missing rate increases: missing rate=10.9%

	α	β	γ
True Values	1	2	-2
CC	1.09 (0.37) [0.34] {93.6%}	2.12 (0.60) [0.59] {96.4%}	-2.26 (0.36) [0.32] {92.0%}
ML	1.03 (0.34) [0.32] {94.8%}	2.07 (0.58) [0.57] {95.4%}	-2.07 (0.29) [0.28] {94.6%}
PPW (Flipped Logistic Model Based)	1.05 (0.34) [0.33] {95.4%}	2.00 (0.55) [0.57] {95.8%}	-2.04 (0.28) [0.28] {95.2%}
PPW (Iteration Based)	1.03 (0.34) [0.33] {95.2%}	2.07 (0.58) [0.60] {96.6%}	-2.07 (0.29) [0.30] {95.4%}
IPW	1.03 (0.36) [0.35] {95.2%}	2.11 (0.61) [0.64] {97.2%}	-2.11 (0.35) [0.36] {95.6%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates

Table 2.9 Comparison of PPW and IPW as the missing rate increases: missing rate=20.4%

	α	β	γ
True Values	1	2	-2
CC	1.12 (0.36) [0.36] {96.2%}	2.11 (0.65) [0.64] {97.0%}	-2.32 (0.38) [0.36] {94.4%}
ML	1.00 (0.32) [0.33] {96.0%}	2.10 (0.63) [0.62] {96.4%}	-2.06 (0.28) [0.29] {96.6%}
PPW (Flipped Logistic Model Based)	1.04 (0.31) [0.34] {97.0%}	1.95 (0.57) [0.59] {96.0%}	-2.00 (0.25) [0.27] {97.4%}
PPW (Iteration Based)	1.00 (0.32) [0.34] {96.4%}	2.09 (0.63) [0.65] {96.6%}	-2.06 (0.28) [0.30] {97.2%}
IPW	1.01 (0.34) [0.37] {97.4%}	2.16 (0.70) [0.72] {97.0%}	-2.12 (0.43) [0.42] {96.8%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates

Table 2.10 Comparison of PPW and IPW as the missing rate increases: missing rate=30.3%

	α	β	γ
True Values	1	2	-2
CC	1.25 (0.42) [0.41] {93.8%}	2.11 (0.74) [0.74] {95.4%}	-2.45 (0.49) [0.44] {91.6%}
ML	1.01 (0.35) [0.35] {96.0%}	2.12 (0.75) [0.70] {94.4%}	-2.08 (0.33) [0.31] {93.6%}
PPW (Flipped Logistic Model Based)	1.07 (0.35) [0.35] {97.8%}	1.88 (0.65) [0.65] {95.0%}	-1.98 (0.28) [0.28] {93.6%}
PPW (Iteration Based)	1.01 (0.35) [0.36] {96.6%}	2.10 (0.74) [0.75] {95.2%}	-2.07 (0.33) [0.33] {95.6%}
IPW	1.04 (0.40) [0.42] {96.4%}	2.23 (0.88) [0.89] {98.2%}	-2.22 (0.52) [0.54] {98.0%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates

Table 2.11 Comparison of PPW and IPW as the missing rate increases: missing rate=39.6%

	α	β	γ
True Values	1	2	-2
CC	1.40 (0.56) [0.50] {90.8%}	2.20 (1.00) [0.93] {97.0%}	-2.61 (0.66) [0.56] {95.2%}
ML	0.99 (0.39) [0.37] {94.6%}	2.18 (0.88) [0.83] {95.0%}	-2.11 (0.38) [0.34] {95.4%}
PPW (Flipped Logistic Model Based)	1.09 (0.38) [0.39] {96.8%}	1.82 (0.75) [0.76] {95.8%}	-1.96 (0.29) [0.29] {95.4%}
PPW (Iteration Based)	0.99 (0.39) [0.40] {95.2%}	2.14 (0.87) [0.89] {96.0%}	-2.10 (0.37) [0.38] {97.2%}
IPW	1.06 (0.52) [0.54] {96.4%}	2.38 (1.13) [1.26] {98.6%}	-2.38 (0.68) [0.76] {97.8%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates

Table 2.12 Comparison of PPW and IPW as the missing rate increases: missing rate=49.9%

	α	β	γ
True Values	1	2	-2
CC	1.75 (1.36) [1.29] {94.4%}	2.46 (1.99) [1.98] {95.2%}	-3.10 (1.82) [0.98] {97.6%}
ML	1.03 (0.46) [0.43] {95.8%}	2.19 (1.12) [1.04] {92.2%}	-2.14 (0.46) [0.40] {93.0%}
PPW (Flipped Logistic Model Based)	1.15 (0.41) [0.45] {98.1%}	1.75 (0.87) [0.98] {95.7%}	-1.94 (0.31) [0.32] {92.8%}
PPW (Iteration Based)	1.04 (0.46) [0.51] {97.2%}	2.11 (1.09) [1.17] {95.8%}	-2.11 (0.44) [0.44] {95.4%}
IPW**	3.91 (30.72) [23.81] {98.0%}	5.75 (32.35) [41.16] {99.0%}	-7.10 (46.29) [47.81] {98.8%}

* Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates

** The dramatic large standard deviation was caused by some extreme values. The minimum values of were recorded as (-1.81, -16.26, -695.24) and maximum values of them are recorded as (497.95, 379.89, -0.86)

The following Table 2.13 gives an explicit side by side comparison across the different levels of missing rates by putting the point estimates and standard deviation (in brackets [ ]) of all five batches together.

Table 2.13 Summary of the PPW and IPW results as the missing rate increases

	Overall Missing Rate	α	β	γ
True Values		1	2	-2
CC	10.9%	1.09 [0.34]	2.12 [0.59]	-2.26 [0.32]
	20.4%	1.12 [0.36]	2.11 [0.64]	-2.32 [0.36]
	30.3%	1.25 [0.41]	2.11 [0.74]	-2.45 [0.44]
	39.6%	1.40 [0.50]	2.20 [0.93]	-2.61 [0.56]
	49.9%	1.75 [1.29]	2.46 [1.98]	-3.10 [0.98]
ML	10.9%	1.03 [0.32]	2.07 [0.57]	-2.07 [0.28]
	20.4%	1.00 [0.33]	2.10 [0.62]	-2.06 [0.29]
	30.3%	1.01 [0.35]	2.12 [0.70]	-2.08 [0.31]
	39.6%	0.99 [0.37]	2.18 [0.83]	-2.11 [0.34]
	49.9%	1.03 [0.43]	2.19 [1.04]	-2.14 [0.40]
PPW (Flipped Logistic Model Based)	10.9%	1.05 [0.33]	2.00 [0.57]	-2.04 [0.28]
	20.4%	1.04 [0.34]	1.95 [0.59]	-2.00 [0.27]
	30.3%	1.07 [0.35]	1.88 [0.65]	-1.98 [0.28]
	39.6%	1.09 [0.39]	1.82 [0.76]	-1.96 [0.29]
	49.9%	1.15 [0.45]	1.75 [0.98]	-1.94 [0.32]
PPW (Iteration Based)	10.9%	1.03 [0.33]	2.07 [0.60]	-2.07 [0.30]
	20.4%	1.00 [0.34]	2.09 [0.65]	-2.06 [0.30]
	30.3%	1.01 [0.36]	2.10 [0.75]	-2.07 [0.33]
	39.6%	0.99 [0.40]	2.14 [0.89]	-2.10 [0.38]
	49.9%	1.04 [0.51]	2.11 [1.17]	-2.11 [0.44]
IPW	10.9%	1.03 [0.35]	2.11 [0.64]	-2.11 [0.36]
	20.4%	1.01 [0.37]	2.16 [0.72]	-2.12 [0.42]
	30.3%	1.04 [0.42]	2.23 [0.89]	-2.22 [0.54]
	39.6%	1.06 [0.54]	2.38 [1.26]	-2.38 [0.76]
	49.9%	3.91 [23.81]	5.75 [41.16]	-7.10 [47.81]

Numbers in each cell reflect mean [mean estimated standard errors] based on 1,000 simulations with sample size of 500.
As missing rate gets larger, IPW becomes more instable

Methods have been proposed to deal with large weights such as weight truncation, where a maximum weight is specified. Sensitivity analysis on the choice of maximum weight should be performed. Instead of using logistic regression in the missingness model, Kang and Schafer (2007) suggest using robit regression (Liu 2004), which is more robust to outliers, therefore less likely to produce very large weights. Cao et al. (2009) propose an enhanced logistic regression model that contains ordinary logistic regression as a special case.

2.3.2.3. Double Robustness of Weighted Estimating Equations

As a semi-parametric method, WEE has its advantage of robustness against misspecification of the models: the model for the missingness indicator, $p(m \mid \text{Observed data})$ as in (2.22) , and the model of the exposure rate, $p(x \mid \text{Observed data})$ as in (2.17). In WEE, only one of the two models needs to be correctly specified. In contrast, the IPW method would rely on the first model assumption and the ML, MI and iterative weighting methods would rely on the second model assumption.

Lipsitz et al. (1999) conducted simulation studies to assess the asymptotic bias in estimating β using the parametric likelihood based method and the semi-parametric WEE method. In their simulation study, they assumed that the above two models, $p(m \mid \text{Observed data})$ and $p(x \mid \text{Observed data})$, took the form of a linear logistic model in the data generation model. In the data analysis, they omitted certain regression terms used in the analysis model to deviate away from the underlying true model. In such a way they assessed how the misspecification made an impact on estimating β. They found that the relative bias tends to be smaller using WEE than that using likelihood based method with the same misspecification of the corresponding model. However, their conclusion is not definitive, due to the broad range of configurations of model misspecification.

Here, simulation studies were conducted to assess the robustness of these methods from a slightly different angle. As an extension to earlier studies, we induced misspecification by letting the functional form of the analysis model deviate away from the data generation model, therefore

the functional form of the model assumption is no longer correctly specified. We examined the simulation results in search of any degradation of the methods under wrong model assumption.

Firstly, the model of generating the missingness indicator was set to be non-logistic. To be more specific, the probability of missingness, $m=1$, took the form of a convex step function on the linear term of the observed data , $\psi_0+\psi_1 y+\psi_2 c$, instead of the logit function as in (2.22). In the data analysis model, a logistic model was assumed blindly.

$$\text{logit}(\Pr(m=1\mid y,c))=\psi_0+\psi_1 y+\psi_2 c+\psi_3 c^2+\psi_{12} yc$$

Figure 2.1 shows how the true step function is compared to the fitted logistic regression. The X-axis represents the value of the linear term $\psi_0+\psi_1 y+\psi_2 c$, and the Y-axis represents the true/predictive probability of $m=1$. The result in Table 2.14 shows that this misspecification of the missingness model did not result in significant failure of the IPW method. This infers that, although the IPW method requires the assumption of the missingness model (the linear logistic regression model), there may still be relatively robustness against misspecification of the missingness model under regular conditions.

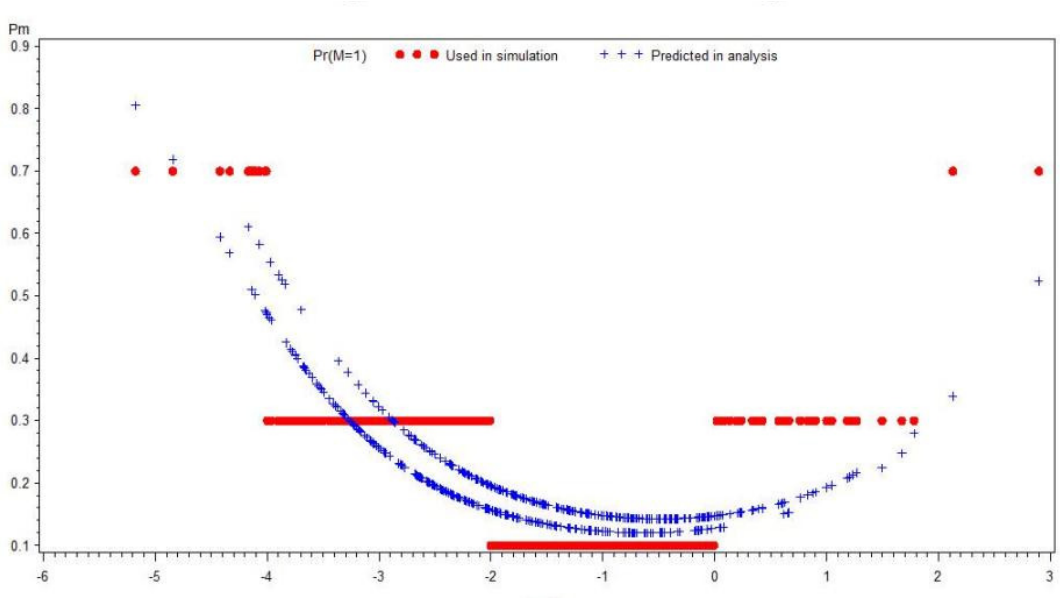

Figure 2.1 Discrepancy of the missingness model in simulation and in analysis. Noticeable discrepancy is found between the true probabilities used in simulation (round dots) and the predicted propensity values used in analysis (crosses)

Table 2.14 Robustness of IPW when the missingness model $(m \mid Y, C)$ is mis-specified

	α	β	γ
True Values	1	2	-1
IPW	-1.05 (0.25) [0.27] {96.5%}	2.07 (0.42) [0.45] {98.0%}	-1.13 (0.12) [0.13] {90.5%}
WEE	-1.02 (0.23) [0.24] {96.0%}	2.02 (0.40) [0.41] {97.5%}	-1.01 (0.10) [0.11] {98.5%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates. Missingness model was mis-specified in analysis, resulting noticeable bias in parameter estimate of C

Secondly, the model of the exposure given the observed data was set to be non-logistic. To be more specific, the probability $X = 1$ took the form of a convex step function on the linear term of the observed data , $\theta_0 + \theta_1 c$, instead of the logit function as in (2.17). However, in the analysis model, we still used the linear term logistic model blindly when fitting the regression model of X on C. That is to say, the conditional probability of X given C is predicted by the model

$$\text{logit}[\Pr(X = 1 \mid C = c)] = \theta_0 + \theta_1 c + \theta_2 c^2$$

This resulted in misspecification of the sub-logistic model, and the parameter estimates of the sub-logistic model were invalid. Figure 2.2 shows how the true quadratic function term is compared to the fitted linear function term in the logistic regression model. The X-axis represents the value of C, and the Y-axis represents the predicted probability of $X = 1$. However, as we can see in Table 2.15, the parameter estimates of the main model still appear reasonable.

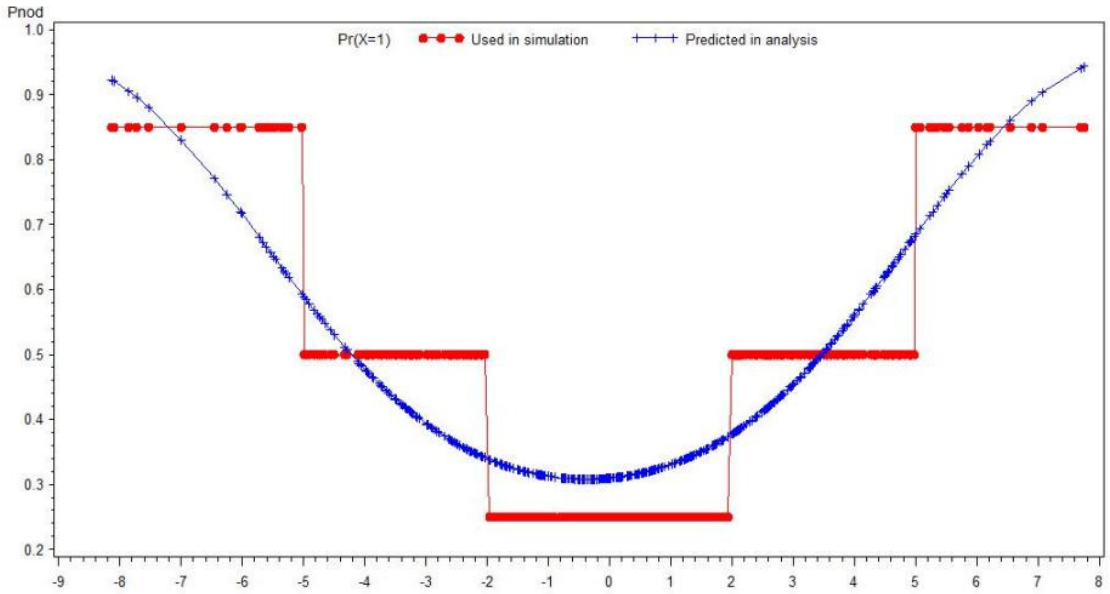

Figure 2.2 Discrepancy of the exposure model in simulation and in analysis. Noticeable discrepancy is found between the true probabilities used in simulation (round dots) and the predicted probabilities used in PPW (crosses)

Table 2.15 Robustness of PPW when the exposure model $(X \mid C)$ is mis-specified

	α	β	γ
True Values	1	2	-1
ML	-1.00 (0.29) [0.28] {93.2%}	2.00 (0.37) [0.36] {95.8%}	-0.96 (0.16) [0.17] {95.4%}
PPW (Iteration Based)	-0.98 (0.29) [0.28] {94.4%}	1.95 (0.36) [0.37] {95.0%}	-0.94 (0.15) [0.17] {94.6%}
WEE	-1.01 (0.29) [0.29] {94.2%}	2.01 (0.37) [0.38] {96.8%}	-0.99 (0.16) [0.18] {96.8%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates. Missingness model was mis-specified in analysis, resulting noticeable bias in parameter estimate of C

Lastly, both of the models for the missingness indicator, $(m \mid Y, C)$, and for the exposure given observed data, $(X \mid C)$, were set up in breach of the logistic model. As we can see in Table 2.16, the results for all methods still appear reasonable. Therefore, some misspecification of the model assumption was quite tolerable in this simulation experiment, and all methods showed good robustness against model misspecification.

Table 2.16 Robustness of PPW and IPW when both models $(X \mid C)$ and $(m \mid Y, C)$ are mis-specified

	α	β	γ
Model Setup	1	2	-1
ML	-0.97 (0.27) [0.28] {96.2%}	1.99 (0.34) [0.35] {95.6%}	-1.04 (0.17) [0.17] {95.4%}
PPW (Iteration Based)	-0.92 (0.25) [0.27] {95.2%}	1.92 (0.32) [0.35] {94.8%}	-1.03 (0.17) [0.17] {96.8%}
WEE	-0.96 (0.27) [0.29] {96.8%}	1.95 (0.33) [0.36] {96.0%}	-1.02 (0.17) [0.17] {97.0%}
IPW	-0.98 (0.28) [0.30] {97.4%}	1.99 (0.35) [0.38] {96.4%}	-1.07 (0.18) [0.18] {97.4%}

Numbers in each cell reflect mean (standard deviation) based on 1,000 simulations with sample size of 500. Values in brackets [ ] are mean estimated standard errors; values in braces { } are 95% confidence interval coverage rates.

2.4. Discussions

2.4.1. Comparison of the Methods

Although the complete case analysis works well in certain situations, this strategy can

obviously cause loss of information. As the fraction of missing data increases, the deletion of all subjects with missing data can be unnecessarily wasteful and quite inefficient. In addition, the complete case method violates intention to treat principles currently widespread in biometric research (Nich and Carroll 2002; Liu and Gould 2002; Hollis 2002). Finally, by excluding entire incomplete records, it might result in biased estimates of the regression coefficients for the covariates $\mathbf{C}$ that are not subject to missing values. When the regression coefficients corresponding to the covariates $\mathbf{C}$ are of interest, complete-case analysis is not appropriate even if the less restrictive MAR assumption holds.

ML is a safe approach if one can correctly specify the likelihood function explicitly. However, if the observed data likelihood in (2.10) does not have a closed form and cannot be factorized, approaches such as the EM algorithm are generally needed to obtain MLEs from (2.10) (Ibrahim, Chen, Lipsitz et al. 2005). A general method for estimation in the presence of missing covariates has been proposed by Ibrahim (1990), who used EM via a method of weights to find the MLEs. Although commonly cited as a different estimation method, the EM algorithm can be considered as an alternative approach that applies maximum likelihood estimation to a general model with additional assumptions on the marginal distribution of the variables subject to missingness. As was pointed out by Little (1988), ideally the analyst should formulate a statistical model for the survey variables under study and the missing-data mechanism and then estimate parameters from the incomplete data by methods such as ML without attempting to fill in the missing values. (Little 1982, Little and Rubin 1987). However, in practice, this may place excessive demands on the analyst interested in the subject matter rather than in specialized statistical methodology. Methods such as the EM algorithm (Dempster, Laird and Rubin 1977) were developed for fitting simultaneous equation models, but few investigators (e.g. epidemiologists) are likely to devote the energy to develop customized algorithms for their missing-data problems.

Multiple imputation is an appealing approach and standard statistical software packages are available. However, there is not a universal routine to regulate how the imputation model is

defined. As it was shown in the simulations here, if one uses the flipped-around logistic model for imputation, problems may occur in extreme cases when the covariates are "sufficiently continuous". Therefore in cases when the imputation model is not appropriate, ML or the iterative PPW approach recommended. Furthermore, as an approach based on simulation, MI yields different results each time it is applied, which usually presents difficulty in interpretation and communication.

Propensity score weighting can lead to estimates with large variance, as discussed in Little (1986). Little (1986) proposed smoothing the weights using empirical Bayesian methods. Robins, Rotnitzky and Zhao (1994) showed that in case that the propensity score model is correct, the IPW approach gives a consistent estimator. However, as the fraction of missing data increases, the accuracy of IPW can decrease significantly.

Under the setting of this chapter, the WEE approach requires specification of a logistic model for the missing mechanism or a sub-logistic model for the exposure X given other covariates $\mathbf{C}$, but not necessarily both. The likelihood type methods (ML via numerical maximization and the EM algorithm) require specification of the sub-logistic model. Therefore, besides correctly specifying the main regression model of interest, both approaches need to correctly specify an additional logistic regression model. The WEE method is a little more robust in the sense that it offers two additional regression models to be specified, and only one of them needs to be correct. Actually, in the simulation studies summarized in Table 2.14-Table 2.16 and the results given by Lipsitz et al. (1999), all approaches appeared quite robust against model misspecification under the settings investigated.

