MODEL ASSESSMENT FOR MODELS WITH MISSING DATA

Xiaolei Zhou

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Hongtu Zhu

Elizabeth Andrews

Shrikant Bangdiwala

Yun Li

Wen Sun

ABSTRACT

Xiaolei Zhou: Model Assessment for Models with Missing Data
(Under the direction of Hongtu Zhu)

Missing data commonly occur in various study setting. In this dissertation, we first investigate three likelihood-based models for missing data in longitudinal studies: mixed effects models, pattern mixture models (PMM), and selection models. Extensive simulations from ten missing mechanisms are performed with the focus on treatment effect. Results suggest that no model consistently performs better than others under various missing data mechanism. However, PMM using the treatment-specific proportion and selection model provide some correction of the estimate compared with mixed-effects model in several missing not at random situations, even when the mechanism of missing data is not exactly the same as the model assumption.

Secondly, we focus on the case deletion diagnostic measures for general linear models (GLMs) with missing covariate data. Cook's distance is one of the most important diagnostic tools to identify influential observations on the parametric models. However, Cook's distance may not be directly comparable because its scale stochastically depends on the degree of the perturbation. We define the degree of perturbation for GLM with missing covariates. Then, we derive the Cook's distance based on likelihood function and compare it to the Cook's distance based on the Q-function used in the EM algorithm for models with missing data. We further develop the scaled Cook's distance in the GLM with missing covariate data, which resolves the size issue of Cook's distance. Simulation data are used to illustrate the size

matters issue in GLM with missing covariates. The applications of scaled Cook's distances in a formal influence analysis are examined in simulations and real data examples.

At last, we examine the connection between case deletion measures and cross validation method for GLM with missing covariates models. Based on such connection, we develop case-deletion model complexity (CMC) measures for quantifying the model complexity and case-deletion information criteria (CIC) for model selection. We develop these new measures and criteria based on the likelihood function and the Q-function, respectively. Some properties of CMC and CIC are investigated. Simulations and real data analysis show that CIC is a valuable tool for analysis of models with missing data.

ACKNOWLEDGMENTS

Foremost, I would like to thank my advisor, Dr. Hongtu Zhu, for his guidance, wisdom, and patience throughout the development of this dissertation. In addition, this dissertation would not have been possible without the encouragement from my colleagues at RTI Health Solutions and the support from my husband, Yong Yang.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the "big data" era, a large amount of data is available and waiting for people to find out what is hidden inside. These data may come from a "true" complicated process. Statisticians use statistical models to interpret data and to approximate the true complicate process. However, the fitted model is nearly always not the true process. How to use statistical tools (diagnostic measures) to detect the discrepancies between fitted model and true process is a very important question for statisticians. We can distinguish two types of discrepancies: i) discrepancy existing between isolated observations (influential points and outliers) and the rest of the observations, and ii) systematic discrepancies between the data and the fitted value obtained from statistical models. The existence of missing data further increases the complexity of model fitting and diagnosis. Although several methods have been developed to handle missing data, it remains a challenging and active field for statisticians.

In this dissertation, we first present literature reviews. Then in Chapter 2, we compare mixed-effects model, pattern-mixture model, and selection model in estimating treatment effect for missing data in longitudinal studies. In Chapter 3, we develop the scaled Cook's distance for generalized linear models with missing covariates. In Chapter 4, we develop the case-deletion information criterion for model selection on generalized linear models with missing covariates.

1.1 Missing Data and Treatment Effect

Missing data commonly occur in longitudinal studies. In clinical trial studies, patient-reported outcomes (PROs), such as health-related quality of life (HRQOL) and symptoms collected via validated questionnaires, often have a higher missing rate compared with clinical outcomes evaluated by physicians. In some severe diseases, such as metastatic breast cancer, it is not uncommon for 15% of patients to be missing HRQOL data even at baseline (Zhou et al., 2009). Additional missingness occurring after baseline further reduces the proportion of patients with data available for analysis. The reasons for and amount of missing HRQOL data in clinical trials depend on the disease and when and how the study is conducted and may not be similar among different treatment groups. Fairclough (2010) listed various reasons why subjects fail to complete HRQOL assessments. Some missing data may be caused by administrative reasons, such as staff forgetting to administer the questionnaire or translation not available in the patient's language. The missing value may also be related to the patient's condition; for example, the patient stated that he or she was too ill to complete the questionnaire. The high missing rate in HRQOL data can also be caused by a self-assessment questionnaire that contains a long series of questions. Once a patient has missed an HRQOL assessment, the retrospective collection of these data is usually impossible.

It is well known that missing data may not only reduce the power to detect change from baseline but, more important, will lead to biased estimates of response when missingness depends on the response. At the end of 2009, the Food and Drug Administration (FDA, 2009) published a guidance for industry in using PROs in clinical trials for label claims. The guidance encourages the study to minimize patients' dropouts and collect PRO data even after patients have discontinued treatment. The study's protocol and statistical analysis plan should describe how missing data will be handled in the analysis. However, the FDA does not consider any single method as preferred regarding statisti-

cal strategies to deal with missing data due to early termination of patients before the planned completion of a trial. European Medicines Agency (2010) guideline on missing data in confirmatory clinical trials concur with this.

In 2010, the Panel on the Handling of Missing Data in Clinical Trials under the National Research Council (2010) published a report with recommendations that will be used not only to the FDA but also to the entire clinical trial community (Little et al., 2012, O'Neill and Temple, 2012). The panel classified four types of approaches to adjust for missing data: complete-case analysis (excluding subjects with missing data from analysis), single imputation methods (such as last observation carried forward or baseline value carried forward, was used in some clinical trials), estimating-equation methods, and methods based on a statistical model. In estimating-equation methods, complete cases are weighted by the inverse of an estimate of the probability of being observed, which may be modeled with the use of observed variables, for example, baseline data. The statistical model based methods includes likelihood function-based models, Bayesian methods, and multiple imputation.

For the missing data problem, the applicability of the different methods is based on a classification of the following missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). If the probability of an observation being missing does not depend on observed or unobserved measurements, then the observation is MCAR. If the probability of an observation being missing depends only on observed measurements, then the observation is MAR. If the probability of an observation being missing depends on unobserved measurements (e.g., patients with a poor outcome score are more likely to miss the assessment), then the observation is MNAR. This type of missing data is also called nonignorable missing data.

Complete-case analysis is based on MCAR. Single imputation methods is arbitrary. Weighted estimating equations and multiple imputation are computationally available,

but they are built on ignorable missingness or MAR (Little et al., 2012, Ali and Siddiqui, 2000). The mixed-effects model, as a likelihood function-based model, is a frequent choice for analyzing continuous outcomes in clinical trials because it uses all available observed data and are valid when missing data are MCAR or MAR. However, when missing data depend on an unobserved outcome (MNAR), the parameter estimates from the mixed-effects model can be biased.

In clinical trial studies, we cannot rule out the MNAR scenario and sometimes it may be more realistic than MAR; thus, it is recommended to assess the robustness of the results by performing sensitivity analysis under the assumption of MNAR. Development of statistical methods to handle MNAR data is a very important and promising area. Ibrahim and colleagues (2005) and Ibrahim and Molenberghs (2009) provided several overview articles of various models for missing data problems. Pattern-mixture model (PMM) and selection model are two major methods to handle missing data under MNAR based on likelihood functions. To account for nonignorable missing data, in addition to random variables in the mixed-effects model, PMM and selection model include an additional random variable for missingness in the likelihood functions (missing pattern in PMM and missingness indicator in selection model). In both models, the random variable for missing mechanism is not independent with the response variable. Pattern-mixture models (Little, 1995) have been used widely to analyze continuous PRO data. A popular PPM for PRO data analysis assumes that the missing mechanism depends only on treatment (Hedeker and Gibbons, 1997). Pauler and colleagues (2003) further provided details on how to estimate overall treatment effect from pattern-specific estimates based on the treatment-specific proportion of the pattern when HRQOL changes linearly over time. The missing mechanism may also depend on other covariates or response. The selection model is based on the assumption of MNAR that whether or not the dependent variable is missing depends directly on the value of the dependent variable at the time of missing. The parameters in selection models can be estimated using the Monte Carlo expectation-

maximization (MCEM) algorithm (Ibrahim et al., 2001, 2005, Ibrahim and Molenberghs, 2009). Both PMM and selection model have been applied using the Bayesian approach (Daniels and Hogan, 2008, Little et al., 2011). Note that the pattern-mixture model and selection model factorizations of the likelihood functions can be used to develop more complex methods of joint modeling of responses and missing data process such as the shared-parameter models (Ibrahim and Molenberghs, 2009, Daniels and Hogan, 2008), where the missingness may depend on the random effect.

All models for handling missing data under MNAR make specific assumptions, which are often untestable, and the statistical results obtained from different MNAR models can be different. Therefore, the models under MNAR are often considered as part of a sensitivity analysis (Molenberghs et al., 2001, 2004, Verbeke et al., 2001b). Although it is well known that the estimates of response obtained from mixed-effects models fitted to MNAR missing data can be biased, the impact of missing data on the estimate of treatment effect (the difference between treatments) is more complex, depending on the proportion of and reason for missing data in all treatment groups. Several studies (Pauler et al., 2003, Michiels et al., 2002, Post et al., 2010) have included treatment effects estimated from mixed-effects model and PMM or selection model using collected data. However, because the true treatment effect is unknown in collected data, it is impossible to evaluate the bias of the estimate using collected data. A simulation study is needed to evaluate the impact of missing data on the estimate of treatment effect.

1.2 Model Assessment

The goal of the diagnostic measures is to assess how well the model fits the data and how robust it is. Residuals (the difference between the observed response and the model-predicted response) provide very important information about the model fitting, not only for the individual observations, but also for the global model fitting. Instead of comparing observed value to model-predicted value, another approach to assess model fitting is the

influence measure. If a minor modification of the model seriously influences key results of an analysis, it will be a cause for concern. On the other hand, if such modifications do not have large impact the results, the model is robust with respect to the induced perturbations (Cook, 1986).

1.2.1 Case Influence Measures

Case deletion measures assess the influence of deleting one or a set of observations from the data on certain statistics. They are often used to identify influential points or the impact of one or a few observations on overall model fitting. Two widely used case deletion measures are Cook's distance (Cook, 1977) and likelihood displacement (Cook and Weisberg, 1982, Cook, 1986). The likelihood displacement measures the difference in log-likelihood when one or a set of observations are removed. The likelihood displacement is defined by

$$LD(i) = 2[l(\hat{\theta}) - l(\hat{\theta}_{[i]})],$$

where θ is a p vector of the parameter of interest, $\hat{\theta}$ is the parameter estimated with full data, $\hat{\theta_{[i]}}$ is the parameter estimates using data with the ith case deleted, and $l(\theta)$ is the log-likelihood function for θ. The Cook's distance measures the impact of deleting one or a set of observations on parameter estimates. The generalized Cook's distance is defined by

$$CD(i) = (\hat{\theta}_{[i]} - \hat{\theta})^\top G(\hat{\theta}_{[i]} - \hat{\theta}),$$

where G is a positive definite matrix, e.g., $-\partial^2_\theta l(\hat{\theta})$. It has been shown that Cook's distance combines information from the studentized residuals and the variance of predicted values for general linear model (Cook, 1977).

Most of the diagnostic measures were originally developed under linear regression models (Cook, 1977, Cook and Weisberg, 1982, Chatterjee and Hadi, 1986), and then for more complicated models, such as generalized linear models (Davison and Tsai,

1992), generalized estimating equations (Preisser and Qaqish, 1996), models for clustered data(Christensen et al., 1992, Banerjee and Frees, 1997, Haslett and Dillane, 2004), and survival data (Weissfeld, 1990, Lin et al., 1993). In addition, considerable research has been conducted to develop case influence measures in Bayesian analysis (Johnson and Geisser, 1983, 1985, Pettit, 1986, Carlin and Carlin, 1991, Gelfand et al., 1992, Weiss and Cook, 1992, Blyth, 1994, Peng and Dey, 1995, Weiss, 1996, Bradlow and Zaslavsky, 1997). Zhu et al. (2010) provided a comprehensive review of various Bayesian case influence measures and their properties. In Bayesian analysis, the influence of individual observations (or a set of observations) is often assessed by comparing the posterior (or predictive) distribution of the full data to the distribution after deleting these observations (case deletion). For example, the Cook's posterior mode distance, denoted by $CP(i)$, quantifies the discrepancy between the posterior mode of θ with and without the ith case (Cook and Weisberg, 1982). The posterior modes of θ for the full sample Y and a subsample $Y_{[i]}$ are defined as $\hat{\theta} = \text{argmax}_\theta \log p(\theta|Y)$ and $\hat{\theta}_{[i]} = \text{argmax}_\theta \log p(\theta|Y_{[i]})$, respectively. $CP(i)$ is given by

$$CP(i) = (\hat{\theta}_{[i]} - \hat{\theta})^\top G_\theta (\hat{\theta}_{[i]} - \hat{\theta}),$$

where G_θ is chosen to be a positive definite matrix. For instance, G_θ can be $-\partial^2_\theta \log p(\theta|Y)$ $= -\partial^2_\theta \log p(Y|\theta) - \partial^2_\theta \log p(\theta)$ evaluated at $\hat{\theta}$. Similarly, the Cook's posterior mean distance, denoted by $CM(i)$, quantifies the discrepancy between the posterior mean of θ with and without the ith case. The posterior means of θ for the full sample Y and a subsample $Y_{[i]}$ are defined as $\tilde{\theta} = \int \theta p(\theta|Y) d\theta$ and $\tilde{\theta}_{[i]} = \int \theta \dot{p}(\theta|Y_{[i]}) d\theta$, respectively. $CM(i)$ is given by

$$CM(i) = (\tilde{\theta}_{[i]} - \tilde{\theta})^\top W_\theta (\tilde{\theta}_{[i]} - \tilde{\theta}),$$

where W_θ is chosen to be a positive definite matrix.

Diagnostic measures have also been developed for models with missing data (Zhu

et al., 2001, 2009, 2012b). The models for missing data usually use the EM algorithm to obtain the maximum likelihood estimates (Lipsitz and Ibrahim, 1996, Ibrahim et al., 1999). In EM algorithm, the MLE of the parameters θ in the complete likelihood function (based on complete data) is obtained through iterations that maximizes the Q-function

$$Q(\theta|\hat{\theta}) = E[l_c(\theta|D_c)|D_o, \hat{\theta}],$$

where D_c being the complete data, D_o being the observed data, and $l_c(\theta|D_c)$ is the complete-data log-likelihood function. Zhu et al. (2001) used Q-function to replace the log likelihood in the likelihood displacement and showed that the analytic results were very similar to those obtained from a classical local influence approach based on the observed data likelihood function. Q-function is also used to obtain Cook's distance for generalized linear model with missing covariates (Zhu et al., 2009). However, to our knowledge, there is no literature to assess how close the Cook's distance obtained from the Q-function compares to that obtained from the classical likelihood function.

Cook's distance is one of the most important diagnostic tools. A large value of Cook's distance indicates that the observation is influential. However, Zhu et al. (2012a) deliberated size matters issue of Cook's distance. Cook's distance may not be directly comparable because the scale of Cook's distance stochastically depends on the degree of the perturbation. Dr. Zhu et al. introduced the scaled Cook's distance to detect the relatively influential subjects in the sense that the Cook's distance is large relative to the degree of perturbation. How the missing data impacts the degree of perturbation has not been evaluated.

In addition to case deletion measures, a broader range of sensitive analyses have been developed to assess the robustness of a model when perturbing the model assumptions and/or individual observations. In frequentist analysis, extensive literature exists on sensitivity analysis for missing data problems (Troxel, 1998, Zhu and Lee, 2001, Little

and Rubin, 2002, Verbeke et al., 2001a, van Steen et al., 2001, Jansen et al., 2003, 2006, Copas and Eguchi, 2005, Daniels and Hogan, 2008, Shi et al., 2009). The general questions are how to introduce an appropriate perturbation and how to assess its influence on a model. In the Bayesian analysis of statistical models with missing data, Zhu et al. (2014b) in their recent paper introduced various perturbations to modeling assumptions and individual observations, and then developed a formal sensitivity analysis to assess theses perturbations.

1.2.2 Criterion-based Model Assessment

To select an optimal model from a pool of statistical models for a given dataset, we often need to consider both goodness of fit and model complexity. A model which balances model fitting and complexity is preferred. To achieve this, various information criteria have been proposed for model comparisons, which incorporate measures of fit and complexity for model choice. Such information criteria include Akaiki Information Criterion (AIC) (Akaike, 1974), Takeuchi Information Criterion (TIC) (Takeuchi, 1976), Generalized Information Criterion (GIC), (Konishi and Kitagawa, 1996), Network Information Criterion (NIC) (Murata et al., 1994) the Bayesian Information Criterion (BIC) (Schwarz et al., 1978, Lv and Liu, 2014, Konishi et al., 2004), the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), Bayesian Predictive Information Criterion (BPIC) (Ando, 2007), and many others. In these criteria, the goodness of fit are based on the deviance component $-2\log p(y|\hat{\theta})$ or the posterior mean of the deviance component $-2E_{\theta|y}\log p(y|\theta)$, while the measures of model complexity vary from simply the number of parameters to more complicated form. (Ibrahim et al., 2008) Recently, Zhu et al. (2014a) proposed a set of Bayesian case-deletion model complexity measure to quantify the effective number of parameters in a given statistical model, which leads to a Bayesian case-deletion information criterion (BCIC) for model comparison.

The above mentioned model selection criteria can be difficult to obtain for models

including missing data, because these model selection criteria depend on the likelihood function based on the observed data. Some research (Garcia et al., 2010, Ibrahim et al., 2008) has been conducted to use the key components of the EM algorithm, such as the Q-function, to develop an easily computable model selection criterion.

CHAPTER 2

COMPARISON OF STATISTICAL MODELS IN ESTIMATING TREATMENT EFFECT FOR MISSING DATA IN LONGITUDINAL STUDIES

2.1 Introduction

The aim of this chapter is to examine the magnitude of the bias and the robustness of mixed-effects models, PMMs, and selection models under different MNAR mechanisms that are at different degrees of perturbation to the model assumptions. Our primary interest is on the estimate of treatment effect, which is the interest of most clinical trials and many observational studies. We also perform sensitivity analyses and simulations to evaluate the robustness of the PMM, selection model, and mixed-effects model on the estimates of treatment effects. The methods discussed in this manuscript and the simulations performed were motivated by the analysis of PROs. However, they are appropriate for the broader problem of missing data. These analyses are the first we know to systematically evaluate and compare the mixed-effects models, PMMs, and selection models using simulation data.

Section 2.2 defines treatment effect, and Section 2.3 reviews the three existing models: mixed-effects models, PMMs, and selection models. In Section 2.4, we describe an extensive simulation study on comparing the three statistical models under several common scenarios of missing-data mechanisms, focusing on the treatment effect. Results are presented in Section 2.5. Section 2.6 provides concluding remarks.

2.2 Treatment Effect

Treatment effect is the primary interest in clinical trials. A large amount of literature has been developed regard the effect of treatments on HRQOL in various disease areas (Zhou et al., 2009, Sherrill et al., 2010). Different analysis methods have been used to evaluate the treatment effect on HRQOL. However, the statistical definition of treatment effect is not always clear or correct, even in some published methods articles.

Depending on the structure of models, treatment effect can be obtained from different regression coefficients. For continuous outcomes, there are two popular types of models based on the form of outcome. One type directly uses the response score as the dependent variable, whereas the other uses change from baseline score as the dependent variable. These models can also be further classified based on whether baseline measure is included. The baseline measure is the one assessed before treatment initiation. We denote this time point as $t = 0$. Table 2.1 summarizes how to obtain the treatment effects for four commonly used models. We define treatment effect as the difference in the expected value of the response scores between treatments after accounting for baseline difference. In models 1 and 2, baseline value is not included in the models as a covariate. Therefore, the treatment effect is obtained by subtracting the baseline difference from the response difference. In models 3 and 4, baseline difference is accounted for by including the baseline value in the model as a covariate.

Table 2.1: Treatment Effect for Four Commonly Used Models

Model	Response Variable	Treatment Effect
Model 1 $E(y_t) = \beta_0+\beta_1\, trt+\beta_2\, t+\beta_3\, trt\times t$, $t \geq 0$	Response score is the response variable, including y_0; y_0 is not a covariate	Difference in $E(y_t)$ at $t > 0$ minus difference in $E(y_0)$ = $E(y_t\|trt = 1) - E(y_t\|trt = 0) - [E(y_0\|trt = 1) - E(y_0\|trt = 0)] = \boxed{\beta_3\, t}$, where $E(y_t\|trt = 1) = \beta_0+\beta_1+(\beta_2+\beta_3)\times t$, $E(y_t\|trt = 0) = \beta_0 + \beta_2\, t$
Model 2 $E(y_t - y_0) = \beta_0+\beta_1\, trt+\beta_2\, t+\beta_3\, trt\times t$, $t > 0$	Change from baseline score is the response variable; y_0 is not a covariate	Difference in $E(y_t)$ at $t > 0$ minus difference in $E(y_0)$ = $E[y_t - y_0\|trt = 1]-E[y_t - y_0\|trt =0] = \boxed{\beta_1+\beta_3\, t}$, where $E(y_t - y_0\|trt = 1) = \beta_0+\beta_1+(\beta_2+\beta_3)t$, $E(y_t - y_0\|trt = 0) = \beta_0 + \beta_2\, t$
Model 3 $E(y_t)=\beta_0+\beta_1\, trt+\beta_2\, t+\beta_3\, trt\times t+\beta_4 y_0$, $t > 0$	Response score after baseline is the response variable; y_0 is a covariate	Difference in $E(y_t)$ at $t > 0$ given a fixed y_0 = $E[y_t\|trt = 1, y_0] - E[y_t\|trt = 0, y_0] = \boxed{\beta_1+\beta_3\, t}$, where $E(y_t\|trt=1,y_0)= \beta_0+\beta_1+(\beta_2+\beta_3)t+\beta_4 y_0$, $E(y_t\|trt = 0, y_0)= \beta_0+\beta_2\, t +\beta_4 y_0$
Model 4 $E(y_t - y_0)=\beta_0+\beta_1 trt+\beta_2 t+\beta_3\, trt\times t+\beta_4 y_0$, $t > 0$	Change from baseline score is the response variable; y_0 is a covariate	Difference in $E(y_t)$ at $t > 0$ given a fixed y_0 = $E[y_t\|trt = 1, y_0] - E[y_t\|trt = 0, y_0]$ $= E[y_t - y_0\|trt = 1, y_0] - E[y_t - y_0\|trt = 0, y_0]$ $= \boxed{\beta_1+\beta_3\, t}$, where $E(y_t - y_0\|trt=1, y_0)=\beta_0+\beta_1+(\beta_2+\beta_3)t+\beta_4 y_0$, $E(y_t - y_0\|trt=0, y_0)=\beta_0+\beta_2 t+\beta_4 y_0$

Note: *trt* is a 0/1 variable indicating treatment. t is used to indicate the time variable and the time index for the dependent variable ; $t = 0$ indicates baseline.

Note that by including the treatment-by-time interaction term, these models allow treatment effect varies by time. In model 1, treatment effect depends only on the coefficient for the $trt \times t$ interaction (i.e., slope difference, β_3). The slope difference obtained from model 1 is the treatment effect at $t = 1$. If this coefficient is 0 or the interaction is not included in the model, there is no treatment effect. However, in the other three models, it is possible to estimate the treatment effect when the $trt \times t$ interaction is not included. In other words, when treatment effect does not vary by time, the $trt \times t$ interaction can be removed from models 2, 3, and 4.

Model 3 and model 4 in Table 2.1 are equivalent. The only difference is that β_4 in model 4 equals β_4 in model 3 minus 1. When using model 4, it is often observed that the change from baseline score ($y_t - y_0, t > 0$) is negatively related to baseline score (i.e., $\beta_4 < 0$). This seems counterintuitive. However, as long as β_4 in model 4 is greater than -1, the response score (y_t) is still positively correlated to y_0. On the other hand, if β_4 in model 4 is less than -1, caution should be used because it indicates a negative correlation between y_t and y_0.

The treatment effects can be obtained from a linear combination of the fixed-effects coefficients. When assuming a linear relationship between the dependent variable and time (*e.g.*, models shown in Table 2.1), researchers sometimes naively compare the difference between treatments in slope and intercept or simply compare the difference in the estimated value of the dependent variable. However, these comparisons are not always appropriate depending on the setting of the model and whether treatment groups are different at baseline. Although in clinical trials patients are randomized into two treatment groups, in some clinical trials, less than 70% of patients complete baseline PRO assessments, and the baseline scores in the two treatment groups are not always comparable. Therefore, to obtain treatment effect, the treatment difference in response scores must be adjusted for baseline difference. Otherwise, the difference between post-treatment scores may be due to the difference at baseline, not to treatment.

In Table 2.1, time t is a continuous variable. However, it can also be generalized to the case that time is a categorical variable, which allows a non-linear relationship between treatment effect and time. In general, the methods described in the rest of this paper to estimate treament effect when missing data exist are applicable to all above model structures.

2.3 Existing Methods

In this section, we describe mixed-effects models, PMMs, and selection models, which are used to estimate treatment effect when missing data exist in the response variable. Let $y_i = (y_{i1}, \ldots, y_{in_i})^\top$, where y_{ij} denotes the response score (or the change from baseline score) of the ith subject on the jth visit for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. In some clinical trials, n_i may be the same for all subjects. For example, in a study to treat irritable bowel syndrome, all patients received 12 weeks of treatment, and adequate relief by patients as an endpoint was reported weekly (Mangel et al., 2008). However, n_i may vary across subjects–for example, in cancer trials where HRQOL is periodically reported by patients until disease progression, death, or withdrawal from the study for toxicity or other reasons (Zhou et al., 2009).

In clinical trial data, the response y_i may contain missing values with nonmonotone patterns–that is, some response values are observed again after a missing value occurs. For example, y_{ik} may be missing and $y_{ik'}$ may be observed for some $k' > k$. In this case, we call y_{ik} intermittent missing. A subject may also have dropout missing, which is the missing value after the last nonmissing value–that is, no response values are observed again after the missing value occurs. One subject may have an intermittent missing or a dropout missing value or both. When data are missing, we write $y_i = (y_{mis,i}, y_{obs,i})$ for convenience, where $y_{mis,i}$ denotes the missing components of y_i, and $y_{obs,i}$ denotes the observed components of y_i.

2.3.1 Mixed-effects Models

Consider the normal mixed-effects model without missing data given by $y_i = X_i\beta + Z_i b_i + \epsilon_i, i = 1, \ldots, N$, where X_i is an $n_i \times p$ matrix of the fixed-effect covariates, and Z_i is an $n_i \times q$ matrix of the random-effect covariates. Both X_i and Z_i are fixed and known, and Z_i is usually a subset of X_i with fewer covariates. In addition, β is a $p \times 1$ vector of unknown regression parameters, b_i is a $q \times 1$ vector of random effects, and ϵ_i is an $n_i \times 1$ vector of errors. It is commonly assumed that all ϵ_i's and b_i's are independent, $b_i \sim N_q(0, G)$ and $\epsilon_i \sim N_{ni}(\theta, \phi I_{ni})$, where G is a $q \times q$ matrix and I_{ni} is an $n_i \times n_i$ identity matrix.