Based on one's belief of the underlying mechanism of how the data were generated and how the missing values were generated, different methods make different model assumptions. When the assumptions of the methods are correct, the methods should yield valid results. When the assumptions of the methods are incorrect, bias might be induced. How the methods perform in terms of the relative bias in estimation is an important question when the assumptions of the methods are incorrect. However, it is hard to make a definitive assessment of the performance of

different methods when their assumptions are incorrect, due to the wide range of configurations of model misspecification. Table 2.17 summarizes the model assumptions made by the different methods studied here.

Table 2.17 Summary of the model assumptions of different methods

	$(m \mid X, C)$	$(X \mid C)$	$(X \mid Y, C)$
ML	No	Yes	No
IPW	Yes	No	No
WEE (Robins 1994, 1995)*	Yes		Yes
WEE (Lipsitz et al. 1999)*	Yes	Yes	
MI	No	No	Yes
PPW w/ "flipped-around" model	No	No	Yes
PPW via iteration	No	Yes	No

* Only one of the two models needs to be correctly specified to get consistent estimate (Double robustness).

2.4.2. Connection between the IPW and the "Flipped-Around" Model

There is an inherent analog between the IPW and MI methods and the PPW approach based on the flipped-around logistic model. A nice discussion can be found in (Little 1988). When the value of a missing variable is imputed, it is equivalent that the corresponding observation with the observed value the same as the imputed value to be replaced by a weighted sample weighted by 1 plus the number of times it appears in the imputation. When there are other covariates $\mathbf{C}$, the matching is not exact on $\mathbf{C}$. In this case, some coarsening of the $\mathbf{C}$ information results from replacing the incomplete records by complete records. If the effects of this coarsening are negligible, then estimates from the three methods are the same. Possible effects that can cause non-negligible coarsening include large variation or abnormal distribution of the continuous components of $\mathbf{C}$.

Chapter 3. SENSITIVITY ANALYSIS FOR DATA NOT MISSING AT RANDOM IN LOGISTIC REGRESSION

3.1. Introduction

Missing values in exposure data are a widespread obstacle in statistical applications in epidemiology. Commonly, the missing at random assumption is made by investigators to ease analysis, which excludes the potential association between the occurrence of missing values and the underlying unobserved values. Investigators can be released from the hook of missing values by making the MAR assumption, and are then able to make use of well-developed methods discussed in the previous chapter, such as the likelihood based method, multiple imputation and the inverse propensity weighting method. However, this assumption is generally not testable against its alternative, Not-Missing-At-Random (NMAR), and quite often is questionable based on intuition or experience. For many types of data collection procedures, the assumption is obviously violated; for example, when subjects are asked about their daily alcohol consumption, missing rates might increase with the daily dose; when collecting data from hospital records, a given treatment is usually well documented, but it is hard to find sufficient information that a treatment is definitely not given; when asking subjects about previous diseases (especially childhood diseases) a suffered disease will be remembered well, but it might not be recalled that a disease is certainly absent.

Much has been discussed regarding statistical analysis in case of NMAR. Likelihood based methods are proposed using EM algorithm to explore the data allowing NMAR, without specific direct assumptions on the missing data mechanism. More discussions can be found in Ibrahim and Lipsitz (1996), Lipsitz, Ibrahim, Chen and Peterson (1999) and Lipsitz, Ibrahim, Chen (1999). However, these methods are actually based on specific assumptions regarding the probability distributions of the variables subject to missing values. The probability distributional form assumed for the unobserved variables actually has strong implications on the underlying missing

data mechanism (Little and Rubin 1987, Kenward 1998). Therefore, these methods should be used with caution. Application of these methods in NMAR cases blindly without checking assumptions on the missing data mechanism could induce bias.

The concept of sensitivity analysis has been used for years. As Little (1982) stated, if the response mechanism is non-ignorable, one can eliminate bias only by constructing "a model that correctly represents the response mechanism". Nordhein (1984) conducted a study on the prevalence of a genetic abnormality with sensitivity analysis by assuming the missing mechanism through the relative risk of missing rates. Vach and Blettner (1995) addressed the importance of sensitivity analysis, and proposed a framework of conducting sensitivity analysis with specification of alternative missing data mechanisms via the odds ratio of missing rates. The EM algorithm or numerical maximization was used for estimation. Molenberghs, Goetghebeur, Lipsitz, and Kenward (1999) pointed out that in contingency table settings, different models of the missing mechanism might give different prediction of the unobserved values, even though they all produce the same fit to the observed data. Therefore, Molenberghs, Kenward, and Goetghebeur (2001) argued that the role of such sensitivity analysis is to supplement information obtained from the MAR model. Besides the methods discussed above, Bayesian methods have been used in missing data problems, such as Ibrahim, Chen and Lipsitz (2002), Chen, Ibrahim, and Shao (2004), Huang, Chen and Ibrahim (2005).

Here, a framework to specify alternative missing mechanisms is proposed to facilitate sensitivity analysis of epidemiology studies with missing data using standard statistical software. To handle a questionable MAR assumption, results under a series of plausible NMAR assumptions are compared with that under the MAR assumption, such that the sensitivity to violations of the MAR assumption can be assessed. Investigators can perform analysis under the MAR assumption as usual, but are strongly encouraged to perform sensitivity analysis to assess the impact of potential departures from MAR. The proposed framework provides a simple approach to such sensitivity analysis by assigning proper weights to the subjects of an expanded data set that is designed to represent the underlying unobserved data set if all values were

observed. The expanded data set and the weights can then be supplied to any standard statistical software packages that accommodate a weighting option (e.g. SAS PROC LOGISTIC), which, as a motivation of this research, could facilitate an accessible routine for epidemiologists in standard research. Closed form formulae for the weights are given, and examples are provided to illustrate the procedures using standard software packages. A framework for such a sensitivity analysis is presented, where different options are provided to examine the violation of MAR in different configurations. The relationship among the different configurations of the missing data mechanism is illustrated. In this chapter I focus on the special case of logistic regression models with one or more binary exposure variables subject to missing data, while the remaining covariates are complete.

3.2. Methods

3.2.1. The No-Covariate Case: Basic Sensitivity Analysis

Consider a simple 2×2 table where Y indicates a binary outcome (e.g. hypertensive versus normal) and X indicates a binary predictor indicating a risk exposure (e.g. smoker versus nonsmoker). Such a dataset can be expressed in the following table.

Table 3.1 Data missing at random in a 2×2 table

$n_{m,y,x}$	Smoker not missing $(X=1,m=0)$	Nonsmoker not missing $(X=0,m=0)$	Smoking status missing $(m=1)$
Hypertensive $(Y=1)$	n_{011}	n_{010}	$n_{11\cdot}$
Normal $(Y=0)$	n_{001}	n_{000}	$n_{10\cdot}$

Define m as the missingness indicator that takes value 1 if the value of X is missing and 0 if it is not missing (note that in some literature, R is used to represent the missing/observed status, with $R=1$ as observed (response) and $R=0$ as missing (non-response)). In this case,

the assumption of MAR defined by Rubin (1976) can be expressed as (2.4) or (2.5).

Suppose one is interested in making inference on the odds ratio

$$OR = \frac{\pi_1 / (1-\pi_1)}{\pi_0 / (1-\pi_0)}, \text{where} \pi_y = \Pr(X = 1 | Y = y)$$

However, the data observed do not support the direct inference of the above π_y. Based on the above data one can only make inference on

$$\pi_y^* = \Pr(X = 1 | m = 0, Y = y) \tag{3.1}$$

without making further assumptions. A simple maximum likelihood estimate can be expressed by the cell counts as

$$\hat{\pi}_y^* = \frac{n_{0y1}}{n_{0y1} + n_{0y0}} \tag{3.2}$$

The methods that are based on the MAR assumption have been introduced in the previous chapter. Here, to conduct sensitivity analysis on π_y, one needs to make assumptions on the missing data mechanism and explore the impact it makes on the estimation of the odds ratio. A sensitivity analysis to examine potential effects of missingness on the statistical inference can be constructed based on a set of assumptions about the underlying missing data mechanism

$$Pm_{yx} = \Pr(m = 1 | Y = y, X = x), \text{ where } \; y = 0,1 \; \text{ and } \; x = 0,1$$

The missing rate is indexed by both Y and X to represent the "differential" NMAR missing mechanism, which allows the occurrence of missing values to depend on both the outcome and the exposure itself. The methodology proposed by Vach and Blettner (1995) accommodates differential NMAR, but their discussions and examples were mainly focused on non-differential NMAR, i.e.,

$$\Pr(m = 1 | Y = y, X = x) = \Pr(m = 1 | X = x).$$

Although the non-differential violation of MAR is a helpful simplifying assumption, it is unlikely to be true in practice (Carpenter, Kenward and White 2007). Such an assumption can be maintained in prospective studies, but the retrospective character of case-control studies makes it

highly questionable. Furthermore, the problem of NMAR in logistic regression is not well manifested in the non-differential NMAR case. As was discussed in the previous chapters, the estimate of the odds ratio associating the exposure and outcome would not be impacted if the missing rate of the exposure depends on itself, but not on the outcome. That is to say, a biased estimate of the odds ratio will only be induced with 'differential' missingness (Jones 1996, Proschan, McMahon, Shih 2001). As a result, without special considerations, it is necessary to assume a non-differential missing data mechanism in logistic regression if MAR is in doubt.

For a data set with the layout as in Table 3.1, an intuitive approach is to reconstruct an artificial data set that represents the underlying complete data set that would have been observed. In the setting of this dissertation, the exposure status of those subjects with X missing can only take two values, 1 or 0. Therefore an expanded data set can be constructed by replacing each of the subjects $(Y = y, \bullet)$ with X missing by the two possible realizations $(Y = y, X = 1)$ and $(Y = y, X = 0)$. A weight can be assigned to each of the two possible realizations, defined by the conditional probability of the occurrence the realization, given the observed information:

$$w_y = \Pr(X = 1 \mid m = 1, Y = y)$$

Therefore an expanded data set can be constructed as in Table 3.2. This data set can be readily supplied to standard statistical software packages with a weighting option for analysis. In the remainder of this section, a framework is proposed for how the weights are formulated in sensitivity analysis.

Table 3.2 The constructed expanded data set with appropriate weights

Observation Status of X	Observation	Number of subjects	Weight
X observed	$(Y=1, X=1)$	n_{011}	1
X observed	$(Y=1, X=0)$	n_{010}	1
X observed	$(Y=0, X=1)$	n_{001}	1
X observed	$(Y=0, X=0)$	n_{000}	1
X missing	$(Y=1, X=1)$	$n_{11.}$	$w_1 = \Pr(X=1 \mid m=1, Y=1)$
X missing	$(Y=1, X=0)$	$n_{11.}$	$1-w_1$
X missing	$(Y=0, X=1)$	$n_{10.}$	$w_0 = \Pr(X=1 \mid m=1, Y=0)$
X missing	$(Y=0, X=0)$	$n_{10.}$	$1-w_0$

The unobservable underlying exposure probability can be written as

$$\begin{aligned}\pi_y &= \Pr(m=1, X=1 \mid Y=y) + \Pr(m=0, X=1 \mid Y=y) \\ &= \Pr(X=1 \mid m=1, Y=y)\Pr(m=1 \mid Y=y) \\ &\quad + \Pr(X=1 \mid m=0, Y=y)\Pr(m=0 \mid Y=y) \\ &= w_y M_y + \pi_y^*(1-M_y)\end{aligned} \tag{3.3}$$

which cannot be directly estimated from the data set. The observable exposure probability can be written as

$$\begin{aligned}\pi_y^* &= \Pr(X=1 \mid m=0, Y=y) \\ &= \frac{\Pr(X=1, m=0 \mid Y=y)}{\Pr(m=0 \mid Y=y)} \\ &= \frac{(1-Pm_{y1})\pi_y}{1-M_y}\end{aligned} \tag{3.4}$$

which can be directly estimated from the data set by $\hat{\pi}_y^* = \dfrac{n_{0y1}}{n_{0y1}+n_{0y0}}$. In the above 2 equations,

an overall missing rate $M_y = \Pr(m=1 \mid Y=y)$ is defined to describe the overall missing probability given Y, and can be readily estimated based on the data as in Table 3.1:

$$\hat{M}_y = \frac{n_{1y\cdot}}{n_{0y1} + n_{0y0} + n_{1y\cdot}}$$

Intuitively, this overall missing rate M_y should be a weighted average of Pm_{y1} and Pm_{y0}, given by

$$\begin{aligned} M_y &= \Pr(X=1, m=1 \mid Y=y) + \Pr(X=0, m=1 \mid Y=y) \\ &= Pm_{y1}\pi_y + Pm_{y0}(1-\pi_y) \end{aligned} \tag{3.5}$$

We note that with a specified missing data mechanism Pm_{yx}, the overall missing rate M_y depends only on π_y from the underlying complete data set. Plugging (3.5) into (3.4) we have

$$\pi_y = \frac{\pi_y^*(1-Pm_{y0})}{1-Pm_{y1}-\pi_y^*(Pm_{y0}-Pm_{y1})} \tag{3.6}$$

In passing we note that, with the assumed values of missing rates specified based on one's experience, equation (3.6) gives a direct way to conduct sensitivity analysis in the no-covariate case, where π_y^* can be directly estimated by the cell counts in Table 3.1 and the Pm_{yx} can be specified.

An important relationship between the missing rates of exposure and non-exposure, implied by the observed overall missing rate, should be noted. By plugging (3.6) into (3.5), we get

$$M_y = \frac{Pm_{y0}(1-Pm_{y1}) + \pi_y^*(Pm_{y1}-Pm_{y0})}{1-Pm_{y1}+\pi_y^*(Pm_{y1}-Pm_{y0})} \tag{3.7}$$

M_y and π_y^* in (3.7) are two conditional probabilities only related to the observed data. Thus maximum likelihood estimates of these parameters obtained from the observed data are unbiased without need of additional information regarding the missing data mechanism. If we plug the MLE of them into (3.7), it yields an equation with only Pm_{y1} and Pm_{y0} as unknowns.

$$\hat{M}_y = \frac{Pm_{y0}(1-Pm_{y1}) + \hat{\pi}_y^*(Pm_{y1} - Pm_{y0})}{1-Pm_{y1} + \hat{\pi}_y^*(Pm_{y1} - Pm_{y0})} \tag{3.8}$$

Without any additional data, the missing mechanism (Pm_{y1}, Pm_{y0}) cannot be completely identified. All plausible paired values of (Pm_{y1}, Pm_{y0}) that satisfy equation (3.8) are possible. It can be easily shown from equation (3.8) that, if one assumes $Pm_{y1} = Pm_{y0}$ (or $Pm_{y1} = \hat{M}_y$), i.e., a MAR missing data mechanism, then it automatically implies that $Pm_{y1} = Pm_{y0} = \hat{M}_y$.

In sensitivity analysis, although there is no information regarding the Pm_{yx}, one should be restricted to specify values of the pair (Pm_{y1}, Pm_{y0}) that satisfy equation (3.8).

Plugging (3.6) and (3.4) into (3.3), we obtain

$$w_y = \frac{\pi_y^* Pm_{y1}(1-Pm_{y0})}{Pm_{y0}(1-Pm_{y1}) + \pi_y^*(Pm_{y1} - Pm_{y0})} \tag{3.9}$$

This gives the weights to be used to reconstruct the expanded data set as introduced above, where π_y^* can be estimated from the cell counts, and (Pm_{y1}, Pm_{y0}) should be specified by the investigator according to (3.7).

There are three ways to proceed from here with sensitivity analysis from different configuration angles. Investigators can choose the one that fits best based on their belief and experience with the potential missing data mechanism.

3.2.1.1. Direct Specification of the Missing Rate

If the maximum likelihood estimates of both M_y and π_y^* are plugged into (3.8), it gives a form of determinate relationship between Pm_{y1} and Pm_{y0}. By assuming any value for Pm_{y1}, the value of Pm_{y0} would be fully determined by the following transformation of (3.8) after plugging in the MLE of M_y and π_y^*.

$$Pm_{y0} = \frac{\hat{M}_y - Pm_{y1}\hat{M}_y - Pm_{y1}\hat{\pi}_y^* + Pm_{y1}\hat{\pi}_y^*\hat{M}_y}{1 - Pm_{y1} - \hat{\pi}_y^* + \hat{\pi}_y^*\hat{M}_y} \tag{3.10}$$

If we use the above expression (3.10) to replace the Pm_{y0} in (3.9), the defined weight as the conditional probability of the occurrence of exposure can be rewritten as

$$w_y = \frac{\pi_y^* Pm_{y1}(1 - M_y)}{(1 - Pm_{y1})M_y} \tag{3.11}$$

By replacing M_y and π_y^* with their estimated values, the weights to be used become dependent only on Pm_{y1}. Therefore if one has previous experience with the missing rates Pm_{y1}, the configuration of the alternative missing data mechanism can be specified using (3.11) as the weights. As it has been pointed out, if one assumes $Pm_{y1} = Pm_{y0}$ (or $Pm_{y1} = \hat{M}_y$), i.e., a MAR missing data mechanism, then it automatically implies that $Pm_{y1} = Pm_{y0} = \hat{M}_y$. We also notice that, by assuming MAR, the form of weight in (3.11) becomes $w_y = \hat{\pi}_y^*$, which is just the weight we used under the assumption of MAR in the previous chapter. If any previous study or experience indicates that the missing rate is higher for subjects with positive exposure (NMAR), then one can specify the missing data mechanism as $Pm_{y1} > Pm_{y0}$ (or $Pm_{y1} > M_y$). It then automatically implies that $Pm_{y1} > \hat{M}_y > Pm_{y0}$ by equation (3.8). Thus the estimate of M_y from the observed data can provide a good benchmark for one to specify the missing data mechanism via Pm_{y1} and Pm_{y0}. This saves researchers from the trouble of making too many assumptions and avoids the possibility where the assumed values of Pm_{y1} and Pm_{y0} are not compatible with the observed data.

3.2.1.2. Specification of the Risk Ratio of the Missing Rates

If there is previous experience or knowledge on the risk ratio of the missing rates (missingness risk ratio, or MRR)

$$MRR_y = \frac{Pm_{y1}}{Pm_{y0}}, \tag{3.12}$$

then one can also conduct sensitivity analysis by specifying MRR_y. By (3.10) (3.11) and (3.12), we can get a quadratic equation regarding w_y:

$$aw_y^{\,2} + bw_y + c = 0,$$

where

$$\begin{cases} a = M_y(1-M_y)(1-MRR_y) \\ b = (1-M_y)(-\pi_y^*(1-M_y)(1-MRR_y)+1-M_yMRR_y) \\ c = -\pi_y^* MRR_y(1-M_y)^2 \end{cases} \tag{3.13}$$

If one specifies $MRR_y = 1$, then there exists a single root $w_y = \pi_y^*$, which turns out to be the weight used under MAR in the previous chapter. If previous knowledge or experience suggests that $MRR_y \neq 1$, then after some simple algebra, one can find that only one root of the quadratic equation,

$$w_y = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \tag{3.14}$$

could fall within the reasonable range [0,1], under reasonable restrictions to the parameters involved in the quadratic function coefficients.

3.2.1.3. Specification of the Odds Ratio of the Missing Rates

When the variation of the missing rate becomes large, relative risks are not appropriate (Vach & Bletter 1995). In such cases, odds ratio of the missingness can be used to specify the alternative missing data mechanism. If there is previous experience or knowledge on the odds ratio of the missing rates (missingness odds ratio, or MOR)

$$MOR_y = \frac{Pm_{y1}/(1-Pm_{y1})}{Pm_{y0}/(1-Pm_{y0})} \tag{3.15}$$

then one can also conduct the sensitivity analysis by specifying MOR_y. By (3.10) (3.11) and

(3.15), we can get a quadratic equation regarding w_y :

$$aw_y^{\,2} + bw_y + c = 0,$$

The quadratic coefficient turns out to be zero, therefore it reduces to a linear term equation with solution

$$w_y = \frac{\pi_y^*(1-M_y)MOR_y}{1-\pi_y^* - M_y + \pi_y^* M_y + \pi_y^* MOR_y - \pi_y^* M_y MOR_y} \tag{3.16}$$

If one specifies $MOR_y = 1$, then it implies that $w_y = \pi_y^*$, which turns out to be the weight used in the MAR case. If previous knowledge or experience suggests that $MOR_y \neq 1$, then one can specify the alternative missing data mechanism via MOR_y and conduct sensitivity analysis using the weighting method.

3.2.2. The Covariate Case

When other covariates $\mathbf{C}$ are present, the above result can be easily generalized. Due to the complexity of multiple missing covariates, we assume that only X is exposed to missingness, while $(Y, \mathbf{C})$ shall be completely observed. To fully specify the missing data mechanism, we allow the missingness probability to depend on the outcome Y and covariates $\mathbf{C}$, as well as the actual status of exposure X . Then the conditional probability $Pm_{y\mathbf{c}x} = \Pr(m=1 \mid Y=y, \mathbf{C}=\mathbf{c}, X=x)$ defines the missing data mechanism, but cannot be directly estimated from the observed data.