Table 2.2 presents the likelihood function of the joint distribution of y_i and b_i for subject i. Upon integration over the random effects, the marginal distribution of y_i is $N_{ni}(X_i\beta, Z_i G Z_i^\top + \phi I n_i)$. When missing data exist in y_i, the standard mixed-effects model assumes that missing data are ignorable and the regression coefficients are estimated using the observed data, $y_{obs,i}$. Specifically, we can write

$$y_{obs,i} = X_{obs,i}\beta + Z_{obs,i} b_i + \epsilon_{obs,i}, i = 1, \ldots, N,$$

where $X_{obs,i}, Z_{obs,i}$, and $\epsilon_{obs,i}$ are the subsets of X_i, Z_i, and ϵ_i, respectively, corresponding to $y_{obs,i}$.

Table 2.2: Examples of Mixed-effects Models, PMMs, and Selection Models

Model	Likelihood Function	Distributions	Note
Mixed	$f(y_i, b_i) = f(y_i \mid b_i; \beta, \varphi) f(b_i; G)$	$y_i = X_i\beta + Z_i b_i + \varepsilon_i$, $i = 1, \ldots, N$ where $b_i \sim N_q(0, G)$ $\varepsilon_i \sim N_{ni}(0, \varphi I_{ni})$ $b_i \perp \varepsilon_i$	
PMM	$f(y_i, b_i, s_i) = f(y_i \mid b_i, s_i; \beta^k, \varphi) f(b_i; G) f(s_i)$, given $s_i \perp b_i$	$s_i \sim multinomial(1, \pi_{i1}, \ldots, \pi_{iK})$ For missing pattern k, $y_i^{(k)} = X_i\beta^{(k)} + Z_i b_i + \varepsilon_i$	s_i is the random variable that subject i belongs to a missing pattern k. π_{ik}, $k = 1, .., K$, is the proportion of subjects in missing pattern k in the true population, which may depend on covariates, such as treatment, or on response.
Selection	$f(y_i, b_i, r_i) = f(r_i \mid y_i) f(y_i \mid b_i; \beta, \varphi) f(b_i; G)$, given $r_i \mid y_i \perp b_i$	$y_i = X_i\beta + Z_i b_i + \varepsilon_i$, $i = 1, \ldots, N$ $r_i = (r_{i1}, \ldots, r_{ini})^{\mathrm{T}}$, $r_{ij} \mid y_{ij} \sim Bernoulli(\pi_{ij})$ $\mathrm{Logit}(\pi_{ij}) = \xi_0 + \xi_1 y_{ij}$	r_{ij} is the random variable that the observation j for subject i is missing. π_{ij} is the proportion of y_{ij} being missing given the value of y_{ij} in the true population. For MNAR, π_{ij} is dependent on y_{ij}, but it can also depend on response variables at other time points or on covariates.

Note: X_i is an $n_i \times p$ matrix of the fixed-effect covariates. Z_i is an $n_i \times q$ matrix of the random-effect covariates. Both X_i and Z_i are fixed and known, and Z_i is usually a subset of X_i with fewer covariates. β is a $p \times 1$ vector of unknown regression parameters, b_i is a $q \times 1$ vector of random effects, and ε_i is an $n_i \times 1$ vector of errors. G is a $q \times q$ matrix and I_{ni} is an $n_i \times n_i$ identity matrix.

2.3.2 Pattern-mixture Models

In addition to the response variables and the random effects, the likelihood function of a PMM includes the missing pattern stratum variable s_i, which is usually assumed independent of b_i. For subject i, which is in pattern k, the likelihood function is given by

$$L_{i,PMM} = f(y_i, b_i, s_i) = f(s_i)f(y_i|b_i, s_i; \beta^k, \phi)f(b_i; G),$$

where $s_i \sim multinomial(1, \pi_{i1}, \ldots, \pi_{iK})$, and $\pi_{i1}, \ldots, \pi_{iK}$ are the proportion of subjects in all K missing patterns in the true population, which may depend on covariates, such as treatment, or on response. Moreover, y_i, b_i, and s_i are assumed independent for different subjects.

Conditional on the kth missing pattern, for subject i, the complete data $y_i = (y_{mis,i}, y_{obs,i})$ are given by

$$y_i^{(k)} = X_i\beta^{(k)} + Z_i b_i + \epsilon_i,$$

where the superscript $^{(k)}$ on $y_i^{(k)}$ indicates that subject i is in missing pattern k, and $\beta^{(k)}$ is a $p \times 1$ vector of unknown regression parameters specifically for pattern k. Other model assumptions are the same as those in the mixed-effects model. The PMM is conducted by simply adding the covariate missing pattern and its interaction terms with other fixed-effect covariates into a mixed-effect model. Parameter estimates can be obtained by standard statistical software such as PROC MIXED in SAS.

The estimated response value given each dropout pattern can be obtained from pattern-specific regression parameters. However, when estimating the response variable for time points after dropout, additional assumptions (restriction) must be made, for example, based on complete case, available case, or neighboring case (Fairclough, 2010). These assumptions are not testable due to lack of data.

The estimation of the response value across all missing patterns can be obtained by

a weighted sum of pattern-specific estimates given by

$$E(y|X) = \sum_{k=1}^{K} E(y|X, pattern = k)\pi_k,$$

where π_k is the proportion of patients with the missing pattern, and X is fixed such as treatment and baseline score.

A PMM assumes that a subject belongs to a specific missing pattern. When performing a PMM analysis, initially, K strata are created based on missing patterns. There are several different ways of defining strata. In practice, the missing pattern strata are often defined based on last visit before or after a certain time point (Hedeker and Gibbons, 1997). However, creating strata based only on last visit makes a strong assumption that missing data within each stratum (i.e., intermittent missing) are ignorable (MCAR, or MAR). Strata have also been defined based on reasons for withdrawal (e.g., death, disease progression, toxicity, or other reason) (Pauler et al., 2003, Post et al., 2010). In summary, the choice of strata (or pattern) is a clinical judgment or is made for computational convenience.

The parameters $\pi_{ik}, k = 1, \ldots, K$, may or may not vary by subpopulation. In HRQOL analysis, two popular assumptions are made on the mechanism, by which data are missing: (i) π_{ik} does not depend on any variable, or (ii) π_{ik} depends on treatment. Under assumption (i), π_{ik} is estimated by the proportion of the kth missing pattern in the whole sample. Under assumption (ii), π_{ik} is estimated by the proportion of the kth missing pattern in each treatment group, separately.

2.3.3 Selection Models

Selection models get the name that observations were selected to be missing. Selection models assume that the complete data, y_i, follow a distribution and the probability of missingness depends on the current value of y_i, which may be missing. The conditional

probability of missingness, π_{ij}, may be assumed to follow a logistic regression in which response may be included as a covariate. In selection models, the complete data $y_i = (y_{mis,i}, y_{obs,i})$ are typically assumed to follow the same model given by

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, i = 1, \ldots, N.$$

The selection model introduces an n_i vector of the missing data mechanism variable $r_i = (r_{i1}, \ldots, r_{i,n_i})^\top$, where $r_{ij} = 1$ indicating that y_{ij} is missing. Conditional on y_i, r_i is commonly assumed to be independent of b_i. Thus, for subject i, the joint distribution of y_i, b_i, and r_i is given by

$$L_{i,SM} = f(y_i, b_i, r_i) = f(r_i|y_i)f(y_i|b_i; \beta, \phi)f(b_i; G),$$

where $r_{ij}|y_{ij} \sim Bernoulli(\pi_{ij})$, and π_{ij} is the proportion of y_{ij} being missing given the value of y_{ij} in the true population. $f(r_i|y_i)$ is often modeled by logistic regression or probit models. For MNAR, π_{ij} is dependent on y_{ij}, but it can also depend on response variables at other time points or on covariates. For different subjects, y_i, b_i, and r_i are assumed independent. Ibrahim and Molenberghs (2009) and Ibrahim and colleagues (2001) described the MCEM algorithm used for parametric estimation in selection models (Ibrahim et al., 2001, Ibrahim and Molenberghs, 2009).

2.4 Simulation

We conducted an extensive simulation study to compare the three existing models handling of missing data under various missing patterns and mechanisms.

2.4.1 Data Generation

We simulated missing data on the response variable according to 10 different missing mechanisms (Table 2.3). These missing mechanisms included MCAR (scenario 1); MAR

(missing depending on treatment [scenario 2], the baseline score [scenario 3], or the response at the previous assessment [scenario 4]); and MNAR (missing depending on response at the current assessment [scenarios 5a and 5b]). We also considered intermittent missing data in scenario 5b. Scenario 6 assumed that the missing mechanism was different between two treatment groups: MCAR in one treatment group and MNAR in the other treatment group. In scenarios 7 to 10, data were generated based on the assumption of PMM. Specifically, subjects first were assigned randomly to one of the four dropout groups: pattern 1, only one visit at time 1; pattern 2, last visit at time 2; pattern 3, last visit at time 3; and pattern 4, last visit at time 4 (completer). In these scenarios, the response variable depended on the dropout group to which a particular subject belonged. In all scenarios, subjects were assigned randomly at a 1:1 ratio to one of the two treatment groups. For each subject, a continuous baseline covariate x_i was generated from $N(0,1)$, and the random-effect variable b_i was generated from $N(0,1)$. The measurement errors $\epsilon_{ij}, i = 1, \ldots, N, j = 1, \ldots, n_i$, were generated independently from $N(0,1)$.

Table 2.3: Missing Data Mechanism and Missing Rate in Simulation Data

Scenario	Missing Mechanism	Setting	Missing Pattern	Average Missing Rate (%) at Each Time Point by Treatment
1	MCAR	20% of patients completing previous visit are missing randomly	Dropout	Trt1: 0/21/36/48 Trt0: 0/20/37/50
2	MAR Missing depends on treatment.	Trt1: additional 30% missing at each visit Trt0: additional 10% missing at each visit	Dropout	Trt1: 0/29/50/65 Trt0: 0/9/19/26
3	MAR Dropout depends on x	Logit(prob(r_{ij} = 1)) = x_i - 3, j = 2,3,4	Dropout	Trt1: 0/25/40/51 Trt0: 0/24/40/51
4	MAR Missing depends on y at the previous visit	Logit(prob(r_{ij} = 1)) = $y_{i,j-1}$ - 3, j = 2,3,4	Dropout	Trt1: 0/35/62/75 Trt0: 0/21/44/58
5a	MNAR Missing depends on y at the current visit	Logit(prob(r_{ij} = 1)) = y_{ij} - 3, j = 2,3,4	Dropout	Trt1: 0/50/69/79 Trt0: 0/36/54/64
5b	MNAR Missing depends on y at the current visit	Logit(prob(r_{ij} = 1)) = y_{ij} - 3, j = 1, 2,3,4	Intermit-tent + dropout	Trt1: 33/50/51/49 Trt0: 21/33/34/35
6	MCAR in trt0 MNAR in trt1	Trt0: 20% of patents completing previous visit are missing randomly Trt1: Logit(prob(r_{ij} = 1)) = y_{ij} - 3, j = 2,3,4	Dropout	Trt1: 0/50/70/80 Trt0: 0/20/36/49
7	MNAR-PMM Dropout pattern does not depend on treatment	15% in pattern1, 20% pattern2, 25% pattern3, 40% pattern4.	Dropout	Trt1: 0/15/35/59 Trt0: 0/16/36/60
8	MNAR-PMM Dropout pattern depends on treatment	Trt0: 25% in each pattern Trt1: 15% in pattern 1, 20% pattern 2, 25% pattern 3, 40% pattern 4.	Dropout	Trt1: 0/15/35/59 Trt0: 0/26/51/76
9	MNAR Dropout pattern does not depend on treatment; treatment effect depends on pattern	15% in pattern 1, 20% pattern 2, 25% pattern 3, 40% pattern 4.	Dropout	Trt1: 0/16/35/60 Trt0: 0/15/34/59
10	MNAR Dropout pattern depends on treatment; treatment effect depends on pattern	Trt0: 25% in each pattern Trt1: 15% in pattern 1, 20% pattern 2, 25% pattern 3, 40% pattern 4.	Dropout	Trt1: 0/16/35/60 Trt0: 0/24/49/74

MAR = missing at random; MCAR = missing completely at random; MNAR = missing not at random; r_{ij} = indicator that y_{ij} is missing; Trt = treatment; x_i = baseline score for subject *i*; y_{ij} = response score for subject *i* at visit *j*.

The simulation data for scenarios 1 to 6 included 100 subjects with four planned postbaseline visits. The response variables were generated using an equation given by

$$y_{ij} = \beta_0 + \beta_1 trt_i + \beta_2 x_i + \beta_3 time_{2,ij} + \beta_4 time_{3,ij} + \beta_5 time_{4,ij} + b_i + \epsilon_{ij},$$

where trt_i is the indicator for treatment (1 = trt 1, 0 = trt 0); x_i is baseline score; and $time_{2,ij}, time_{3,ij}$, and $time_{4,ij}$ are indicator variables for visit time corresponding to times 2, 3, and 4 (time 1 is the reference level). By treating time as a categorical variable, we allow the relationship between the response score and time to be nonlinear, which is common in HRQOL data. We generated two sets of data–one with and one without treatment effect. In the first set, we set the treatment parameter as 1. In the second set, we set the treatment parameter as 0. In both sets, the parameters other than treatment were set as 1, indicating that response score increased at the first two visits and then remained stable after that.

For scenarios 7 through 10, a total of 200 subjects were randomly assigned to one of the four dropout patterns: pattern 1, only one visit at time 1; pattern 2, last visit at time 2; pattern 3, last visit at time 3; and pattern 4, last visit at time 4 (completer). The proportion of each pattern was the same across treatments in scenarios 7 and 9 and was treatment specific in scenarios 8 and 10. The response variables were generated by

$$\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 trt_i + \beta_2 x_i + \beta_3 time_{2,ij} + \beta_4 time_{3,ij} + \beta_5 time_{4,ij} \\
&+ \beta_6 pattern_{1,i} + \beta_7 pattern_{2,i} + \beta_8 pattern_{3,i} \\
&+ \beta_{15} time_{2,ij} \times trt_i + \beta_{16} time_{3,ij} \times trt_i + \beta_{17} time_{4,ij} \times trt_i \\
&+ \beta_{12} time_{2,ij} \times pattern_{2,i} + \beta_{13} time_{2,ij} \times pattern_{3,i} + \beta_{14} time_{3,ij} \times pattern_{3,i} \\
&+ \beta_9 trt_i \times pattern_{1,i} + \beta_{10} trt_i \times pattern_{2,i} + \beta_{11} trt_i \times pattern_{3,i} \\
&+ \beta_{18} time_{2,ij} \times pattern_{2,i} \times trt_i + \beta_{19} time_{2,ij} \times pattern_{3,i} \times trt_i \\
&+ \beta_{20} time_{3,ij} \times pattern_{3,i} \times trt_i + b_i + \epsilon_{ij},
\end{aligned}$$

where $pattern_{1,i}, pattern_{2,i}$, and $pattern_{3,i}$ were indicator variables for dropout pattern corresponding to patterns 1, 2, and 3 (pattern 4, completer, is the reference group). The pattern parameters were set as $\beta_6 = 3, \beta_7 = 2$, and $\beta_8 = 1$, that is, patients in the earlier dropout groups had higher (worse) response scores. The time-by-treatment interaction parameters were set as $\beta_{15} = 2, \beta_{16} = 1.5$, and $\beta_{17} = 1$. The pattern-by-time interactions involve only three, rather than nine, parameters ($= 3 \times 3$) because, except for the completer group (pattern 4), the responses after dropout could not be observed or included in analysis. Therefore, there was no need to generate the value of those responses in our simulation. For the same reason, the treatment-by-time-by-pattern three-way interactions also involved only three parameters ($\beta_{18}, \beta_{19}, \beta_{20}$). The treatment effect in scenarios 9 and 10 depended on dropout pattern ($\beta_1 = 1, \beta_9 = \beta_{10} = \beta_{11} = 1, \beta_{18} = \beta_{19} = \beta_{20} = 1$), whereas in scenarios 7 and 8, the treatment effect did not depend on dropout pattern ($\beta_1 = 1, \beta_9 = \beta_{10} = \beta_{11} = 0, \beta_{18} = \beta_{19} = \beta_{20} = 0$). We also generated data with no treatment effect ($\beta_1 = 0, \beta_9 = \beta_{10} = \beta_{11} = 0, \beta_{15} = \beta_{16} = \beta_{17} = 0, \beta_{18} = \beta_{19} = \beta_{20} = 0$). In all scenarios, we set $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 1$.

Data were generated using SAS for Windows (Version 9.2). Table 2.3 summarizes the average missing rate from the simulated data by treatment group and time point. For example, in scenario 1, there were no missing data at time 1, and the missing rate at time 2 was 21% among the trt 1 group and 20% among the trt 0 group.

2.4.2 Analysis of the Simulated Data

For each simulation scenario, three models, including a mixed-effects model, a PMM, and a selection model, described as follows were fitted to estimate the treatment effect.

Analysis for scenarios 1 to 6 (treatment effect does not vary by time)

For the mixed-effects model, the analysis included all observed visits. The models included treatment, baseline score, and discrete time as fixed effects and a random intercept.

Specifically, we specified

$$y_{ij} = \beta_0 + \beta_1 trt_i + \beta_2 x_i + \beta_3 time_{2,ij} + \beta_4 time_{3,ij} + \beta_5 time_{4,ij} + b_i + \epsilon_{ij}.$$

The treatment effect was estimated by β_1. Parameter estimates and the estimates of the standard error were obtained using PROC MIXED (SAS for Windows, Version 9.2).

For the PMM analysis, subjects in the simulated data first were grouped into four dropout patterns based on the last visit. The sample proportion of the dropout groups was calculated for the overall sample and for each treatment group separately. The proportion was calculated using the number of subjects in each pattern group divided by the total number of subjects. The analysis models included fixed effects and the random intercept, as follows:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 trt_i + \beta_2 x_i + \beta_3 time_{2,ij} + \beta_4 time_{3,ij} + \beta_5 time_{4,ij} \\ &+ \beta_6 pattern_{1,i} + \beta_7 pattern_{2,i} + \beta_8 pattern_{3,i} \\ &+ \beta_9 trt_i \times pattern_{1,i} + \beta_{10} trt_i \times pattern_{2,i} + \beta_{11} trt_i \times pattern_{3,i} \\ &+ b_i + \epsilon_{ij}. \end{aligned}$$

The parameters were estimated using PROC MIXED based on all observed visits. Then, using the model described previously, two estimates of the overall treatment effect were obtained, one using the overall proportion of the dropout pattern and the other using the treatment-specific proportion of the dropout pattern. The variance of the treatment effects was obtained using the delta methods (Pauler et al., 2003).

The selection model contained the same fixed and random effects as the mixed model. In addition, the probability of missingness was modeled by using logistic regression with the current response variable as the covariate. For dropout missingness, only the first missing visit was informative, and the missing data after that were not informative.

Therefore, the missing data after the first missing were not included in the logistic regression. Analysis was performed using R software.

Analysis for scenarios 7 to 10 (treatment effect varied by time)

In this set of scenarios, we allowed treatment effect to vary by time by adding treatment by time interaction. For the mixed-effects model and selection model analyses, we added the time-by-treatment interactions (3 terms: $time_2 \times trt; time_3 \times trt; time_4 \times trt$) as fixed covariates. The treatment effects were estimated for each time point.

For PMM, to obtain pattern-specific estimates, the model was specified in the same way as data were generated. Note that only three time-by-pattern interaction terms and, correspondingly, three time-by-pattern-by-treatment interaction terms were included in the model. Other time-by-pattern and time-by-pattern-by-treatment interactions were not estimable because of the missing pattern. The estimates of overall treatment effect were obtained using the overall proportion of dropout pattern and the treatment-specific proportion of dropout pattern as described previously.

2.5 Results

For scenarios 1 to 6, which do not include time-by-treatment interaction effect, Table 2.4 presents the simulation results for the estimates from the mixed-effects model, PMM, and selection model. One estimate from the PMM was based on the overall proportion of dropout groups, and the other was based on the treatment-specific proportion of dropout groups.

Although the PMM and the selection model were designed to handle MNAR data, both provided unbiased estimates of treatment effect (0.96 to 1.01) when missingness did not depend on response. This was the case for scenario 1, in which, at each time point, 20% of subjects completing previous visit were missing randomly; scenario 2, in which the probability of missingness depended on treatment; and scenario 3, in which

the probability of missingness response depended on the baseline score. As expected, the mixed-effects model also provided unbiased estimates of treatment effect in these scenarios.

When missingness depended on the observed response at the previous time point (scenario 4) but not the current time point, the mixed-effects model provided an unbiased estimate (1.03) as expected. However, the estimates from the PMM were biased, especially when the PMM estimates were based on the overall proportion of dropout groups. The estimate from the selection model was also biased. This suggested that the selection model might be sensitive to the parametric form of the missing mechanism, which was not testable from the data.

When missingness at a visit depended on the current response (scenarios 5a and 5b), the selection model built on such a missing mechanism provided an unbiased estimate. Estimates of treatment from both the mixed-effects model and PMM were biased. This was observed when the missing data were related to dropout or were intermittent. The magnitude of the bias was the smallest for the PMM based on treatment-specific proportion when missing data contained only dropout and was the largest for the PMM using the overall proportion of dropout groups.

In scenario 6, the missing mechanism was different between treatment groups; 20% of data were randomly missing among those who completed the previous visit in one treatment group, while the missingness depended on the current response in the other treatment group. In this scenario, the estimates from the selection model and the estimates from the PMM based on the treatment-specific proportion were better than those from the mixed-effect model and the PMM based on the overall proportion.

Table 2.4: Average Treatment Effect and Its Standard Error Estimated from the Mixed-Effects Model, the Pattern-Mixture Model, and the Selection Model for Scenarios 1 to 6 (true treatment effect = 1)

(a) Point estimate of treatment effect

Scenario	Mixed	PMM (ov[a])	PMM (tr[b])	Selection Model
1 MCAR	1.00	1.00	1.01	1.00
2 MAR (trt)	0.96	0.96	0.96	1.00
3 MAR (x)	0.99	1.00	0.99	0.98
4 MAR (y_{t-1})	1.03	0.73[c]	1.08[c]	1.07[c]
5a MNAR (y_t)	0.92[c]	0.77[c]	0.94[c]	0.97
5b MNAR (y_t) Intermittent missing	0.79[c]	0.71[c]	0.79[c]	0.99
6 MCAR in trt0, MNAR(y_t) in trt1	0.80[c]	0.73[c]	0.90[c]	0.91[c]

(b) Standard error

	Mixed		PMM (ov)[a]		PMM (tr)[b]		Selection Model	
Scenario	**Emp. SE**	**SE est.**	**Emp. SE**	**SE est.**	**Emp. SE**	**SE est.**	**Emp. SE**	**SE est.**
1 MCAR	0.24	0.24	0.23	0.25[c]	0.24	0.25[c]	0.22	0.24[c]
2 MAR (trt)	0.27	0.24[c]	0.30	0.28[c]	0.26	0.25[c]	0.26	0.24[c]
3 MAR (x)	0.24	0.24[c]	0.25	0.26[c]	0.25	0.26[c]	0.21	0.24[c]
4 MAR (y_{t-1})	0.25	0.24[c]	0.22	0.22	0.27	0.27	0.21	0.27[c]
5a MNAR (y_t)	0.24	0.24	0.25	0.24[c]	0.25	0.26[c]	0.23	0.25[c]
5b MNAR (y_t) Intermittent missing	0.22	0.22[c]	0.21	0.23[c]	0.22	0.23[c]	0.20	0.23[c]
6 MCAR in trt0, MNAR(y_t) in trt1	0.22	0.25[c]	0.26	0.28[c]	0.23	0.26[c]	0.21	0.24[c]

Emp = Empirical; est = estimate; MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random; PMM pattern-mixture model; SE = standard error; trt = treatment.

[a] Treatment effects estimated using proportion of dropout in overall sample (ov).

[b] Treatment effects estimated using treatment-specific proportion of dropout (tr).

[c] The 95% confidence interval does not cover the true value.

Note: The average point estimate was obtained using the mean of the point estimates from the 100 simulations. The empirical SE was obtained using the standard deviation of the point estimates from the 100 simulations. The average SE estimate was obtained using the mean of the 100 SEs from the 100 simulations. The percent bias was calculated using the SE with three decimal places.

Table 2.5: Average Treatment Effect and Its Standard Error Estimated from the Mixed-Effects Model, the Pattern-Mixture Model, and the Selection model for Scenarios 7 to 10

(a) **Point estimate of treatment effect**

Scenario	Time	True trt eff.	Mixed model	PMM (ov[a])	PMM (tr[b])	Selection model
7 PMM; dropout does not depend on trt, trt effect does not depend on pattern	1	1	1.02	1.00	1.02	1.01
	2	3	3.02	3.01	3.02	3.01
	3	2.5	2.52	2.51	2.51	2.51
	4	2	2.05	2.05	2.07	2.03
8 PMM; dropout depends on trt, trt effect does not depend on pattern	1	0.6	0.63	1.01[c]	0.63	0.62
	2	2.55	2.54	3.01[c]	2.58	2.48[c]
	3	2.1	2.08	2.52[c]	2.14	2.01[c]
	4	1.6	1.67	2.05[c]	1.67	1.55
9 PMM; dropout does not depend on trt, trt effect depends on pattern	1	1.6	1.64	1.62	1.64	1.61
	2	4.05	4.15[c]	4.05	4.09	4.11
	3	3.35	3.54[c]	3.36	3.39	3.50[c]
	4	2.6	2.87[c]	2.65	2.67	2.82[c]
10 PMM; dropout depends on trt, trt effect depends on pattern	1	1.2	1.23	1.70[c]	1.23	1.20
	2	3.6	3.65	4.15[c]	3.63	3.57
	3	2.95	3.08[c]	3.45[c]	2.98	2.99
	4	2.2	2.46[c]	2.72[c]	2.26	2.33[c]

(b) **Standard error**

		Mixed Model		PMM (ov[a])		PMM (tr[b])		Selection Model	
	Time	Emp. SE	SE est.	Emp. SE	SE est.	Emp. SE	SE est.	Emp. SE	SE est.
7 PMM; dropout does not depend on trt, trt effect does not depend on pattern	1	0.25	0.26[c]	0.21	0.20[c]	0.24	0.26[c]	0.25	0.38[c]
	2	0.27	0.28[c]	0.19	0.22[c]	0.27	0.29[c]	0.25	0.45[c]
	3	0.32	0.30[c]	0.27	0.26[c]	0.33	0.31[c]	0.30	0.45[c]
	4	0.34	0.34	0.30	0.32[c]	0.34	0.36[c]	0.33	0.55[c]
8 PMM; dropout depends on trt, trt effect does not depend on pattern	1	0.25	0.26[c]	0.21	0.21[c]	0.25	0.26[c]	0.23	0.34[c]
	2	0.29	0.28[c]	0.21	0.24[c]	0.27	0.30[c]	0.28	0.44[c]
	3	0.35	0.31[c]	0.28	0.29[c]	0.35	0.33[c]	0.34	0.49[c]
	4	0.40	0.37[c]	0.35	0.36[c]	0.39	0.40[c]	0.41	0.56[c]
9 PMM; dropout does not depend on trt, trt effect depends on pattern	1	0.28	0.31[c]	0.20	0.21[c]	0.28	0.28[c]	0.25	0.36[c]
	2	0.35	0.32[c]	0.22	0.24[c]	0.34	0.34	0.32	0.44[c]
	3	0.38	0.34[c]	0.28	0.28[c]	0.36	0.34[c]	0.35	0.50[c]
	4	0.35	0.38[c]	0.34	0.33[c]	0.37	0.37	0.33	0.62[c]
10 PMM; dropout depends on trt, trt effect depends on pattern	1	0.28	0.31[c]	0.21	0.22[c]	0.28	0.28	0.25	0.54[c]
	2	0.36	0.32[c]	0.23	0.26[c]	0.33	0.34[c]	0.32	0.58[c]
	3	0.40	0.35[c]	0.30	0.31	0.38	0.36[c]	0.37	0.62[c]
	4	0.38	0.41[c]	0.35	0.37[c]	0.40	0.42[c]	0.37	0.77[c]

Emp = Empirical; est = estimate; PMM = pattern-mixture model; SE = standard error; trt = treatment.

[a] Treatment effects estimated using proportion of dropout in overall sample (ov).

[b] Treatment effects estimated using treatment-specific proportion of dropout (tr).

[c] The 95% confidence interval does not cover the true value.

Note: The average point estimate was obtained using the mean of the point estimates from the 100 simulations. The empirical SE was obtained using the standard deviation of the point estimates from the 100 simulations. The average SE estimate was obtained using the mean of the 100 SEs from the 100 simulations. The percent bias was calculated using the SE with three decimal places.

For scenarios 7 to 10, data were generated from the preidentified dropout patterns, and the treatment effects varied by time; therefore, the treatment effects were estimated at each time point. Table 2.5 presents the simulation results. The PMM estimates based on the treatment-specific proportion were unbiased in all four scenarios as expected.

The PMM estimates using overall proportion were unbiased when the dropout distribution did not depend on treatment (scenarios 7 and 9), but they were biased when the dropout distribution depended on treatment (scenarios 8 and 10). When the dropout distribution did not depend on treatment, the estimates based on overall proportion were more efficient (smaller standard error [SE]) than those based on the treatment-specific proportion. When the treatment effects depended on the dropout pattern (scenarios 9 and 10), estimates obtained from the mixed-effects model was biased at later time points. However, the mixed-effects model provided unbiased estimates of treatment effect in the scenarios in which treatment effect did not depend on dropout pattern (scenarios 7 and 8), even though the response depended on dropout pattern. The selection model provided biased results at some time points, when either dropout depended on treatment or treatment effect depended on dropout pattern, although the magnitude of bias was small (<10%). Tables 2.4(b) and 2.5(b) present the simulation results on SE estimates. Except for a few cases, the SEs from the mixed-effects model, the PMM, and the selection model are all biased (the 95% confidence interval does not cover the true value.)

When there was no true treatment effect (i.e., parameters for treatment were 0), the bias was less than 0.10 for all models, except for scenario 6, in which treatments had different missing mechanisms. In scenario 6, the bias from PMM estimates using overall proportion (-0.16) was larger than that from the mixed model (-0.12), selection model (-0.13), and PMM using treatment-specific proportion (-0.10).

2.6 Conclusions

Our simulation results suggest that the treatment effect defined as difference in expected value of the response variable given the same baseline was unbiased when the model assumption on missing mechanism holds. Pattern-mixture model provides good estimates when subjects are from different missing patterns. The selection model provides good estimates when the probability of missingness is dependent on the value of response variable. When the model assumption does not hold, the estimate is biased. There is not a model that consistently performs better than others. However, the PMM using the treatment-specific proportion and the selection model provide some correction of the estimate compared with the mixed-effects model in several MNAR situations, even when the mechanism of missing data is not exactly the same as the model assumption.

To obtain the overall treatment effect from the PMM, the treatment-specific proportion of missing pattern should be used in most cases. The overall treatment effect obtained based on the overall proportion of the pattern should be used only when the dropout proportion does not depend on treatment. In this situation, the estimate is equivalent to the estimate based on the treatment-specific proportion but is more efficient.

For selection models, when the probability of missingness depends on the previous response, the estimates from the selection model used in analysis, which specify that the probability of missingness depends on current response in the logit model, are biased. However, this will be improved if the logit model includes the previous response as a covariate.

CHAPTER 3

DIAGNOSTIC MEASURES FOR GENERALIZED LINEAR MODELS WITH MISSING COVARIATES

3.1 Introduction

Missing data commonly occur in various study setting, in both observational studies and clinical trials. Methods for handling missing data strongly depend on the model assumptions. Diagnostic measures provide useful information to measure impact of observations on the parametric models. Cook's distance is one of the most important diagnostic tools and has been expended to the models with missing data with the log likelihood replaced by the Q-function. However, to our knowledge, there is no literature to assess how close the Cook's distance obtained from the Q-function compares to that obtained from the classical likelihood function. A large value of Cook's distance indicates that the observation is influential. However, Cook's distance may not always be directly comparable. Zhu et al. (2012a) deliberated size matters issue of Cook's distance and showed that the scale of Cook's distance stochastically depends on the degree of the perturbation. Zhu, et al. illustrated this in mixed effect models. How the missing data impact the degree of perturbation and the Cook's distance has not been evaluated.

The aim of this chapter is to deliberate the Cook's distance and degree of perturbation for generalized linear models (GLMs) with missing covariates, and to develop scaled Cook's distance for GLMs with missing covariates. In Section 3.2, we review the GLMs with missing covariates. In Section 3.3, we defined the degree of perturbation for GLMs with missing covariates. In Section 3.4, we derive the Cook's distance based on observed likelihood function and compare it to the Cook's distance based on $Q-$function. In

Section 3.5, we illustrate our development of the scaled Cook's distance in the GLM with missing covariate data. In Section 3.6, we performed simulation studies for multiple scenarios. Final remarks are presented at last.

3.2 Generalized Linear Model with Missing Covariate Data

Consider n independent observations (x_i, z_i, y_i), $i = 1, \ldots, N$, where y_i denotes the response variable of the ith subject, x_i is a p_1-dimensional vector of completely observed covariates, and z_i is a p_2-dimensional vector of partially observed covariates. We write $z_i = (z_{m,i}, z_{o,i})$ for convenience, where $z_{m,i}$ denotes the missing components of z_i, and $z_{o,i}$ denotes the observed components of z_i. Let r_i be a p_2-dimensional random vector, whose k-th component, r_{ik}, equals 1 if z_{ik} is missing for subject i, and 0 if z_{ik} is missing, where z_{ik} is the k-th component of z_i. Under the NMAR setting, we specify the joint distribution of (x_i, z_i, r_i, y_i) for each i. We further decompose $f(x_i, z_i, r_i, y_i)$ into a product of three conditional distributions as follows (Little and Schluchter, 1985; Little and Rubin, 2002 Ibrahim, 1990; Lipsitz and Ibrahim, 1996; Ibrahim and Lipsitz, 1996):

$$f(x_i, z_i, r_i, y_i) = f(y_i|x_i, z_i)f(x_i, z_i)f(r_i|x_i, z_i, y_i).$$

The model involves three levels of assumptions: i) For generalized linear models, including general linear model, logistic regression, probit regression, Poisson regression, and gamma regression, y_i given (x_i, z_i) has a density in the exponential family with parameters for regression coefficients $\beta = (\beta_1, \ldots, \beta_p), p = p_1 + p_2$, and the scale parameter τ. ii) We need to specify the distribution of $f(z_i, x_i)$ with parameter α, although it is not necessary to specify a distribution for x_i because x_i's are completely observed. iii) The missing data mechanism $f(r_i|x_i, z_i, y_i)$ is commonly specified using logistic regression models for the binary variables r_{ij}, with parameters for regression coefficients ξ.

The EM algorithm has been a popular technique for obtaining the maximum likeli-

hood estimates (MLE) of $\eta = (\beta, \tau, \alpha, \xi)$ in GLMs with missing covariate data, where the MLE $\hat{\eta}$ is the maximizer of the Q-function $Q(\eta|\hat{\eta}) = E[l_c(\eta|D_c)|D_o, \hat{\eta}]$ with $D_c = \{(y_i, x_i, z_i, r_i) : i = 1, \ldots, n\}$ being the complete data, and $D_o = \{(y_i, x_i, z_{o,i}, r_i) : i = 1, \ldots, n\}$ being the observed data, and $D_m = \{z_{m,i} : i = 1, \ldots, n\}$ being the missing data. At the sth iteration of the EM iteration, given $\eta^{(s)}$, the E-step involves evaluating the Q-function, given by

$$\begin{aligned}
Q(\eta|\eta^{(s)}) &= E[l_c(\eta|D_c)|D_o, \eta^{(s)}] \\
&= \sum_{i=1}^{n} \int \log[f(y_i|x_i, z_i; \beta, \tau)] f(z_{m,i}|x_i, z_{o,i}, r_i, y_i; \eta^{(s)}) dz_{m,i} \\
&+ \sum_{i=1}^{n} \int \log[f(x_i, z_i; \alpha)] f(z_{m,i}|x_i, z_{o,i}, r_i, y_i; \eta^{(s)}) dz_{m,i} \\
&+ \sum_{i=1}^{n} \int \log[f(r_i|x_i, z_i, y_i; \xi)] f(z_{m,i}|x_i, z_{o,i}, r_i, y_i; \eta^{(s)}) dz_{m,i} \\
&= \sum_{i=1}^{n} Q_{1,i}(\beta, \tau|\eta^{(s)}) + \sum_{i=1}^{n} Q_{2,i}(\alpha|\eta^{(s)}) + \sum_{i=1}^{n} Q_{3,i}(\xi|\eta^{(s)}) \\
&= Q_1(\beta, \tau|\eta^{(s)}) + Q_2(\alpha|\eta^{(s)}) + Q_3(\xi|\eta^{(s)}), \qquad (3.1)
\end{aligned}$$

where $l_c(\eta|D_c) = \log f(D_c|\eta)$ is the complete-data log-likelihood function. In E-step, the missing value (of the random variables) is replace with the estimated value (a function of parameter estimates) at sth iteration. The M-step consists of maximizing $Q_1(\beta, \tau|\eta^{(s)})$, $Q_2(\alpha|\eta^{(s)})$, and $Q_3(\xi|\eta^{(s)})$, separately. In M-step, the estimated value is used as "observed" value of the missing data in MLE for complete data.

When subject i contains missing data, we take expecation of the complete-data log-

likelihood function with respect to $z_{m,i}$ given $x_i, z_{o,i}, r_i, y_i, \eta^{(s)}$ noted as follows:

$$\begin{aligned} Q_{1,i}(\beta, \tau|\eta^{(s)}) &= E_{z_{m,i}}\{\log[f(y_i|x_i, z_{m,i}; \beta, \tau)]\}, \\ Q_{2,i}(\alpha|\eta^{(s)}) &= E_{z_{m,i}}\{\log[f(x_i, z_{m,i}; \alpha)]\}, \\ Q_{3,i}(\xi|\eta^{(s)}) &= E_{z_{m,i}}\{\log[f(r_i|x_i, z_{m,i}, y_i, ; \xi)]\}. \end{aligned}$$

Note that when subject i does not contain missing data, they reduce to

$$\begin{aligned} Q_{1,i}(\beta, \tau|\eta^{(s)}) &= \log[f(y_i|x_i, z_{o,i}; \beta, \tau)], \\ Q_{2,i}(\alpha|\eta^{(s)}) &= \log[f(x_i, z_{o,i}; \alpha)], \\ Q_{3,i}(\xi|\eta^{(s)}) &= \log[f(r_i|x_i, z_{o,i}, y_i, ; \xi)]. \end{aligned}$$

Example 1. Consider the example for the general linear model defined below:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \epsilon_i, \tag{3.2}$$

where the ϵ_i's are independent and identically distributed (i.i.d.), $\epsilon_i \sim N(0, \tau)$, and $z_i \sim N(\mu_z, \tau_z)$ for $i = 1, \ldots, n$. Because x_i is completely observed for $i = 1, \ldots, n$, we do not need to specify its distribution. The covariate z_i may be missing for some cases. We assumed MAR for z_i as follows:

$$\text{logit}\,[\text{prob}(r_i = 1|x_i, z_i, y_i)] = \xi_0 + \xi_1 x_i, \tag{3.3}$$

where $r_i = 1$ when z_i is missing. Let $X_i^\top = (1, x_i, z_i), \beta^\top = (\beta_1, \beta_2, \beta_3), X_{12,i}^\top = (1, x_i), \beta_{12}^\top = (\beta_1, \beta_2)$, and $\xi^\top = (\xi_1, \xi_2)$. The complete-data log-likelihood function for

subject i is given by

$$
\begin{aligned}
l_{1,i} &= \log\left(f(y_i|x_i, z_i; \beta, \tau)\right) = -0.5\log(2\pi) - 0.5\log\tau - 0.5\frac{(y_i - X_i^\top\beta)^2}{\tau} \\
&= -0.5\log(2\pi) - 0.5\log\tau - 0.5\frac{(y_i - X_{12,i}^\top\beta_{12} - z_i\beta_3)^2}{\tau}, \\
l_{2,i} &= \log\left(f(z_i; \alpha)\right) = -0.5\log(2\pi) - 0.5\log\tau_z - 0.5\frac{(z_i - \mu_z)^2}{\tau_z}, \\
l_{3,i} &= \log\left(f(r_i|x_i, z_i, y_i; \xi)\right) = r_i X_{12,i}^\top\xi - \log[1 + \exp(X_{12,i}^\top\xi)].
\end{aligned}
$$

In this MAC setting for general linear model, we show that the $f(z_{m,i}|x_i, z_{o,i}, r_i, y_i; \eta^{(s)})$ in equation (3.1) can be written as follows:

$$
f(z_{m,i}|x_i, r_i, y_i; \eta^{(s)}) = f(z_{m,i}|y_i, x_i; \eta^{(s)}) \quad \propto \quad f(z_{m,i}; \eta^{(s)}) f(y_i|x_i, z_{m,i}; \eta^{(s)}).
$$

This is a product of two likelihood functions of the normal distribution. It can be shown that $z_{m,i}|x_i, r_i, y_i \sim N(\mu_{zm,i}, \tau_{zm})$ with

$$
\mu_{zm,i} = \left(\frac{\beta_3^2}{\tau} + \frac{1}{\tau_z}\right)^{-1}\left(\frac{\beta_3(y_i - X_{12,i}^\top\beta_{12})}{\tau} + \frac{\mu_z}{\tau_z}\right) \text{ and } \tau_{zm} = \left(\frac{\beta_3^2}{\tau} + \frac{1}{\tau_z}\right)^{-1}.
$$

We define S_{obs} as the subset of subjects with observed z_i and S_{mis} as the subset of subjects with missing z_i. $I(.)$ is the indicator function. Since $E_{z_{m,i}}\{z_{m,i}\} = \mu_{zm,i}$ and

$E_{z_{m,i}}\{z_{m,i}^2\} = \mu_{zm,i}^2 + \tau_{zm}$, we have

$$
\begin{aligned}
Q_1(\beta,\tau|\eta^{(s)}) &= \sum_{i=1}^{N} Q_{1,i}(\beta,\tau|\eta^{(s)}) \\
&= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\tau - \frac{1}{2\tau}\sum_{i=1}^{N}(y_i - X_{12,i}^{\top}\beta_{12})^2 \\
&\quad -\frac{1}{2\tau}\sum_{i=1}^{N}(-2)(y_i - X_{12,i}^{\top}\beta_{12})[z_i I(i\in S_{obs}) + \mu_{zm,i}^{(s)} I(i\in S_{mis})]\beta_3 \\
&\quad -\frac{1}{2\tau}\sum_{i=1}^{N}[z_i^2 I(i\in S_{obs}) + (\mu_{zm,i}^{(s)2} + \tau_{zm}^{(s)}) I(i\in S_{mis})]\beta_3^2, \\
Q_2(\alpha|\eta^{(s)}) &= \sum_{i=1}^{N} Q_{2,i}(\alpha|\eta^{(s)}) \\
&= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\tau_z \\
&\quad -\frac{1}{2\tau_z}\sum_{i=1}^{N}[z_i^2 I(i\in S_{obs}) + (\mu_{zm,i}^{(s)2} + \tau_{zm}^{(s)}) I(i\in S_{mis})] \\
&\quad -\frac{1}{2\tau_z}\sum_{i=1}^{N}(-2)[z_i I(i\in S_{obs}) + \mu_{zm,i}^{(s)} I(i\in S_{mis})]\mu_z - \frac{1}{2\tau_z}N\mu_z^2, \\
Q_3(\xi|\eta^{(s)}) &= \sum_{i=1}^{N} l_{3,i} = \sum_{i=1}^{N}\{r_i X_{12,i}^{\top}\xi - \log[1+\exp(X_{12,i}^{\top}\xi)]\},
\end{aligned}
$$

which is the log likelihood function of the regular logistic regression with nonmissing data (under MAR).

The estimate of ξ can be simply obtained from the regular logistic regression analysis. For other parameters, the M-step maximizes $Q_1(\beta,\tau|\eta^{(s)})$ and $Q_2(\alpha|\eta^{(s)})$. The closed forms are available for $(\beta_{12},\beta_3,\tau,\mu_z,\tau_z)$. Let J be the $N\times 1$ matrix with value 1 for all elements, y be the $N\times 1$ matrix of response variables, and X_{12} be the $N\times 2$ design matrix of known covariates. To simplify the notation, we write $Z^{(s)}$ as the $N\times 1$ design matrix of covariate with missing value replaced with estimated value $\mu_{zm,i}^{(s)}$, and define

$A^{(s)}, B^{(s)}$ as follows:

$$A^{(s)} = \sum_{i=1}^{N} [z_i^2 I(i \in S_{obs}) + (\mu_{zm,i}^{(s)2} + \tau_{zm}^{(s)}) I(i \in S_{mis})], B^{(s)} = (b_{ij}),$$

where b_{ij}'s are given by

$$b_{ij} = \begin{cases} [z_i I(i \in S_{obs}) + \mu_{zm,i}^{(s)} I(i \in S_{mis})][z_j I(j \in S_{obs}) + \mu_{zm,j}^{(s)} I(j \in S_{mis})], \text{when } i \neq j, \\ z_i^2 I(i \in S_{obs}) + (\mu_{zm,i}^{(s)2} + \tau_{zm}^{(s)}) I(i \in S_{mis}), \text{when } i = j. \end{cases}$$

To calculate $(\beta_{12}^{(s+1)}, \beta_3^{(s+1)})$, we take the first derivative of Q_1,

$$\begin{cases} \frac{\partial Q_1}{\partial \beta_{12}} = \tau^{-1}\{X_{12}^\top (y - X_{12}\beta_{12}) - X_{12}^\top Z^{(s)} \beta_3\} = 0 \\ \frac{\partial Q_1}{\partial \beta_3} = \tau^{-1}\{Z^{(s)\top}(y - X_{12}\beta_{12}) - A^{(s)}\beta_3\} = 0. \end{cases}$$

This yields

$$\beta_{12}^{(s+1)} = \{X_{12}^\top [I - A^{(s)-1} B^{(s)}] X_{12}\}^{-1} X_{12}^\top [I - A^{(s)-1} B^{(s)}] y,$$

$$\beta_3^{(s+1)} = A^{(s)-1} Z^{(s)\top} (y - X_{12}\beta_{12}^{(s+1)}).$$

Similarly, other parameter estimates can be easily obtained as follows:

$$\begin{aligned} \tau^{(s+1)} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - X_{12,i}^\top \beta_{12}^{(s+1)})^2 - \frac{2}{N}\sum_{i=1}^{N}(y_i - X_{12,i}^\top \beta_{12}^{(s+1)})[z_i I(i \in S_{obs}) \\ &+ \mu_{zm,i}^{(s)} I(i \in S_{mis})]\beta_3^{(s+1)} + \frac{1}{N}\sum_{i=1}^{N}[z_i^2 I(i \in S_{obs}) + (\mu_{zm,i}^{(s)2} + \tau_{zm}^{(s)}) I(i \in S_{mis})]\beta_3^{(s+1)2} \\ &= \frac{1}{N}[y - X_{12}\beta_{12}^{(s+1)}]^\top [y - X_{12}\beta_{12}^{(s+1)}] - \frac{2}{N}[y - X_{12}\beta_{12}^{(s+1)}]^\top Z^{(s)}\beta_3^{(s+1)} \\ &+ \frac{1}{N} A^{(s)} \beta_3^{(s+1)2}, \end{aligned}$$

$$\begin{aligned}\mu_z^{(s+1)} &= \frac{\sum_{i\in obs} z_i + \sum_{i\in mis} \mu_{zm,i}^{(s)}}{N} = \frac{J^\top Z^{(s)}}{N},\\ \tau_z^{(s+1)} &= \frac{A^{(s)}}{N} - \mu_z^{(s+1)2}.\end{aligned}$$

3.3 Degree of Perturbation

We use $P(I|M)$ to denote the degree of perturbation introduced by deleting the subset I for the fitted model M. Critchley et al. (2001) used the Euclidean geometry to quantify the size of perturbation for one-sample problems. However, it cannot be easily generated for relatively complex data structures, such as longitudinal data. Recently, Zhu et. al (2012) proposed $P(I|M)$ based on the Kullback-Leibler distance. The proposed $P(I|M)$ has the following four desired principle properties: i) non-negativity, ii) uniqueness, iii) monotonicity, and iv) additivity.

Degree of perturbation is derived below in a more general setting. Consider the probability function of a random vector $Y^\top = (Y_1^\top, \ldots, Y_n^\top)$, denoted by $p(Y|\theta)$, where $\theta = (\theta_1, \ldots, \theta_q)^\top$ is the set of parameters. A subscript '$[I]$' denotes the relevant quantity with all observations in I deleted. Let $Y_{[I]}$ be a subsample of Y with $Y_I = \{Y_i : i \in I\}$ deleted and $p(Y_{[I]}|\theta)$ be its probability function. We define the maximum likelihood estimators of θ for the full sample Y and a subsample $Y_{[I]}$ as $\hat{\theta} = \text{argmax}_\theta \log p(Y|\theta)$ and $\hat{\theta}_{[I]} = \text{argmax}_\theta \log p(Y_{[I]}|\theta)$, respectively.

We can write $p(Y|\theta) = p(Y_{[I]}|\theta)p(Y_I|Y_{[I]}, \theta)$. Consider the following model for characterizing the deletion of Y_I given by

$$p(Y|\theta, I) \equiv p(Y_{[I]}|\theta)p_0(Y_I|Y_{[I]}),$$

where $p_0(Y_I|Y_{[I]})$ is a fixed conditional density of Y_I given $Y_{[I]}$ independent of θ. Zhu et. al. suggest setting $p_0(Y_I|Y_{[I]}) = p(Y_I|Y_{[I]}, \theta_*)$, where θ_* is the true value of θ under M. When M is correctly specified, $p(Y_I|Y_{[I]}, \theta_*)$ is the true data generator for Y_I given $Y_{[I]}$.

The Kullback-Leibler distance between $p(Y|\theta)$ and $p(Y|\theta, I)$, denoted by $KL(Y, \theta|I)$ can be written as follows:

$$KL(Y,\theta|I) = \int p(Y|\theta) \log\left(\frac{p(Y|\theta)}{p(Y|\theta, I)}\right) dY = \int p(Y|\theta) \log\left(\frac{p(Y_I|Y_{[I]}, \theta)}{p(Y_I|Y_{[I]}, \theta_*)}\right) dY.$$

When Y_I is independent of $Y_{[I]}$, $KL(Y, \theta|I)$ reduces to

$$KL(Y,\theta|I) = \int p(Y_I|\theta) \log\left(\frac{p(Y_I|\theta)}{p(Y_I|\theta_*)}\right) dY_I.$$

Finally, $P(I|M)$ is defined as the weighted Kullback-Leibler distance between $p(Y|\theta)$ and $p(Y|\theta, I)$ as follows:

$$P(I|M) = \int KL(Y,\theta|I) p(\theta|\theta_*, \Sigma_*) d\theta, \tag{3.4}$$

where $p(\theta|\theta_*, \Sigma_*)$ is a Gaussian prior for unknown θ with mean θ_* and positive definite covariance matrix Σ_* (e.g., the Fisher information matrix). Furthermore, if a particular set of components of θ is of interest and other components are treated as nuisance parameters, we may fix these nuisance parameters in their true value.

Here, we apply general definition (3.4) to derive degree of perturbation for GLM with missing covariates as below. For GLM with missing covariates, $Y_i = (x_i, z_i, r_i, y_i)$, we can write the Kullback-Leibler distance as

$$\begin{aligned}
KL(\eta|I) &= \int\int\int\int \prod_{i\in I} f(x_i, z_i, r_i, y_i|\eta) \log\left(\frac{\prod_{i\in I} f(x_i, z_i, r_i, y_i|\eta)}{\prod_{i\in I} f(x_i, z_i, r_i, y_i|\eta_*)}\right) \prod_{i\in I} dx_i dz_i dr_i dy_i \\
&= \int\int\int\int \prod_{i\in I} f(x_i, z_i, r_i, y_i|\eta) \left(\sum_{i\in I} \left[\log \frac{f(x_i, z_i, r_i, y_i|\eta)}{f(x_i, z_i, r_i, y_i|\eta_*)}\right]\right) \prod_{i\in I} dx_i dz_i dr_i dy_i \\
&= \sum_{i\in I} E_{Y_i}\left[\log \frac{f(x_i, z_i, r_i, y_i|\eta)}{f(x_i, z_i, r_i, y_i|\eta_*)}\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned} P(I|M) &= \int KL(\eta|I)p(\eta|\eta_*)d\eta \\ &= \sum_{i\in I} E_\eta E_{Y_i}\left[\log\frac{f(x_i,z_i,r_i,y_i|\eta)}{f(x_i,z_i,r_i,y_i|\eta_*)}\right] = \sum_{i\in I} P(i|M). \end{aligned}$$

For GLM with missing covariates, the degree of perturbation is different for subjects with observed data and subjects with missing data. When subject i is observed, we have

$$\begin{aligned} P(i|M) &= E_\eta E_{Y_i}\left[\log\frac{f(x_i,z_i,r_i,y_i|\eta)}{f(x_i,z_i,r_i,y_i|\eta_*)}\right] \\ &= E_\eta E_{Y_i}\left[\log\frac{f(y_i|x_i,z_i,\beta,\tau)}{f(y_i|x_i,z_i,\beta_*,\tau_*)} + \log\frac{f(x_i,z_i|\alpha)}{f(x_i,z_i|\alpha_*)} + \log\frac{f(r_i|x_i,z_i,y_i,\xi)}{f(r_i|x_i,z_i,y_i,\xi_*)}\right]. \end{aligned}$$

When subject i has missing covariates, we can only observe (x_i, r_i, y_i), and therefore have

$$P(i|M) = E_{\beta,\tau,\xi}E_{x_i,r_i,y_i}\left[\log\frac{\int f(x_i,z_i,r_i,y_i|\eta)dz_i}{\int f(x_i,z_i,r_i,y_i|\eta_*)dz_i}\right].$$

This is further illustrated using the Example 1.