Similarly to the previous section, one can reconstruct an expanded data set (or augmented data set) that represents the underlying complete data set that would have been observed. To be more specific, each complete record $(y, x, \mathbf{c})$ is kept as is, and assigned weight 1 $(y, x, \mathbf{c}, wt = 1)$. The exposure status of those subjects with X missing can only take two values, 1 or 0. Therefore an expanded data set can be constructed with each of the subjects $(m=1, Y=y, \mathbf{C}, X \text{ is missing})$ replaced by the two possible realizations $(m=1, Y=y, \mathbf{C}, X=1)$ and $(m=1, Y=y, \mathbf{C}, X=0)$. A weight $w_{y\mathbf{c}}$ is assigned to the first realization, and a weight

$1-w_{y\mathbf{c}}$ is assigned to the second, where

$$w_{y\mathbf{c}} = \Pr(X=1 \mid m=1, Y=y, \mathbf{C}=\mathbf{c}).$$

Then the expanded data set with the weights can be supplied to any standard statistical software. Practically, one is maximizing a weighted log-likelihood function

$$l(\alpha,\beta,\gamma) = \sum_{i=1}^{n} \{ I_i l_{yx\mathbf{c}}(\alpha,\beta,\gamma) + (1-I_i)[l_{y,X=1,\mathbf{c}}(\alpha,\beta,\gamma) w_{y\mathbf{c}} + l_{y,X=0,\mathbf{c}}(\alpha,\beta,\gamma)(1-w_{y\mathbf{c}})] \}$$

where $I_i = 1$ if the case is completely observed; $l_{yx\mathbf{c}}$ is the log-likelihood contribution of a complete record $(y,x,\mathbf{c})$; and $l_{y,X=x,\mathbf{c}}$ is the log-likelihood contribution of a reconstructed record $(y,X=x,\mathbf{c})$, $x=0,1$, and $w_{y\mathbf{c}}$ is the weight defined above. The reconstructed expanded data set together with the corresponding weights are indicated in Table 3.3.

Table 3.3 The structure of the constructed expanded data set with appropriate weights

Observation Status of X	Observed data	Reconstructed data	Weight
Observed	$(y,x,\mathbf{c})$	$(y,x,\mathbf{c})$	1
Missing	$(y,\bullet,\mathbf{c})$	$(y,x=1,\mathbf{c})$	$w_{y\mathbf{c}} = \Pr(x=1 \mid m=1, y, \mathbf{c})$
		$(y,x=0,\mathbf{c})$	$1-w_{y\mathbf{c}}$

The equations in section 1 can be generalized as follows. The observable exposure rate from the complete cases is

$$\begin{aligned}
\pi^{*}_{y\mathbf{c}} &= \Pr(X=1 \mid m=0, y, \mathbf{c}) \\
&= \frac{\Pr(X=1, m=0 \mid y, \mathbf{c})}{\Pr(m=0 \mid y, \mathbf{c})} \\
&= \frac{\Pr(m=0 \mid y, \mathbf{c}, X=1)\Pr(X=1 \mid y, \mathbf{c})}{\Pr(m=0 \mid y, \mathbf{c})} \\
&= \frac{(1-Pm_{y\mathbf{c}})\pi_{y\mathbf{c}}}{1-M_{y\mathbf{c}}}
\end{aligned}$$

which can be directly estimated from the data set. As X is a binary variable, intuitively, we can fit a logistic model to the complete cases as in (2.33).

$$\text{logit}[\Pr(X=1|\, y,\mathbf{c}))] = \alpha^* + \beta^* y + \gamma^{*\prime}\mathbf{c} \tag{3.17}$$

Then $\pi^*_{y\mathbf{c}}$ can be easily estimated as

$$\hat{\pi}^*_{y\mathbf{c}} = [1+\exp^{-1}(\hat{\alpha}^* + \hat{\beta}^* y + \hat{\gamma}^{*\prime}\mathbf{c})]^{-1}.$$

Define the exposure rate of interest as

$$\begin{aligned}\pi_{y\mathbf{c}} &= \Pr(X=1|\, y,\mathbf{c}) \\ &= \Pr(m=1, X=1|\, y,\mathbf{c}) + \Pr(m=0, X=1|\, y,\mathbf{c}) \\ &= \Pr(X=1|\, m=1, y,\mathbf{c})\Pr(m=1|\, y,\mathbf{c}) \\ &\quad + \Pr(X=1|\, m=0, y,\mathbf{c})\Pr(m=0|\, y,\mathbf{c}) \\ &= w_{y\mathbf{c}} M_{y\mathbf{c}} + \pi^*_{y\mathbf{c}}(1-M_{y\mathbf{c}})\end{aligned},$$

which cannot be directly estimated from the data set.

In the above 2 equations, an overall missing rate $M_{y\mathbf{c}} = \Pr(m=1|\, Y=y, \mathbf{C}=\mathbf{c})$ is defined to describe the overall missing probability given Y and $\mathbf{C}$, and can be readily estimated from the data via another logistic model

$$\begin{aligned}\text{logit}(M_{y,\mathbf{c}}) &= \text{logit}[\Pr(m=1|\, y,\mathbf{c})] \\ &= \psi_0 + \psi_1 y + \psi_2{}'\mathbf{c} + \psi_{12}{}' y\mathbf{c}\end{aligned} \tag{3.18}$$

Intuitively, this overall missing rate $M_{y\mathbf{c}}$ should be a weighted average of $Pm_{y\mathbf{c}1}$ and $Pm_{y\mathbf{c}0}$, given by

$$\begin{aligned}M_{y\mathbf{c}} &= \Pr(X=1, m=1|\, y,\mathbf{c}) + \Pr(X=0, m=1|\, y,\mathbf{c}) \\ &= \Pr(m=1|\, y,\mathbf{c}, X=1)\Pr(X=1|\, y,\mathbf{c}) + \Pr(m=1|\, y,\mathbf{c}, X=0)\Pr(X=0|\, y,\mathbf{c}) \\ &= Pm_{y\mathbf{c}1}\pi_{y\mathbf{c}} + Pm_{y\mathbf{c}0}(1-\pi_{y\mathbf{c}})\end{aligned}$$

where $Pm_{y\mathbf{c}1} = \Pr(m=1|\, y,\mathbf{c}, X=1)$ and $Pm_{y\mathbf{c}0} = \Pr(m=1|\, y,\mathbf{c}, X=0)$. Then we have

$$\pi_{y\mathbf{c}} = \frac{\pi^*_{y\mathbf{c}}(1-Pm_{y\mathbf{c}0})}{1-Pm_{y\mathbf{c}1} - \pi^*_{y\mathbf{c}}(Pm_{y\mathbf{c}0} - Pm_{y\mathbf{c}1})}$$

Then we can get an important relationship that reveals how the overall missing rate is determined

by the underlying missing data mechanism.

$$M_{yc} = \frac{Pm_{yc0}(1-Pm_{yc1}) + \pi^*_{yc}(Pm_{yc1} - Pm_{yc0})}{1 - Pm_{yc1} + \pi^*_{yc}(Pm_{yc1} - Pm_{yc0})} \tag{3.19}$$

All plausible paired values of (Pm_{yc1}, Pm_{yc0}) that satisfy (3.19) are possible. Without additional information, the actual conditional probability that defines the missing mechanism cannot be fully identified. We can write the weights that are to be assigned to reconstruct the data set as follows,

$$w_{yc} = \frac{\pi^*_{yc} Pm_{yc1}(1-Pm_{yc0})}{Pm_{yc0}(1-Pm_{yc1}) + \pi^*_{yc}(Pm_{yc1} - Pm_{yc0})},$$

where π^*_{yc} can be estimated from the cell counts, and (Pm_{yc1}, Pm_{yc0}) should be specified by the investigator. Similar to the no covariate case, there are three ways to specify the alternative missing mechanism.

3.2.2.1. Direct Specification of the Missing Rate

If the maximum likelihood estimates of both M_{yc} and π^*_{yc} are plugged into above equation, it gives a form of determinate relationship between Pm_{yc1} and Pm_{yc0}. By assuming any value for Pm_{yc1}, the value of Pm_{yc0} would be fully determined by the following transformation after plugging in the estimated values of M_{yc} and π^*_{yc}.

$$Pm_{yc0} = \frac{\hat{M}_{yc} - Pm_{yc1}\hat{M}_{yc} - Pm_{yc1}\hat{\pi}^*_{yc} + Pm_{yc1}\hat{\pi}^*_{yc}\hat{M}_{yc}}{1 - Pm_{yc1} - \hat{\pi}^*_{yc} + \hat{\pi}^*_{yc}\hat{M}_{yc}} \tag{3.20}$$

If one assumes $Pm_{yc1} = Pm_{yc0}$ (or $Pm_{yc1} = \hat{M}_{yc}$), ie. a MAR missing data mechanism, then it automatically implies that $Pm_{yc1} = Pm_{yc0} = \hat{M}_{yc}$. In passing we note that we can also write down the relationship between the conditional probability of missingness and the missingness odds ratio. It turns out that Pm_{yc1} is the roots of a quadratic equation of π^*, M_{yc}, MOR_{yc}.

$$aPm_{yc1}^2 + bPm_{yc1} + c = 0$$

where

$$\begin{cases} a = 1 - \pi^*_{yc} - M_{yc} + \pi^*_{yc} MOR_{yc} + \pi^*_{yc} M_{yc} + M_{yc} MOR_{yc} - \pi^*_{yc} M_{yc} MOR_{yc} \\ b = -1 + \pi^*_{yc} + M_{yc} - \pi^*_{yc} M_{yc} - \pi^*_{yc} MOR_{yc} - 2M_{yc} MOR_{yc} + \pi^*_{yc} M_{yc} MOR_{yc} \\ c = M_{yc} MOR_{yc} \end{cases}$$

There are two roots

$$Pm_{yc1} = \frac{M_{yc} MOR_{yc}}{1 - \pi^*_{yc} - M_{yc} + \pi^*_{yc} MOR_{yc} + \pi^*_{yc} M_{yc} + M_{yc} MOR_{yc} - \pi^*_{yc} M_{yc} MOR_{yc}} \qquad (3.21)$$

or $Pm_{yc1} = 1$

Only the first root is reasonable. Therefore, given the MLE of $\hat{\pi}^*_{yc}$ and $\hat{M}_{yc}$, the conditional probabilities of missing has a one-to-one functional form of the missingness odds ratio MOR_{yc}. In other words, given the observable $\hat{\pi}^*_{yc}$ and $\hat{M}_{yc}$ from the observed data set, the missing data mechanism cannot be determined, without additional information. All missing data mechanisms (Pm_{yc1}, Pm_{yc0}) that satisfy (3.19) are possible. After specification of the missingness odds ratio MOR_{yc}, the missing data mechanism becomes determined as in (3.21) and (3.20).

Similarly to the result in section 3.2.1, if one specifies the missing probability Pm_{yc1}, then the weights can be calculated through the following equation

$$w_{yc} = \frac{\hat{\pi}^*_{yc} Pm_{yc1} (1 - \hat{M}_{yc})}{(1 - Pm_{yc1}) \hat{M}_{yc}} \qquad (3.22)$$

.

We also notice that, by assuming MAR, the form of weight in (3.22) becomes $w_{yc} = \hat{\pi}^*_{yc}$, which is quite intuitive and the same as that used by other methods (for instance, multiple imputation) under the MAR assumption. If any previous study or experience indicates that the missing rate is higher when the exposure status is positive (NMAR), then one can assume $Pm_{yc1} > Pm_{yc0}$ (or $Pm_{yc1} > \hat{M}_{yc}$), and it automatically implies that $Pm_{yc1} > \hat{M}_{yc} > Pm_{yc0}$ by equation (3.7). Thus the estimate of M_{yc} from the available data can provide a good benchmark

for one to make assumptions about the values of Pm_{yc1} and Pm_{yc0}. This saves researchers from the trouble of making too many assumptions and avoids the possibility that the assumed values of Pm_{yc1} and Pm_{yc0} are not compatible with the observed data.

3.2.2.2. Specification of the Risk Ratio of the Missing Rates

It is possible that there is previous experience or knowledge on the risk ratio of the missing rates (missingness risk ratio, or MRR),

$$MRR_{yc} = \frac{Pm_{yc1}}{Pm_{yc0}}. \tag{3.23}$$

Then one can conduct sensitivity analysis by specifying MRR_{yc}. By (3.10) (3.11) and (3.12), we can get a quadratic equation regarding w_{yc}:

$$aw_{yc}^2 + bw_{yc} + c = 0,$$

where

$$\begin{cases} a = M_{yc}(1-M_{yc})(1-MRR_{yc}) \\ b = (1-M_{yc})(-\pi_{yc}^*(1-M_{yc})(1-MRR_{yc}) + 1 - M_{yc}MRR_{yc}) \\ c = -\pi_{yc}^* MRR_{yc}(1-M_{yc})^2 \end{cases}$$

If one specifies $MRR_{yc} = 1$, then there exists a single root $w_{yc} = \pi_{yc}^*$. If previous knowledge or experience suggests that $MRR_{yc} \neq 1$, then after some simple algebra, one can find that only one root of the quadratic equation,

$$w_{yc} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \tag{3.24}$$

could fall within the feasible range [0, 1], under reasonable restrictions to the parameters involved in the quadratic equation coefficients.

3.2.2.3. Specification of the Odds Ratio of the Missing Rates

If there is previous experience or knowledge on the odds ratio of the missing rates (missingness odds ratio, or MOR)

$$MOR_{y\mathbf{c}} = \frac{Pm_{y\mathbf{c}1} / (1 - Pm_{y\mathbf{c}1})}{Pm_{y\mathbf{c}0} / (1 - Pm_{y\mathbf{c}0})},$$

then the weights can be calculated through the following equation

$$w_{y\mathbf{c}} = \frac{\pi^*_{y\mathbf{c}}(1 - M_{y\mathbf{c}})MOR_{y\mathbf{c}}}{1 - \pi^*_{y\mathbf{c}} - M_{y\mathbf{c}} + \pi^*_{y\mathbf{c}}M_{y\mathbf{c}} + \pi^*_{y\mathbf{c}}MOR_{y\mathbf{c}} - \pi^*_{y\mathbf{c}}M_{y\mathbf{c}}MOR_{y\mathbf{c}}} \tag{3.25}$$

Again, if one specifies $MOR_{y\mathbf{c}} = 1$, then it implies $w_{y\mathbf{c}} = \pi^*_{y\mathbf{c}}$, which turns out to be the weight used in the MAR case.

The parameters used in the calculation of weights include the complete case exposure probability $\pi^*_{y\mathbf{c}}$, the overall missing probability $M_{y\mathbf{c}}$, and the specified $Pm_{y\mathbf{c}}$ or $MRR_{y\mathbf{c}}$ or $MOR_{y\mathbf{c}}$. With covariate $\mathbf{C}$, the MLE of $\pi^*_{y\mathbf{c}}$ can be readily obtained assuming a logistic model of X on Y and $\mathbf{C}$ with the complete cases; whilst the MLE of $M_{y\mathbf{c}}$ can be readily reached by fitting a logistic model of the missing indicator m on Y and $\mathbf{C}$ with all the observed data. Then with the MLE's and the specified values, the weights can be calculated and standard statistical software packages can be used to carry out the rest of the analysis as usual.

3.2.3. Standard Error Estimation

Due to the fact that the weights in the proposed method are constructed with estimated probabilities, their sampling variability should be considered in a proper estimation of standard error. Resampling-based methods like the bootstrap (Efron and Tibshirani 1993) or jackknife (Hinkley 1983) are recommended to properly account for such variability, and we recommend the jackknife over bootstrapping due to a lower propensity for numerical problems (e.g., see Lyles and Lin, 2010). For each leave-one-out sample from the original observed data, we re-calculate the weights and re-fit the weighted logistic regression using standard software. The jackknife

standard error is calculated based on the re-fitted estimate as in equation (1.3).

3.3. Examples

3.3.1. No Covariate Case

Lyles and Allen (2002) used an artificially augmented data set as an illustration of their proposed maximum likelihood method, based on an example from Rosner (1995). Data from 13,465 women in a breast cancer study are presented in Table 3.4. The disease status (Cancer positive or not) and the risk exposure (dichotomized age of first birth: ≥ 30 or < 30) were recorded in a 2×2 table. Suppose hypothetically that there were 300 additional cases and 2,500 additional controls for whom the risk exposure was missing.

Table 3.4 Data from a case-control study of the association between breast cancer and age at first birth*

Age at first birth (years)	Disease status		
	Case ($D^{\dagger} = 1$)	Control ($D = 0$)	Total
$\geq 30 (E^{\dagger} = 1)$	683	1,498	2,181
$< 30 (E = 0)$	2,537	8,747	11,284
Total	3,220	10,245	13,465

*Source: MacMahon et al. (1970). From Fundamentals of Biostatistics, 4th edition, by B. Rosner © 1995. Reprinted with permission of Brooks/Cole an imprint of the Wadsworth Group, a division of Thomson Learning. Fax 800 730-2215.

† D, disease status; E, exposure status.

With the hypothetical additional cases and controls with missing exposure status, it suggests that the missingness probabilities for cases and controls are 0.085 and 0.196, respectively, or $\hat{M}_1 = 0.085, \hat{M}_0 = 0.196$. If one assumes that the exposure is missing at random, then, using the observed cells in Table 3.4, one can obtain that $\hat{\pi}_1^* = 683 / 3,220 = 0.212$ (standard error (SE),

0.0072) and $\hat{\pi}_0^* = 1,498/10,245 = 0.146$ (SE, 0.0035). The complete-case analysis gives an estimate of the odds ratio $\widehat{OR}^* = 1.57$, with a 95 percent confidence interval of (1.42, 1.74). In contrast, there might be evidence suggesting a NMAR missing data mechanism. Suppose that if a subject has breast cancer $(D=1)$ and her age at first birth is greater than 30 $(E=1)$, then the probability is 0.105 that her exposure status is missing; whilst if a subject does not have breast cancer $(D=0)$ and her age at first birth is greater than 30 $(E=1)$, then the probability is 0.315 that the exposure status is missing. That is to say that $Pm_{11} = 0.105, Pm_{01} = 0.315$. Apparently, the missingness probabilities depend on the true value of the exposure, as well as the outcome, which is a differential NMAR missing data mechanism. To get estimates under this NMAR assumption, one can reconstruct an expanded data set as in Table 3.2. Using the formula (3.11), one can get $w_1 = 0.266, w_0 = 0.276$. Then any standard statistical software package can be used to perform analysis on the data set with weights as in Table 3.2. Here with PROC LOGISTIC in SAS, we obtain that the estimate of the odds ratio is 1.34, with a 95 percent confidence interval of [1.21, 1.48].

If there is no direct information on the missingness probability, but the risk ratio of the missingness probabilities is available, then one can specify the missingness risk ratio. Suppose that the MRR for cases is 1.31, and the MRR for controls is 1.83. One can reconstruct an expanded data set as in Table 3.2. Then by the formulas (3.13) and (3.14), one can also obtain the weights $w_1 = 0.266, w_0 = 0.276$.

If there is information on the odds ratio of the missingness probabilities, then one can specify the missingness odds ratio. Suppose that the MOR for cases is 1.35, and the MRR for controls is 2.21. One can reconstruct an expanded data set as in Table 3.2. Then by the formulas (3.16) one can obtain the weights $w_1 = 0.266,$ and $w_0 = 0.276$.

In the application of the three different configurations of the alternative missing data mechanism, the values of the corresponding parameters were set so that they represent the same

magnitude of deviation from MAR; therefore they result in the same value of weights.

If it is of interest to conduct analysis on the sensitivity of the odds ratio estimate to the violation of MAR, one can make a series of possible alternative NMAR assumptions and compare the results under these assumptions against that under MAR. Assuming a differential missing mechanism, it allows that the missingness rates of exposure can deviate from MAR differently in cases and controls. First we set the missing data mechanism of cases to be MAR, and let the missing data mechanism of controls to deviate away from MAR. This can assess the sensitivity to violation of MAR for the controls. That is, for $Y=1$, keep the missing probability of X at the overall missing rate, which implies MAR for the subjects with $Y=1$. The missing mechanism for $Y=1$ is specified by $Pm_{11}=0.085$ or $MRR_1=1$ or $\log(MOR_1)=0$. The sensitivity of violation to MAR for subjects with $Y=0$ is assessed. The result is shown in Table 3.5. We can also implement the same idea the other way around to assess the sensitivity to violation of MAR for the cases. That is, for $Y=0$, keep the missing probability of X at the overall missing rate, which implies MAR for the subjects with $Y=0$. The missing mechanism for $Y=0$ is specified by $Pm_{01}=0.196$ or $MRR_0=1$ or $\log(MOR_0)=0$. The sensitivity of violation to MAR for subjects with $Y=1$ is assessed. The result is shown in Table 3.6. The point estimate and the 95% confidence interval were displayed in Figure 3.1.

Table 3.5 Sensitivity analysis of MAR assumption for the Controls

Pm_{01}	MRR_0	$\log(MOR_0)$	Estimated Odds Ratio [95% Confidence Interval]
0.0052	0.023	-4	2.01 [1.82, 2.22]
0.014	0.063	-3	1.99 [1.80, 2.20]
0.036	0.167	-2	1.94 [1.75, 2.14]
0.090	0.425	-1	1.81 [1.64, 2.01]
0.196	1	0	1.57 [1.42, 1.74]
0.346	2.12	1	1.23 [1.11, 1.36]
0.482	4.13	2	0.92 [0.82, 1.02]
0.564	9.32	3	0.73 [0.66, 0.81]
0.601	22.37	4	0.64 [0.58, 0.72]

* Numbers in each cell reflect mean and 95% confidence interval in [ ].

** Missing data in Cases were assumed MAR. Alternative missing data mechanisms were specified for the Control group to assess sensitivity to violation of MAR.

*** The missingness probability Pm_{01}, missingness risk ratio MRR_0 and log-missingness odds ratio $\log(MOR_0)$ in each row are mutually determined as discussed in Section 3.2.1. These three columns are displayed here to demonstrate how the three ways of specification of alternative missing data mechanisms are unified when the overall missing rate $\hat{M}_0$ is estimated from available data.

Table 3.6 Sensitivity analysis of MAR assumption for the Cases

Pm_{11}	MRR_1	$\log(MOR_1)$	Estimated Odds Ratio [95% Confidence Interval]
0.0021	0.020	-4	1.41 [1.28, 1.56]
0.0058	0.055	-3	1.42 [1.28, 1.57]
0.015	0.148	-2	1.43 [1.30, 1.59]
0.038	0.392	-1	1.48 [1.33, 1.63]
0.085	1	0	1.57 [1.42, 1.74]
0.156	2.45	1	1.74 [1.57, 1.93]
0.226	5.95	2	1.95 [1.76, 2.17]
0.270	14.93	3	2.11 [1.90, 2.35]
0.291	39.01	4	2.20 [1.97, 2.45]

* Numbers in each cell reflect mean and 95% confidence interval in [ ].