Example 1 (continue). Let $X_i^\top = (1, x_i, z_i), \beta^\top = (\beta_1, \beta_2, \beta_3), X_{12,i}^\top = (1, x_i), \beta_{12}^\top = (\beta_1, \beta_2)$, and $\xi^\top = (\xi_1, \xi_2)$. When subject i is observed, we have

$$\begin{aligned} P(i|M) &= E_\eta E_{Y_i}\left[\log\frac{f(y_i|x_i,z_i,\beta,\tau)}{f(y_i|x_i,z_i,\beta_*,\tau_*)} + \log\frac{f(x_i,z_i|\alpha)}{f(x_i,z_i|\alpha_*)} + \log\frac{f(r_i|x_i,\xi)}{f(r_i|x_i,\xi_*)}\right] \\ &= E_\eta E_{z_i}E_{y_i|z_i}\left[\log f(y_i|z_i,\beta,\tau) - \log f(y_i|z_i,\beta_*,\tau_*)\right] \\ &\quad + E_\eta E_{z_i}\left[\log f(z_i|\mu_z,\tau_z) - \log f(z_i|\mu_{z*},\tau_{z*})\right] \\ &\quad + E_\xi E_{r_i}\left[\log f(r_i|\xi) - \log f(r_i|\xi_*)\right] \\ &= P_1(i|M) + P_2(i|M) + P_3(i|M). \end{aligned}$$

Specifically, we have

$$
\begin{aligned}
P_1(i|M) &= E_\eta E_{z_i} E_{y_i|z_i}\left[\log\frac{f(y_i|x_i,z_i,\beta,\tau)}{f(y_i|x_i,z_i,\beta_*,\tau_*)}\right] \\
&= E_\eta E_{z_i} E_{y_i|z_i}\{-0.5\log(2\pi)-0.5\log\tau-0.5\frac{(y_i-X_i^\top\beta)^2}{\tau} \\
&\quad -\left[-0.5\log(2\pi)-0.5\log\tau_*-0.5\frac{(y_i-X_i^\top\beta_*)^2}{\tau_*}\right]\} \\
&= 0.5E_\eta[\log(\tau_*/\tau)]+0.5E_\eta E_{z_i}E_{y_i|z_i}\{\frac{(y_i-X_i^\top\beta_*)^2}{\tau_*}-\frac{(y_i-X_i^\top\beta)^2}{\tau}\} \\
&= 0.5E_\eta[\log(\tau_*/\tau)]+0.5E_\eta E_{z_i}E_{y_i|z_i}\left[\frac{(y_i-X_i^\top\beta_*+y_i-X_i^\top\beta)(X_i^\top\beta-X_i^\top\beta_*)}{\tau_*}\right] \\
&\quad +0.5E_\eta\frac{\tau}{\tau_*}-0.5E_\eta\frac{\tau}{\tau} \\
&= 0.5E_\eta[\log(\tau_*/\tau)]+0.5\frac{X_{12,i}^\top E_\eta[(\beta_{12}-\beta_{12*})(\beta_{12}-\beta_{12*})^\top]X_{12,i}}{\tau_*} \\
&\quad -\frac{E_\eta[X_{12,i}^\top(\beta_{12}-\beta_{12*})(\beta_3-\beta_{3*})\mu_z]}{\tau_*}+0.5\frac{E_\eta[(\tau_z+\mu_z^2)(\beta_3-\beta_{3*})^2]}{\tau_*},
\end{aligned}
$$

$$
\begin{aligned}
P_2(i|M) &= E_\eta E_{z_i}\left[\log\frac{f(z_i|\mu_z,\tau_z)}{f(z_i|\mu_{z*},\tau_{z*})}\right] \\
&= E_\eta E_{z_i}\{-0.5\log(2\pi)-0.5\log\tau_z-0.5\frac{(z_i-\mu_z)^2}{\tau_z} \\
&\quad -\left[-0.5\log(2\pi)-0.5\log\tau_{z*}-0.5\frac{(z_i-\mu_{z*})^2}{\tau_{z*}}\right]\} \\
&= 0.5E_\eta[\log(\tau_{z*}/\tau_z)]+0.5E_\eta E_{z_i}\{\frac{(z_i-\mu_{z*})^2}{\tau_{z*}}-\frac{(z_i-\mu_z)^2}{\tau_z}\} \\
&= 0.5E_\eta[\log(\tau_{z*}/\tau_z)]+0.5E_\eta E_{z_i}\{\frac{(z_i-\mu_{z*})^2}{\tau_{z*}}-\frac{(z_i-\mu_z)^2}{\tau_{z*}} \\
&\quad +\frac{(z_i-\mu_z)^2}{\tau_{z*}}-\frac{(z_i-\mu_z)^2}{\tau_z}\} \\
&= 0.5E_\eta[\log(\tau_{z*}/\tau_z)]+0.5E_\eta E_{z_i}\left[\frac{(z_i-\mu_{z*}+z_i-\mu_z)(\mu_z-\mu_{z*})}{\tau_{z*}}\right] \\
&\quad +0.5E_\eta\frac{\tau_z}{\tau_{*z}}-0.5E_\eta\frac{\tau_z}{\tau_z} \\
&= 0.5E_{\tau_z}[\log(\tau_{z*}/\tau_z)]+0.5\frac{E_{\mu_z}[(\mu_z-\mu_{z*})^2]}{\tau_{z*}},
\end{aligned}
$$

$$
\begin{aligned}
P_3(i|M) &= \int E_{r_i}\left[\log\frac{f(r_i|x_i,\xi)}{f(r_i|x_i,\xi_*)}\right]p(\xi|\xi_*)d\xi \\
&= E_\xi E_{r_i}\{r_i X_{12,i}^\top\xi - \log[1+\exp(X_{12,i}^\top\xi)] \\
&\quad -r_i X_{12,i}^\top\xi_* + \log[1+\exp(X_{12,i}^\top\xi_*)]\} \\
&= E_\xi \log\{\frac{1+\exp(X_{12,i}^\top\xi_*)}{1+\exp(X_{12,i}^\top\xi)}\} + E_\xi[(X_{12,i}^\top\xi - X_{12,i}^\top\xi_*)E_{r_i}(r_i)] \\
&= E_\xi \log\{\frac{1+\exp(X_{12,i}^\top\xi_*)}{1+\exp(X_{12,i}^\top\xi)}\} + E_\xi\{X_{12,i}^\top(\xi-\xi_*)\frac{\exp(X_{12,i}^\top\xi)}{1+\exp(X_{12,i}^\top\xi)}\} \\
&= E_\xi \log\frac{(1-\pi)}{(1-\pi_*)} + E_\xi[X_{12,i}^\top(\xi-\xi_*)\pi], \qquad (3.5)
\end{aligned}
$$

where $\pi = \text{logit}^{-1}(X_{12,i}^\top\xi)$ and $\pi_* = \text{logit}^{-1}(X_{12,i}^\top\xi_*)$.

When subject i has missing covariates, to calculate $P(i|M)$, we first derive the marginal likelihood function as below:

$$
\begin{aligned}
\int f(x_i,z_i,r_i,y_i|\eta)dz_i &= \int f(y_i|x_i,z_i,\beta,\tau)f(z_i,\alpha)f(r_i|x_i,z_i,\xi)dz_i \\
&= f(r_i|x_i,\xi)\int f(y_i|x_i,z_i,\beta,\tau)f(z_i,\alpha)dz_i,
\end{aligned}
$$

where $\int f(y_i|x_i,z_i,\beta,\tau)f(z_i,\alpha)dz_i$ is the marginal distribution of y_i, which follows a normal distribution with mean $\mu_{yi} = E(y_i) = E_{z_i}[E(y_i|z_i)] = X_{12,i}\beta_{12} + \beta_3\mu_z$ and variance $\tau_y = Var(y_i) = E_{z_i}[Var(y_i|z_i)] + Var_{z_i}[E(y_i|z_i)] = \tau + \beta_3^2\tau_z$. Therefore, we have

$$
\begin{aligned}
\log\left[\int f(x_i,z_i,r_i,y_i|\eta)dz_i\right] &= -0.5\log(2\pi) - 0.5\log\tau_y - 0.5\frac{(y_i-\mu_{yi})^2}{\tau_y} \\
&\quad +r_i X_{12,i}^\top\xi - \log[1+\exp(X_{12,i}^\top\xi)],
\end{aligned}
$$

and

$$
\begin{aligned}
P(i|M) &= E_\eta E_{y_i}\{-0.5\log\tau_y - 0.5\frac{(y_i-\mu_{yi})^2}{\tau_y} - [-0.5\log\tau_{y,*} - 0.5\frac{(y_i-\mu_{yi,*})^2}{\tau_{y,*}}]\} \\
&\quad +E_\xi E_{r_i}\{r_i X_{12,i}^\top\xi - \log[1+\exp(X_{12,i}^\top\xi)] - \left[r_i X_{12,i}^\top\xi_* - \log[1+\exp(X_{12,i}^\top\xi_*)]\right]\} \\
&= P_4(i|M) + P_3(i|M),
\end{aligned}
$$

where $P_3(i|M)$ is the same as (3.5) and $P_4(i|M)$ is given by

$$\begin{aligned} P_4(i|M) &= 0.5E_\eta[\log(\tau_{y*}/\tau_y)] + 0.5\frac{E_\eta[(\mu_{yi}-\mu_{y*})^2]}{\tau_{y*}} \\ &= 0.5E_\eta[\log(\frac{\tau_* + \beta_{3*}^2\tau_{z*}}{\tau + \beta_3^2\tau_z})] + 0.5\frac{E_\eta\{[X_{12,i}^\top(\beta_{12}-\beta_{12*}) + \beta_3\mu_z - \beta_{3*}\mu_{z*}]^2\}}{\tau_* + \beta_{3*}^2\tau_{z*}}. \end{aligned}$$

The degree of perturbation for GLM with missing covariates has a much more complicated form than the model without missing data. As shown in Example 1, when subject i is observed, the degree of perturbation includes three components: $P_1(i|M)$ for the model on response y_i, $P_2(i|M)$ for the model on covariate z_i, and $P_3(i|M)$ for the model on missing mechanism r_i. In $P_1(i|M)$, the first two components are similar to what are in the general linear models with fixed covariates ($X_{12,i}$). In addition, $P_1(i|M)$ includes the component for the random covariates (z_i) and the component involving the product of the fixed covariates and the random covariates. When subject i has missing covariate z_i, the degree of perturbation includes the same component $P_3(i|M)$ for missing mechanism and the different component $P_4(i|M)$ for response. When subjects do not have missing data, the degree of perturbation is related to the variance of β_{12} and μ_z, while the degree of perturbation for subjects with missing data is not related to the variance of μ_z. In both cases, the degree of perturbation is a function of observed covariate X_{12}. This will be further illustrated in the simulation study.

3.4 Cook's Distance

Cook's distance and many other deletion diagnostics measure the distance between the maximum likelihood estimators of θ with and without observations in set I. A subscript $[I]$ denotes the relevant quantity with all observations in I deleted. Cook's distance for I, denoted by CD_I, can be defined as follows:

$$CD_I = (\hat{\theta}_{[I]} - \hat{\theta})^\top G(\hat{\theta}_{[I]} - \hat{\theta}),$$

where $\hat{\theta}$ is the MLE for the full sample, $\hat{\theta}_{[I]}$ is the MLE for the subsample with all observations in I deleted, and G is a positive definite matrix. When the interest is on a subset of θ or a linear combination of θ, say $L^\top\theta$, the partial influence of the subset I on $L^\top\hat{\theta}$, denoted by $CD(I|L)$, can be defined as

$$CD(I|L) = (\hat{\theta}_{[I]} - \hat{\theta})^\top L\{L^\top G^{-1} L\}^{-1} L^\top (\hat{\theta}_{[I]} - \hat{\theta}).$$

For GLM with missing covariates models, we can calculate Cook's distance based on the log-likelihood function (l) of the observed data or the Q-function used in the EM algorithm (i.e., $E[l_c(D_c, \theta)|D_o, \hat{\theta}]$). We denote the $l-$based Cook's distance as CD and the $Q-$based Cook's distance as QCD. The MLE from maximizing the observed likelihood function or maximizing the Q-function are equivalent. Specifically,

$$\hat{\theta} = \text{argmax}\, l(D_o, \theta) = \text{argmax}\, E[l_c(D_c, \theta)|D_o, \hat{\theta}].$$

However, as shown below, Q-based MLE for $\hat{\theta}_{[I]}$ is different from l-based estimate, because Q-based MLE is conditional on the MLE $\hat{\theta}$ from the full sample:

$$\hat{\theta}^{CD}_{[I]} = \text{argmax}\, l_{[I]}(D_{o[I]}, \theta),\ \hat{\theta}^{QCD}_{[I]} = \text{argmax}\, E[l_c(D_{c[I]}, \theta)|D_o, \hat{\theta}],$$

where $D_{c[I]}$ is a subsample, in which all observations in I are deleted from D_c. Because $\hat{\theta}_{[I]}$ is needed for every case, to reduce the total computational burden, we use the one-step approximation $\hat{\theta}^1_{[I]}$ of $\hat{\theta}_{[I]}$ (Zhu et al., 2001, 2009) based on Taylor expansion. For the subsample $D_{c[I]}$, we define $Q_{[I]}(\hat{\theta}|\hat{\theta})$ as $Q_{[I]}(\hat{\theta}|\hat{\theta}) = E[l_c(D_{c[I]}, \theta)|D_o, \hat{\theta}]|_{\theta=\hat{\theta}}$, where the expectation is taken with respect to $f(D_m|D_o, \hat{\theta})$. For l-based and Q-based estimates, we can write

$$\hat{\theta}^{1CD}_{[I]} = \hat{\theta} + \{-\partial^2_\theta l(D_o|\hat{\theta})\}^{-1} \partial_\theta l_{[I]}(D_{o[I]}|\hat{\theta}), \tag{3.6}$$

$$\hat{\theta}^{1QCD}_{[I]} = \hat{\theta} + \{-\partial^2_\theta Q(\hat{\theta}|\hat{\theta})\}^{-1}\partial_\theta Q_{[I]}(\hat{\theta}|\hat{\theta}).$$

The G matrix is chosen to be a positive definite matrix as below for $l-$based and $Q-$based Cook's distance, respectively:

$$G^{CD} = -\partial^2_\theta l(D_o|\hat{\theta}),\ G^{QCD} = -\partial^2_\theta Q(\hat{\theta}|\hat{\theta}).$$

Therefore, we can approximate CD and QCD as

$$CD_I = \partial_\theta l_{[I]}(D_{o[I]}|\hat{\theta})^\top\{-\partial^2_\theta l(D_o|\hat{\theta})\}^{-1}\partial_\theta l_{[I]}(D_{o[I]}|\hat{\theta}),$$

$$QCD_I = \partial_\theta Q_{[I]}(\hat{\theta}|\hat{\theta})^\top\{-\partial^2_\theta Q(\hat{\theta}|\hat{\theta})\}^{-1}\partial_\theta Q_{[I]}(\hat{\theta}|\hat{\theta}).$$

Continuing with Example 1, we derive the CD and QCD for general linear model with missing covariates, which will be used later in simulation study to compare these two quantities.

Example 1 (continued). We can write the log likelihood function as

$$
\begin{aligned}
l &= \sum_{i\in S_{obs}} \{-0.5\log(2\pi) - 0.5\log\tau - 0.5\frac{(y_i - X_i^\top\beta)^2}{\tau} \\
&\quad -0.5\log(2\pi) - 0.5\log\tau_z - 0.5\frac{(z_i-\mu_z)^2}{\tau_z} \\
&\quad +r_i X_{12,i}^\top\xi - \log[1+\exp(X_{12,i}^\top\xi)]\} \\
&+ \sum_{i\in S_{mis}} \{-0.5\log(2\pi) - 0.5\log\tau_y - 0.5\frac{(y_i-\mu_{yi})^2}{\tau_y} \\
&\quad +r_i X_{12,i}^\top\xi - \log[1+\exp(X_{12,i}^\top\xi)]\} \\
&= -\sum_{i\in S_{obs}} \{0.5\log\tau + 0.5\frac{(y_i - X_i^\top\beta)^2}{\tau} + 0.5\log\tau_z + 0.5\frac{(z_i-\mu_z)^2}{\tau_z}\} \\
&\quad -\sum_{i\in S_{mis}} \{0.5\log(\tau+\beta_3^2\tau_z) + 0.5\frac{(y_i - X_{12,i}^\top\beta_{12} - \beta_3\mu_z)^2}{(\tau+\beta_3^2\tau_z)}\} \\
&\quad +\sum_{i=1}^{N}\{r_i X_{12,i}^\top\xi - \log[1+\exp(X_{12,i}^\top\xi)]\} + constant \\
&= l_{12}(y, z_{obs}, x|\beta,\tau,\mu_z,\tau_z) + l_3(r,x|\xi).
\end{aligned}
$$

Note that the log-likelihood function (l) can be decomposed to two parts, with each involves a distinct subset of the parameters. G^{CD} can be written as a block diagnal matrix with two blocks. Therefore, the diagnostic measure CD_i can be decomposed as the sum of two diagnostic measures. With the one-step approximation, we can write

$$CD_i = CD_{i,12} + CD_{i,3},$$

where $CD_{i,12}$ and $CD_{i,3}$ are, respectively, given by

$$
\begin{aligned}
&CD_{i,12} = \\
&\partial_{(\beta,\tau)} l_{12[i]}(y_{[i]}, x_{[i]}, z_{obs[i]}|\hat\beta,\hat\tau)^\top \{-\partial^2_{(\beta,\tau)} l_{12}(y,x,z_{obs}|\hat\beta,\hat\tau)\}^{-1} \partial_{(\beta,\tau)} l_{12[i]}(y_{[i]}, x_{[i]}, z_{obs[i]}|\hat\beta,\hat\tau),
\end{aligned}
$$

$$CD_{i,3} = \partial_\xi l_{3[i]}(r_{[i]}, x_{[i]}|\hat\xi)^\top \{-\partial^2_\xi l_3(r,x|\hat\xi)\}^{-1} \partial_\xi l_{3[i]}(r_{[i]}, x_{[i]}|\hat\xi).$$

As shown in (3.1), the Q-function can be decomposed into three components as follows:

$$Q(\beta, \tau, \alpha, \xi|\eta^{(s)}) = Q_1(\beta, \tau|\eta^{(s)}) + Q_2(\alpha|\eta^{(s)}) + Q_3(\xi|\eta^{(s)}).$$

Therefore, the diagnostic measure QCD_i can be decomposed as the sum of three diagnostic measures. With the one-step approximation, we can write

$$QCD_i = QCD_{i,1} + QCD_{i,2} + QCD_{i,3},$$

where

$$QCD_{i,1} = \partial_{(\beta,\tau)} Q_{1[i]}(\hat{\beta}, \hat{\tau}|\hat{\eta})^\top \{-\partial^2_{(\beta,\tau)} Q_1(\hat{\beta}, \hat{\tau}|\hat{\eta})\}^{-1} \partial_{(\beta,\tau)} Q_{1[i]}(\hat{\beta}, \hat{\tau}|\hat{\eta}),$$

$$QCD_{i,2} = \partial_\alpha Q_{2[i]}(\hat{\alpha}|\hat{\eta})^\top \{-\partial^2_\alpha Q_2(\hat{\alpha}|\hat{\eta})\}^{-1} \partial_\alpha Q_{2[i]}(\hat{\alpha}|\hat{\eta}),$$

$$QCD_{i,3} = \partial_\xi Q_{3[i]}(\hat{\xi}|\hat{\eta})^\top \{-\partial^2_\xi Q_3(\hat{\xi}|\hat{\eta})\}^{-1} \partial_\xi Q_{3[i]}(\hat{\xi}|\hat{\eta}).$$

In the MAR case, $QCD_{i,3}$ is the same as $CD_{i,3}$. The difference between CD_i and QCD_i lies on the difference between $CD_{i,12}$ and $QCD_{i,1} + QCD_{i,2}$. Let τ and τ_z be nuisance parameters. To calculate $CD_{i,12}$ and $QCD_{i,1} + QCD_{i,2}$, we derive the first derivatives and second derivatives of the log-likelihood function (l) and the Q-functions as follows, which will be used later in the simulation study.

For CD, we have

$$
\begin{aligned}
\frac{\partial l}{\partial \beta_{12}} &= \sum_{i\in obs} \frac{(y_i - X_{12,i}^\top \beta_{12} - z_i\beta_3)X_{12,i}}{\tau} + \sum_{i\in mis} \frac{(y_i - X_{12,i}^\top \beta_{12} - \mu_z\beta_3)X_{12,i}}{\tau + \beta_3^2\tau_z},\\
\frac{\partial l}{\partial \beta_3} &= \sum_{i\in obs} \frac{(y_i - X_{12,i}^\top \beta_{12} - z_i\beta_3)z_i}{\tau} + \sum_{i\in mis} \{-\frac{\beta_3\tau_z}{\tau + \beta_3^2\tau_z}\\
&\quad + \frac{(y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)\mu_z}{\tau + \beta_3^2\tau_z} + \frac{(y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)^2\beta_3\tau_z}{(\tau + \beta_3^2\tau_z)^2}\},\\
\frac{\partial l}{\partial \mu_z} &= \sum_{i\in obs} \frac{z_i - \mu_z}{\tau_z} + \sum_{i\in mis} \frac{(y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)\beta_3}{\tau + \beta_3^2\tau_z},\\
\frac{\partial^2 l}{\partial \beta_{12}^2} &= -\sum_{i\in obs} \frac{X_{12,i}^\top X_{12,i}}{\tau} - \sum_{i\in mis} \frac{X_{12,i}^\top X_{12,i}}{\tau + \beta_3^2\tau_z},\\
\frac{\partial^2 l}{\partial \beta_{12}\partial \beta_3} &= -\sum_{i\in obs} \frac{z_i X_{12,i}}{\tau} - \sum_{i\in mis} \{\frac{\mu_z X_{12,i}}{\tau + \beta_3^2\tau_z} + \frac{2\beta_3\tau_z(y_i - X_{12,i}^\top \beta_{12} - \mu_z\beta_3)X_{12,i}}{(\tau + \beta_3^2\tau_z)^2}\},\\
\frac{\partial^2 l}{\partial \beta_{12}\partial \mu_z} &= -\sum_{i\in mis} \frac{\beta_3 X_{12,i}}{\tau + \beta_3^2\tau_z},\\
\frac{\partial^2 l}{\partial \beta_3^2} &= -\sum_{i\in obs} \frac{z_i z_i}{\tau} - \sum_{i\in mis} \{\frac{\tau_z + \mu_z\mu_z}{\tau + \beta_3^2\tau_z} + \frac{2\beta_3\tau_z[-\beta_3\tau_z + (y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)\mu_z]}{(\tau + \beta_3^2\tau_z)^2}\\
&\quad + \frac{-(y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)^2\tau_z + 2(y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)\mu_z\beta_3\tau_z}{(\tau + \beta_3^2\tau_z)^2}\\
&\quad + \frac{(y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)^2 4\beta_3^2\tau_z^2}{(\tau + \beta_3^2\tau_z)^3}\},\\
\frac{\partial^2 l}{\partial \beta_3\partial \mu_z} &= \sum_{i\in mis} \{\frac{y_i - X_{12,i}^\top \beta_{12} - 2\beta_3\mu_z}{\tau + \beta_3^2\tau_z} - \frac{2(y_i - X_{12,i}^\top \beta_{12} - \beta_3\mu_z)\beta_3^2\tau_z}{(\tau + \beta_3^2\tau_z)^2}\},\\
\frac{\partial^2 l}{\partial \mu_z^2} &= -\sum_{i\in obs} \frac{1}{\tau_z} - \sum_{i\in mis} \frac{\beta_3^2}{\tau + \beta_3^2\tau_z}.
\end{aligned}
$$

For QCD, we have

$$\begin{aligned}
\frac{\partial Q_1}{\partial \beta_{12}} &= -\frac{1}{2\tau}(-2)X_{12}^{\top}(y - X_{12}\beta_{12}) - \frac{1}{2\tau}2X_{12}^{\top}Z^{(s)}\beta_3, \\
\frac{\partial Q_1}{\partial \beta_3} &= -\frac{1}{2\tau}(-2)Z^{(s)\top}(y - X_{12}\beta_{12}) - \frac{1}{2\tau}2A^{(s)}\beta_3, \\
\frac{\partial Q_2}{\partial \mu_z} &= \frac{1}{\tau_z}\sum_{i=1}^{N}[z_i I(i \in obs) + \mu_{zm,i}^{(s)} I(i \in mis)] - \frac{N\mu_z}{\tau_z}, \\
\frac{\partial^2 Q_1}{\partial \beta_{12}^2} &= -\frac{X_{12}^{\top}X_{12}}{\tau}, \ \frac{\partial^2 Q_1}{\partial \beta_{12}\beta_3} = -\frac{X_{12}^{\top}Z^{(s)}}{\tau}, \ \frac{\partial^2 Q_1}{\partial \beta_3^2} = -\frac{A^{(s)}}{\tau}, \ \frac{\partial^2 Q_2}{\partial \mu_z^2} = -\frac{N}{\tau_z}.
\end{aligned}$$

3.5 Scaled Cook's Distance

Following (Zhu et al., 2012a), we introduce the scaled Cook's distance for GLM with missing covariates as below.

The scaled Cook's distance for matching mean and standard deviation is defined based on the log-likelihood function and the Q-function, respectively, as follows:

$$\begin{aligned}
SCD(I) &= (CD(I) - E[CD(I)|M])/Std[CD(I)|M], \\
SQCD(I) &= (QCD(I) - E[QCD(I)|M])/Std[QCD(I)|M],
\end{aligned}$$

where the expectation is taken with respect to M, the current model fitted to the data; and I is the subset that we would like to assess the influence. This scaled Cook's distance measures the standardized influential level of the subset I when M is true. A large value of SCD or SQCD indicates that the subset I is relatively influential.

We use the parametric bootstrap method to compute $E[CD(I)|M]$, $Std[CD(I)|M]$, $E[QCD(I)|M]$ and $Std[QCD(I)|M]$ as follows:

Step 1. We use $\widehat{M} = \{f(x_i, z_i, r_i, y_i; \hat{\eta})\}$ to approximate the model $M = \{f(x_i, z_i, r_i, y_i; \eta)\}$, generate a random sample Y^s from $f(x_i, z_i, r_i, y_i; \hat{\eta})$, and then calculate $CD(I)^s$ and $QCD(I)^s$ for subset I.

Step 2. By repeating this process S times, we can obtain a sample $\{CD\{I\}^s : s =$

$1, \ldots, S\}$ and then we use its empirical mean $\overline{CD(I)} = \sum_{s=1}^{S} CD(I)^s/S$ to approximate $E[CD(I)|M]$, and use its empirical standard deviation to approximate $Std[CD(I)|M]$. Similarly, we approximate $E[QCD(I)|M]$ and $Std[QCD(I)|M]$ using the empirical mean and standard deviation from the sample $\{QCD\{I\}^s : s = 1, \ldots, S\}$.

Moreover, we calculate the two probabilities as follows:

$$P_r(I) = \sum_{s=1}^{S} I(CD(I)^s \leq CD(I))/S,\ QP_r(I) = \sum_{s=1}^{S} I(QCD(I)^s \leq QCD(I))/S,$$

where $I(.)$ is an indicator function of a set. Because $CD(I)^s$ can be regarded as the "true" Cook's distance when the model is true, a large value of $P_r(I)$ (or $QP_r(I)$) indicates I as influential.

3.6 Simulation Studies Using One Dataset

We conducted simulation studies to compare the Cook's distance and scaled Cook's distance based on the log-likelihood function and Q-function. We first simulated data for the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$, where the ϵ_i's are i.i.d. and $\epsilon_i \sim N(0, \tau), x_i \sim N(0, 1), z_i \sim N(\mu, \tau_z), i = 1, \ldots, n$. We set $n = 100$ subjects, $\beta_0 = \beta_1 = \beta_2 = 1, \tau = \tau_z = 1$, and $\mu = 0$. The response y_i and the covariate x_i are completely observed for $i = 1, \ldots, n$, but the covariate z_i may be missing for some subjects.