** Missing data in Controls were assumed MAR. Alternative missing data mechanisms were specified for the Cases to assess sensitivity to violation of MAR.

*** The missingness probability Pm_{11}, missingness risk ratio MRR_1 and log-missingness odds ratio $\log(MOR_1)$ in each row are mutually determined as discussed in Section 3.2.1. These three columns are displayed here to demonstrate how the three ways of specification of alternative missing data mechanisms are unified when the overall missing rate $\hat{M}_1$ is estimated from available data.

In Figure 3.1, it is apparent that the violation of MAR of the controls makes more impact on the estimation of the odds ratio. The estimated odds ratio and its 95% confidence interval are above 1, inferring that it is statistically significant that a subject with positive risk exposure (first birth age greater than 30) would be more likely to have breast cancer. However, if the missing

data mechanism works in such a way that subjects in the control group with positive risk exposure are more likely to have missing values in their exposure status, then the above statistically significant conclusion could become insignificant. It might even become statistically significant in the opposite direction if the magnitude of the deviation from MAR as described above is strong enough. However, the deviation from MAR for the cases does not make as much impact. Reasons for the more sensitivity in the controls include, but are not limited to, the larger number of subjects and higher missing rate in that group, which leave more room for the impact of the missing data mechanism on the estimation. Therefore, the sensitivity analysis here raised an important alert for investigator to closely examine the missingness or the exclusion criteria of the control group.

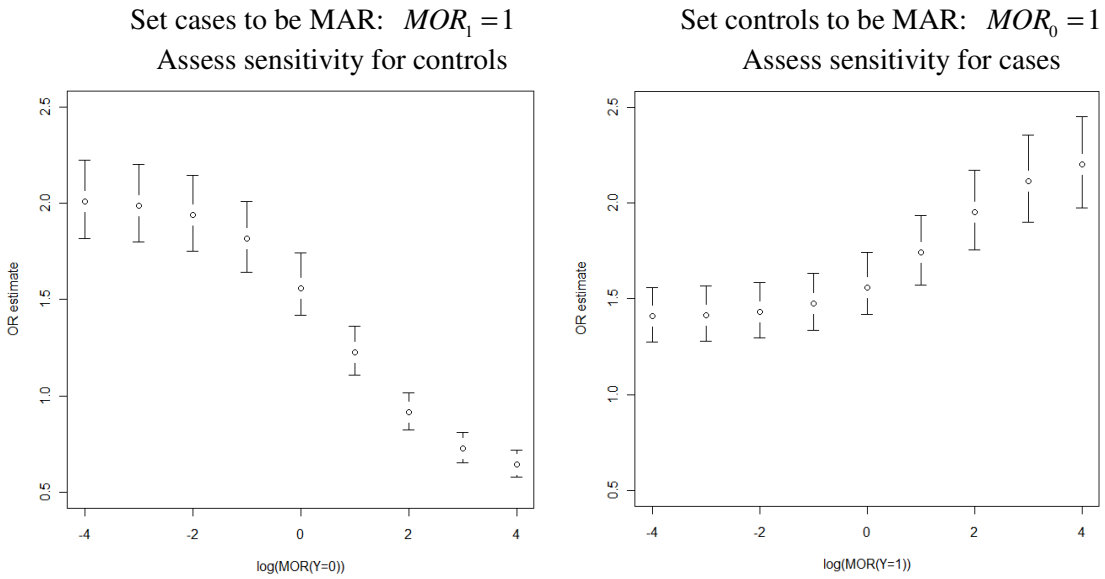

Figure 3.1 Point estimate of odds ratio with 95% confidence interval for a series of combinations of alternative missing data mechanisms as in Table 3.5 and Table 3.6

A contour plot in Figure 3.2 was drawn to make further investigation on the sensitivity to violation of MAR. In the right lower corner beyond the blue line (corresponds to $\widehat{OR} = 1$), the specified missing data mechanism in that area would change the direction of the estimation. For instance, if it is MAR for the cases, and the missing data mechanism for the controls is NMAR

with $\log(MOR_0) = 1.69$ or higher, then the estimated OR becomes less than 1. This corresponds to ($Pm_{11} = Pm_{10} = 0.085$, $Pm_{01} \geq 0.45$, $Pm_{00} \leq 0.13$) and ($MRR_1 = 1$, $MRR_0 \geq 3.45$).

We know that under the MAR assumption, the complete-case analysis gives an estimate of the odds ratio $\widehat{OR}^{*} = 1.57$, with a 95 percent confidence interval of (1.42, 1.74). Therefore, we conclude that when assuming MAR there is a significant positive association between the occurrence of breast cancer and late first birth at level 0.05. It is of interest to investigate the impact of the possible NMAR on the conclusion of the statistical inference. Therefore, all the missingness scenarios resulting the lower 95 percent confidence limit equal or less than 1 was drawn in Figure 3.3. The paired values of $\{\log(MOR_1), \log(MOR_0)\}$ within the right lower shaded corner of the curve will result in a 95 percent confidence interval containing 1. Therefore, it leads to failure of rejection of the null hypothesis in the hypothesis testing on the association between the occurrence of breast cancer and late first birth at level 0.05. If there is suspicion that the missing data mechanism might be within that area, then one should be cautious when making the conclusion that there is a significant positive association between the occurrence of breast cancer and late first birth. To be more specific, this area includes, but not limited to, all the settings on the line for $\{\log(MOR_1) = 0, \log(MOR_0) \geq 1.36\}$. This corresponds to ($Pm_{11} = Pm_{10} = 0.085$, $Pm_{01} \geq 0.40$, $Pm_{00} \leq 0.15$) and ($MRR_1 = 1$, $MRR_0 \geq 2.74$).

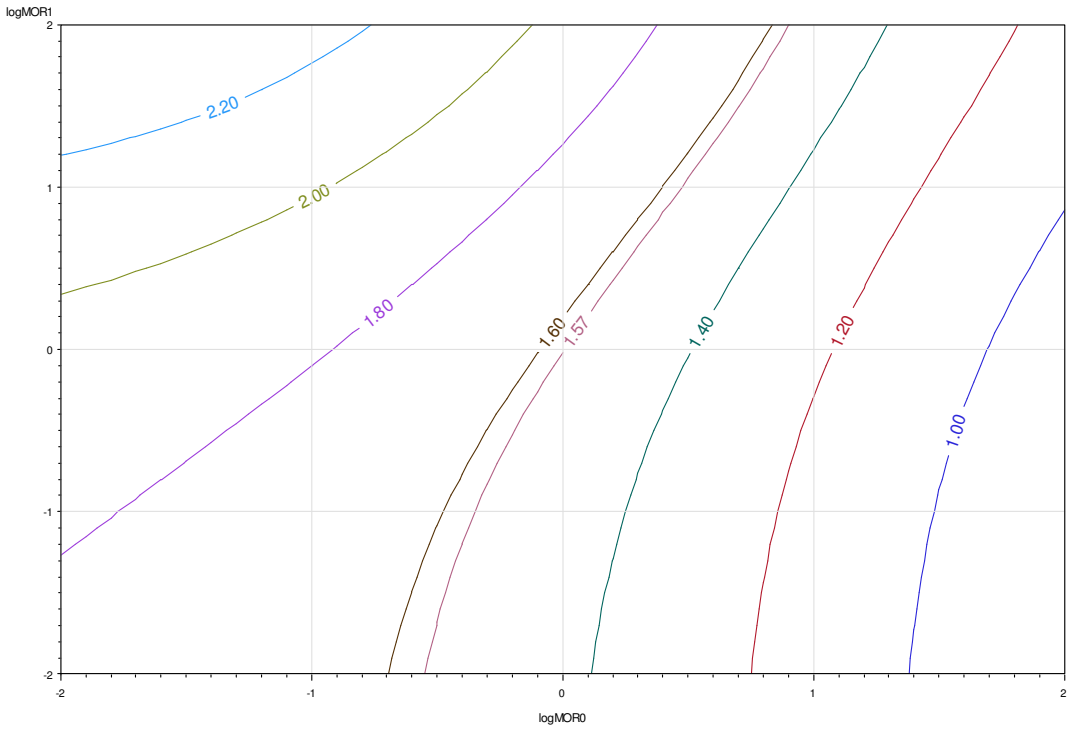

Figure 3.2 Contour plot of estimated odds ratio (labeled in each line) by a variety of combinations of MOR_0 and MOR_1 demonstrates the sensitivity to violation of MAR. The line crossing the original point (0, 0) represents all the scenarios resulting estimated OR being 1.57, which is the same as assuming MAR.

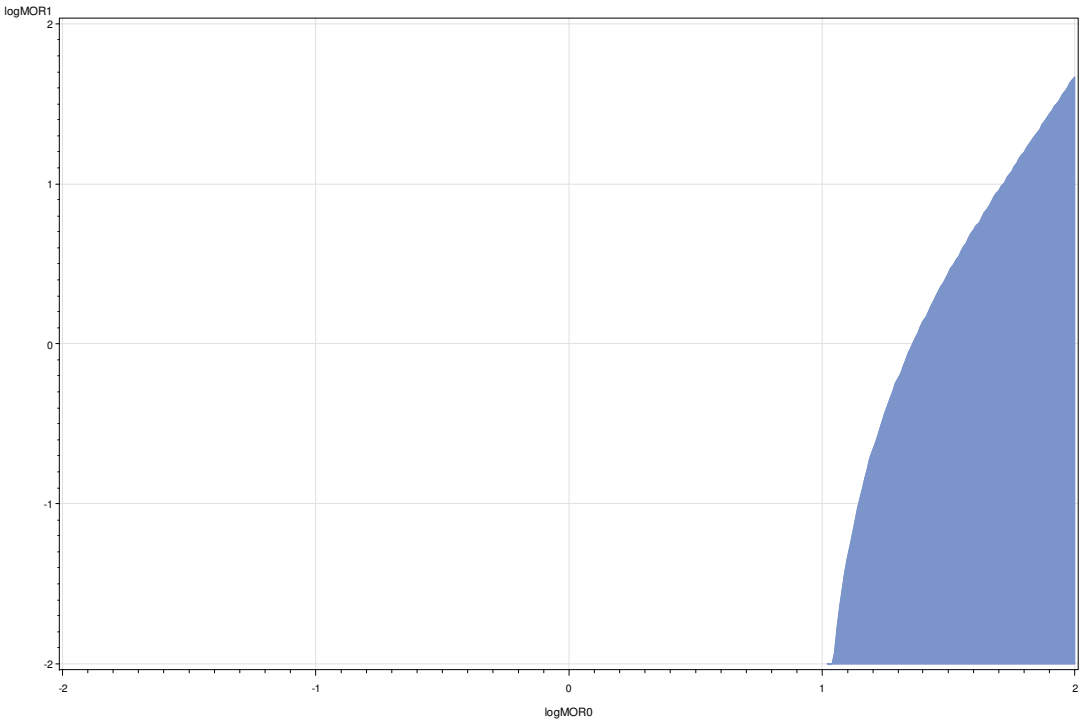

Figure 3.3 The shaded area represents all scenarios resulting failure to reject H_0: OR=1

3.3.2. Covariate Case

The National Alzheimer's Coordinating Center (NACC) collected uniform data sets from approximately thirty Alzheimer's Disease Centers (ADCs) that covered information on patients' demographics, results of clinical testing and assessment, medical and family history, and diagnoses. Steenland et al. (2010) conducted a cross-sectional analysis of the data set via polytomous logistic regression (generalized logits model) to determine which variables were most important in diagnoses of Normal, Mild Cognitive Impairment (MCI) or Alzheimer's Disease (AD). Subjects with cognitive impairment but not MCI were excluded, resulting 8,495 subjects in their analysis. Missing values were found in many variables. Complete case analysis was performed, and then multiple imputation under the missing at random assumption was also conducted to address the missing data problem.

In this example, an illustrative data set based on the above study was used to demonstrate the application of the proposed method on sensitivity analysis. We recommend assuming that only one variable is subject to NMAR at a time, and the other variables are missing at random, because it could help to pinpoint the source of impact compared to allowing multiple variables NMAR simultaneously. For instance, we were interested in sensitivity analysis on the variable Wechsler Adult Intelligence Scale (WAIS), for which the overall missing rate was 11.55%. The variable WAIS took values of positive integers, with higher values indicating a better cognitive ability; therefore it was dichotomized for illustration as an indicator of whether the subject's WAIS value was above or below its corresponding norm. For simplicity of illustration, we assumed that all the other variables were completely observed. The missing values in the other variables were imputed using imputation method under the MAR assumption (PROC MI, SAS Institute), and treated as known. The result in this example is only for demonstration of the proposed method, without intention of making any inferences on the original study by Steenland et al (2010).

There were 8,495 subjects in the illustrative data set, with variable WAIS subject to missing values in 981 subjects, leaving 7,514 complete cases. Table 3.7 summarizes the prevalence of missing values of the illustrative data set. In Table 3.7, we observe much higher missing rate in

the AD group, whilst the missing rates are close in the Normal and MCI groups. One important reason for missing values in WAIS is due to severe impairment of the subject's recognition, which results in failure to finish the test. Such failure is recorded as missing values, but actually infers low WAIS score. Therefore the missing data mechanism is suspiciously NMAR.

Table 3.7 Summary of the prevalence of missing values in the National Alzheimer's Coordinating Center data (Steenland et al., 2010)

Cognitive impairment levels	No. of subjects	WAIS missing	Missing %
Normal	4241	400	9.43
MCI	2198	203	9.24
AD	2056	378	18.39
Total	8495	981	11.55

The original study result was presented in Table 5 in Steenland et al. (2010) where multiple imputation was used to address the missing data problem under the assumption of MAR. In this example, the same polytomous logistic regression model was considered to compare the result for demonstration. However, to apply the proposed method, the predictor variable WAIS was dichotomized based on a standard norm such that $X=1$ if WAIS is higher than norm (less severe disease status), $X=0$ otherwise. Missing values in other variables are assumed MAR, and imputed. With the diagnosis result as the nominal outcome at three levels (Normal, MCI and AD), the polytomous logistic regression, taking MCI as the reference group, can be written as follows,

$$\begin{cases} \log[\Pr(\text{AD})/\Pr(\text{MCI})] = \alpha_1 + \beta_1 x + \gamma_1' \mathbf{c} \\ \log[\Pr(\text{Normal})/\Pr(\text{MCI})] = \alpha_2 + \beta_2 x + \gamma_2' \mathbf{c} \end{cases}$$

where $\mathbf{c}$ represents 17 other controlling covariates. Naïve analysis using only the complete cases were performed (Table 3.8).

Table 3.8 Summary of the complete case analysis on the National Alzheimer's Coordinating Center data

Variable	MCI versus normal			AD versus MCI		
	OR	χ^2	p-value	*OR*	χ^2	p-value
Clinician-reported decline	19.23	585.00	<0.0001	5.55	21.77	<0.0001
CDR sum of boxes (per unit)	2.17	65.83	<0.0001	2.45	286.79	<0.0001
Consensus versus single-clinician diagnosis	3.27	100.91	<0.0001	0.67	9.12	0.003
MMSE (per unit)	0.85	31.26	<0.0001	0.85	56.99	<0.0001
Logical memory delayed (per unit)	0.93	38.32	<0.0001	0.94	14.63	0.0001
Category Fluency Test (per unit)	0.95	53.42	<0.0001	0.96	16.82	<0.0001
Education (per year of schooling)	1.09	28.97	<0.0001	1.07	15.64	<0.0001
Hachinski Ischemia score (per unit)	0.97	0.56	0.45	0.76	40.83	<0.0001
Boston Naming Test (per unit)	0.93	25.25	<0.0001	0.96	12.25	0.0005
Trail Making Test B (per 10 units)	1.01	40.03	<0.0001	1.005	27.91	<0.0001
Age (per year)	0.99	4.73	0.03	0.98	8.02	0.005
Race (white vs. black/Hispanic)	1.30	4.01	0.05	1.98	17.02	<0.0001
First-degree relative demented	1.14	1.88	0.17	1.53	13.73	0.0002
Trail Making Test A (per 10 units)	0.991	9.15	0.003	0.997	1.89	0.17
Depression	0.69	5.87	0.02	0.71	5.21	0.02
FAQ (per 5 units)	1.02	0.60	0.44	1.03	5.46	0.02
Geriatric Depression Scale (per unit)	0.99	0.36	0.55	0.98	2.80	0.09
WAIS (above norm vs. not)	0.74	4.46	0.03	1.06	0.15	0.69

7,514 subjects included. Missing values in variables other than WAIS were filled in by simple imputation assuming MAR.

The result from the CC analysis indicated that subjects with below-norm WAIS were more likely to be diagnosed as MCI (vs. Normal), and the association is statistically significant; whilst subjects with above-norm WAIS are slightly more likely to be diagnosed as AD (vs. MCI), and the association is not statistically significant. With the observation of the sizable missing values in the AD subjects, we were interested in exploring the impact of the missing data on the above statistical inference.

One potential reason for missing values to be recorded for WAIS is that the subject is severely cognitive impaired, thus not able to finish the test. Therefore it is suspected that the occurrence of missingness of WAIS is associated with low WAIS values, which results in NMAR. This is especially of concern for the subjects in AD group, whose average WAIS tends to be lower. We conducted sensitivity analysis on the missing values in WAIS in AD subjects. The missing data mechanism was specified in three different ways, namely the directly specification of the missing rates, specification of the missingness relative risk, and specification of the missingness odds ratio. In specification of the missingness odds ratio, compared to the examples in Vach and Blettner (1995), we allowed "differential" NMAR, so that subjects with diagnoses of Normal and MCI were assumed to be MAR, whilst only subjects with diagnosis of AD might be NMAR in such a way that subjects with below-norm WAIS were more likely to incur missing values. For simplicity, we also assumed that the missing data mechanism would not depend on the other controlling variables, such as demographics and etc. With the diagnostic result considered as a nominal variable (represented by dummy variables I_{AD} and I_{Normal}), the estimable exposure rate by complete cases, as modeled by (3.17) can be written as

$$\text{logit}(\pi^{*}_{y\mathbf{c}}) = \phi_0 + \phi_1 I_{\text{AD}} + \phi_2 I_{\text{Normal}} + \phi_3' \mathbf{c}\,.$$

The overall missing rate as modeled by (3.18) can be written as,

$$\text{logit}(M_{y\mathbf{c}}) = \psi_0 + \psi_1 I_{\text{AD}} + \psi_2 I_{\text{Normal}} + \psi_3' \mathbf{c}\,.$$

The magnitude of the deviation from MAR was increased at different levels, and the estimated odds ratio and relative statistics at each level are summarized in Table 3.9. Apparently,

the differential NMAR could make dramatic change on the estimated OR for AD vs. MCI. However, the chi-square test for the OR is not significant even when the logarithm missingness ratio takes extreme value as -2.0. Therefore the association between the diagnosis (AD vs. MCI) and WAIS testing is not significant even under extreme differential NMAR condition.

As discussed above, the missing values in AD subjects are more likely to be associated with lower WAIS. From the missing data pattern in Table 3.7, if we assume that the around 9% missing rate in the Normal and MCI groups represents some background MAR or MCAR due to data collection, it is intuitive to speculate that the surplus proportion of missing values beyond the background is caused by failure to finish the test. Taking the average missing rate in the Normal and MCI groups, 9.36%, to induce MAR subjects in the AD group, it results in 91 subjects had above norm WAIS while 82 for below norm. We assume that the surplus 205 subjects were actually all below norm WAIS. In this way, the odds ratio of missingness for the AD is 0.286.

Table 3.9 Sensitivity analysis of violation to MAR in AD group, assuming MAR for Normal and MCI groups

Missing odds ratio (log(OR)) for AD subjects	MCI versus normal			AD versus MCI		
WAIS above norm vs. not	OR	χ^2	p-value	OR	χ^2	p-value
1 (0)	0.74	4.64	0.03	1.04	0.06	0.80
0.82 (-0.2)	0.74	4.65	0.03	1.01	0.01	0.92
0.67 (-0.4)	0.74	4.66	0.03	0.99	0.0032	0.95
0.55 (-0.6)	0.74	4.68	0.03	0.97	0.048	0.83
0.45 (-0.8)	0.73	4.70	0.03	0.95	0.15	0.70
0.37 (-1)	0.73	4.72	0.03	0.92	0.31	0.58
0.29 (-1.251)	0.73	4.74	0.03	0.90	0.59	0.44
0.22 (-1.5)	0.73	4.78	0.03	0.87	0.97	0.33
0.14 (-2.0)	0.73	4.86	0.03	0.82	2.00	0.16

3.3.3. Monte Carlo Sensitivity Analysis

In previous examples, the missing data mechanism was specified by investigator and assumed fixed. Previous literature suggested considering additional variability due to uncertainty about these parameters in traditional sensitivity analysis on misclassification (Fox, Lash and Greenland, 2005; Chu, Wang, Cole and Greenland, 2006; Orsini et al. 2008; and Gustafson, Le, Saskin 2001). The same idea was borrowed here to facilitate a Monte Carlo sensitivity analysis to consider additional variability in specification of the missing data mechanism. This is implemented by specifying a underlying density for the parameters define the missing data mechanism and applying imputation-like or Bayesian methods.

We use again the Alzheimer disease example for illustration. As discussed in Section 3.3.2, we observe a much higher missing rate in the AD group (Table 3.7), whilst the missing rates are close in the Normal and MCI groups. We know that one important reason for missing values in WAIS is due to severe impairment of the subject's recognition, which results in failure to finish the test. Such failure is recorded as missing values, but actually infers low WAIS score. Therefore the missing data mechanism is suspiciously NMAR, and analysis based on MAR assumption would induce bias. We know that missing values in a data set can come from different sources and form a fixture of MCAR, MAR and NMAR. Based on the knowledge and previous experience from epidemiology studies, it is reasonable that the shared missingness pattern (missing rate at about 9.36% on average) in the Normal and MCI groups is assumed to be MAR, and represents some background rate of missing values recorded due to unrelated reasons. Furthermore, it is also arguable that in the higher missing rate in AD group, the surplus of missing values beyond this MAR background are from the NMAR source where failure of finishing the test was recorded as missing.

The total number of subjects with missing WAIS in AD group is 378. We allocate them to the two categories of WAIS level based on the above argument. This results in a missing data mechanism as described in Table 3.10. The estimated missing odds ratio is $MOR_{AD} = 0.286$, which yields estimated logarithm transformed missing odds ratio $\log(MOR_{AD})$ to be -1.251 ,

with standard error 0.130.

Table 3.10 Reallocated Contingency Table for AD Group

		WAIS is Missing		Total
		No	Yes	
WAIS	Equal or below norm	795	287	1,082
	Above norm	883	91	974

378 subjects with WAIS missing were reallocated. This table represents the case where these 378 subjects are reallocated in such a way that missingness beyond the MAR background missingness is all among the low WAIS group (worse disease status).

Suppose that we wish to summarize uncertainty about the specification of MOR_{AD} by assuming that $\log(MOR_{AD})$ derive from a normal distribution with mean -1.251 and standard deviation 0.130. We assume MAR for Normal and MCI groups, therefore the missing data mechanism is ignorable for these two groups, and we do not need to impose a random distribution on their missing odds ratios. We produce one point estimate for the odds ratio, together with an interval estimate that simultaneously takes account of the variability in the original data and the postulated systematic variability of the missing data mechanism. We accomplished this as follows. First, we independently selected 100 values of $\log(MOR_{AD})$ randomly from the assumed normal distribution. For each value, we computed the estimated logarithm OR for AD vs. MCI and MCI vs. Normal groups, via the proposed weighting method, together with its jackknife standard error. We then generated 100 random draws from a normal distribution with mean and standard deviation matching that estimated logarithm OR, and its associated standard error. Therefore we obtained the OR distribution for each fixed specification of missing data mechanism. Repeating this process for each $\log(MOR_{AD})$ and pooling the results produced a histogram of 100×100 OR values, which is depicted Figure 3.4. The 2.5^{th}, 50^{th}, and 97.5^{th} percentiles of this distribution are 0.680, 0.895, and 1.186, respectively, which produces a median OR estimate of 0.895, with approximate 95% confidence limits of (0.680, 1.186). These may be contrasted with the

traditional sensitivity analysis at $\log(MOR_{AD}) = -1.251$, which produces OR estimate of AD versus MCI of 0.896 (0.679, 1.184).

Another reasonable prior distribution for $\log(MOR_{AD})$ is a triangular distribution with peak at -1.251 (Table 3.10), minimum at -6.061 (reallocate all but one subject with missing value to low WAIS group) and maximum at 0 (Figure 3.5). In this case, we assume that the missingness in AD group happens at least MAR as that in Normal and MCI groups, but can deviate away from MAR in the direction toward more missing values for low WAIS subjects up to the extreme case where all but one missing value are from low WAIS.

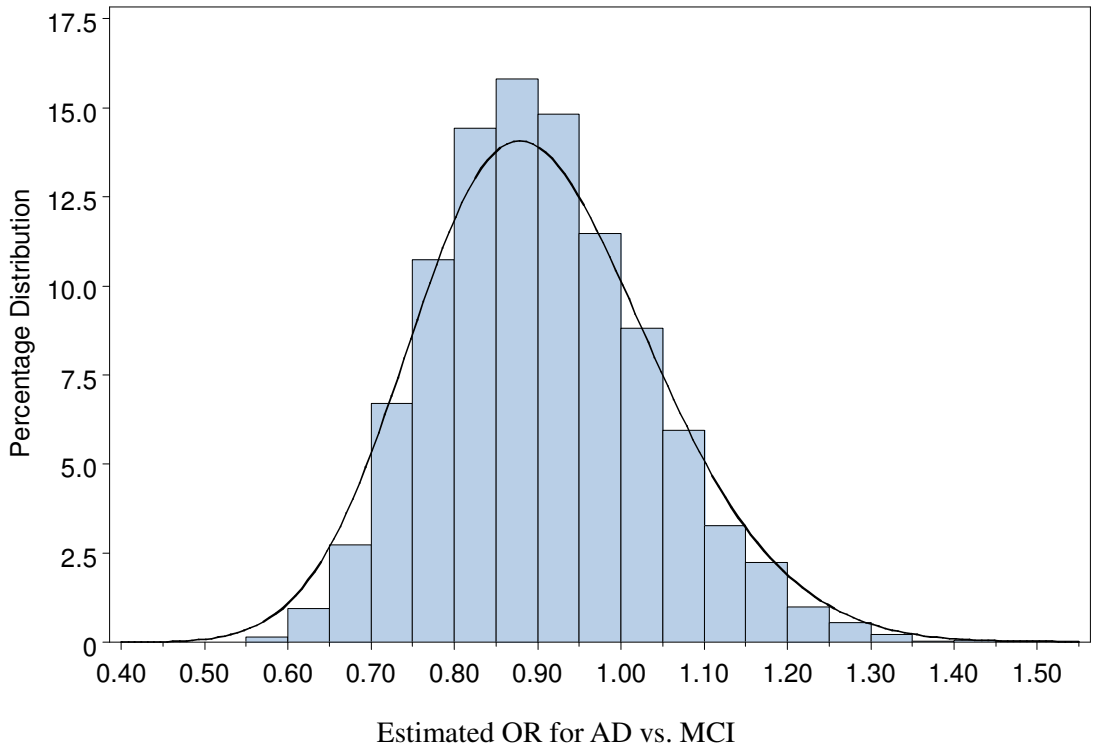

Figure 3.4 Histogram of posterior odds ratio AD vs. MCI with empirical kernel smoothing with prior of $\log(MOR_{AD})$ from normal distribution

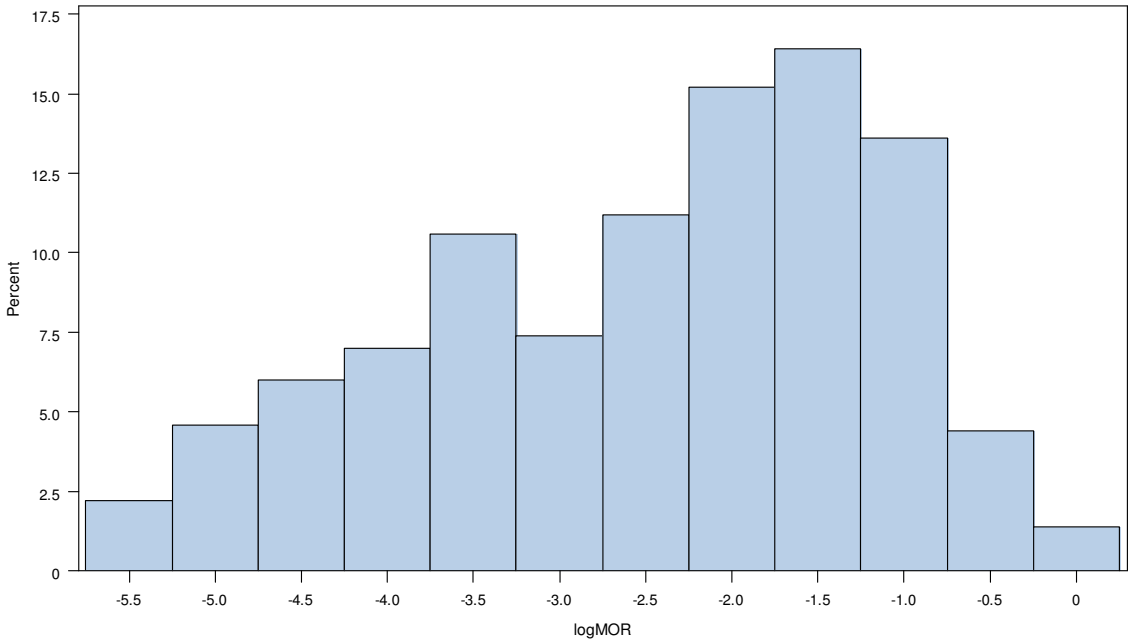

Figure 3.5 Histogram of triangular prior distribution of $\log(MOR_{AD})$ with peak at -1.251 minimum at -6.061 and maximum at 0

Suppose that we wish to summarize uncertainty about the specification of MOR_{AD} by assuming that $\log(MOR_{AD})$ derive from such a triangular distribution. Again, we assume MAR for Normal and MCI groups, therefore the missing data mechanism is ignorable for these two groups, and we do not need to impose a random distribution on their missing odds ratios. This time, we independently selected 1000 values of $\log(MOR_{AD})$ randomly from the assumed triangular distribution. Repeating the processes described as above and we obtained a histogram of 1000×100 OR values, which is depicted Figure 3.6. The 2.5^{th}, 50^{th}, and 97.5^{th} percentiles of this distribution are 0.548, 0.790, and 1.141, respectively, which produces a median OR estimate of 0.790, with approximate 95% confidence limits of (0.548, 1.141). These may be contrasted with the traditional sensitivity analysis at $\log(MOR_{AD}) = -1.254$, which produces OR estimate of AD versus MCI of 0.896 (0.679, 1.184).

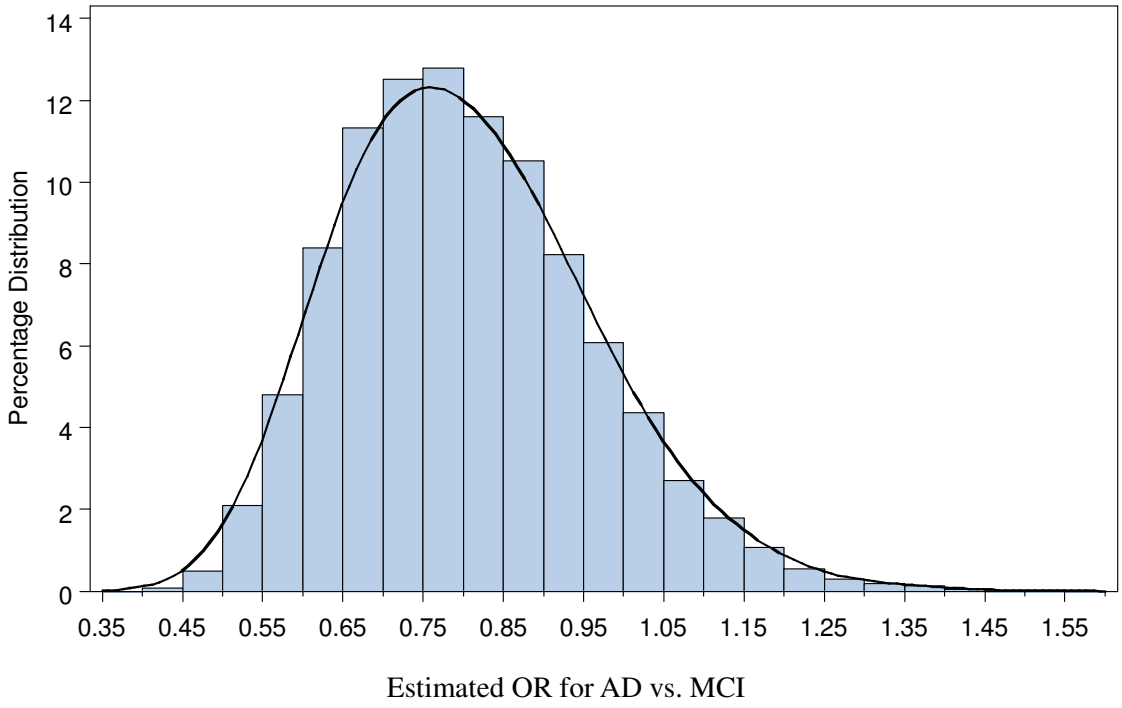

Figure 3.6 Histogram of posterior odds ratio AD vs. MCI with empirical kernel smoothing with prior of $\log(MOR_{AD})$ from triangular distribution

3.4. Simulations

The rationale of sensitivity analysis is that quite often in reality, there is no information available to make inference on the underlying missing data mechanism. Therefore it is not the purpose of sensitivity analysis to discover or correct any bias or improve efficiency, but rather to assess how much stake we put on the assumption of MAR in our statistical inference. Therefore simulation studies were conducted here to assess the ability of the proposed method to capture the underlying deviation from MAR. We generate a random binomial variable C_1 from Bin(10, 0.4), and a random normal variable C_2 from N(3,1). Then a binary variable X was generated by a logistic model from these two to induce possible correlation between the predictors. A binary outcome variable Y was generated by a logistic model with the above three predictors

$$\text{logit}[\Pr(y=1 \mid x, c_1, c_2)] = \alpha + \beta x + \gamma_1 c_1 + \gamma_2 c_2 .$$

Missingness was induced by generating an missing indicator from a logistic model

$$\text{logit}[\Pr(m=1 \mid y, x, c_1, c_2)] = \psi_0 + \psi_1 y + \psi_2 x + \psi_3 c_1 + \psi_4 c_2 + \psi_{12} yx .$$

Therefore the probability of X missing is dependent on its underlying value and the outcome and their interaction. This induces a NMAR case, and the result of complete case analysis would be biased. Sensitivity analysis was conducted by specification of alternative missing data mechanism through the missingness odds ratio. From the form of the model by which the missingness was generated, the true underlying missingness odds ratio is

$$\text{logit}(MOR_y) = \psi_2 + \psi_{12} y \tag{3.26}$$

and $\psi_2 = 0.5, \psi_{12} = -1.0$. In specification of the alternative missing data mechanism, suppose we have no information on the values of ψ_2 and ψ_{12}, but we suspect that the missing values are differentially NMAR. To examine the sensitivity of the parameter estimates to the violation of MAR, we specify a series of alternative missing data mechanisms through the missingness odds ratio by specifying paired values to (ψ_2, ψ_{12}). The result in each setting is shown in Table 3.11. We can see that the estimate for β is sensitive to the missing data mechanism. When the missing data mechanism is correctly specified, the point estimate displays minimal bias. The average Jackknife standard error is very close to the empirical standard deviation of the point estimate, and yields satisfying CI coverage. In sensitivity analysis, even though we could not correctly specify the underlying unknown missing data mechanism, we could still see how the parameter estimates change according to different deviation from MAR. When the specified missing data mechanism gets close to the true underlying one $\psi_2 = 0, \psi_{12} = -1.0$, the parameter estimate gets pretty close to the designated values.

Table 3.11 Summary of simulation study with a random binomial and a random normal covariate

Coefficients	X $\beta = 1$	C_1 $\gamma_1 = 0.5$	C_2 $\gamma_2 = -0.3$
Complete Case Analysis	1.20 (0.24) [0.24] {86.8%}	0.50 (0.09) [0.09] {95.2%}	-0.30 (0.12) [0.12] {94.4%}
With True Missing Data Mechanism $\psi_2 = 0.5, \psi_{12} = -1.0$	1.00 (0.24) [0.25] {95.6%}	0.50 (0.08) [0.08] {95.2%}	-0.30 (0.11) [0.11] {95.2%}
$\psi_2 = 0, \psi_{12} = 0.5$	1.30 (0.24) [0.24] {77.8%}	0.50 (0.08) [0.08] {95.4%}	-0.28 (0.11) [0.11] {95.4%}
$\psi_2 = 0, \psi_{12} = 0$ (MAR)	1.21 (0.24) [0.25] {86.0%}	0.50 (0.08) [0.08] {95.4%}	-0.29 (0.11) [0.11] {95.2%}
$\psi_2 = 0, \psi_{12} = -0.5$	1.10 (0.24) [0.25] {93.4%}	0.49 (0.08) [0.08] {95.2%}	-0.30 (0.11) [0.11] {95.2%}
$\psi_2 = 0, \psi_{12} = -1.0$	0.99 (0.24) [0.24] {95.4%}	0.49 (0.08) [0.08] {95.4%}	-0.31 (0.11) [0.11] {95.0%}
$\psi_2 = 0, \psi_{12} = -1.5$	0.87 (0.24) [0.24] {92.6%}	0.49 (0.08) [0.08] {95.2%}	-0.32 (0.11) [0.11] {95.6%}

Numbers in each cell reflect mean (standard deviation) based on 500 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces { } are 95 per cent confidence interval coverage rates.

3.5. Discussions

3.5.1. Connection Between the Three Ways to Specify Alternative Missing Mechanism

The proposed method provides a framework to specify alternative missing data mechanism in sensitivity analysis. The mutually deterministic relationship between common ways of specification is demonstrated, therefore different means to sensitivity analysis are unified. In section 3.2, three options are illustrated for investigators to choose from, based on what information is at available, to specify the alternative missing data mechanisms. Alternative missing data mechanisms can be specified through the missingness probability, the missingness risk ratio and the missingness odds ratio. The weights to be used can be calculated directly from the closed form as given. On the other hand, given the observable exposure rate $\pi^*_{y\bar{c}}$ and the overall missingness probability $M_{y\bar{c}}$, the missingness probabilities $(Pm_{y\bar{c}1}, Pm_{y\bar{c}0})$, the missingness risk ratio $MRR_{y\bar{c}}$ and the missingness odds ratio $MOR_{y\bar{c}}$ are mutually determined. It was pointed out by Vach and Bletter (1995), specification of the risk ratio is not appropriate when the missing probability becomes large. However, with the observation of the relationship among these three, there is equal preference in choosing any of them as long as it fits with the information at hand on the missing data mechanism.

3.5.2. Extensions

An advantage of the proposed method is that extensions to more complex scenarios are conceptually straightforward. Suppose the outcome is a categorical variable, then the models used in prediction of the weights, (3.17) and (3.18), can still be used directly. As shown in the second example, the method can be easily extended to polytomous logistic regression with nominal outcomes. For other types of outcome, one can just specify a canonical link function, and consider the generalized linear model

$$E(Y) = \mu = g^{-1}(\alpha + \beta x + \gamma' \mathbf{c})$$

and work with the corresponding likelihood to construct a weighted log-likelihood.

The proposed method can also be extended to conduct sensitivity analysis with categorical exposure. Suppose the exposure can take k distinct values, then the record $(y,\bullet,\mathbf{c})$ can be replaced by k records $(y,x_1,\mathbf{c}),\ldots,(y,x_k,\mathbf{c})$, with weights $w_{y\mathbf{c}i}$, $i = 1,\ldots,k$, where

$$w_{y\mathbf{c}i} = \frac{\dfrac{Pm_{y\mathbf{c}x_i}}{1-Pm_{y\mathbf{c}x_i}} \times \pi^*_{y\mathbf{c}i}}{\dfrac{M_{y\mathbf{c}}}{1-M_{y\mathbf{c}}}} = \frac{\dfrac{Pm_{y\mathbf{c}x_i}}{1-Pm_{y\mathbf{c}x_i}} \times \pi^*_{y\mathbf{c}i}}{\displaystyle\sum_{j=1}^{k} \frac{Pm_{y\mathbf{c}x_j}}{1-Pm_{y\mathbf{c}x_j}} \pi^*_{y\mathbf{c}j}} = \frac{MOR_{y\mathbf{c}i} \times \pi^*_{y\mathbf{c}i}}{\pi^*_{y\mathbf{c}1} + \displaystyle\sum_{j=2}^{k} MOR_{y\mathbf{c}j} \pi^*_{y\mathbf{c}j}},$$

$\pi_{y\mathbf{c}i} = \Pr(X = x_i \mid y,\mathbf{c})$ is the exposure rate of the i-th distinct value, and $MOR_{y\mathbf{c}j} = \dfrac{Pm_{y\mathbf{c}x_j} / (1-Pm_{y\mathbf{c}x_j})}{Pm_{y\mathbf{c}x_1} / (1-Pm_{y\mathbf{c}x_1})}$ is the missingness odds ratio of the j-th distinct value.

Chapter 4. JOINT MODEL FOR LOGISTIC REGRESSION WITH MISSING DATA AND REASSESSMENT DESIGN

4.1. Introduction

Common approaches for dealing with missing data are based on the missing at random (MAR) assumption. When the MAR assumption is questionable, the parameter estimates from these approaches might be badly biased. Sensitivity analysis has been discussed by Vach and Blettner (1995) and in Chapter 3 to assess the severity of the impact of NMAR via artificial specification of the missing data mechanism artificially. When only responses are missing and covariates are completely observed, the problem is easier. Chambers and Welsh (1993) discussed non-ignorable non-response for log-linear models and obtained expressions for the likelihood function based on the observed data. However, in the presence of covariates, obtaining a manageable and computable expression for the likelihood based on the observed data is highly unlikely, and thus the EM algorithm proves to be a very powerful and necessary tool for this problem. Likelihood-based methods were proposed by imposing a distributional assumption on the missing data mechanism and modeling the joint likelihood (Baker and Laird 1988; Ibrahim and Lipsitz 1999).

An alternative to the sensitivity analysis and likelihood based methods based on distributional assumptions is to obtain direct information on the missing records. It has been proposed that a second wave of data collection (reassessment) is performed on a portion of the subjects with missing values in the original data set in the study design, and that an extra effort may be made to recover information for those individuals. For example, reassessment is made by telephone or interview when the initial survey is by mail (Hansen and Hurwitz 1946); or reassessment may involve a cash incentive to convert nonparticipants to participation (Crwaford, Johnson, and Laird 1993); or when participants fail to respond to a mailed survey then information may be collected at a later examination (Glyn, Laird, and Rubin 1986).

Lyles and Allen (2003) proposed an inference scheme when reassessment data (section 1.2.10) is available in statistical analysis of a 2×2 table. Likelihood functions were derived under different scenarios of missing data mechanisms, and estimation was achieved by numerical routines in SAS IML. Unlike the case-control (or cohort) study in which exposure is assessed subsequent to determining disease status (or vice versa), in the cross-sectional design both exposure and disease status may well be subject to missing values. The missingness of exposure may depend on disease or exposure status or both, and likewise for the missingness of disease information. Further, missingness of exposure and disease may be associated. In this chapter, this method is generalized to incorporate covariates, which can be categorical or continuous or a mixture of both. A logistic regression model is considered to describe the association between the disease and the exposure, with control of the covariates. The information from the reassessment data could be incorporated into analysis to aid identification of the missing data mechanism. Estimation is achieved by joint modeling of the logistic regression model of interest and the missing data mechanism. Maximization of the joint log-likelihood function and evaluation the associated Hessian can be obtained via a built-in Quasi-Newton optimization routine in SAS IML. Simulation studies are conducted to assess the performance of the proposed method, and to provide a guideline for designing appropriate reassessment schemes in practice.