We consider the following three missing mechanisms:

i) The missingness is only dependent on x_i (MAR). We set z_i as missing if $r_i = 1$. The missingness variable r_i follows a Bernoulli distribution with $\text{logit}(\text{prob}(r_i = 1)) = \xi_0 + \xi_1 x_i$. We set $\xi_0 = -1.5$, and $\xi_1 = 1.0$ to obtain an average missingness fraction of 20%.

ii) The missingness is dependent on both x_i and z_i (MNAR). The missingness variable r_i follows a Bernoulli distribution with $\text{logit}(\text{prob}(r_i = 1)) = \xi_0 + \xi_1 x_i + \xi_2 z_i$. We set $\xi_0 = -1.5$, $\xi_1 = 1.0$, and $\xi_2 = 1.0$. The missing rate is approximately 20%.

iii) The missingness is dependent on both x_i and z_i (MNAR) with a larger missing rate. The missingness variable r_i follows a Bernoulli distribution with logit(prob($r_i = 1$)) $= \xi_0 + \xi_1 x_i + \xi_2 z_i$. We set $\xi_0 = -1.5$, $\xi_1 = 1.0$, and $\xi_2 = 15$. The missing rate is approximately 40%.

For each missing mechanism, we generated data to assess how diagnostic measures work in four scenarios i) no outlier, ii) having an outlier(s) in the covariate with missing data (adding outliers in the z domain), iii) having an outlier(s) in the response variable (adding outliers in the y domain), and iv) having an outlier(s) in the covariate without missing data (adding outliers in the x domain). For each scenario with outliers, we started with the simple scenario which included only one outlier by replacing z_{100} with $z_{100} + 5$, replacing y_{100} with $y_{100} + 5$, and replacing x_{100} with $x_{100} + 5$, respectively. Then, we assess the scenario including multiple outliers. When the outliers are all in the same directions, it is possible that there may be a confounding effect that increases the impact of outliers on the model. In the other hand, when the outliers are in the opposite directions, their impact on the model may be diminished. To assess how diagnostic measures perform when multiple outliers exist in the GLM with missing covariates, we modified the last two observations indexed by [99] and [100]. We added 5 to both observations [99] and [100] to get two outliers in the same directions. To get two outliers in the opposite directions, we added 5 to observation [100] and subtracted 5 to observation [99]. Summary of simulation scenarios are presented in Table 3.1.

Table 3.1: Summary of Simulation Scenarios

No.	Description	Missing Mechanism	Outlier
MAR scenario			
1	Missing mechanism is dependent on X only	logit(prob(r_i=1)) = $\xi_0 + \xi_1 x_i$, where r_i=1 when z_i is not observed. We set ξ_0 = -1.5, and ξ_1 =1.0. (~20% missing)	none
2a	Missing mechanism is dependent on X only, plus outlier(s) in Z domain		Change z_{100} to z_{100} + 5.0
2b			Change z_{100} to z_{100} + 5.0 Change z_{99} to z_{99} + 5.0
2c			Change z_{100} to z_{100} + 5.0 Change z_{99} to z_{99} - 5.0
3a	Missing mechanism is dependent on X only, plus outlier(s) in Y domain		Change y_{100} to y_{100} + 5.0
3b			Change y_{100} to y_{100} + 5.0 Change y_{99} to y_{99} + 5.0
3c			Change y_{100} to y_{100} + 5.0 Change y_{99} to y_{99} - 5.0
4a	Missing mechanism is dependent on X only, plus outlier(s) in X domain		Change x_{100} to x_{100} + 5.0
4b			Change x_{100} to x_{100} + 5.0 Change x_{99} to x_{99} + 5.0
4c			Change x_{100} to x_{100} + 5.0 Change x_{99} to x_{99} - 5.0
MNAR scenario			
5	all missing mechanism is dependent on Z	logit(prob(r_i=1)) = $\xi_0 + \xi_1 x_i + \xi_2 z_i$, where r_i=1 when z_i is not observed. We set ξ_0 = -1.5, and ξ_1 =1.0, ξ_2 =1.0. (~20% missing)	none
6a	all missing mechanism is dependent on Z, plus outlier(s) in Z domain		Change z_{100} to z_{100} + 5.0
6b			Change z_{100} to z_{100} + 5.0 Change z_{99} to z_{99} + 5.0
6c			Change z_{100} to z_{100} + 5.0 Change z_{99} to z_{99} - 5.0
7a	all missing mechanism is dependent on Z, plus outlier(s) in Y domain		Change y_{100} to y_{100} + 5.0
7b			Change y_{100} to y_{100} + 5.0 Change y_{99} to y_{99} + 5.0
7c			Change y_{100} to y_{100} + 5.0 Change y_{99} to y_{99} - 5.0
8a	all missing mechanism is dependent on Z, plus outlier in X domain		Change x_{100} to x_{100} + 5.0
8b			Change x_{100} to x_{100} + 5.0 Change x_{99} to x_{99} + 5.0
8c			Change x_{100} to x_{100} + 5.0 Change x_{99} to x_{99} - 5.0
9	all missing mechanism is dependent on Z	logit(prob(r_i=1)) = $\xi_0 + \xi_1 x_i + \xi_2 z_i$, where r_i=1 when z_i is not observed. We set ξ_0 = -1.5, and ξ_1 =1.0, ξ_2 =15. (~40% missing)	none
10a	all missing mechanism is dependent on Z, plus outlier(s) in Z domain		Change z_{100} to z_{100} + 5.0
10b			Change z_{100} to z_{100} + 5.0 Change z_{99} to z_{99} + 5.0
10c			Change z_{100} to z_{100} + 5.0 Change z_{99} to z_{99} - 5.0
11a	all missing mechanism is dependent on Z, plus outlier(s) in Y domain		Change y_{100} to y_{100} + 5.0
11b			Change y_{100} to y_{100} + 5.0 Change y_{99} to y_{99} + 5.0
11c			Change y_{100} to y_{100} + 5.0 Change y_{99} to y_{99} - 5.0
12a	all missing mechanism is dependent on Z, plus outlier in X domain		Change x_{100} to x_{100} + 5.0
12b			Change x_{100} to x_{100} + 5.0 Change x_{99} to x_{99} + 5.0
12c			Change x_{100} to x_{100} + 5.0 Change x_{99} to x_{99} - 5.0

To analyze these data, we used the general linear model with missing covariate described previously in Example 1. This model is based on the assumption of MAR. We obtained $\hat{\beta}, \hat{\tau}, \hat{\alpha}, \hat{\xi}$ using the EM algorithm. We calculated CD and QCD for deleting each subject using the one-step approximation. We also calculated the scaled CD and QCD based on empirical mean and standard deviation from the parametric bootstrap sample. Correspondingly, we calculated the degree of perturbation for each subject with and without missing data. Note that, as shown previously, the component related to ξ is the same for CD and QCD ($CD_{i,3} = QCD_{i,3}$). The degree of perturbation related to ξ (3.5) is the same for subjects with and without missing data. In addition, in practice, the primary interest is often on β. Therefore, for comparison purpose, we did not include the component involving ξ in our analysis.

Figure 3.1 presents the results for the scenario with no outlier. Both the CD and QCD are small for all subjects. QCD is very close to CD when missing mechanism is correctly specified (missing data is MAR; the top plots). They are also very close when the model is misspecified (missing data is MNAR; the middle plots show the scenario with a missing rate of approximately 20%, and the bottom plots show the scenario with a missing rate of approximately 40%).

Figure 3.1: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with No Outlier. The three plots in the top are from the scenario of MAR (Scenario 1). The three plots in the middle are from the scenario of MNAR (Scenario 5). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 9).

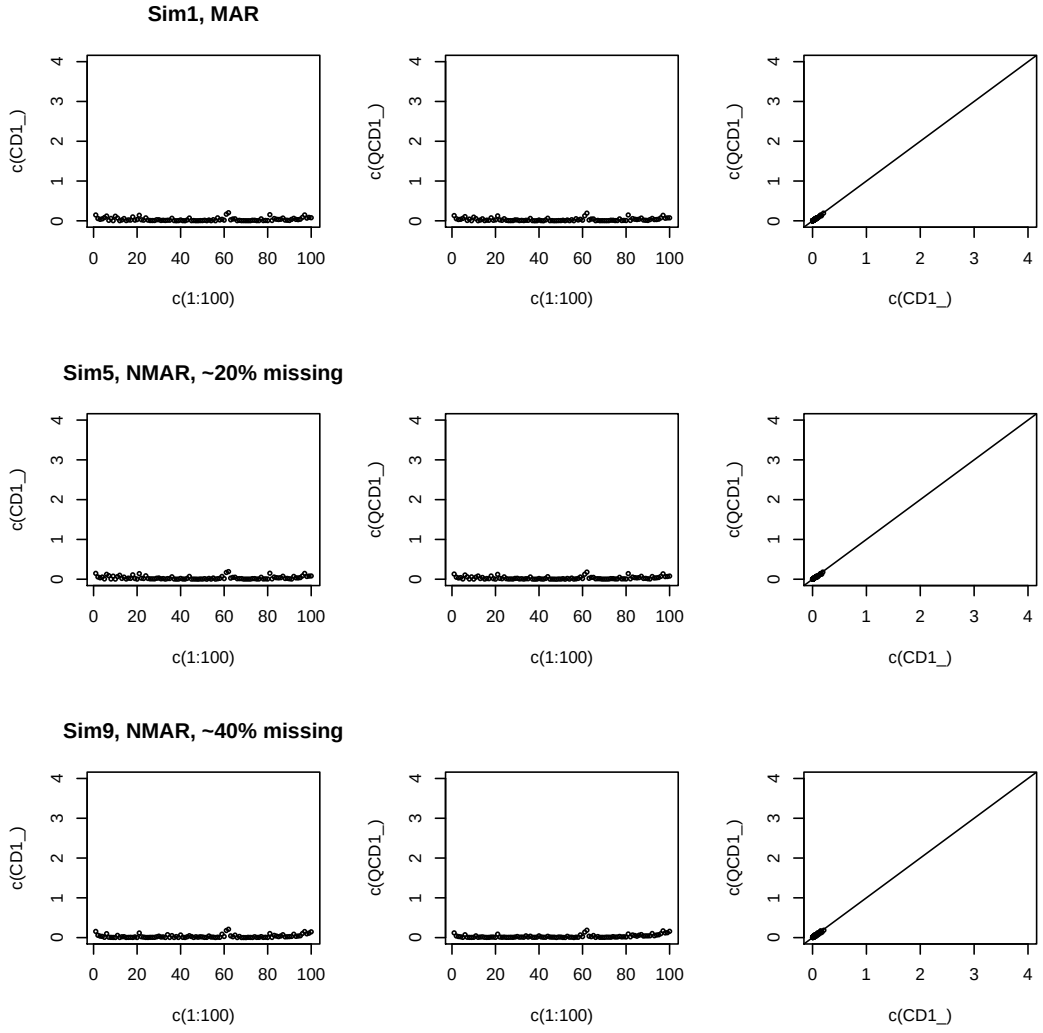


Figure 3.2 presents the results with one outlier in the covariate with missing data, z domain (Scenario 2a). The outlier, observation [100], has a large value in CD (1.8). In the scenario with two outliers in the same direction into the simulation data in z domain (Figure 3.3, Scenario 2b), both outliers have larger CD comparing to other observations (1.3 for observation [99] and 1.2 for observation [100]). However, the magnitude of CD for both observations is smaller than that in Scenario 2a. In the scenario with two outliers in the opposite directions in the z domain (Figure 3.3, Scenario 2c), the two outliers have different CD (2.1 and 0.9) even though the perturbation we added are the same. When the true missing mechanism is MNAR (Scenarios 6a, 6b, 6c), the results are similar to the Scenarios 2a, 2b, and 2c, in which the missing mechanism is correctly specified. However, when the true missing mechanism is MNAR and the missing rate is approximately 40% (Scenarios 10a, 10b, and 10c), notable differences are seen in the value of CD comparing to Scenarios 2a, 2b, and 2c. In all these scenarios, QCD is very close to CD except for the outliers in the scenarios with misspecified missing mechanism and a large amount of missing data (Scenarios 10a, 10b, and 10c).

Figures 3.5 to 3.7 display results with outliers in the response, y domain. The outliers have larger CD comparing to other observations. In the scenario with only one outlier, we see a larger CD in the scenario with misspecified missing mechanism and a large amount of missing data (Scenarios 11a) comparing to Scenarios 3a and 7a in Figure 3.5. In the scenarios with two outliers in the same direction (Figure 3.6), the magnitude of CD for each observation is smaller than that in the scenario with one outlier. However, in the scenarios with two outliers in the opposite direction (Figure 3.7), the magnitude of CD for observation [99] is larger than that in the scenarios with only one outlier. The CD is also sensitive with the missing mechanism and missing rate when the outliers are in the response domain.

Figure 3.2: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with One Outlier in z Domain. The three plots in the top are from the scenario of MAR (Scenario 2a). The three plots in the middle are from the scenario of MNAR (Scenario 6a). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 10a).

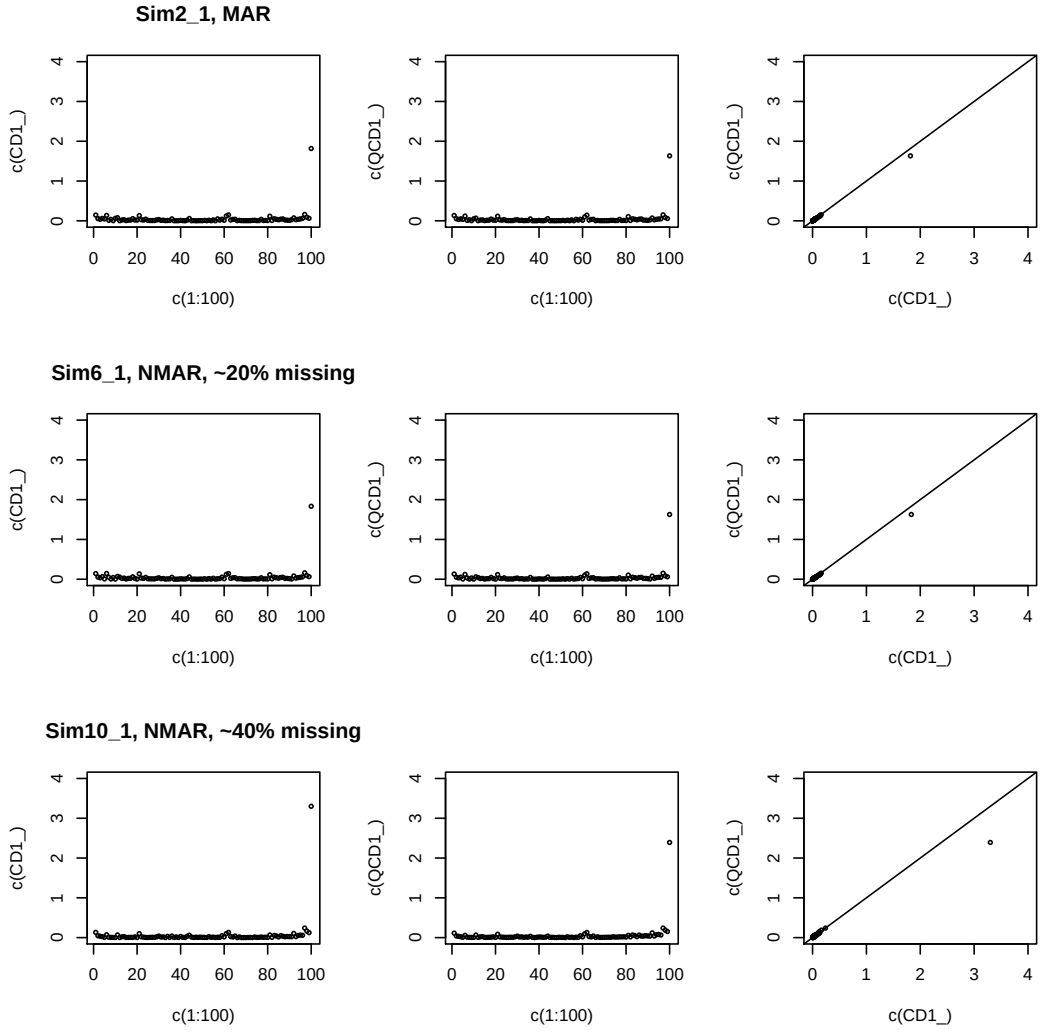


Figure 3.3: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with Two Outliers in z Domain in Same Direction. The three plots in the top are from the scenario of MAR (Scenario 2b). The three plots in the middle are from the scenario of MNAR (Scenario 6b). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 10b).

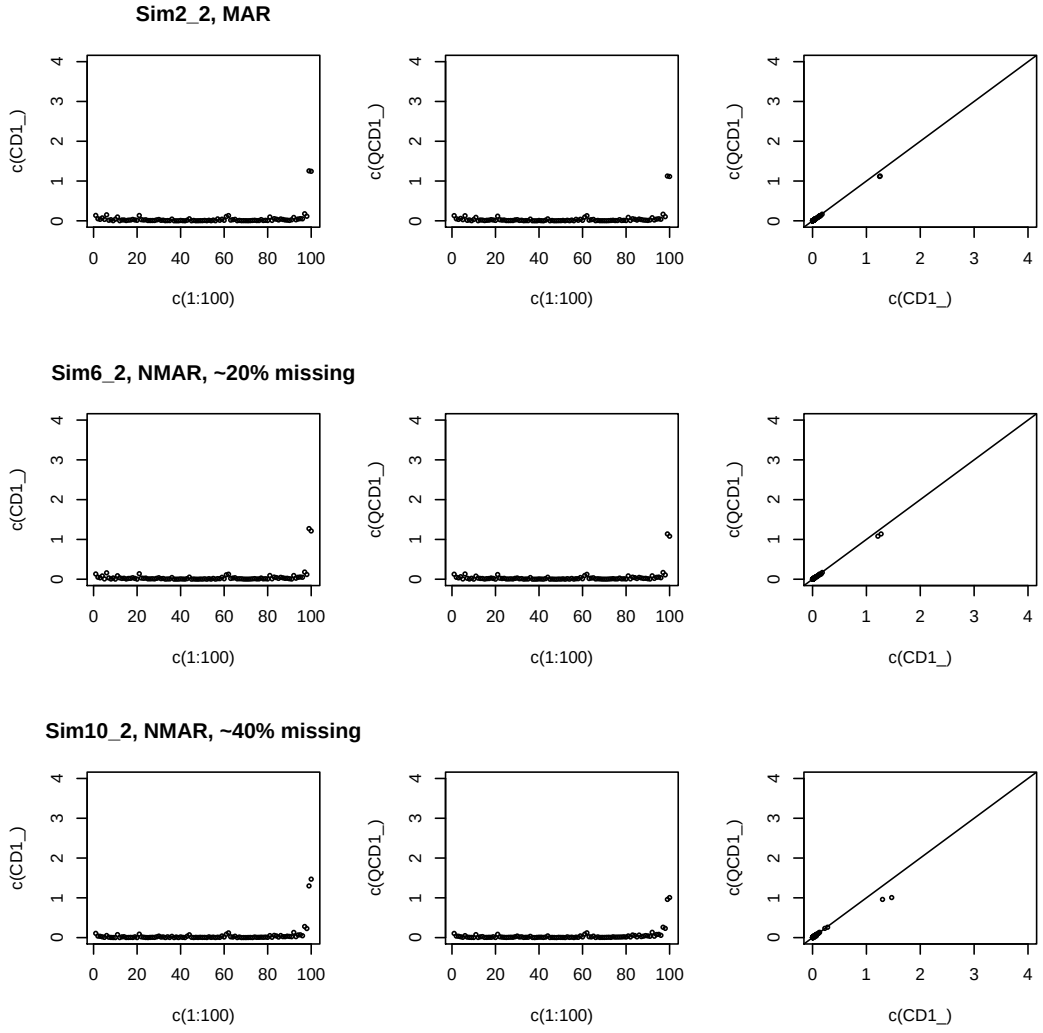


Figure 3.4: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with Two Outliers in z Domain in Opposite Direction. The three plots in the top are from the scenario of MAR (Scenario 2c). The three plots in the middle are from the scenario of MNAR (Scenario 6c). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 10c).

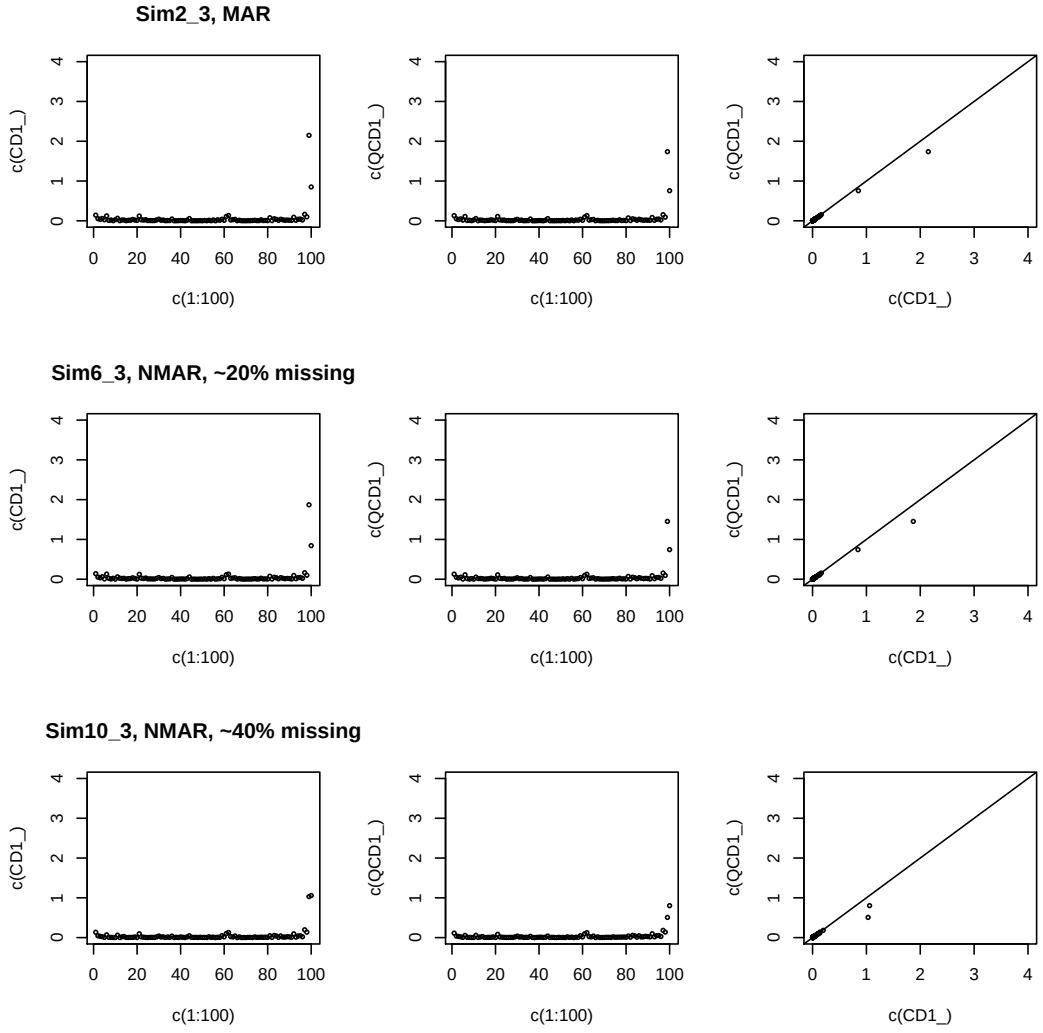


Figure 3.5: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with One Outlier in y Domain. The three plots in the top are from the scenario of MAR (Scenario 3a). The three plots in the middle are from the scenario of MNAR (Scenario 7a). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 11a).

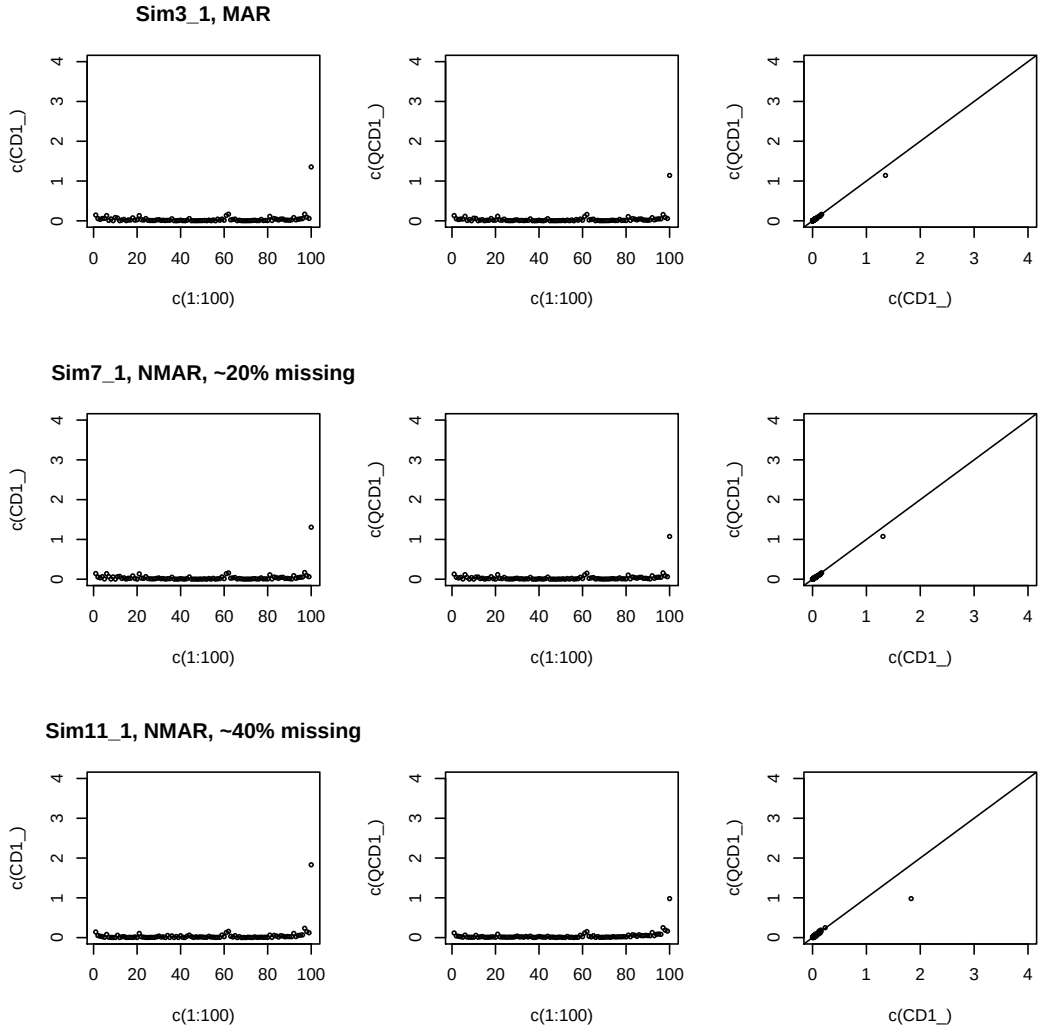


Figure 3.6: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with Two Outliers in y Domain in Same Direction. The three plots in the top are from the scenario of MAR (Scenario 3b). The three plots in the middle are from the scenario of MNAR (Scenario 7b). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 11b).

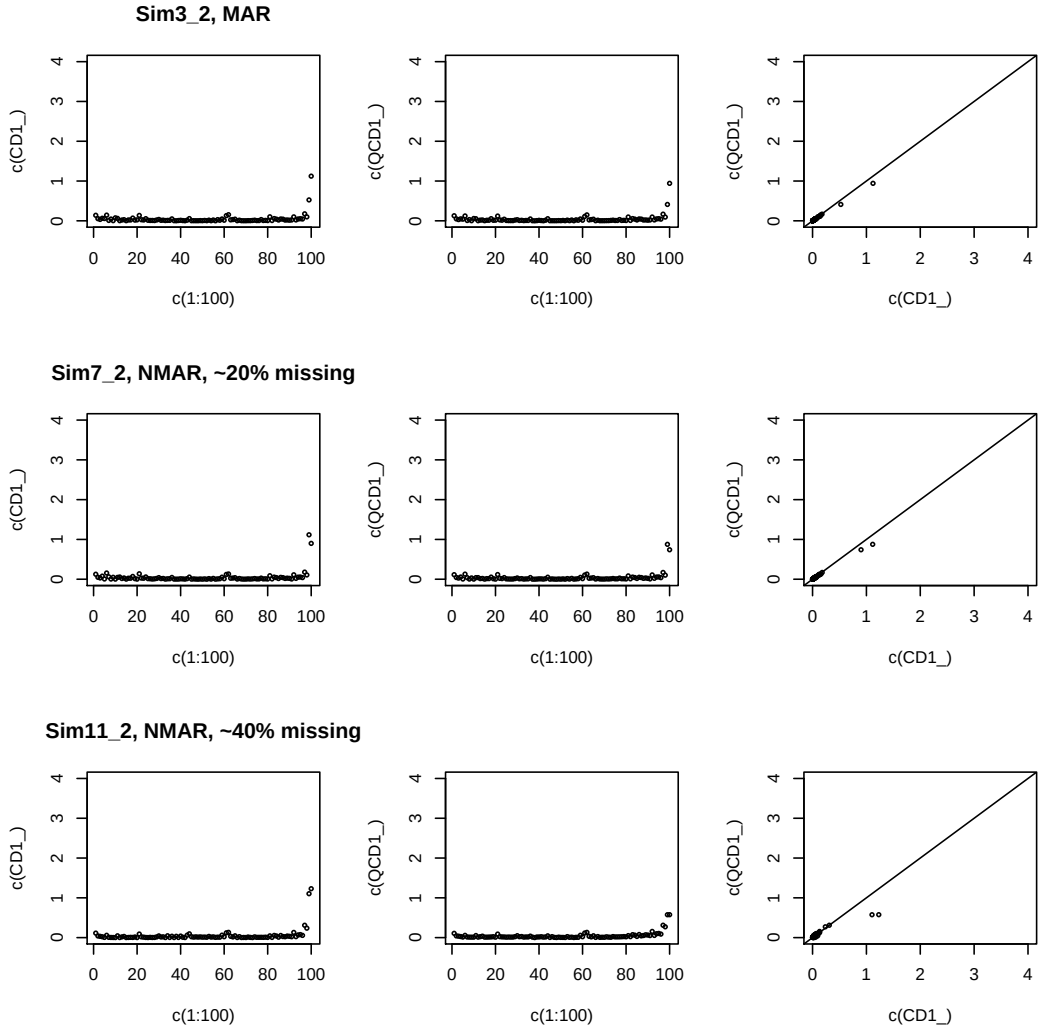
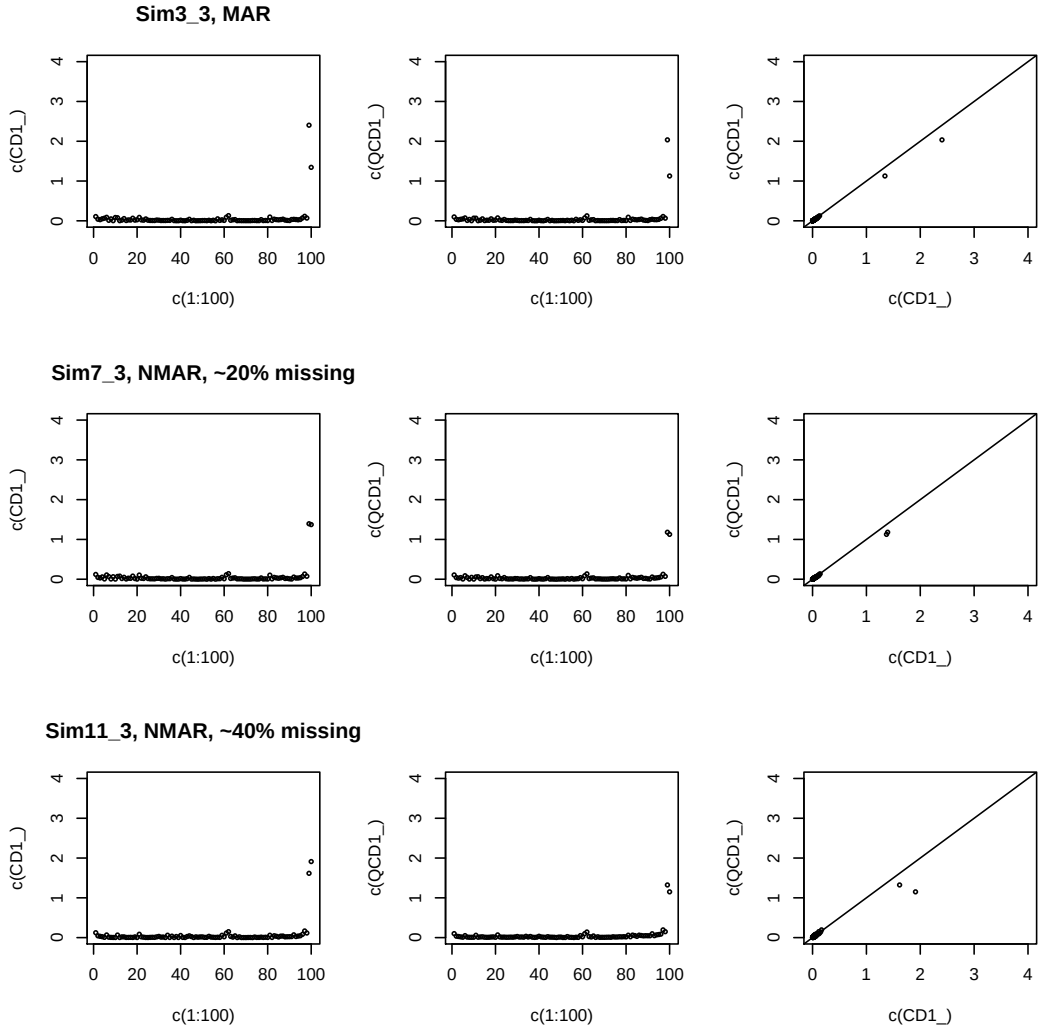


Figure 3.7: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with Two Outliers in y Domain in Opposite Direction. The three plots in the top are from the scenario of MAR (Scenario 3c). The three plots in the middle are from the scenario of MNAR (Scenario 7c). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 11c).



Figures 3.8 to 3.10 display results with outliers in the covariate without missing data, x domain. The outliers in x domain also have larger CD comparing to other observations. In the scenarios with two outliers, in same or opposite directions, the magnitude of CD for each outlier is smaller than that in the scenario with only one outlier. It appears that the CD is not sensitive to missing mechanism and missing rate when the outliers are in the covariate without missing data.

For each scenario we examined, we also present the index plots for QCD and the scatter plot of CD and QCD. In all these scenarios, QCD is very close to CD for non-outliers regardless of missing mechanism and missing rate. For outliers, QCD is slightly smaller than CD, but still sufficient to identify influential observations.

Figure 3.8: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with One Outlier in x Domain. The three plots in the top are from the scenario of MAR (Scenario 4a). The three plots in the middle are from the scenario of MNAR (Scenario 8a). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 12a).

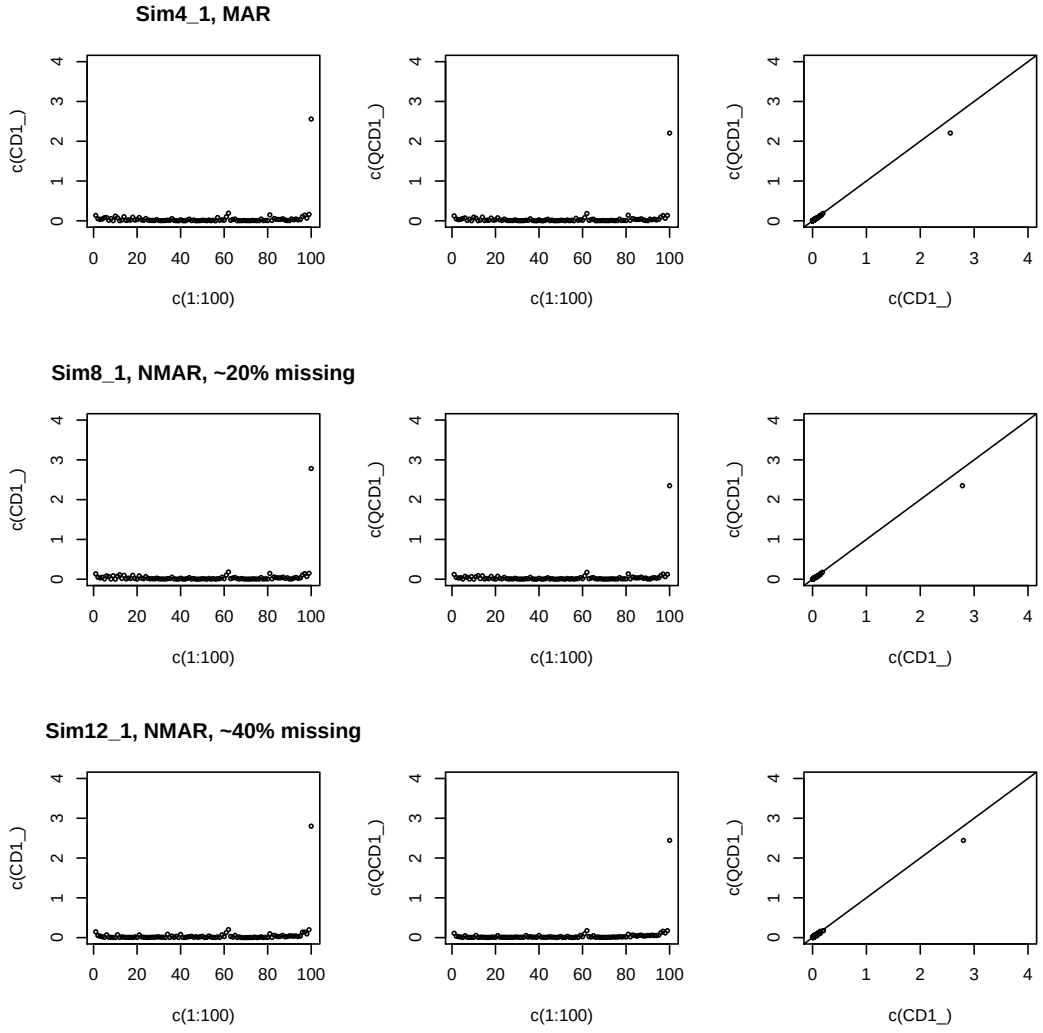


Figure 3.9: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with Two Outliers in x Domain in Same Direction. The three plots in the top are from the scenario of MAR (Scenario 4b). The three plots in the middle are from the scenario of MNAR (Scenario 8b). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 12b).

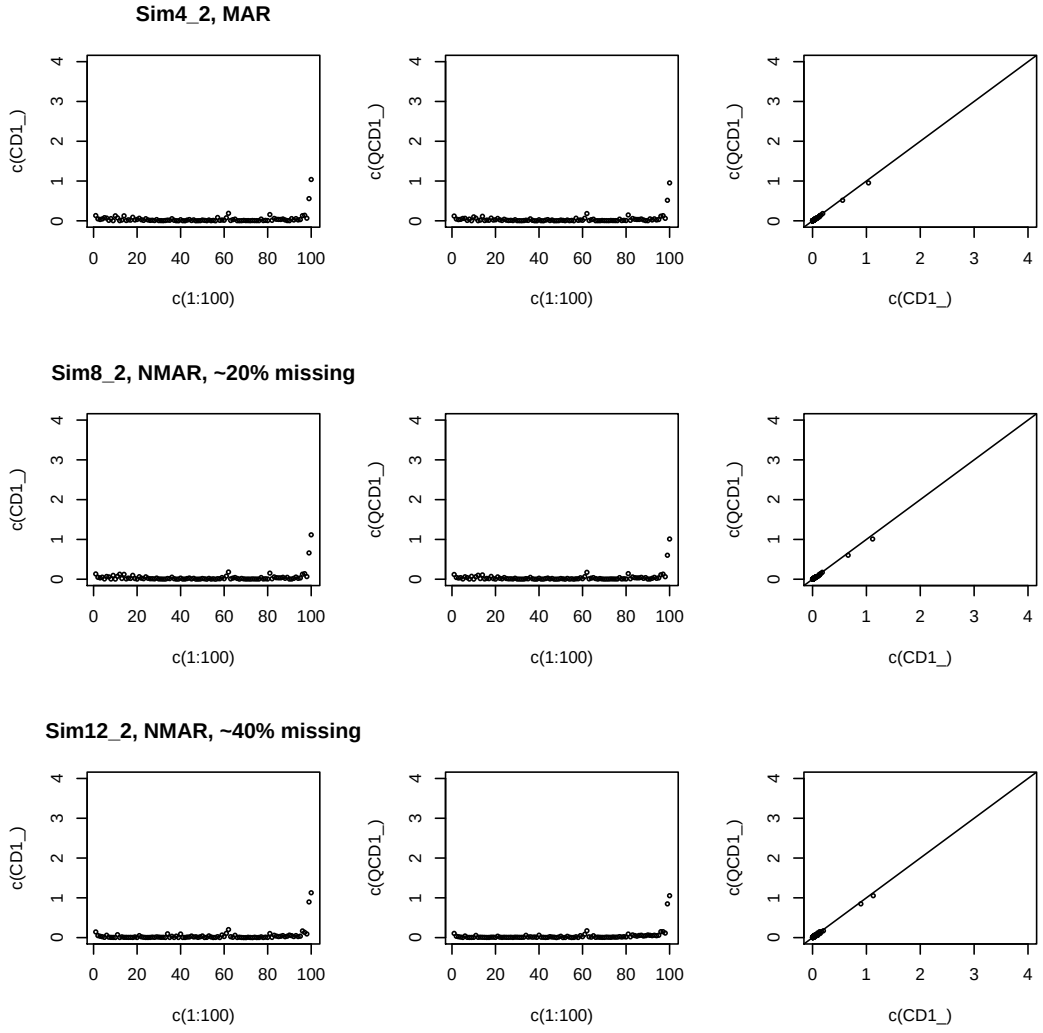
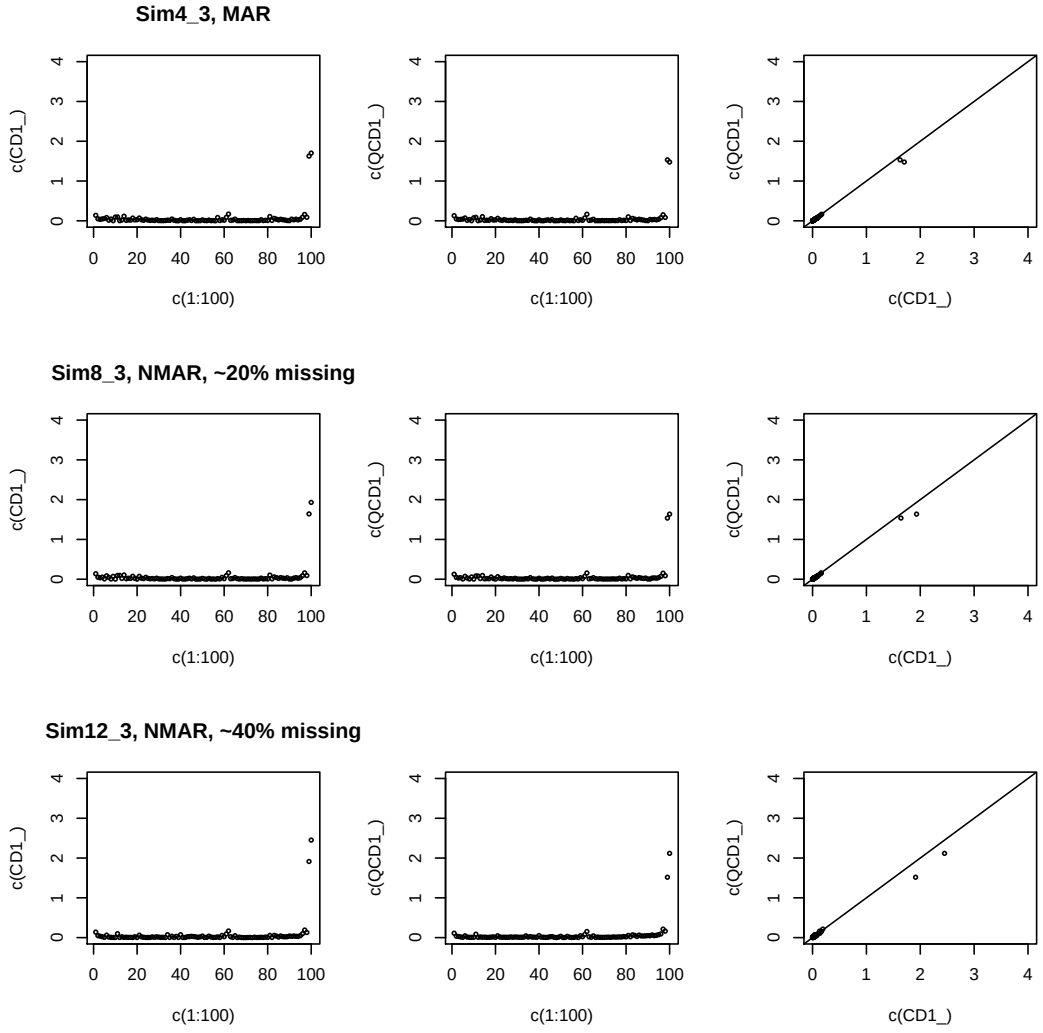


Figure 3.10: Index Plots and Scatter Plots of CD and QCD from the Simulation Data with Two Outliers in x Domain in Opposite Direction. The three plots in the top are from the scenario of MAR (Scenario 4c). The three plots in the middle are from the scenario of MNAR (Scenario 8c). The three plots in the bottom are from the scenario of MNAR with a greater missing rate (Scenario 12c).



3.7 Additional Simulation Studies

To further assess the performance of CD, QCD, scaled CD/QCD and their relationship with degree of perturbation when the covariate has missing value, additional simulations were conducted. For scenario 1, which does not include outliers, we generated 100 sample data with fixed x and r, and calculated CD, QCD, scaled CD/QCD, Pr, QPr, and degree of perturbation. First, we generated the box plots for CD, scaled CD, and Pr of each observation (Figure 3.11). The CDs are always positive, and the mean CDs (green dots) for all observations are close to zero. The scaled CD distributed around zero. The mean Pr is around 0.5 in the case without outliers. The $Q-$function based approximation QCD is slightly smaller than CD in means, but the mean scaled QCD and QPr are very close to scaled CD and Pr.

The mean degree of perturbation by x plot is shown in Figure 3.12. It is clearly shown that the degree of perturbation is a function of x. In addition, subjects with missing z have a much smaller degree of perturbation. Correspondingly, the subjects with missing data are likely to have a smaller CD. The mean CD by mean degree of perturbation plot shows a positive correlation between CD and degree of perturbation, while the scaled CD accounted for this correlation as shown in the mean scaled CD by mean degree of perturbation plot.

We also conducted the simulation that the missing mechanism also depends on z (Scenario 5). Results are similar to Scenario 1.

Figure 3.11: Box Plots for CD, Scaled CD, Pr and Scatter Plots with Q-based Approximation. Results from 100 Simulation Samples for Scenario 1. Green dots indicates means. Red bars (dots) are observed, and blue bars (dots) are subjects with missing z.

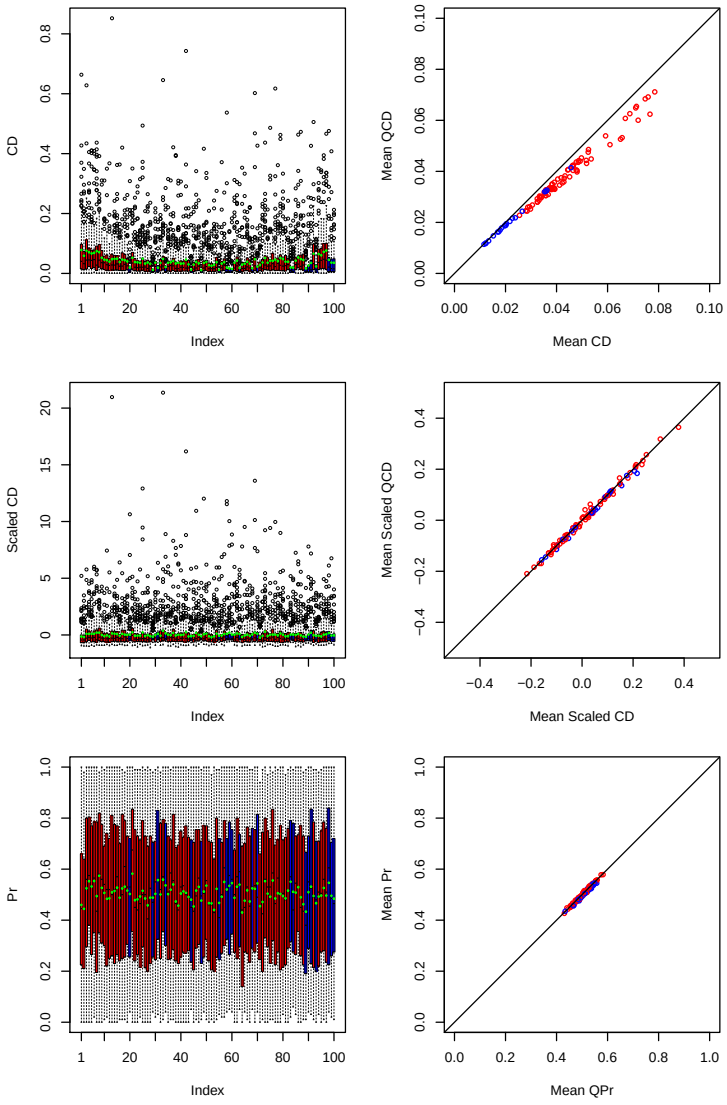
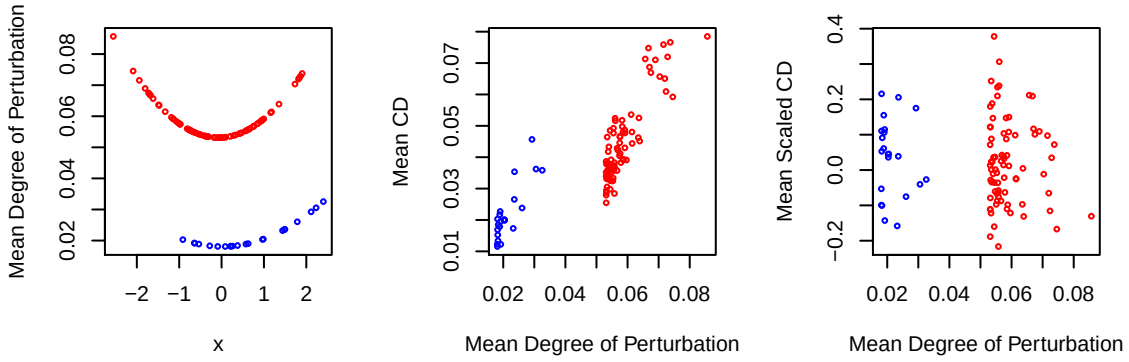


Figure 3.12: Scatter Plots of Mean Degree of Perturbation with x, Mean CD, and Mean Scaled CD. Results from 100 Simulation Samples for Scenario 1. Red dots are observed subjects, and blue dots are subjects with missing z.



Then, we further evaluated the scenarios with outliers similar to the scenarios 2, 3, 4 for correctly specified missing mechanism and 6, 7, 8 for misspecified missing mechanism. For each scenario, we added 5 outliers to assess how diagnostic measures perform when multiple outliers exist in the GLM with missing covariates. We evaluated two scenarios: a) all outliers are in the same direction by adding 5 to the value; and b) outliers are in the opposite direction by adding or subtracting 5 to the value. We ordered the simulated data by x before adding the outliers. We chose one subject with a close to median x and observed z, two subjects with a close to first quartile of x and two subjects with a close to third quartile of x. Because the probability of being missing on z increased as x increased, we chose the two subjects near the first quartile of x with observed z and chose the two subjects near the third quartile of x with missing z. To generate the outliers, we replace the selected subjects with the value plus 5, except that for scenario b, the outlier near the first quartile with a larger index was replaced by the value subtracting 5 and same for the subject near the third quartile with a larger index. For each of the scenarios 2, 3, 4, 100 sample data were generated with the same x and r (the missingness

of z). For scenarios 6, 7, 8, the sample data also had the same z because the missingness depends on z.

Selected results are presented in Figures 3.13 to 3.18. In these index box plots, red bars (dots) are for observed subjects, and blue bars (dots) are for subjects with missing z. Green dots indicates means. When adding outliers in z domain (Figure 3.13), the three outliers had outstanding value of CD, scaled CD, and Pr. The other two outliers are not observed and therefore are not influential. When the missing mechanism is correctly specified, all non-outliers have a small CD or scaled CD. However, when the missing mechanism is misspecified, some non-outliers, with missing z or observed z have a CD or scaled CD larger than other subjects. This is more obvious in the Pr box plot. The same results are seen when the outliers are in the different direction (Figure 3.14).

Figure 3.13: Box Plots for CD, Scaled CD, and Pr. Results from 100 Simulation Samples for Scenarios 2a (MAR, left panels) and 6a (MNAR, right panels) - 5 Outliers in z Domain in the Same Direction. Green dots indicates means. Red bars (dots) are observed, and blue bars (dots) are subjects with missing z.