4.2. Methods

4.2.1. Outcome or Exposure Missing in Logistic Regression with Reassessment Data

We consider a reassessment mechanism which randomly selects a portion of the subjects with missing exposure X values to be enrolled in a second wave of sampling. We will assume that missing data are recovered for those subjects selected. We let R_i be the indicator for whether the missing value of a subject is reassessed or not. For a subject with X missing, let p_{Ri} be the conditional probability of being selected for reassessment. The reassessment

mechanism can be described by the following conditional probability.

$$\Pr(R_i = 1 \mid y_i, x_i, \mathbf{c}_i, m_i) = \begin{cases} 0 & \text{if } m_i = 0, \text{ i.e., the exposure is observed} \\ p_{Ri} & \text{if } m_i = 1, \text{ i.e., the exposure is missing} \end{cases} \quad (4.1)$$

The reassessment indicator is allowed to be dependent on the outcome Y and controlling covariates $\mathbf{C}$, but given these, it is assumed independent of the variable X which is subject to missing values, i.e., we assume $p_{Ri} = \Pr(R_i = 1 \mid y_i, \mathbf{c}_i)$. This assumption can be termed as "reassessment at random", analogous to the terminology "missing at random". Similarly, if the conditional probability of being reassessed is not dependent on any variable, i.e. $p_{Ri} = \Pr(R_i = 1)$, it is termed as "reassessment completely at random". We will focus on reassessment at random and consider reassessment completely at random as a special case. If the reassessment indicator is dependent on X given the outcome Y and controlling covariates $\mathbf{C}$, it becomes "reassessment not at random", and is not discussed here.

We consider a logistic model as in (2.1). We have 3 types of possible observations. Specifically, a subject with X observed contributes the term

$$\begin{aligned} p(y_i, x_i, m_i = 0, R_i = 0 \mid \mathbf{c_i}) &= \Pr(R_i = 0 \mid y_i, x_i, \mathbf{c_i}, m_i = 0) \times \Pr(m_i = 0 \mid y_i, x_i, \mathbf{c_i}) \times \\ &\quad p(y_i \mid x_i, \mathbf{c_i}) p(x_i \mid \mathbf{c_i}) \\ &= 1 \times \Pr(m_i = 0 \mid y_i, x_i, \mathbf{c_i}) \times \\ &\quad p(y_i \mid x_i, \mathbf{c_i}) p(x_i \mid \mathbf{c_i}) \end{aligned},$$

A subject with X missing originally, but reassessed and found to have $X_i = x_i$, contributes the term

$$\begin{aligned} p(y_i, x_i, m_i = 1, R_i = 1 \mid \mathbf{c}_i) &= \Pr(R_i = 1 \mid y_i, x_i, \mathbf{c}_i, m_i = 1) \times \Pr(m_i = 1 \mid y_i, x_i, \mathbf{c}_i) \times \\ &\quad p(y_i \mid x_i, \mathbf{c}_i) p(x_i \mid \mathbf{c}_i) \\ &= p_{Ri} \times \Pr(m_i = 1 \mid y_i, x_i, \mathbf{c}_i) \times \\ &\quad p(y_i \mid x_i, \mathbf{c}_i) p(x_i \mid \mathbf{c}_i) \end{aligned}$$

A subject with X missing originally, but not reassessed, contributes the term

$$
\begin{aligned}
p(y_i, m_i = 1, R_i = 0 \mid \mathbf{c}_i) &= \Pr(R_i = 0 \mid y_i, \mathbf{c}_i, m_i = 1) \times \Pr(Y_i = y_i, m_i = 1 \mid \mathbf{c}_i) \\
&= (1 - p_{Ri}) \times \sum_{x=0}^{1} \big[\Pr(m_i = 1 \mid y_i, X_i = x, \mathbf{c}_i) \times \\
&\qquad p(y_i \mid x, \mathbf{c}_i)\, p(x \mid \mathbf{c}_i) \big]
\end{aligned}
$$

The conditional probability of missingness can be modeled by a logistic model, e.g., as follows

$$
\text{logit}[\Pr(m = 1 \mid y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_2 x + \boldsymbol{\phi}_3' \mathbf{c} + \phi_{12} yx + \boldsymbol{\phi}_{23}' x\mathbf{c} + \boldsymbol{\phi}_{13}' y\mathbf{c} + \boldsymbol{\phi}_{123}' yx\mathbf{c} \tag{4.2}
$$

The conditional probability $\Pr(Y = y \mid x, \mathbf{c})$ is the main model of interest, which can be modeled as in (2.1) and as restated here:

$$
\text{logit}[\Pr(Y = 1 \mid x, \mathbf{c})] = \beta_0 + \beta_1 x + \boldsymbol{\beta}_2' \mathbf{c}\,, \tag{4.3}
$$

The conditional probability of exposure rate given other covariates is modeled by a sub-logistic model as in (2.17), e.g.,

$$
\text{logit}[\Pr(X = 1 \mid \mathbf{C} = \mathbf{c})] = \theta_0 + \boldsymbol{\theta}_1' \mathbf{c} \tag{4.4}
$$

Therefore, assuming ordering of the index i to reflect observations of the respective types, the log-likelihood is

$$
\begin{aligned}
l(\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} l(\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}; y_i, x_i, m_i, R_i \mid \mathbf{c}_i) \\
&= \sum_{i=1}^{n_{cc}} \log[\, p(y_i, x_i, m_i = 0, R_i = 0 \mid \mathbf{c}_i)] \\
&\quad + \sum_{i=n_{cc}+1}^{n_{cc}+n_R} \log[\, p(y_i, x_i, m_i = 1, R_i = 1 \mid \mathbf{c}_i)] \\
&\quad + \sum_{i=n_{cc}+n_R+1}^{n_{cc}+n_R+n_{\text{missing}}} \log[\, p(y_i, m_i = 1, R_i = 0 \mid \mathbf{c}_i)] \\
&= \sum_{i=1}^{n_{cc}} \log[\Pr(m_i = 0 \mid y_i, x_i, \mathbf{c}_i; \boldsymbol{\phi})] + \log[\, p(y_i \mid x_i, \mathbf{c}_i; \boldsymbol{\beta})] + \log[\, p(x_i \mid \mathbf{c}_i; \boldsymbol{\theta})] \\
&\quad + \sum_{i=n_{cc}+1}^{n_{cc}+n_R} \log(p_{Ri}) + \log[\Pr(m_i = 1 \mid y_i, x_i, \mathbf{c}_i; \boldsymbol{\phi})] + \log[\, p(y_i \mid x_i, \mathbf{c}_i; \boldsymbol{\beta})] + \log[\, p(x_i \mid \mathbf{c}_i; \boldsymbol{\theta})] \\
&\quad + \sum_{i=n_{cc}+n_R+1}^{n_{cc}+n_R+n_{\text{missing}}} \log(1 - p_{Ri}) + \log\left[\sum_{x=0}^{1} \Pr(m_i = 1 \mid y_i, X_i = x, \mathbf{c}_i; \boldsymbol{\phi})\, p(y_i \mid X_i = x, \mathbf{c}_i; \boldsymbol{\beta})\, p(X_i = x \mid \mathbf{c}_i; \boldsymbol{\theta}) \right]
\end{aligned} \tag{4.5}
$$

where n_{cc} is the number of subjects with exposure observed, n_R is the number of subjects with exposure originally missing but recovered by reassessment, n_{missing} is the number of subjects with exposure originally missing and not selected for reassessment, and we have

$n_{cc}+n_R+n_{\text{missing}}=n$. The terms containing p_{Ri} are factorized separately in (4.5) under the assumption of "reassessment at random". Following the analogous argument to the "ignorable" missingness, if the reassessment is at random and the parameters τ involved in modeling $p_{Ri}=\Pr(R_i=1\mid y_i,\mathbf{c}_i,\tau)$ are distinct from the set of parameters (ϕ,β,θ), then the terms $\sum_{i=n_{cc}+1}^{n_{cc}+n_R}\log(p_R)$ and $\sum_{i=n_{cc}+n_R+1}^{n}\log(1-p_R)$ can be omitted in the maximization of the log-likelihood with regard to (ϕ,β,θ). Therefore, we can bypass modeling the reassessment mechanism under the reassessment at random assumption. The log-likelihood simplifies to

$$\begin{aligned} l(\phi,\beta,\theta) &= \sum_{i=1}^{n} l(\gamma; y_i, x_i, m_i, R_i \mid \mathbf{c}_i) \\ &\propto \sum_{i=1}^{n_{cc}} \left\{ \log[\Pr(m_i=0\mid y_i,x_i,\mathbf{c}_i;\phi)] + \log[p(y_i\mid x_i,\mathbf{c}_i;\beta)] + \log[p(x_i\mid \mathbf{c}_i;\theta)] \right\} \\ &\quad + \sum_{i=n_{cc}+1}^{n_{cc}+n_R} \left\{ \log[\Pr(m_i=1\mid y_i,x_i,\mathbf{c}_i;\phi)] + \log[p(y_i\mid x_i,\mathbf{c}_i;\beta)] + \log[p(x_i\mid \mathbf{c}_i;\theta)] \right\} \\ &\quad + \sum_{i=n_{cc}+n_R+1}^{n} \left\{ \log\left[\sum_{x=0}^{1} \Pr(m_i=1\mid y_i, X_i=x, \mathbf{c}_i;\phi)\, p(y_i\mid X_i=x,\mathbf{c}_i;\beta)\, p(X_i=x\mid \mathbf{c}_i;\theta) \right] \right\} \end{aligned} \tag{4.6}$$

Similarly with outcome Y (and not X) subject to missing values originally and recovered by reassessment, the log-likelihood can be derived as

$$\begin{aligned} l(\phi,\beta,\theta) &= \sum_{i=1}^{n} l(\gamma; y_i, x_i, m_i, R_i \mid \mathbf{c}_i) \\ &\propto \sum_{i=1}^{n_{cc}} \left\{ \log[\Pr(m_i=0\mid y_i,x_i,\mathbf{c}_i;\phi)] + \log[p(y_i\mid x_i,\mathbf{c}_i;\beta)] + \log[p(x_i\mid \mathbf{c}_i;\theta)] \right\} \\ &\quad + \sum_{i=n_{cc}+1}^{n_{cc}+n_R} \left\{ \log[\Pr(m_i=1\mid y_i,x_i,\mathbf{c}_i;\phi)] + \log[p(y_i\mid x_i,\mathbf{c}_i;\beta)] + \log[p(x_i\mid \mathbf{c}_i;\theta)] \right\} \\ &\quad + \sum_{i=n_{cc}+n_R+1}^{n} \left\{ \log\left[\sum_{y=0}^{1} \Pr(m_i=1\mid Y_i=y, x_i, \mathbf{c}_i;\phi)\, p(Y_i=y\mid x_i,\mathbf{c}_i;\beta)\, p(x_i\mid \mathbf{c}_i;\theta) \right] \right\} \end{aligned} \tag{4.7}$$

where m_i is the missingness indicator for disease status, R_i is the reassessment indicator, $p_{Ri}=\Pr(R_i=1\mid x_i,\mathbf{c}_i,\tau)$ is the reassessment rate independent of Y given X and $\mathbf{C}$, n_{cc} is the number of subjects with disease status observed, n_R is the number of subjects with disease

status originally missing but recovered by reassessment, n_{missing} is the number of subjects with disease status originally missing and not selected for reassessment, and we have $n_{cc} + n_R + n_{\text{missing}} = n$.

4.2.2. Outcome and Exposure Missing in Logistic Regression with Reassessment Data

Suppose both the disease status and exposure is subject to missing values, and reassessment is conducted on both. This is a direct extension of "Case 5" considered by Lyles and Allen (2003). We allow the missing data mechanism to be dependent on the outcome, the exposure and other covariates. In addition, the missingness of exposure and disease status are also allowed to vary interactively. The reassessment subjects are assumed to be randomly chosen within subjects with missing values independent of the value of Y and X. Let m_{D_i} be the missingness indicator for disease, and m_{E_i} for exposure. Let R_{D_i} be the reassessment indicator for disease, and R_{E_i} for exposure. If all the values are observed, the complete data distribution for a subject i is given by

$$\begin{aligned} p(y_i, x_i, m_{D_i}, m_{E_i}, R_{D_i}, R_{E_i} \mid \mathbf{c}_i) &= p(R_{D_i} \mid y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}, R_{E_i})\, p(R_{E_i} \mid y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}) \\ &\quad \times p(m_{D_i} \mid y_i, x_i, \mathbf{c}_i, m_{E_i})\, p(m_{E_i} \mid y_i, x_i, \mathbf{c}_i) \\ &\quad \times p(y_i \mid x_i, \mathbf{c}_i)\, p(x_i \mid \mathbf{c}_i) \\ &= p(R_{D_i} \mid \mathbf{c}_i, m_{D_i})\, p(R_{E_i} \mid \mathbf{c}_i, m_{E_i}) \\ &\quad \times p(m_{D_i} \mid y_i, x_i, \mathbf{c}_i, m_{E_i})\, p(m_{E_i} \mid y_i, x_i, \mathbf{c}_i) \\ &\quad \times p(y_i \mid x_i, \mathbf{c}_i)\, p(x_i \mid \mathbf{c}_i) \end{aligned} \tag{4.8}$$

Or equivalently,

$$\begin{aligned} p(y_i, x_i, m_{D_i}, m_{E_i}, R_{D_i}, R_{E_i} \mid \mathbf{c}_i) &= p(R_{D_i} \mid y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}, R_{E_i})\, p(R_{E_i} \mid y_i, x_i, \mathbf{c}_i, m_{D_i}, m_{E_i}) \\ &\quad \times p(m_{E_i} \mid y_i, x_i, \mathbf{c}_i, m_{D_i})\, p(m_{D_i} \mid y_i, x_i, \mathbf{c}_i) \\ &\quad \times p(y_i \mid x_i, \mathbf{c}_i)\, p(x_i \mid \mathbf{c}_i) \\ &= p(R_{D_i} \mid \mathbf{c}_i, m_{D_i})\, p(R_{E_i} \mid \mathbf{c}_i, m_{E_i}) \\ &\quad \times p(m_{E_i} \mid y_i, x_i, \mathbf{c}_i, m_{D_i})\, p(m_{D_i} \mid y_i, x_i, \mathbf{c}_i) \\ &\quad \times p(y_i \mid x_i, \mathbf{c}_i)\, p(x_i \mid \mathbf{c}_i) \end{aligned} \tag{4.9}$$

The second equality of both equations holds because of the design of the reassessment

mechanism by which subjects are selected at random given their missingness. To be more specific, we assume that the conditional probability of being reassessed is not dependent on the underlying value of Y or X given the other variables $\mathbf{C}$. This assumption is an extension of the "reassessment at random" in section 4.3.1.

The two ways of factorization of the joint density represent different mechanisms for how the missingness of exposure and disease interact with each other. If all the covariates are categorical, it might be possible to consider a saturated model for the joint distribution of the two missingness indicators, and the two factorizations will be equivalent. In practice where there are continuous covariates, one can choose either one of them and try to approximate the saturated model by additional polynomial terms and careful model selection. We show in simulation studies that in common circumstances, the two factorizations appear to be approximately equivalent.

Adopting the first factorization of the joint density, we denote the conditional probabilities of the missingness as follows:

$$Pm^{D}_{y,x,\mathbf{c},m_E} = \Pr(m_D = 1 \mid y, x, \mathbf{c}, m_E = 1),$$
$$Pm^{D}_{y,x,\mathbf{c},\underline{m_E}} = \Pr(m_D = 1 \mid y, x, \mathbf{c}, m_E = 0), \text{ and}$$
$$Pm^{E}_{y,x,c} = \Pr(m_E = 1 \mid y, x, \mathbf{c}).$$

The reassessment mechanism can be described by the following conditional probabilities.

$$\begin{aligned} \Pr(R_{Di} = 1 \mid \mathbf{c}_i, m_{Di}) &= \begin{cases} 0 & \text{if } m_{Di} = 0, \text{ i.e., the disease is observed} \\ p^{D}_{Ri} & \text{if } m_{Di} = 1, \text{ i.e., the disease is missing} \end{cases} \\ \Pr(R_{Ei} = 1 \mid \mathbf{c}_i, m_{Ei}) &= \begin{cases} 0 & \text{if } m_{Ei} = 0, \text{ i.e., the exposure is observed} \\ p^{E}_{Ri} & \text{if } m_{Ei} = 1, \text{ i.e., the exposure is missing} \end{cases} \end{aligned} \tag{4.10}$$

With missing values and the reassessment mechanism as we define above, we can categorize the subjects based on different missingness and reassessment patterns. There are in total nine possible categories that a subject could fall within. For instance, a subject with both Y and X observed contributes the term

$$\begin{aligned} p(y_i, x_i, m_{D_i} = 0, m_{E_i} = 0, R_{D_i} = 0, R_{E_i} = 0 \mid \mathbf{c}_i) &= \Pr(R_{D_i} = 0 \mid \mathbf{c}_i, m_{D_i} = 0) \Pr(R_{E_i} = 0 \mid \mathbf{c}_i, m_{E_i} = 0) \\ &\quad \times \Pr(m_{D_i} = 0 \mid y_i, x_i, \mathbf{c}_i, m_{E_i} = 0) \Pr(m_{E_i} = 0 \mid y_i, x_i, \mathbf{c}_i) \\ &\quad \times p(y_i \mid x_i, \mathbf{c}_i) p(x_i \mid \mathbf{c}_i) \\ &= 1 \times 1 \times (1 - Pm^{D}_{y,x,c,\underline{m_E}})(1 - Pm^{E}_{y_i, x_i, \mathbf{c}_i}) \\ &\quad \times p(y_i \mid x_i, \mathbf{c}_i) p(x_i \mid \mathbf{c}_i) \end{aligned}$$

The likelihood contribution for a subject in each of the nine categories is summarized in Table 4.1. They correspond to the 25 types of observation described in Table II in Lyles and Allen (2003) if we enumerate the values of Y and X in each category. The log-likelihood can be constructed by looking up the likelihood contribution of each subject from Table 4.1.

Table 4.1 Likelihood contribution of a subject in each of the nine categories

Observation Categories	Likelihood Contribution
Y and X both observed	$(1-Pm^{D}_{y,x,c,\_\_m_E})\times(1-Pm^{E}_{y,x,\mathbf{c}})\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})$
Y observed X missing, reassessed	$p^{E}_{R}\times(1-Pm^{D}_{y,x,c,m_E})\times Pm^{E}_{y,x,\mathbf{c}}\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})$
Y observed X missing not reassessed	$(1-p^{E}_{R})\times\sum_{x=0}^{1}[(1-Pm^{D}_{y,x,c,\_\_m_E})\times Pm^{E}_{y,x,\mathbf{c}}\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})]$
Y missing, reassessed X observed	$p^{D}_{R}\times Pm^{D}_{y,x,c,\_\_m_E}\times(1-Pm^{E}_{y,x,\mathbf{c}})\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})$
Y missing not reassessed X observed	$(1-p^{D}_{R})\times\sum_{y=0}^{1}[Pm^{D}_{y,x,c,\_\_m_E}\times(1-Pm^{E}_{y,x,\mathbf{c}})\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})]$
Y, X both missing both reassessed	$p^{D}_{R}\times p^{E}_{R}\times Pm^{D}_{y,x,c,m_E}\times Pm^{E}_{y,x,\mathbf{c}}\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})$
Y, X both missing Y reassessed, X not	$p^{D}_{R}\times(1-p^{E}_{R})\times\sum_{x=0}^{1}[Pm^{D}_{y,x,c,m_E}\times Pm^{E}_{y,x,\mathbf{c}}\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})]$
Y, X both missing X reassessed, Y not	$(1-p^{D}_{R})\times p^{E}_{R}\times\sum_{y=0}^{1}[Pm^{D}_{y,x,c,m_E}\times Pm^{E}_{y,x,\mathbf{c}}\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})]$
Y, X both missing both not reassessed	$(1-p^{D}_{R})\times(1-p^{E}_{R})\times\sum_{y=0}^{1}\sum_{x=0}^{1}[Pm^{D}_{y,x,c,m_E}\times Pm^{E}_{y,x,\mathbf{c}}\times p(y\mid x,\mathbf{c})\,p(x\mid\mathbf{c})]$

The conditional probabilities $\Pr(Y=1\mid x,\mathbf{c},\beta)$ and $\Pr(X=1\mid\mathbf{c},\theta)$ are modeled as in (4.3) and (4.4). The terms containing $p_R{}^D$ and p^E_R are factorized separately in Table 4.1 under the assumption of "reassessment at random". Again as in section 4.3.1, if the reassessment is at

random and the parameters τ^D and τ^E involved in modeling $p_R{}^D = \Pr(R_D = 1 | \mathbf{c}, \tau^D)$ and $p_R{}^E = \Pr(R_E = 1 | \mathbf{c}, \tau^E)$ are distinct from the set of parameters (ϕ, β, θ), then the terms containing $p_R{}^D$ and p_R^E can be omitted in the maximization of the log-likelihood with respect to (ϕ, β, θ). Therefore, we can bypass modeling the reassessment mechanism under the above assumption. On the other hand, from a study design perspective, the models for $p_R{}^D$ and p_R^E should be chosen with careful consideration of efficiency. The two conditional probabilities that define the missing data mechanism can be modeled by a pair of logistic regression models

$$\begin{cases} \text{logit}[\Pr(m_D = 1 | y, x, \mathbf{c}, m_E)] = \phi_0^D + \phi_1^D y + \phi_2^D x + \phi_3^D{}' \mathbf{c} + \phi_4^D m_E + (\text{additional terms}) \\ \text{logit}[\Pr(m_E = 1 | y, x, \mathbf{c})] = \phi_0^E + \phi_1^E y + \phi_2^E x + \phi_3^E{}' \mathbf{c} + (\text{additional terms}) \end{cases} \tag{4.11}$$

Alternatively, the missingness model for the other representation can be defined similarly:

$$\begin{cases} \text{logit}[\Pr(m_E = 1 | y, x, \mathbf{c}, m_D)] = \phi_0^E + \phi_1^E y + \phi_2^E x + \phi_3^E{}' \mathbf{c} + \phi_4^E m_E + (\text{additional terms}) \\ \text{logit}[\Pr(m_D = 1 | y, x, \mathbf{c})] = \phi_0^D + \phi_1^D y + \phi_2^D x + \phi_3^D{}' \mathbf{c} + (\text{additional terms}) \end{cases} \tag{4.12}$$

where the "additional terms" refers to interaction terms involving some or all of the predictors and higher order terms of continuous covariate C when they are needed. The model selection regarding which terms to be included is discussed in Section 4.2.4.