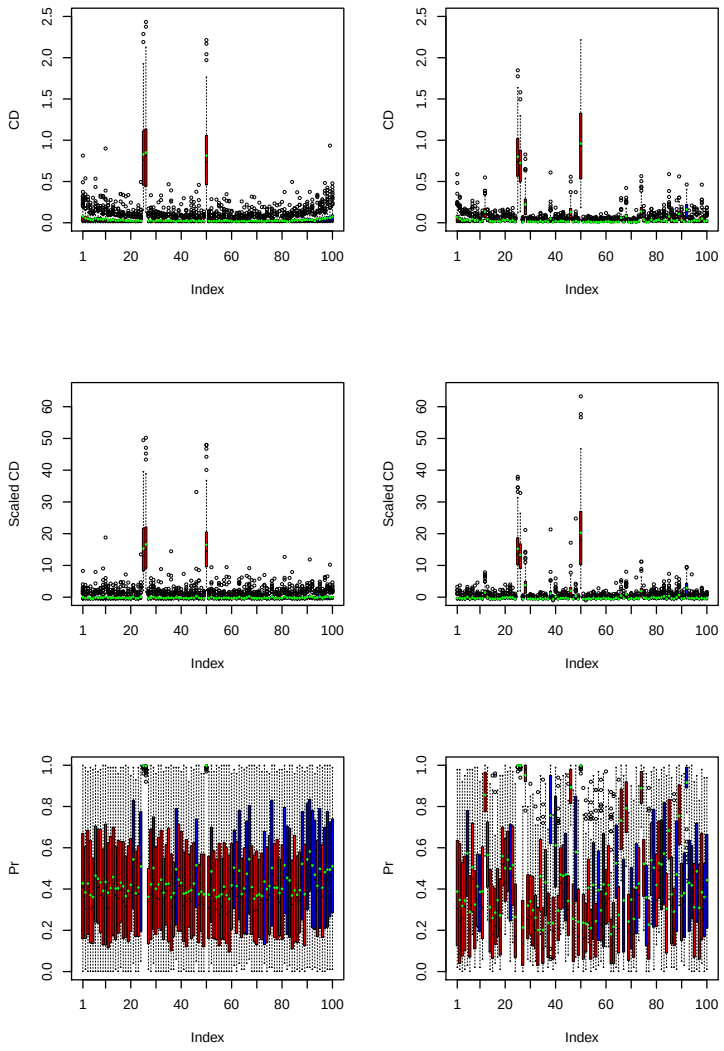
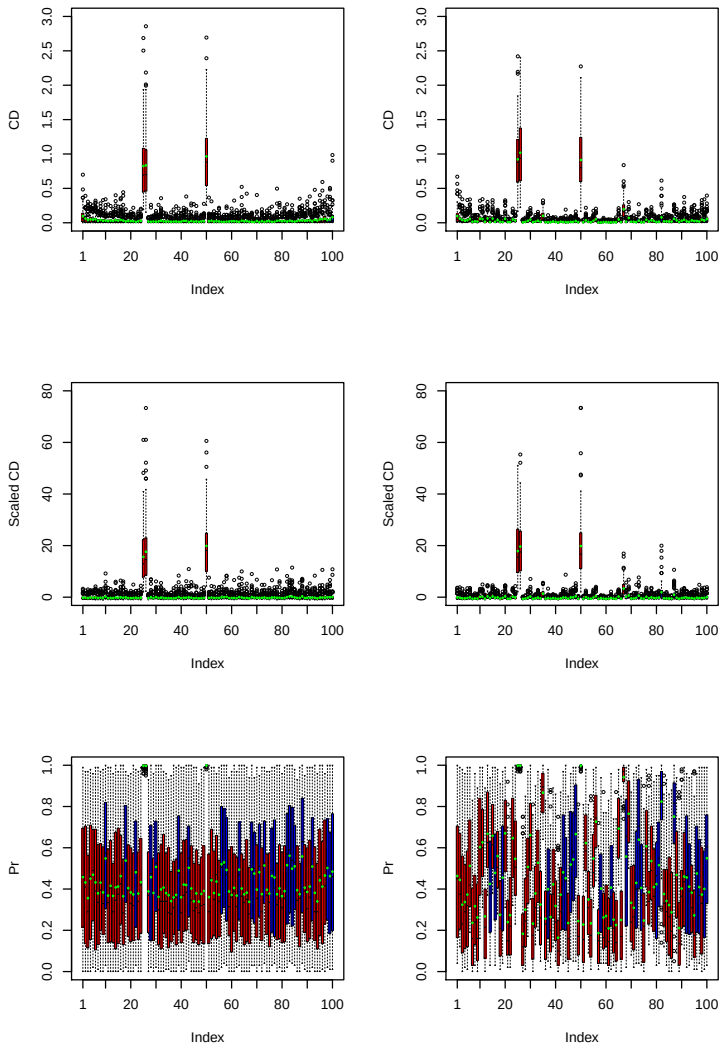


Figure 3.14: Box Plots for CD, Scaled CD, and Pr. Results from 100 Simulation Samples for Scenarios 2b (MAR, left panels) and 6b (MNAR, right panels) - 5 Outliers in z Domain in the Opposite Direction. Green dots indicates means. Red bars (dots) are observed, and blue bars (dots) are subjects with missing z.



When adding outliers in y domain (Figure 3.15), in addition to the three outliers with observed z (red bars), the two outliers with missing z (blue bars) also have an outstanding value of CD, scaled CD, and Pr. As seen for Scenarios 2b and 6b, when the missing mechanism is correctly specified, all non-outliers have a small CD or scaled CD. However, when the missing mechanism is misspecified, some non-outliers, have a CD or scaled CD larger than other subjects, even similar to the true outliers (e.g. subject 43 has a similar CD to the true outlier subject 50). In addition, the magnitude of the diagnostic measures for the true outliers can be larger (e.g., subject 81) or smaller (e.g., subject 50) than the value when the missing mechanism is correctly specified, although they are still much larger than most non-outliers. The same results are seen when the outliers are in the different direction (Figure 3.16).

When adding outliers in x domain (Figure 3.17), all five outliers have an outstanding value of CD, scaled CD, and Pr. However, the influence of the two outliers with missing z (blue bars) are much smaller. As seen for Scenarios 2b and 6b (outliers in y domain), when the missing mechanism is correctly specified, all non-outliers have a small CD or scaled CD. However, when the missing mechanism is misspecified, some non-outliers, have a CD or scaled CD larger than other subjects, even similar to the true outliers. In addition, the magnitude of the diagnostic measures for the true outliers can be different from the value when the missing mechanism is correctly specified, and in some cases the specified outliers are not very influential. The similar results are seen when the outliers are in the different direction (Figure 3.18).

Figure 3.15: Box Plots for CD, Scaled CD, and Pr. Results from 100 Simulation Samples for Scenarios 3a (MAR, left panels) and 7a (MNAR, right panels) - 5 Outliers in y Domain in the Same Direction. Green dots indicates means. Red bars (dots) are observed, and blue bars (dots) are subjects with missing z.

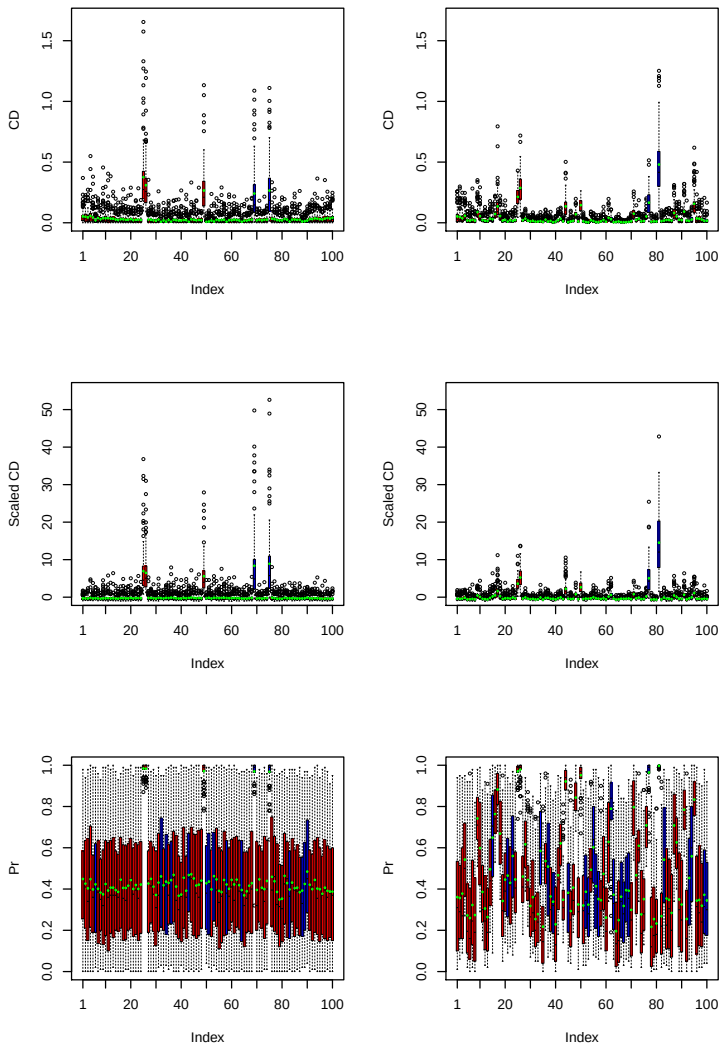


Figure 3.16: Box Plots for CD, Scaled CD, and Pr. Results from 100 Simulation Samples for Scenarios 3b (MAR, left panels) and 7b (MNAR, right panels) - 5 Outliers in y Domain in the Opposite Direction. Green dots indicates means. Red bars (dots) are observed, and blue bars (dots) are subjects with missing z.

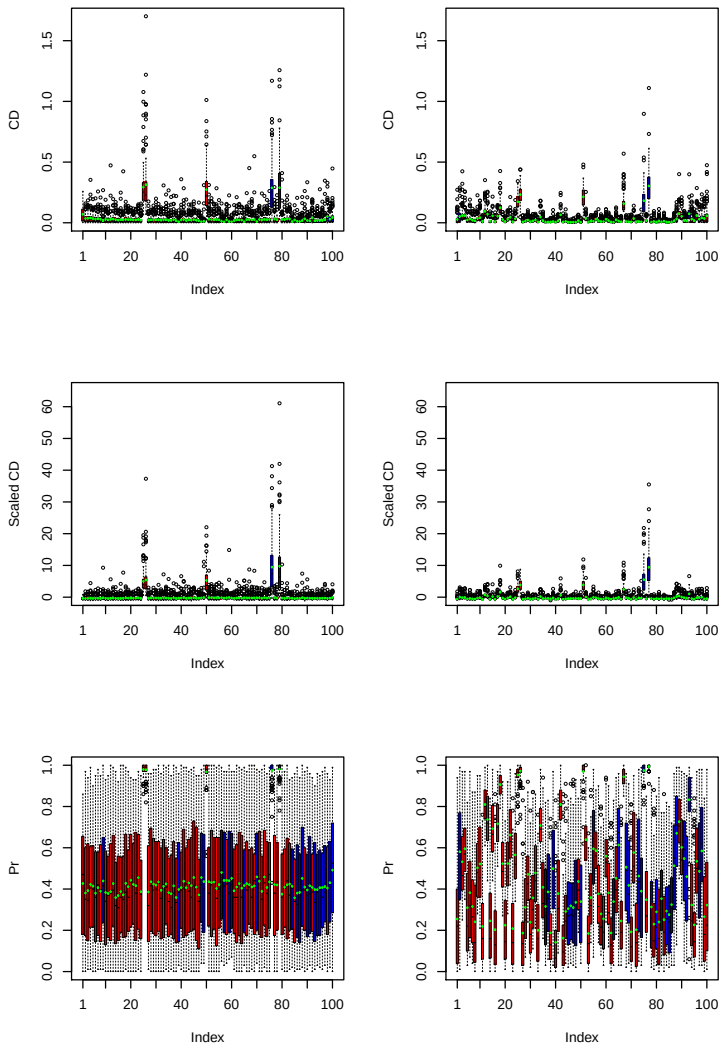


Figure 3.17: Box Plots for CD, Scaled CD, and Pr. Results from 100 Simulation Samples for Scenarios 4a (MAR, left panels) and 8a (MNAR, right panels) - 5 Outliers in x Domain in the Same Direction. Green dots indicates means. Red bars (dots) are observed, and blue bars (dots) are subjects with missing z.

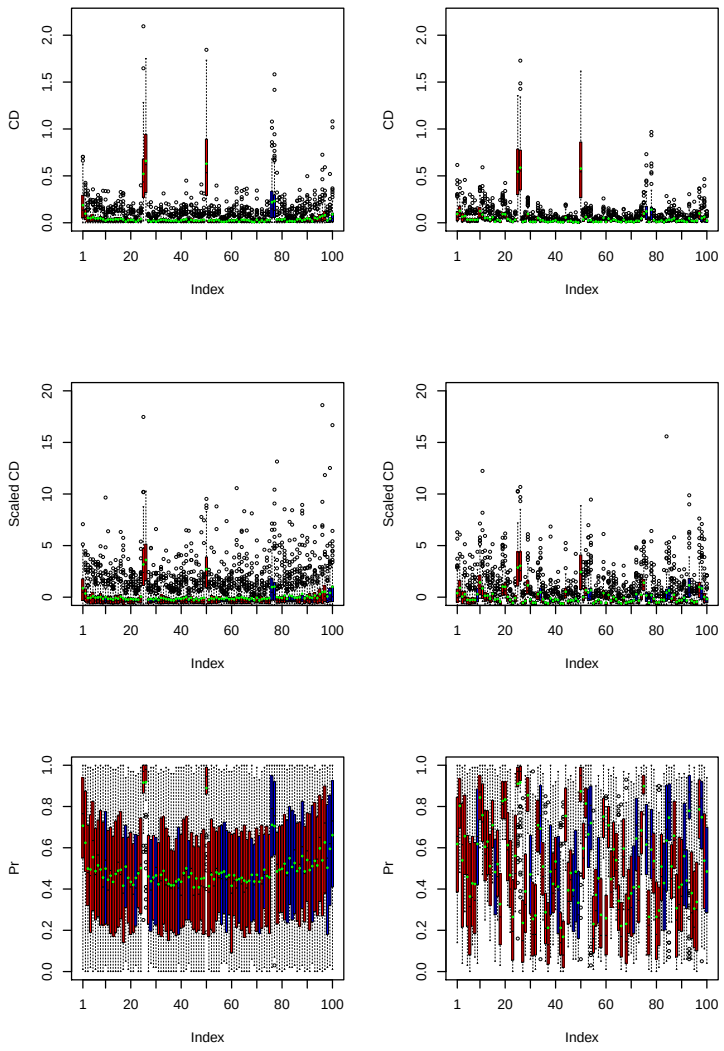
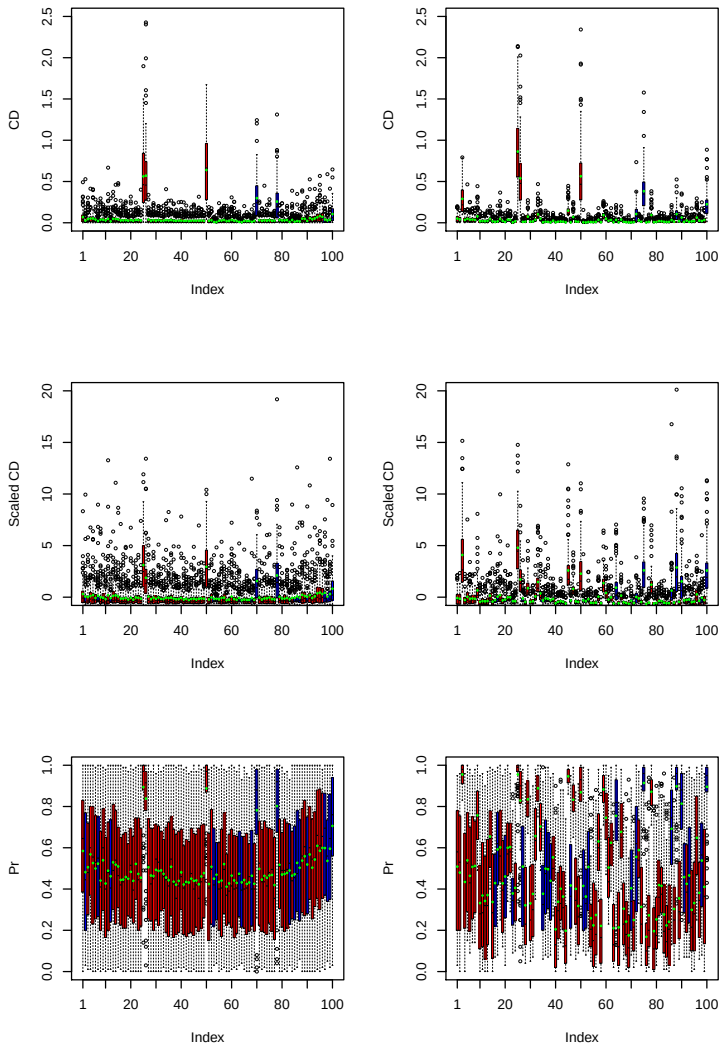


Figure 3.18: Box Plots for CD, Scaled CD, and Pr. Results from 100 Simulation Samples for Scenarios 4b (MAR, left panels) and 8b (MNAR, right panels) - 5 Outliers in x Domain in the Opposite Direction. Green dots indicates means. Red bars (dots) are observed, and blue bars (dots) are subjects with missing z.



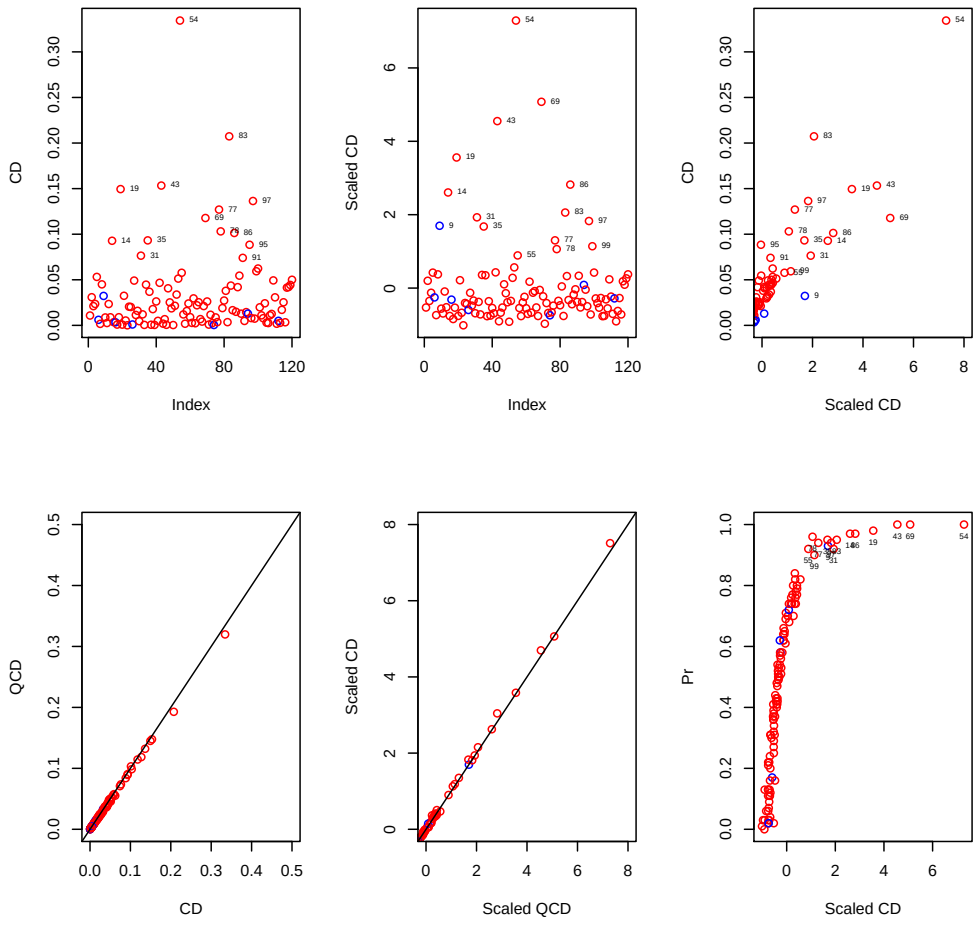
3.8 Real Data Examples

3.8.1 National Survey Cholesterol Data

For illustration purpose, we applied the GLM with missing covariates to a random sample of a publically available US national survey data. This data contains the lab value of 120 subjects for the total cholesterol value, HDL, and LDL. There is no missing data in total cholesterol and HDL (high-density lipoprotein cholesterol, also called "good" cholesterol), but has 6% missing in LDL (low-density lipoprotein cholesterol, also called "bad" cholesterol). We consider a general linear model with total cholesterol value as the response variable and HDL and LDL as the covariates. We assume the missingness of LDL depends on HDL, which is supported by the logistic regression analysis. We calculated $CD(I)$ for each subject and the degree of perturbation. We computed the scaled Cooks distance using 100 bootstrap samples, and then $P_r(I)$. The results from the GLM with missing covariates model are shown in Figure 3.11, where we label observations with relatively large CD and/or scaled CD. A positive correlation between CD and scaled CD is observed, but there are some discrepancies between them. For instance, Subject 83 has the second largest CD, but is not outstanding in scaled CD index plot. It is also interesting to see that Subject 9 with a missing LDL (and therefore smaller degree of perturbation) has a small CD, but relatively large scaled CD. The value of QCD (based on the Q-function) in this sample is similar to CD (based on observed likelihood function). Same is true for scaled QCD and scaled CD. One common question for identifying influential observations is how to interpret the value of CD or scaled CD. What cut off value should be considered as influential. For CD, the value greater than 1 or 4 divided by number of observations (which is 0.03 for our sample) are often used to identify influential observations. However, those criteria can be either too restrict or too loosened. The value of Pr provides an interpretation of the scaled CD. For example, as shown in the Pr vs. scalded CD plot, a value greater than 0.88 in scaled CD is corresponding to

a greater than 0.85 probability that the scaled CD from sample is larger than the scaled CD from the parametric bootstrapping sample. A larger value of the scaled CD is linked to a greater probability of the observed value is greater than the "true" value given that the fitted model is the true model.

Figure 3.19: Results from Cholesterol Data



3.8.2 Liver Cancer Data

As another example, we considered data on 191 patients from two Eastern Cooperative oncology Group clinical trials (Ibrahim et al., 1999). We are interested in how the number of cancerous liver nodes (y) when entering the trials is predicted by four other baseline

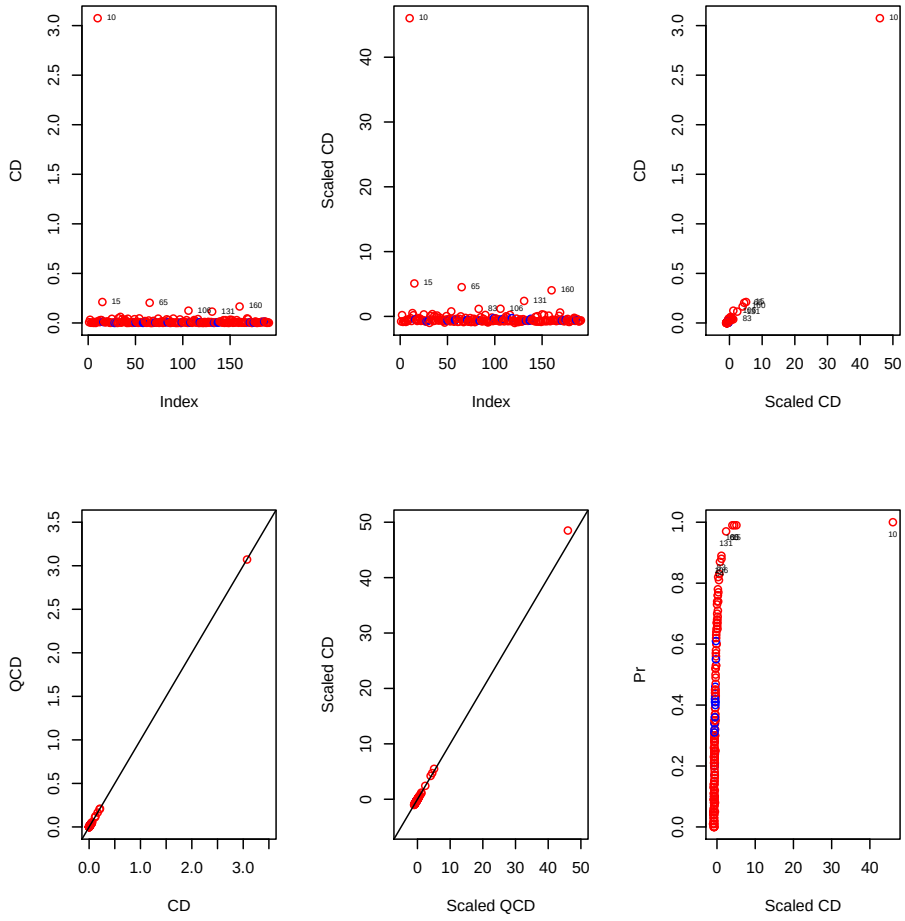
characteristics: age (in years, x_1), body mass index (kg/m^2, x_2), associated jaundice (yes or no, x_3), and time since diagnosis of the disease (in weeks). The time since diagnosis variable has missing data and is skewed. After we took the logarithm transformation on it, the distribution is approximately normal.

We used a general linear model $y_i|x_i, z_i, \beta \sim N(\beta(x_i, z_i)^\top, \tau)$, where $x_i = (x_{i1}, x_{i2}, x_{i3})$, z_i =logarithm of time since diagnosis, and $z_i \sim N(\mu_z, \tau_z)$. We further modeled the missingness of z_i ($r = 1$ if missing, and $r = 0$ ifobserved) by logistic regressions. We assumed the missing covariates are MAR and calculated the MLE of parameters using the EM algorithm. We calculated CD, QCD, scaled CD, scaled QCD, and P_r for each subject. The results are shown in Figure 3.20. The value of QCD and scaled QCD are very close to the CD and scaled CD, respectively. This is the same as what are seen in simulations and the cholesterol data example. All these diagnostic measures and P_r suggest that subject 10 is an influential point and worth further investigation.

3.9 Conclusions

In summary, we have derived the Cook's distance for the GLM with missing covariates based on both the observed likelihood function and the $Q-$function used in EM algorithm. We have defined the degree of perturbation for GLM with missing covariates and demonstrated that the degree of perturbation is more complicated for the model with missing covariates. Subjects without missing data have a larger degree of perturbation than those with a missing covariate given the same x. We have further derived the scaled Cook's distance to adjust for the degree of perturbation. We examined the performance of these diagnostic measures in various scenarios with correctly specified or misspecified missing mechanism. We used simulation data to illustrate the size matters issue in general linear model with missing covariates. In addition, our simulation results suggest that for general linear model with missing covariates, QCD is very close to CD for the "good" data points, even when the sample included outliers or the misspecified

Figure 3.20: Results from Liver Cancer Data



missing mechanism. For the outliers, QCD is close to or slightly smaller than CD, but it is generally sufficient to the purpose to identify influential observations. The largest differences between QCD and CD are found when the missing mechanism is misspecified with a large amount of missing data. The missing mechanism has an impact on the diagnostic measures. When the missing mechanism is misspecified, in our example, some non-outliers have a CD or scaled CD larger than other non-outliers. We applied the proposed method to two real data examples. The results demonstrated the similarity between QCD and CD and illustrated the proposed model diagnostic measures are valuable in real data analyses with missing data.

CHAPTER 4

INFORMATION CRITERIA FOR GENERALIZED LINEAR MODELS WITH MISSING COVARIATES

4.1 Introduction

The aim of this chapter is to select an optimal model from a pool of statistical models for a given dataset. One needs to consider both goodness of fit and model complexity. A model which balances model fitting and complexity is preferred. To achieve this, various information criteria, such as Akaiki Information Criterion (AIC) or Bayesian Information Criterion (BIC), have been proposed for model comparisons. In both AIC and BIC, deviance, $-2\log p(Y|\hat{\theta})$, is used to measure the goodness of fit. The penalty term for model complexity is set as $2p$ in AIC and $p\log(n)$ in BIC, respectively, where p is the number of parameters and n is the number of observations.