4.2.3. Estimation

Statistical inference can be conducted via joint modeling the model of interest and the model of the missing data mechanism. We have found maximization of the joint log-likelihood function to be feasible using a built-in Quasi-Newton routine available in SAS IML. Regarding estimation of the variance-covariance matrix in conjunction with the log-likelihood, one could analytically derive the observed information matrix; however, a very close approximation to it can be obtained as the Hessian matrix of the maximized log-likelihood usually provided by numerical maximization procedures. Taking advantage of such available computational tools helps to enhance the accessibility of the methods for practical use.

As discussed in Section 4.2.1 and 4.2.2, the model for reassessment is omitted from

maximization of the likelihood under "reassessment at random" assumption. On the other hand, from a study design perspective, the values chosen for p_R^D and p_R^E clearly have effect on efficiency of the estimation.

4.2.4. Model Selection and Testing Not-Missing-At-Random

The model selection is important on determining which terms to include in the main model (4.3) and the missingness model (4.2), (4.11) or (4.12). With missing data in practice, the main effects model is not an adequate approximation to the missing data mechanism. The form of the missingness model can actually make dramatic impact on the inference of the main model. Furthermore, the missing data mechanism is usually not 'testable' for MCAR or MAR against NMAR, although such a hypothesis is of important interest in most statistical analysis facing missing data. Without reassessment data, one has to rely on untestable parametric assumptions of the missing data mechanism, which makes the statistical inference vulnerable to bias. Ibrahim, Lipsitz and Chen (1999) proposed a likelihood based method via EM algorithm, which provides a way to conduct sensitivity analysis on the possibility of NMAR via hypothesis testing. It can be viewed as an approximation to a true missing data mechanism that we cannot test.

With the reassessment design when missing data are expected, it becomes feasible now to conduct model selection and hypothesis testing on NMAR with the proposed method. We can use a step-up approach in constructing the suitable model by initially including only the main effects, and then additional terms can be added sequentially. We can then use the likelihood ratio or Akaike information criterion (AIC) to evaluate the fit of each model. This is shown below by means of simulation and example.

4.3. Simulations

Simulation studies were conducted to assess the performance of the proposed method. The parameter estimates and 95% confidence interval coverage rate were compared to that of the CC analysis, multiple imputation under MAR assumption. Performance of the likelihood ratio test

was assessed. Data were generated under MAR to examine the type I error of the proposed method, and then under NMAR to examine the power. A set of simulations were conducted to verify the consistence of the two factorizations (4.11) and (4.12) of the missingness model when both outcome and exposure are subject to missing values.

4.3.1. Comparison of Methods with NMAR X

When the binary exposure is subject to missing values, simulation studies were conducted to assess the point estimate and the associated standard errors. A random covariate was first generated from $N(0,1)$. Then exposure was generated from a logistic model as in (4.4). The disease status was then generated by model (4.3). Missing values were produced by a missingness indicator that follows model (4.2), where up to three way interactions involving X were considered. The interaction terms involving both Y and X are necessary to induce bias into parameter estimates, as discussed in Chapter 2. The reassessment indicator were generated by a logistic model with Y and C as the predictors. The underlying true exposure was then recovered if the indicator was TRUE.

The overall missing rate is 23.4%. For the set of subjects with missing X, the overall reassessment rate is 32.5%. 1000 simulations were performed, each with sample size 1000. The result is summarized in Table 4.2. The CC analysis produced biased estimates for the coefficients related to both X and C due to NMAR. MI based on the complete cases produced similar result to that by the CC analysis. MI based on the combined set of complete cases and reassessed cases reduced the bias and improved the efficiency upon the previous two methods thanks to the additional information recovered from the reassessed cases, but the bias in the point estimates is still noticeable. When the proposed method was applied under the MAR assumption by forcing the coefficients in (4.2) related to X to be zero, i.e. $\phi_2 = \phi_{12} = \phi_{23} = \phi_{123} = 0$, the result is very close to that from MI using the combined data set. In this case the proposed method reduced to common likelihood method under MAR. Finally, with a general model for the missing data mechanism as in (4.2), the bias of the point estimates diminished to minimal; the mean standard

error approached to the empirical standard deviation; and the coverage rate of the 95% confidence interval come close to its designated value. We also confirmed with this simulation that under the "reassessment at random" assumption, we can omit modeling the reassessment model, as in (4.6).

Table 4.2 Comparison of methods with X not missing-at-random

	Intercept 0	X 1	C -0.5
Full Information Data	-0.002 (0.103) [0.102] {95.0%}	1.010 (0.166) [0.161] {94.6%}	-0.504 (0.046) [0.045] {95.3%}
CC Analysis	-0.213 (0.121) [0.117] {54.6%}	1.204 (0.194) [0.189] {80.8%}	-0.571 (0.054) [0.054] {76.1%}
MI (CC)	-0.073 (0.114) [0.111] {89.8%}	1.204 (0.195) [0.189] {82.1%}	-0.528 (0.049) [0.048] {91.4%}
MI (CC + Reassessed Cases)	-0.051 (0.109) [0.107] {91.5%}	1.143 (0.182) [0.178] {87.3%}	-0.520 (0.048) [0.047] {93.0%}
Joint Modeling Under MAR	-0.051 (0.109) [0.106] {91.4%}	1.143 (0.182) [0.177] {86.7%}	-0.520 (0.048) [0.047] {92.8%}
Joint Modeling NMAR, w/o model R	-0.003 (0.116) [0.113] {95.4%}	1.013 (0.204) [0.194] {94.1%}	-0.505 (0.048) [0.047] {94.0%}
Joint Modeling NMAR, w/ model R	-0.003 (0.116) [0.113] {95.4%}	1.013 (0.204) [0.195] {94.1%}	-0.505 (0.048) [0.047] {94.0%}

Numbers in each cell reflect mean (standard deviation) based on 500 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

In the joint model under MAR, the missingness model was specified as

$$\text{logit}[\Pr(m=1 \mid y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_3 c + \phi_{13} yc \,.$$

In the joint model under NMAR, the missingness model was specified as

$$\text{logit}[\Pr(m=1 \mid y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_2 x + \phi_3 c + \phi_{12} yx + \phi_{23} xc + \phi_{13} yc + \phi_{123} yxc \,,$$

which is the same as the data generating model. The difference between the likelihood functions at the maximum were used to construct a likelihood ratio test at level 0.05 for MAR, i.e., $H_0 : \phi_2 = \phi_{12} = \phi_{23} = \phi_{123} = 0$. With 1000 sample size, the rejection rate is 26.0%, i.e., with the effect size of NMAR at the setup of this simulation, the power is around 26.0%. The estimates and standard errors for the parameters in the missingness model are displayed in Table 4.3. The mean estimate and standard error were used to construct a Wald test for significance of each effect. It appears that the missingness model as set up is NMAR, but not strongly deviated away from MAR. Therefore the likelihood ratio test on MAR versus NMAR yield relatively low power and the Wald test also failed on rejecting H_0 on average.

Table 4.3 Parameter estimates of the missingness model

Effect	True Value	Mean Estimate	Mean Std. Err.	Wald χ^2	P-Value
Intercept	-2.00	-2.10	0.39	29.04	<0.0001
Y	1.00	1.09	0.45	5.89	0.015
X	1.00	1.06	0.62	2.89	0.089
C	0.20	0.20	0.20	0.98	0.321
Y*X	-1.00	-1.07	0.70	2.34	0.126
X*C	-0.10	-0.10	0.25	0.15	0.695
Y*C	0.15	0.15	0.23	0.44	0.509
Y*X*C	0.05	0.05	0.29	0.03	0.855

Table 4.4 Comparison of methods with X not missing-at-random at greater magnitude

	Intercept 0	X 1	C -0.5
Full Information Data	0.001 (0.105) [0.102] {94.3%}	1.003 (0.170) [0.161] {94.1%}	-0.502 (0.046) [0.045] {94.9%}
CC Analysis	-0.083 (0.111) [0.108] {86.7%}	1.354 (0.190) [0.182] {51.6%}	-0.489 (0.050) [0.050] {94.5%}
MI (CC)	-0.117 (0.106) [0.103] {79.9%}	1.350 (0.190) [0.182] {52.6%}	-0.541 (0.048) [0.047] {87.2%}
MI (CC + Reassessed Cases)	-0.081 (0.105) [0.103] {86.2%}	1.239 (0.183) [0.175] {71.2%}	-0.528 (0.047) [0.046] {91.3%}
Joint Modeling Under MAR	-0.081 (0.105) [0.103] {86.3%}	1.238 (0.182) [0.174] {71.2%}	-0.528 (0.047) [0.046] {91.4%}
Joint Modeling NMAR, w/o model R	0.005 (0.113) [0.109] {91.3%}	0.998 (0.197) [0.187] {91.8%}	-0.500 (0.047) [0.046] {93.7%}
Joint Modeling NMAR, w/ model R	0.005 (0.113) [0.109] {91.3%}	0.998 (0.197) [0.187] {91.8%}	-0.500 (0.047) [0.046] {93.7%}

Numbers in each cell reflect mean (standard deviation) based on 500 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

In a second simulation, the values of are chosen so that the magnitude of deviation from MAR is larger. As expected, the CC analysis and other methods assuming MAR produce more bias, whilst the proposed method can dramatically reduce the bias to minimal (Table 4.4). The difference between the likelihood functions at the maximum were used to construct a likelihood

ratio test at level 0.05 for MAR, i.e., $H_0 : \phi_2 = \phi_{12} = \phi_{23} = \phi_{123} = 0$. With 1000 sample size, the rejection rate is 67.2%, i.e., with the effect size of NMAR at the setup of this simulation, the power is around 67.2%.

4.3.2. Comparison of Methods with MAR X

Simulation studies were conducted to assess the point estimate and the associated standard errors when X is MAR. The full information data without any missing values were generated as described in Section 4.3.1. Missing values were produced by a missingness indicator that follows model (4.2) with $\phi_2 = \phi_{12} = \phi_{23} = \phi_{123} = 0$. The reassessment indicator were generated by a logistic model with Y and C as the predictors. The underlying true exposure was then recovered if the indicator was TRUE. The result was summarized in Table 4.5.

The overall missing rate is 20.7%, with sample size 1000. For the set of subjects with missing X, the overall reassessment rate is 36.7%. 1000 simulations were performed. As the data were generated under MAR, CC analysis and other methods based on MAR assumption all produced minimal bias in estimate of coefficient of X, whilst CC analysis produced biased estimate for coefficient of C. This is due to the induced NMAR for C because of the correlation between X and C as discussed in Chapter 2. The empirical efficiency of MI based on complete cases is the same as that of the CC analysis, which means that when X is MAR, MI on complete cases does not help in gaining additional efficiency as also discussed in Chapter 2. MI and the proposed joint modeling method based on the combined data with complete cases and reassessed cases managed to gain efficiency thanks to the additional information. The joint modeling method lost some efficiency after the NMAR was taken into consideration.

Table 4.5 Comparison of method under MAR

	Intercept 0	*X* 1	*C* -0.5
Full Information Data	-0.002 (0.105) [0.102] {95.2%}	1.011 (0.161) [0.161] {95.1%}	-0.505 (0.045) [0.045] {95.1%}
CC Analysis	-0.219 (0.121) [0.117] {53.5%}	1.009 (0.185) [0.183] {94.4%}	-0.580 (0.054) [0.053] {69.7%}
MI (CC)	-0.001 (0.116) [0.112] {93.6%}	1.009 (0.186) [0.183] {94.4%}	-0.504 (0.047) [0.047] {94.9%}
MI (CC + Reassessed Cases)	-0.003 (0.110) [0.107] {95.1%}	1.011 (0.171) [0.174] {95.2%}	-0.505 (0.046) [0.046] {95.0%}
Joint Modeling Under MAR	-0.003 (0.110) [0.107] {94.8%}	1.011 (0.170) [0.173] {95.5%}	-0.505 (0.046) [0.046] {94.9%}
Joint Modeling NMAR, w/o model R	0.000 (0.115) [0.112] {94.7%}	1.004 (0.182) [0.188] {95.2%}	-0.503 (0.047) [0.047] {94.6%}
Joint Modeling NMAR, w/ model R	0.000 (0.115) [0.112] {94.8%}	1.004 (0.183) [0.188] {95.2%}	-0.503 (0.047) [0.047] {94.6%}

Numbers in each cell reflect mean (standard deviation) based on 500 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

The difference between the likelihood functions at the maximum were used to construct a likelihood ratio test at level 0.05 for MAR, i.e., $H_0 : \phi_2 = \phi_{12} = \phi_{23} = \phi_{123} = 0$. With 1000 sample size, the rejection rate is now 5.0%, i.e., with the MAR at the setup of this simulation, the type I error is around 5.0%, which is as designated.

4.3.3. Both Outcome and Exposure NMAR with Reassessment

We conducted simulation studies to examine the performance of the proposed method when both outcome and exposure are subject to missing values and are both reassessed as designed. In this case, it is interesting to examine whether the two ways of factorization of the joint missingness model (4.11) and (4.12) yield the same results.

In the first set of simulation, the continuous covariate C was generated from a normal distribution. The missingness indicator was generated by model (4.11) without interaction and higher order terms. The overall missing rate is 21.6% for Y, and 17.7% for X. The reassessment rate is 11.6% for Y and 7.9% for X. In the model fitting, both (4.11) and (4.12) were applied, but without interaction and higher order terms. The parameter estimates of the main model are summarized in Table 4.6.

In Table 4.6, we see that when the underlying missing data mechanism is (4.11) without interaction and higher order terms, the joint modeling via (4.11) and (4.12) yield very close results. Therefore, in this case, the investigator can choose either form to construct the likelihood to the best of their knowledge or experience.

Table 4.6 Parameter estimates of the main model when both outcome and exposure could be NMAR and reassessed

	Intercept 0	*X* 1	*C* -0.5
Full Information Data	-0.004 (0.092) [0.094] {95.7%}	1.010 (0.144) [0.144] {94.3%}	-0.503 (0.074) [0.073] {95.3%}
CC Analysis	-0.212 (0.113) [0.113] {54.2%}	1.360 (0.181) [0.178] {47.8%}	-0.512 (0.091) [0.090] {94.5%}
Multiple Imputation	-0.215 (0.109) [0.110] {51.8%}	1.359 (0.183) [0.179] {49.6%}	-0.523 (0.087) [0.086] {93.5}
Joint Modeling (4.11)	-0.015 (0.172) [0.177] {94.7%}	1.037 (0.297) [0.302] {93.2%}	-0.501 (0.088) [0.088] {95.0%}
Joint Modeling (4.12)	-0.011 (0.174) [0.178] {94.6%}	1.028 (0.295) [0.304] {93.0%}	-0.499 (0.088) [0.088] {95.0%}

Numbers in each cell reflect mean (standard deviation) based on 500 simulated data sets. Values in brackets [] are mean estimated standard errors; values in braces {} are 95 per cent confidence interval coverage rates.

4.4. Example

Higher pre-pregnancy body mass index (BMI) is associated with increased risk of neural tube defects (NTDs) and possibly other negative birth outcomes in the offspring. The mechanism for this association remains uncertain. Lower maternal folate level has been implicated in the etiology of NTDs in general. Therefore, it is of interest to investigate the association of BMI with folate level (Mojtabai 2004). This example examines the association of BMI with folate level in adult women using data from a cross-sectional survey of the U.S. population (National Health and Nutrition Examination Survey (NHANES), 1999–2008), after the 1998 U.S. folate fortification program of cereal products. Better understanding of the association of BMI and folate distribution

and metabolism has important public health implications. The study tests the hypothesis that higher BMI is associated with lower serum levels of folate after controlling for age and race/ethnicity.

Of the 51,623 participants in NHANES 1999-2008, 11,834 were non-pregnant women aged 20 and above, with the serum folate level available. Age and race/ethnicity information were complete. Subject's BMI was obtained by questionnaire and also by body examination. In this example, we consider the BMI obtained by questionnaire as the first wave of sampling, and the missing BMI values were then recovered by the record in the body examination as the reassessment measure. This results in an overall missing rate of 10.4% in the first wave of sampling (1,226 subjects), and among these with missing values, 37.7% were reassessed by body examination (462 subjects).

The association of BMI with serum folate was assessed by logistic regression model in which dichotomized serum folate was the dependent variable of interest and BMI was the independent variable of interest. The analyses controlled for the effect of age, race/ethnicity. BMI were categorized into two categories: less than 30 kg/m^2 and equal to and above 30 kg/m^2. Serum folate level were dichotomized by the 75^{th} percentile (19.7 ng/mL) of the target population.

The NHANES 1999-2008 used a stratified multistage probability sampling design to survey U.S. household civilian populations. The complex sampling design of both NHANES samples requires the use of weights and specific design elements to make the samples representative of the U.S. population and to derive correct standard errors for estimates. However, for simplicity of demonstration of the proposed method, we do not use the weights in this example.

We performed complete case analysis, multiple imputation under the MAR assumption and the proposed likelihood method. The complete case analysis and the multiple imputation method were implemented as introduced in Chapter 2. The likelihood method was implement as introduced in Section 4.3.1 with the missing data mechanism modeled in three ways. First the missing data mechanism was set as MAR by setting the regression coefficients related to X to

zero in (4.2).

$$\text{logit}[\Pr(m=1 \mid y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \boldsymbol{\phi}_3' \mathbf{c} \quad (4.2')$$

Secondly, the missing data mechanism was allowed to be dependent on BMI (i.e., NMAR) but only the interaction between serum folate and BMI was considered.

$$\text{logit}[\Pr(m=1 \mid y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_2 x + \boldsymbol{\phi}_3' \mathbf{c} + \phi_{12} yx \quad (4.2'')$$

Lastly, the missing data mechanism was allowed to be dependent on BMI (i.e., NMAR) and up to three way interactions involving BMI were allowed.

$$\text{logit}[\Pr(m=1 \mid y, x, \mathbf{c}, \boldsymbol{\phi})] = \phi_0 + \phi_1 y + \phi_2 x + \boldsymbol{\phi}_3' \mathbf{c} + \phi_{12} yx + \boldsymbol{\phi}_{23}' x\mathbf{c} + \boldsymbol{\phi}_{123}' yx\mathbf{c} \quad (4.2)$$

The values of the likelihood functions at MLE were recorded for each model, therefore a likelihood ratio test was formulated to test the hypothesis that the missing data mechanism is MAR, and how complex the missing data mechanism needs to be modeled. The parameter estimates of the main model (4.3) are summarized in Table 4.7.

The likelihood ratio test on (4.2'') versus (4.2') yields $\chi^2 = 29.26$ with degree of freedom equal to two ($p<0.0001$); whilst the likelihood ratio test on (4.2) versus (4.2') yields $\chi^2 = 134.29$ with degree of freedom equal to seven ($p<0.0001$). The likelihood ratio tests are significant, indicating that the missingness is not at random. The estimates of the regression coefficients in the missingness model in (4.2) were summarized in Table 4.8. We conclude that the missing data mechanism in this study appears to be dependent on the missing values. Consideration of NMAR is needed. However, it is also interesting to notice that the joint modeling does not yield meaningfully different estimated odds ratio regarding BMI in Table 4.6 unless a relatively comprehensive missingness model with up to three way interactions was considered. It is then interesting to check whether all the three way interactions are needed. If we fit a simpler three-way interaction model by omitting the insignificant terms in Table 4.8, we can conduct a likelihood ratio test to compare the simpler and comprehensive models, which yields a Chi-squared test statistic 48.48 with degree of freedom of three ($p<0.0001$). Therefore, we recommend keeping all these terms in the missingness model, although some are not significant

in Wald test.

Table 4.7 The Parameter Estimates by Different Methods

	BMI	Age	Race
CC Analysis	-0.404 (0.052) [0.67] {p<0.0001}	0.033 (0.001) [1.03] {p<0.0001}	0.701 (0.050) [2.02] {p<0.0001}
CC + Reassessed Cases	-0.379 (0.051) [0.68] {<0.0001}	0.033 (0.001) [1.03] {<0.0001}	0.696 (0.049) [2.01] {<0.0001}
MI (CC)	-0.409 (0.053) [0.66] {<0.0001}	0.026 (0.001) [1.03] {<0.0001}	0.639 (0.047) [1.89] {<0.0001}
MI (CC + Reassessed Cases)	-0.392 (0.050) [0.68] {<0.0001}	0.026 (0.001) [1.03] {<0.0001}	0.643 (0.047) [1.90] {<0.0001}
Joint Modeling MAR (4.2')	-0.394 (0.079) [0.67] {<0.0001}	0.025 (0.001) [1.03] {<0.0001}	0.652 (0.047) [1.92] {<0.0001}
Joint Modeling NMAR (4.2'')	-0.395 (0.052) [0.67] {<0.0001}	0.025 (0.001) [1.03] {<0.0001}	0.635 (0.047) [1.89] {<0.0001}
Joint Modeling NMAR (4.2)	-0.549 (0.053) [0.577] {<0.0001}	0.026 (0.001) [1.03] {<0.0001}	0.636 (0.047) [1.89] {<0.0001}

Dichotomized serum folate level was considered as the dependent variable in a logistic regression, with BMI, age and race/ethnicity as the independent variables. The overall missing rate is 10.4%, and reassessment rate 37.7%. The values in each cell represent the estimated logarithm odds ratio, standard error in (), odds ratio in [], and p-value in {}.

Table 4.8 Parameter Estimates of the Missingness Model

Effect	Estimate	Std. Error	Chi-Square	P-value
Intercept	-3.953	0.207	365.06	<0.0001
Serum Folate (SF)	-0.860	0.173	24.70	<0.0001
BMI	-0.850	0.388	4.785	0.0287
Age	0.035	0.004	82.16	<0.0001
Race	-1.000	0.143	48.81	<0.0001
BMI*SF	1.105	0.853	1.68	0.195
BMI*Age	0.017	0.007	5.90	0.0152
BMI*Race	-0.920	0.504	3.333	0.0679
BMI*SF*Age	-0.008	0.013	0.357	0.550
BMI*SF*Race	-1.624	0.333	23.80	<0.0001
BMI*Age*Race	0.035	0.007	23.95	<0.0001

As an illustration, we are interested to see whether the missing data could make more impact if the overall missing rate is larger. We artificially induce additional missing values based on the missing data mechanism estimated from the joint modeling (Table 4.8). To achieve a higher overall missing rate, we adjusted the intercept term, but keep the other estimated regression coefficients as in Table 4.8. Therefore the overall missing mechanism is only inflated by a constant but the association structure with the related effects remains unchanged. The reassessment rate is targeted at 37.7% on the induced missing subjects. In the data set with artificially induced missing values, the missing rate is 25.1%, and the reassessment rate is 37.3%. We repeated the analyses and the results are summarized in Table 4.9. The results by CC analysis and MI based on complete cases now deviated away noticeably from that by the analysis on the combined data with complete and reassessed cases. The result from later also deviated noticeably from that by the joint modeling with consideration of NMAR, even if the simplest (4.2") model was considered. Therefore, the impact of NMAR on the parameter estimate emerges when the

overall missing rate is large. The result from using (4.2) remains close to that in the original data set, which means that we were able to replicate the correct missing data mechanism in the original data and also to make correction upon it with the proposed method.

Table 4.9 Parameter Estimates by Different Methods

	BMI	Age	Race
CC Analysis	-0.176 (0.056) [0.84] {0.0016}	0.040 (0.001) [1.04] {<0.0001}	0.655 (0.054) [1.93] {<0.0001}
CC + Reassessed Cases	-0.269 (0.052) [0.76] {<0.0001}	0.036 (0.001) [1.04] {<0.0001}	0.679 (0.051) [1.97] {<0.0001}
MI (CC)	-0.204 (0.054) [0.82] {<0.0001}	0.025 (0.001) [1.03] {<0.0001}	0.657 (0.048) [1.93] {<0.0001}
MI (CC + Reassessed Cases)	-0.289 (0.054) [0.75] {<0.0001}	0.025 (0.001) [1.03] {<0.0001}	0.649 (0.047) [1.91] {<0.0001}
Joint Modeling MAR (4.2')	-0.300 (0.052) [0.74] {<0.0001}	0.026 (0.001) [1.03] {<0.0001}	0.653 (0.047) [1.92] {<0.0001}
Joint Modeling NMAR (4.2")	-0.369 (0.056) [0.69] {<0.0001}	0.026 (0.001) [1.03] {<0.0001}	0.659 (0.048) [1.93] {<0.0001}
Joint Modeling NMAR (4.2)	-0.498 (0.055) [0.61] {<0.0001}	0.026 (0.001) [1.03] {<0.0001}	0.643 (0.047) [1.90] {<0.0001}

The values in each cell represent the estimated logarithm odds ratio, standard error in (), odds ratio in [], and p-value in {}.

The likelihood ratio test on (4.2'') versus (4.2') yields $\chi^2 = 44.69$ with degree of freedom equal to two (p<0.0001); whilst the likelihood ratio test on (4.2) versus (4.2') yields $\chi^2 = 236.14$ with degree of freedom equal to seven (p<0.0001). We conclude that the missing data mechanism in this study appears to be dependent on the missing values. Consideration of NMAR is needed. The likelihood ratio test results in a higher χ^2, thus we are able to pick up more evidence in rejecting that the missing data in this example is MAR.

4.5. Discussion

Statistical inference is highly dependent on the assumption of the missing data mechanism. Common approaches are based on MAR blindly, or with sensitivity analysis, but none could give definitive conclusions, nor valid hypothesis testing on the validity of the assumption. With reassessment data, it is feasible to perform such hypothesis tests. The proposed method makes it possible to test NMAR via a likelihood ratio test based on the data. In the simulation studies in Section 4.3.1 and 4.3.2, we see that the likelihood ratio test for NMAR provides valid empirical type I error at the designated level, and the power of the test increases as the magnitude of NMAR gets larger, with the other factors fixed. The power of the proposed test is dependent of the magnitude of NMAR, sample size, missing rate and reassessment rate. However, it is hard to quantify the magnitude of deviation from MAR, therefore we cannot draw a quantitative conclusion on the power of this test, nor can we give a closed form for power/sample size calculation. On the other hand, it is not clear how to characterize the set of all estimable parameters for this class of models given a certain choice of covariates. This issue of estimatility arises often in non-ignorable response models as pointed out by Baker and Laird (1988). The likelihood ratio test proposed here serves as a valid tool for model selection.

The reassessment scheme in study design proposed here is important. As pointed out by Glyn, Laird, and Rubin 1993, none of the approaches mentioned in Section 4.1 can be relied on to obtain information from a completely random sample of nonparticipants, and nonresponse to

reassessment is often high. Nevertheless, the proposed method allows certain dependence of the non-response to reassessment on observable variables. As long as the nonresponse is independent on the underlying value after conditioning on the other variables, namely "Reassessment at Random", incorporating these follow-up data into estimates will lead to reduced bias.

As it has been pointed out by Lyles and Allen (2003), the type of reassessment that we discuss is focused on studies in which missing data occur naturally, compared to the two-stage designs proposed in other previous literature (Breslow and Cain 1988; Flanders and Greenland 1991; Zhao and Lipsitz 1992). Therefore, it is only the distribution of reassessment given missingness that is under the control of the investigator.

Chapter 5. SUMMARY AND FUTURE RESEARCH

5.1. Summary

This dissertation explores methods to deal with missing data in statistical analysis of logistic regression. The disease status (outcome) and risk exposure (predictor of interest, a binary indicator) could be subject to missing data separately or together. The research question of interest is to identify the association between the disease status and the risk exposure, with best available consideration of the potential impact of the missing data on the estimation.

The first research topic was focused on providing a better accessible approach to investigators if the assumption of missing at random was imposed. This assumption is extensively used in practice, and many methods have been proposed in this case. However, due to lack of available statistical software, many are not easily accessible to investigators. We explore the impact of missing data in disease status and/or risk exposure. We found that the subjects with disease status missing contribute no information to the association of interest. Omitting these subjects from analysis induces no bias or loss of efficiency. When the risk exposure is missing for some subjects, we proposed a weighting method which constructs a weighted log-likelihood using the conditional distribution of the subject-to-missing-data variable given the observed data. Two estimation approaches were proposed, each of which has its advantage and drawback. The first approach makes use of a "flipped-around" logistic model with risk exposure as the outcome and disease status as the predictor of interest, controlled for other variables. This approach is easy to implement with standard statistical software, and shares common properties with multiple imputation. However, due to the conflict of the "flipped-around" model with the original model when the controlling covariates are "sufficiently continuous", this approach, as well as multiple imputation, can induce bias in parameter estimates. The second approach rewrites the weight by Bayes rule into a function, the variables of which are parameters from the original model and a sub-logistic model with risk exposure as the outcome and controlling covariates as the predictors. The expectation-maximization (EM) algorithm was used for parameter estimation. The parameters in the sub-logistic model were estimated prior to the EM algorithm therefore the

M-step was simplified. The conditional expectation in the E-step was written as a weighted log-likelihood with closed form, thus the E-step was also simplified. This second approach avoided the problem of the "flipped-around" model, therefore is recommended when the first approach and multiple imputation are not appropriate. However, it requires investigators to code an iterative program.

The first topic is based on the assumption of missing at random, which is subjectively imposed based on previous knowledge and experience and often not testable. If the assumption fails apart, the above methods produce biased parameter estimates. It is then important to assess how sensitive the results are to the violation of the assumption of missing at random. In the second topic, we proposed a framework of sensitivity analysis for such purpose. Alternative missing data mechanisms are specified, and the result from each specified scenario is compared to that from MAR to assess the bias of parameter estimates induced at each level of deviation from MAR. The specification of alternative missing data mechanism can be made through three ways, namely conditional probabilities of being missing, the missingness risk ratio and the missingness odds ratio. These three terms are not mutually deterministic in general, but we found a relationship by making use of the overall missing rate as a tie among them. The overall missing rate can be consistently estimated with the at-hand data set without dealing with missing data problem. Therefore it provides a useful means to guiding the specification of alternative missing data mechanism and avoids unrealistic specifications. Simulation results show that the proposed method succeeds in detecting the direction and magnitude of bias in parameter estimates even if the specification of the alternative missing data mechanism is not completely correct.

The first two topics are based on a common study design where when missing data occur, no attempts are made to collect additional information and statistical methods are aimed to make best use of available data. The assumption of missing at random becomes the key but is vulnerable sometimes and untestable in general. In the third research topic, we explore the reassessment design, where a second wave of sampling is made in attempt to recover a small portion of the missing data in the original wave. We construct a joint model of the original model of interest and

the model of missing data mechanism, where the second one allows for non-ignorable missingness. The estimation is carried out by numerical maximization of the joint likelihood and the standard errors are estimated via a close approximation by the Hessian matrix. We demonstrate that when the reassessment is at random, the model of reassessment can be omitting from the likelihood without harming the estimation. We recommend likelihood ratio test be used for model selection. By this means, it can be used to facilitate hypothesis testing on the assumption of missing at random, which is of great concern in many practical applications.

5.2. Future Research

The first possible extension is to generalize the sensitivity analysis to allow for multiple variables subject to missing data. In practice, it is likely that multiple variables are missing and the occurrences of missing data between variables are mutually dependent. The specification of alternative missing data mechanisms can be generalized for such need.

The reassessment study design proposed here assumes perfect response in the reassessment sample. However, it may not be realistic in practice. The proposed method could be extended to accommodate certain missingness in the reassessment sample. The current study design method also assumes that there is no misclassification in the first wave of sampling. Considering that the first wave of sample is usually conducted in a much larger scale with less expensive sampling instruments, data from the first wave are likely to be exposed to mis-classification/mis-measurement. In these studies, reassessment data might be obtained for the non-missing subjects as well. For example, in the NHANES data in Chapter 4, the same subject with BMI obtained by questionnaire might also be invited for physical body examination. In this case, the reassessment data also serves as a golden standard for the target measure and can be used for correction of mis-classification/mis-measurement. The proposed method can be extended to accommodate such needs.

All the topics focus on binary risk exposure, but the methodology proposed here is not limited to so. It is interesting to generalize the closed form of weights proposed here to

accommodate categorical risk exposure, but application on continuous risk exposure might be difficult because of the complexity in integration and may require numerical integration algorithms.

BIBLIOGRAPHY

Abraham, W. T., & Russell, D. W. (2004). Missing data : a review of current methods and applications in epidemiological research. *Current Opinion in Psychiatry*, 315-321. doi:10.1097/01.yco.0000133836.34543.7e

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological methods and Research*, *28*(3), 301–309. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.2404&rep=rep1&type=pdf

Baker, S G. (1996). The analysis of categorical case-control data subject to nonignorable nonresponse. *Biometrics*, *52*(1), 362-9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8934603

Baker, Stuart G, Ko, C.-W., & Graubard, B. I. (2003). A sensitivity analysis for nonrandomly missing categorical data arising from a national health disability survey. *Biostatistics (Oxford, England)*, *4*(1), 41-56. doi:10.1093/biostatistics/4.1.41

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics*, *43*(4), 951-73. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3427178

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, *91*(433), 14-28. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12155399

Breslow, N. E., & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11-20. doi:10.1093/biomet/75.1.11

Breslow, N. E., & Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics*, *34*(1), 100–105. JSTOR. Retrieved from http://www.jstor.org/stable/2529594

Cao, W., Tsiatis, A. a, & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, *96*(3), 723-734. doi:10.1093/biomet/asp033

Carpenter, J. R., Kenward, M. G., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 571-584. doi:10.1111/j.1467-985X.2006.00407.x

Carroll, R. J., Wang, S., & Wang, C. Y. (1995). Prospective Analysis of Logistic Case-Control Studies. *Journal of the American Statistical Association*, *90*(429), 157. doi:10.2307/2291139

Chen, M.-H., Ibrahim, J. G., & Shao, Q.-M. (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*(466), 421-438. doi:10.1198/016214504000000368

Chen, Q., & Ibrahim, J. G. (2006). Semiparametric models for missing covariate and response data in regression models. *Biometrics*, (March), 177-184. doi:10.1111/j.1541-0420.2005.00438.x

Chen, Q., Ibrahim, J. G., Chen, M.-H., & Senchaudhuri, P. (2008). Theory and inference for regression models with missing responses and covariates. *Journal of multivariate analysis*, *99*(6), 1302–1331. Elsevier. doi:10.1016/j.jmva.2007.08.009

Consentino, F., & Claeskens, G. (2008). Missing covariates in logistic regression, estimation and distribution selection. *Open Access publications from*. Retrieved from http://www.econ.kuleuven.ac.be/public/ndbaf45/papers/ConsentinoClaeskensP2.pdf

Dempster, A. P., Laird, N. M., Rubin, D. B., & others. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. Royal Statistical Society. Retrieved from http://www.ams.org/leavingmsn?url=http://links.jstor.org/sici?sici=0035-9246(1977)39:1<1:MLFIDV>2.0.CO;2-Z&origin=MSN

D'Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*(19), 2265–2281. John Wiley & Sons. Retrieved from http://www3.interscience.wiley.com/journal/10008041/abstract

D'Agostino, R. B., & Rubin, D. B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association*, *95*(451), 749. doi:10.2307/2669455

Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. *Journal of the American Statistical Association*, *89*(426), 463- 475. American Statistical Association. Retrieved from http://www.questia.com/PM.qst?a=o&se=gglsc&d=5002214766

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Retrieved from http://books.google.com/books?hl=en&lr=&id=gLlpIUxRntoC&oi=fnd&am

p;pg=PR14&dq=An+introduction+to+the+bootstrap&ots=A6CzZ6QaA4&sig=MYl2VN9h43Lw7oHiyzBOrFkkgMk

Elashoff, M., & Ryan, L. (2004). An EM Algorithm for Estimating Equations. *Journal of Computational and Graphical Statistics*, *13*(1), 48-65. doi:10.1198/1061860043092

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*(5), 713. doi:10.1177/0013164401615001

Flanders, W. D., & Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in medicine*, *10*(5), 739-47. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2068427

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups. *Journal of the American Statistical Association*, *88*(423), 984. doi:10.2307/2290790

Gong, G., & Samaniego, F. (1981). Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*. Retrieved from http://www.jstor.org/stable/2240854

Greenland, S. (2001). Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk analysis : an official publication of the Society for Risk Analysis*, *21*(4), 579-83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11726013

Greenland, Sander. (1996). Basic methods for sensitivity analysis of biases. *International journal of epidemiology*, *25*(6), 1107-16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9027513

Greenland, Sander, & Finkle, W. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, *142*(12), 1255. Oxford Univ Press. Retrieved from http://aje.oxfordjournals.org/cgi/content/abstract/142/12/1255

Horton, N., & Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, *2802*(98), 37-50. Retrieved from http://smm.sagepub.com/cgi/content/abstract/8/1/37

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*(411), 765. doi:10.2307/2290013

Ibrahim, J. G., & Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, *52*(3), 1071–1078. JSTOR. Retrieved from http://www.jstor.org/stable/2533068

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, *55*(2), 591–596. John Wiley & Sons. Retrieved from http://www3.interscience.wiley.com/journal/119061984/abstract

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & A.H., H. (2005). Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*. Retrieved from http://d.wanfangdata.com.cn/NSTLQK_NSTL_QK8945168.aspx

Ibrahim, J. G., Lipsitz, S. R., & Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(1), 173–190. John Wiley & Sons. Retrieved from http://www3.interscience.wiley.com/journal/119099481/abstract

Institute, S. (1999). *SAS/IML User's Guide, Version 8. Cary, NC: SAS Institute Inc*. SAS Institute. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:SAS/IML+User's+Guide#0

Lipsitz, S. R., & Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, *83*(4), 916. Biometrika Trust. Retrieved from http://biomet.oxfordjournals.org/cgi/content/abstract/83/4/916

Lipsitz, S. R., Ibrahim, J. G., & Zhao, L. P. (1999). A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood. *Journal of the American Statistical Association*, *94*(448), 1147- 1160. American Statistical Association. Retrieved from http://www.questia.com/PM.qst?a=o&se=gglsc&d=5002342393

Lipsitz, S. R., Ibrahim, J. G., Chen, M.-H., & Peterson, H. (1999). Non-ignorable missing covariates in generalized linear models. *Statistics in medicine*, *18*(17-18), 2435–2448. John Wiley & Sons. Retrieved from http://www3.interscience.wiley.com/journal/63500927/abstract

Little, R.J.A., & Rubin, D. B. (1987). Statistical analysis with missing data. Wiley New York: Retrieved from http://www.gbv.de/dms/ilmenau/toc/33682193X.PDF

Little, R.J.A., & Rubin, D. B. (2002). Statistical analysis with missing data, Second Edition. Wiley New York: Retrieved from http://www.gbv.de/dms/ilmenau/toc/33682193X.PDF

Little, Roderick J. A. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, *87*(420), 1227- 1237. Retrieved from http://www.jstor.org/stable/2290664

Little, Roderick J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician*, *37*(3), 218–220. JSTOR. Retrieved from http://www.jstor.org/stable/2683374

Liu, C. (2006). Robit Regression : A Simple Robust Alternative to Logistic and Probit Regression, 1-23.

Lyles, R.H., & Lin, J. (2010). Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in medicine*, *29*(22), 2297–2309. Wiley. doi:10.1002/sim.3971

Lyles, Robert H, & Allen, A. S. (2002). Estimating crude or common odds ratios in case-control studies with informatively missing exposure data. *American journal of epidemiology*, *155*(3), 274-81. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11821253

Lyles, Robert H, & Allen, A. S. (2003). Missing data in the 2× 2 table: patterns and likelihood-based analysis for cross-sectional studies with supplemental sampling. *Statistics in Medicine*, *534*(May 2002), 517-534. doi:10.1002/sim.1348

Miller, R. G. (1974, April). The Jackknife--A Review. *Biometrika*. doi:10.2307/2334280

Mojtabai, R. (2004). Body mass index and serum folate in childbearing age women. *European journal of epidemiology*, *19*(11), 1029-36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15648596

Nordheim, E. V. (1984). Inference from nonrandomly missing categorical data: an example from a genetic study on Turner's syndrome. *Journal of the American Statistical Association*, *79*(388), 772–780. JSTOR. Retrieved from http://www.jstor.org/stable/2288707

Prentice, R., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*(3), 403-411. Retrieved from http://biomet.oxfordjournals.org/cgi/content/abstract/66/3/403

Proschan, M. A., McMahon, R. P., Shih, J. H., Hunsberger, S. A., Geller, N. L., Knatterud, G., & Wittes, J. (2001). Sensitivity analysis using an imputation method for missing binary data in clinical trials. *Journal of statistical planning and inference*, *96*(1), 155–165. Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0378375800003323

Robins, J. M., & Rotnitzky, A. (1995, March). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*. doi:10.2307/2291135

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical*, *89*(427), 846- 866. Retrieved from http://www.questia.com/PM.qst?a=o&se=gglsc&d=5002215339

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, *90*(429), 106. doi:10.2307/2291134

Rosenbaum, P. R. (1984). Conditional Permutation Tests and the Propensity Score in Observational Studies. *Journal of the American Statistical Association*, *79*(387), 565. doi:10.2307/2288402

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. Retrieved from http://www.jstor.org/stable/2288398

Rosner, B. (1995). *Fundamentals of biostatistics* (4th ed.). Belmont CA: Wadsworth Publishing Company. Retrieved from http://books.google.com/books?hl=en&lr=&id=9FXZZRBtVeUC&oi=fnd&pg=PR15&dq=Fundamentals+of+biostatistics&ots=NLp0846Sv2&sig=TDi23ojyTmNtdMHJNT45aJm198w

Rotnitzky, A., & Robins, J. M. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in medicine*, *16*(1), 81–102. John Wiley & Sons. Retrieved from http://www3.interscience.wiley.com/journal/9474/abstract

Rotnitzky, A., & Wypij, D. (1994). A Note on the Bias of Estimators with Missing Data. *Biometrics*, *50*(4), 1163. doi:10.2307/2533454

Royston, P. (2005). Multiple imputation of missing values: update. *Stata Journal*, *5*(2), 188. STATA PRESS. Retrieved from http://oregonstate.edu/~acock/growth-curves/royston_SJ_paper_nearly_final.pdf

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. doi:10.1093/biomet/63.3.581

Rubin, D. B. (1996, June). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*. doi:10.2307/2291635

Steenland, K., & Greenland, S. (2004). Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American journal of epidemiology*, *160*(4), 384-92. doi:10.1093/aje/kwh211

Steenland, K., Macneil, J., Bartell, S., & Lah, J. (2010). Analyses of diagnostic patterns at 30 Alzheimer's disease centers in the US. *Neuroepidemiology*, *35*(1), 19-27. doi:10.1159/000302844

Vach, W., & Blettner, M. (1995). Logistic regression with incompletely observed categorical covariates-investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*, *14*(12), 1315-29. Retrieved from http://www3.interscience.wiley.com/journal/114131514/abstract

Wang, C. Y., Wang, S., Zhao, L. P., & Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, *92*(438), 512–525. JSTOR. Retrieved from http://www.jstor.org/stable/2965700

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, *14*(4), 1261-1295. doi:10.1214/aos/1176350142

Zhao, L. P., & Lipsitz, S. R. (1992). Designs and analysis of two-stage studies. *Statistics in medicine*, *11*(6), 769-82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1594816

Zhao, L. P., Lipsitz, S. R., & Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics*, *52*(4), 1165–1182. JSTOR. Retrieved from http://www.jstor.org/stable/2532833