Recently, (Zhu et al., 2014a) developed a new measure of model complexity based on Bayesian case deletion measures, called Bayesian Case-deletion Model Complexity (BCMC), and then further proposed a Bayesian Case-deletion Information Criterion (BCIC). Motivated from BCMC and BCIC, we use those new case deletion measures developed in Chapter 3 to construct various new case-deletion model complexities (CMCs) and case-deletion information criteria (CIC) for generalized linear models (GLM) with missing covariates.

4.2 Method

4.2.1 Case Deletion Measures

Consider a probability function of observed data $Y = (Y_1, \cdots, Y_n)^T$, denoted by $p(Y|\theta)$, where $\theta = (\theta_1, \cdots, \theta_p)^T$ is a $p \times 1$ vector in an open subset Θ of R^p. Cook's distance and other case-deletion diagnostics measure the distance between the maximum likelihood estimators (MLEs) of θ with and without observations in a set I. The set I may contain one observation or multiple observations in data. A subscript $[I]$ denotes the relevant quantity with all observations in I deleted. Cook's distance for I, denoted by $CD(I)$, can be defined as follows:

$$CD(I) = (\hat{\theta}_{[I]} - \hat{\theta})^\top G(\hat{\theta}_{[I]} - \hat{\theta}),$$

where $\hat{\theta}$ is the MLE for the full sample, $\hat{\theta}_{[I]}$ is the MLE for the subsample with all observations in I deleted, and G is a positive definite matrix. When our primary interest is to make inference on a linear combination of θ, say $L^\top\theta$, the partial influence of the subset I on $L^\top\hat{\theta}$, denoted by $CD(I|L)$, may be defined as

$$CD(I|L) = (\hat{\theta}_{[I]} - \hat{\theta})^\top L\{L^\top G^{-1} L\}^{-1} L^\top(\hat{\theta}_{[I]} - \hat{\theta}).$$

For GLM with missing covariates models, as shown in Chapter 3, we can calculate Cook's distance based on either the log-likelihood function (ℓ) of the observed data or the Q-function. We can approximate CD and QCD as

$$CD^1(I) = \partial_\theta l_{[I]}(D_{o[I]}|\hat{\theta})^\top\{-\partial_\theta^2 l(D_o|\hat{\theta})\}^{-1}\partial_\theta l_{[I]}(D_{o[I]}|\hat{\theta}), \tag{4.1}$$

$$QCD^1(I) = \partial_\theta Q_{[I]}(\hat{\theta}|\hat{\theta})^\top\{-\partial_\theta^2 Q(\hat{\theta}|\hat{\theta})\}^{-1}\partial_\theta Q_{[I]}(\hat{\theta}|\hat{\theta}). \tag{4.2}$$

4.2.2 Cross Validation and Model Complexity

Cross-validation (CV) is a popular strategy for carrying out model selection (Stone, 1974, Arlot et al., 2010). The primary idea behind CV is to repeatedly split data into a training sample and a validation sample multiple times in order to estimate the "risk" of each model. The training sample is primarily used for training, whereas the validation sample is used for estimating the risk of a given model. Finally, CV selects an optimal model with the smallest estimated risk from a pool of candidate models. This data splitting heuristics are valid for a wide range of data generating processes.

Case deletion measure and cross-validation method share the same strategy of splitting the data into two subsamples. Specifically, to calculate case deletion measure, one divides the data into a given set Y_S and the remaining set, $Y_{[S]}$, and then quantifies the influential level of the set S. In contrast, the CV method divides the data into two subsamples including a training sample $Y_{[S]}$ for model fitting and a validation sample Y_S for assessing "risk". To calculate the risk of a given model, one has to carry out data splitting many times, and the final validation result is averaged over all splittings.

For GLM with missing covariates, our measure of cross validation is based on the predictive distribution of the observed data $l(D_{o,S}|\tilde{\theta}_{[S]})$, where $\tilde{\theta}_{[S]}$ is an estimate of θ based on the training set $D_{o,[S]}$. Let $S_1, \ldots, S_{n_B}$ be a sequence of non-empty proper subsets of $\{1, \cdots, n\}$ corresponding to our data splitting scheme, where n_B is an integer. For example, $n_B = n$ for the leave-one-out CV and $n_B = n!/[m!(n-m)!]$ for the exhaustive leave-m-out CV where $n \geq m \geq 1$. The CV estimator of the risk based on $I_S = \{S_k\}_{1 \leq k \leq n_B}$ is defined as

$$\mathrm{CVR}(I_S) = -n_B^{-1} \sum_{S_k \in I_S} l(D_{o,[S_k]}|\tilde{\theta}_{[S_k]}). \tag{4.3}$$

Similarly, we define another CV measure based on the Q-function as follows

$$\text{QCVR}(I_S) = -n_B^{-1} \sum_{S_k \in I_S} Q_{S_k}(\tilde{\theta}_{[S_k]} | \hat{\theta}_{[S_k]}). \tag{4.4}$$

In addition to sharing the same strategy of splitting the data, we can establish a formal connection between the case deletion measures (CD and QCD) and their corresponding CV measures (CVR and QCVR). We obtain the following theorems, whose detailed proof can be found in the Appendix.

THEOREM 1. *Under Assumptions C1-C4 in the Appendix, $CVR(I_S)$ and $QCVR(I_S)$ have the following asymptotic expansions:*

$$CVR(I_S) = -n_B^{-1} \sum_{S_k \in I_S} l(D_{o,[S_k]} | \hat{\theta}) + n_B^{-1} \sum_{S_k \in I_S} CD^1(S_k; \hat{\theta})[1 + o_p(1)], \tag{4.5}$$

$$QCVR(I_S) = -n_B^{-1} \sum_{S_k \in I_S} Q_{S_k}(\hat{\theta} | \hat{\theta}) + n_B^{-1} \sum_{S_k \in I_S} QCD^1(S_k; \hat{\theta})[1 + o_p(1)]. \tag{4.6}$$

Theorem 1 shows a direct connection between $\text{CVR}(I_S)$ and $\text{CD}^1(I_S)$, and between $\text{QCVR}(I_S)$ and $\text{QCD}^1(I_S)$. Based on this, we define case-deletion model complexity measures CMC and QCMC, which are based on the observed likelihood and the Q-function, respectively:

$$\text{CMC}(I_S) = n n_S^{-1} n_B^{-1} \times \sum_{S_k \in I_S} \text{CD}(S_k; \tilde{\theta}). \tag{4.7}$$

$$\text{QCMC}(I_S) = n n_S^{-1} n_B^{-1} \times \sum_{S_k \in I_S} \text{QCD}(S_k; \tilde{\theta}). \tag{4.8}$$

4.2.3 Case-deletion Information Criterion

Based on the development of CMC, we develop a model selection criterion, called case-deletion information criterion (CIC), to select an optimal model from a pool of candidate models $\{M_l : l = 1, \cdots, L\}$ for the same dataset. Specifically, for model M_l and the deletion set I_S, CIC is defined as

$$\text{CIC}(I_S, M_l) = -2 \sum_{S_k \in I_S} l(D_{o,[S_k]}|\hat{\theta}(M_l), M_l) + (n_B n_S/n) C_n(I_S, \hat{\theta}(M_l), M_l), \tag{4.9}$$

where $\hat{\theta}(M_l)$ is the maximum likelihood estimator of θ under model M_l. Moreover, $C_n(I_S, \hat{\theta}(M_l), M_l)$ is a penalty term, which is a variation of CMC and a function of the data, the deletion set I_S, and an estimator of $\theta(M_l)$. We choose an optimal model, denoted by M_{opt}, which minimizes $\text{CIC}(I_S, M_l)$, as follows:

$$M_{opt}(I_S) = \text{argmin}_{M_l : 1 \le l \le L} \text{CIC}(I_S, M_l). \tag{4.10}$$

When deleting one observation at a time (leave-one-out deletion), CIC for model M_l is equivalent to

$$CIC(M_l) = -2l(D_o|\hat{\theta}(M_l), M_l) + C_n(\hat{\theta}(M_l), M_l).$$

The first term is the deviance, which is the same as AIC and BIC. The penalty term C_n is different for different criteria. For CIC, we define it based on CMC. Two popular choices of C_n are the AIC-type penalty and the BIC-type penalty. For the AIC-type penalty, $C_n = C_0 \times \text{CMC}(I_S)$, where C_0 is a bounded positive scalar. In practice, similar to AIC, it is common to set $C_0 = 2$ following Akaike's heuristics (Arlot et al., 2010). For the BIC-type penalty, $C_n = C_{0,n} \times \text{CMC}(I_S)$ with $\lim_{n\to\infty} C_{0,n} = \infty$. Similar to BIC, $C_{0,n}$ is often set as $\log(n)$ or other functions of a higher order (Birgé and Massart,

2007). Therefore, CIC can be regarded as a generalization of the existing model selection criteria. Just as the well-known results on the asymptotic equivalence between CV and AIC (Stone 1977), Theorem 1 shows the asymptotic equivalence between CVR and CIC with AIC-type penality.

Because different deletion sets lead to slightly different $\mathrm{CIC}(I_S, M_l)$ for all l, it is possible that $M_{opt}(I_S)$ may vary across I_S. However, when we consider the leave-m-out (LMO) deletion, we are able to obtain an invariant property of $M_{opt}(I_S)$.

THEOREM 2. *Assume that Y_is' are independent and $C_n(I_S, \hat{\theta}(M_l), M_l) = \hat{C}_{0,n} \times CMC(I_S)$, where $\hat{C}_{0,n}$ is independent of I_S and M_l, but it may depend on n. We have the following results.*

(i) For the leave-m-out CV, we have $CIC(I_{LMO}, M_l) = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} CIC(I_{LOO}, M_l)$ *and $M_{opt}(I_{LMO}) = M_{opt}(I_{LOO})$ for any m that $n \geq m \geq 1$.*

(ii) If $CIC(I_{LOO}, M_{opt}(I_{LOO})) - CIC(I_{LOO}, M_l) >> O_p(n_B n^{-3/2})$ for all $M_l \neq M_{opt}(I_{LOO})$, Assumption C5 holds, and we use $CD^1(I_S)$ to approximate $CMC(I_S)$, then $M_{opt}(I_{LMO}) = M_{opt}(I_{LOO})$ in probability 1 for any m that $n \geq m \geq 1$.

Theorem 2 shows that $\mathrm{CIC}(I_S, M_l)$ and $M_{opt}(I_S)$ are invariant under different exhaustive splitting. Detailed proof can be found in the Appendix.

Similarly, we define QCIC for Q-function based case-deletion information criterion as follows.

$$\mathrm{QCIC}(I_S, M_l) = -2 \sum_{S_k \in I_S} Q_{S_k}(\hat{\theta}(M_l) | \hat{\theta}(M_l)) + (n_B n_S / n) C \times QCMC(I_S, \hat{\theta}(M_l)), \quad (4.11)$$

where $C = C_0$ or $C_{0,n}$.

4.3 Simulation

We conducted simulation studies to investigate the finite sample performance of CIC and compare CIC with AIC and BIC in GLM with missing covariates. Furthermore, we

assessed the performance of the Q-function based criterion QCIC.

Simulated datasets and candidate models are described as follows. Simulated datasets were generated from a general linear model with a missing covariate. Specifically, we consider the following true model given by $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \epsilon_i$, where the ϵ_i's are independent and identically distributed (i.i.d.), $\epsilon_i \sim N(0, \tau), x_i \sim N(0, 1)$, and $z_i \sim N(\mu_z, \tau_z)$ for $i = 1, \ldots, n$. The covariate x_i is completely observed for $i = 1, \ldots, n$. The covariate z_i may be missing for some cases. We assumed MAR for z_i as $\text{logit}\left[\text{prob}(r_i = 1|x_i, z_i, y_i)\right] = \xi_0 + \xi_1 x_i$, where $r_i = 1$ when z_i is missing. We set $n = 100$ subjects, $\beta_1 = \beta_2 = \beta_3 = 1, \tau = \tau_z = 1$, $\mu_z = 0, \xi_0 = -1.5$, and $\xi_1 = 1$. An additional continuous covariate v_i was simulated from $N(0, 1)$, and a categorical covariate w_i was simulated from $Bernoulli(0.5)$. We consider five candidate models as follows:

M1 (true model): $y_i|x_i, z_i \sim N(\beta_1 + \beta_2 x_i + \beta_3 z_i, \tau)$;

M2: $y_i|z_i \sim N(\beta_1 + \beta_2 z_i, \tau)$;

M3: $y_i|x_i, v_i, z_i \sim N(\beta_1 + \beta_2 x_i + \beta_3 v_i + \beta_4 z_i, \tau)$;

M4: $y_i|x_i, w_i, z_i \sim N(\beta_1 + \beta_2 x_i + \beta_3 w_i + \beta_4 z_i, \tau)$;

M5: $y_i|x_i, v_i, w_i, z_i \sim N(\beta_1 + \beta_2 x_i + \beta_3 v_i + \beta_4 w_i + \beta_5 z_i, \tau)$.

We generated 1000 simulated datasets from M1 and then calcualted AIC, BIC, CIC, and QCIC for the five candidate models. For CIC, we calculated both CICA, that is the CIC with AIC-type penalty, $C_n = 2 \times CMC$, and CICB, that is the CIC with BIC-type penalty, $C_n = \log(n) \times CMC$. Similarly, we calculated QCICA and QCICB based on Q-function.

Table 4.1 shows the number of times out of the 1000 simulations that each rank was achieved for M1, the true model. The columns in Table 4.1 correspond to the rankings of AIC and BIC. The rows of Table 4.1 correspond to the proposed criteria CICA, CICB and their variations based on the Q-function. CICA is highly concordant with AIC (94.6%

concordant). CICB is highly concordant with as BIC (96.1% concordant). In contrast, QCICA has only 46.2% concordant rate with CICA, and QCICB has 69.9% concordant rate with CICB.

Table 4.1: Comparison of Ranks of the True Model M1 from Various Model Selection Criteria in GLM with Missing Covariates

		AIC						BIC					
	Rank	1	2	3	4	5		Rank	1	2	3	4	5
CICA	1	660	14				CICB	1	779	11			
	2	20	186	4				2	24	166	2		
	3	1	8	85	2			3		2	13		
	4			4	14			4				3	
	5				1	1		5					
QCICA	1	410	125	54	13		QCICB	1	682	155	14	1	
	2	102	34	16	2			2	87	17	1	2	
	3	136	40	18	2	1		3	30	7			
	4	33	9	5				4	4				
	5							5					

Table 4.2 shows the number of times out of the 1000 simulations that each rank was achieved for each model for all model selection criteria. M1 got ranked as number one 674 times by CICA and 790 times by CICB, which is similar to AIC (681 times) and BIC (803 times), respectively. The Q-based criterion QCICA ranked M1 as number one for the least times (602 times). However, the other Q-based criterion QCICB correctly ranked M1 as number one 852 times, which is the highest among all model selection criteria. M2 is the misspecified model, which missed an important covariate x. For all selection criteria, the model M2 was ranked last most of the time, but still in great than 10% of the simulations it was ranked as number one by CICA, CICB, as well as AIC and BIC. AIC-type criteria have a fewer percentage ranking number one than BIC-type criteria. Notably, the Q-based criteria, both QCICA and QCICIB, ranked M2 as number five for all the simulations. M3, M4, and M5 include additional covariates that are not in the true model. Most of the time, selection criteria ranked M5 (least parsimonious model)

as number 4, more favorable than the misspecified model M2, but less favorable than M3 and M4. Models with an additional covariate, either continuous (M3) or categorical (M4), have a rank of 2 or 3 most of the time for all selection criteria.

Table 4.2: Comparison of Ranks for M1 to M5 from Various Model Selection Criteria in GLM with Missing Covariates

Model	Rank	AIC	CICA	QCICA	BIC	CICB	QCICB
M1 (true model)	**1**	**681**	**674**	**602**	**803**	**790**	**852**
	2	208	210	154	179	192	107
	3	93	96	197	15	15	37
	4	17	18	47	3	3	4
	5	1	2				
M2 (misspecified model)	1	111	101		138	145	
	2	15	13		56	47	
	3	32	37		39	39	
	4	22	19		74	71	
	5	**820**	**821**	**1000**	**693**	**698**	**1000**
M3 (with additional continuous covariate)	1	95	102	177	25	27	71
	2	**334**	**338**	**358**	**364**	**399**	**475**
	3	**389**	**379**	**291**	**459**	**431**	**379**
	4	165	164	174	147	137	75
	5	17	17		5	6	
M4 (with additional categorical covariate)	1	100	99	164	37	32	73
	2	**372**	**365**	**301**	**354**	**394**	**381**
	3	**358**	**361**	**350**	**457**	**439**	**471**
	4	156	160	185	145	128	75
	5	14	15		7	7	
M5 (least parsimonious model)	1	13	15	57	2	1	4
	2	71	74	187	7	8	37
	3	128	127	162	48	58	113
	4	**640**	**639**	**594**	**648**	**644**	**846**
	5	148	145		295	289	

These results indicate that CICs perform reasonably well for model selection in GLM with missing covariates. CICA is highly concordant with AIC, and CICB is highly concordant with BIC. BIC-type criteria (BIC and CICB) performs better than AIC-type criteria (AIC and CICA). The Q-based criteria are moderately concordant with the likelihood function-based criteria. However, they also perform reasonably well for model selection, especially the BIC-type QCICB.

4.4 Real Data Analysis

To illustrate our proposed methods, we considered data on 191 patients from two Eastern Cooperative oncology Group clinical trials as mentioned in Chapter 3 (Ibrahim et al., 1999). We are interested in how the number of cancerous liver nodes (y) when entering the trials is predicted by four other baseline characteristics: age (in years, x_1), body mass index ($\text{kg/m}^2, x_2$), associated jaundice (yes or no, x_3), and time since diagnosis of the disease (in weeks). The time since diagnosis variable has missing data and is skewed. After we took the logarithm transformation on it, the distribution is approximately normal.

We used a general linear model $y_i|x_i, z_i, \beta \sim N(\beta(x_i, z_i)^\top, \tau)$, where $x_i = (x_{i1}, x_{i2}, x_{i3})$, z_i =logarithm of time since diagnosis, and $z_i \sim N(\mu_z, \tau_z)$. We further modeled the missingness of z_i ($r = 1$ if missing, and $r = 0$ if observed) by logistic regressions. We assumed the missing covariates are MAR and calculated the MLE of parameters using the EM algorithm. We applied the proposed CIC method in addition to the existing model selection criteria to illustrate the application of CIC.

Table 4.3 shows the values of AIC, BIC, and four CIC measurements as well as the ranks of four candidate models (listed below) for each criterion.

M1: $y_i|x_i, z_i \sim N(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \beta_5 z_i, \tau)$;

M2: $y_i|x_i, z_i \sim N(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \beta_5 x_{i1} x_{i3} + \beta_6 z_i, \tau)$;

M3: $y_i|x_i, z_i \sim N(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \beta_5 x_{i2} x_{i3} + \beta_6 z_i, \tau)$;

M4: $y_i|x_i, z_i \sim N(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \beta_7 z_i, \tau)$.

The best model selected by different criteria is consistent, which is the model including only the main effects (M1). M4 was ranked as worst by all criteria. The two models ranked 2-3 are slightly off, however, with almost indistinguishable values. Same as the simulation results, the values of CICA and CICB are very close to AIC and BIC,

Table 4.3: Model Selection Results of Liver Cancer Data

Candidate Model	AIC (Rank)	CICA (Rank)	QCICA (Rank)	BIC (Rank)	CICB (Rank)	QCICB (Rank)
M1	1066.49 (1)	1066.06 (1)	1081.34 (1)	1086.00 (1)	1084.89 (1)	1099.86 (1)
M2	1067.75 (2)	1067.62 (3)	1082.84 (3)	1090.52 (2)	1090.17 (3)	1105.07 (3)
M3	1068.39 (3)	1067.21 (2)	1082.55 (2)	1091.15 (3)	1088.06 (2)	1103.09 (2)
M4	1069.72 (4)	1068.48 (4)	1083.73 (4)	1095.74 (4)	1092.49 (4)	1107.42 (4)

Table 4.4: Model Selection Results of Liver Cancer Data, Excluding One Outlier

Candidate Model	AIC (Rank)	CICA (Rank)	QCICA (Rank)	BIC (Rank)	CICB (Rank)	QCICB (Rank)
M1	1063.31 (1)	1061.53 (1)	1076.00 (1)	1082.79 (1)	1078.13 (1)	1092.21 (1)
M2	1065.30 (3)	1063.07 (3)	1077.54 (3)	1088.03 (3)	1082.17 (3)	1096.26 (3)
M3	1065.02 (2)	1062.00 (2)	1076.22 (2)	1087.75 (2)	1079.82 (2)	1093.67 (2)
M4	1067.02 (4)	1063.52 (4)	1077.74 (4)	1093.00 (4)	1083.82 (4)	1097.65 (4)

respectively. The values of CIC based on Q-function (QCICA and QCICB) are slightly off from that based on the likelihood function (CICA and CICB), but result in the same rank in this liver cancer data. Table 4.4 shows the results when excluding the outlier (Observation 10) identified in Chapter 3. All criteria gave the same ranking when the outlier was removed.

4.5 Conclusions

We have examined the connection between case deletion measures and cross validation method for GLM with missing covariates models. Based on such connection, we have developed CMC measures for quantifying the model complexity and CICs for model selection. We have developed these new measures and criteria based on the likelihood function and the Q-function used in the EM algorithm for models with missing data. Some properties of CMC and CIC are investigated. Simulations and real data analysis show that CIC is a valuable tool for analysis of models with missing data.

APPENDIX: ASSUMPTIONS AND PROOFS

We need some notation. We use $||\cdot||$ to denote the Euclidean norm of a vector or a matrix and use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the largest and smallest eigenvalues of a symmetric matrix A, respectively. We use the mathematical symbols (e.g., $O(N^{-1})$) and the stochastic-order symbols, such as $O_p(1)$, $o_p(1)$, and $O_p(N^{-1})$ throughout. We define $F_Q(\theta) = \partial_\theta E[\ell_c(\theta|D_c)|D_o, \hat{\theta}]$, $F_{Q,[S]}(\theta) = \partial_\theta E[\ell_c(\theta|D_{c,[S]})|D_o, \hat{\theta}]$, $F(\theta) = \partial_\theta \ell(Y|\theta)$, and $F_{[S]}(\theta) = \partial_\theta \ell(Y_{[S]}|\theta)$, where $\ell(Y|\theta) = \log p(Y|\theta)$ and $\ell(Y_{[S]}|\theta) = \log p(Y_{[S]}|\theta)$.

The following assumptions are needed to facilitate the technical details, although they are not the weakest possible conditions. Because we develop all results for general parametric models, we only assume several high-level assumptions for the observed-data log-likelihood function as follows. It is assumed that similar conditions hold for the complete-data log-likelihood function.

Assumption C1. $\hat{\theta}$ and $\hat{\theta}_{[S]}$ for all S are consistent estimates of $\theta_* \in \Theta^o$.

Assumption C2. Let $\Delta(\theta) = \theta - \theta_*$ and suppose

$$\begin{aligned}
\log p(Y|\theta) &= \log p(Y|\theta_*) + \Delta(\theta)^T F_N(\theta_*) - 0.5\Delta(\theta)^T J_N(\theta_*)\Delta(\theta)[1 + o_p(1)] \quad \text{and} \\
\log p(Y_{[S]}|\theta) &= \log p(Y_{[S]}|\theta_*) + \Delta(\theta)^T F_{N,[S]}(\theta_*) - 0.5\Delta(\theta)^T J_{N,[S]}(\theta_*)\Delta(\theta)[1 + o_p(1)]
\end{aligned}$$

uniformly for all $\theta \in B(\theta_*, \delta_0/\sqrt{N}) = \{\theta : \sqrt{N}||\theta - \theta_*|| \le \delta_0\}$. Moreover, $N^{-1/2}F_N(\theta_*) = O_p(1)$, $N^{-1/2}F_{N,[S]}(\theta_*) = O_p(1)$, $\max_{S\in I_S} \sup_{\theta,\theta' \in B(\theta_*, N^{-1/2}\delta_0)} ||J_{N,[S]}(\theta) - J_{N,[S]}(\theta')|| = o_p(N)$,

$$\begin{aligned}
0 &< \inf_{\theta \in B(\theta_*, \delta_0 N^{-1/2})} \lambda_{\min}(n^{-1}J_N(\theta)) \le \sup_{\theta \in B(\theta_*, \delta_0 N^{-1/2})} \lambda_{\max}(N^{-1}J_N(\theta)) < \infty, \text{ and} \\
0 &< \min_{S\in I_S} \inf_{\theta \in B(\theta_*, \delta_0 N^{-1/2})} \lambda_{\min}(N^{-1}J_{N,[S]}(\theta)) \le \max_{S\in I_S} \sup_{\theta \in B(\theta_*, \delta_0 N^{-1/2})} \lambda_{\max}(N^{-1}J_{N,[S]}(\theta)) < \infty.
\end{aligned}$$

Assumption C3. Assume that for small $\delta_0 > 0$, if $N_S \leq N_0$, a fixed constant, then

$$\max_{S \in I_S} \sup_{\theta \in B(\theta_*, \delta_0)} ||\partial_\theta \log p(Y_S|Y_{[S]}, \theta)|| = O_p(1) \text{ and } \max_{S \in I_S} \sup_{\theta \in B(\theta_*, \delta_0)} ||\partial_\theta^2 \log p(Y_S|Y_{[S]}, \theta)|| = o_p(N).$$

Assumption C4. $\lim_{N_{I_S} \to \infty} N_B^{-1} E[K_N(I_S|\theta_*)] = K_*(I_S)$ and $\lim_{N \to \infty} N^{-1} E[J_N(\theta_*)] = J_*$, where $K_N(I_S|\theta) = n_B^{-1} \sum_{S_k \in I_S} [\partial_\theta \log p(Y_S|Y_{[S]}, \theta)^{\otimes 2}]$ and the expectation is taken with respect to the true data generator. Moreover, for a small $\delta_0 > 0$, we have

$$\sup_{\theta \in B(\theta_*, \delta_0)} ||K_N(I_S|\theta) - E[K_N(I_S|\theta)]|| = o_p(1) \text{ and } \sup_{\theta \in B(\theta_*, \delta_0)} ||J_N(I_S|\theta) - E[J_N(I_S|\theta)]|| = o_p(1).$$

Assumption C5. Each component of $N_B^{-1}\sqrt{N}\{K_N(I_S|\theta_*) - E[K_N(I_S|\theta_*)]\}$ is asymptotically tight.

Proof of Theorem 1. It follows from Assumptions C1-C3 that we can expand $l(D_{o,[S_k]}|\tilde{\theta}_{S_k})$ at $\hat{\theta}$ for each S and obtain

$$\sum_{S_k \in I_S} l(D_{o,[S_k]}|\tilde{\theta}_{S_k}) = \sum_{S_k \in I_S} l(D_{o,[S_k]}|\hat{\theta}) + \sum_{S_k \in I_S} \partial_\theta l(D_{o,[S_k]}|\hat{\theta})^T (\tilde{\theta}_{[S_k]} - \hat{\theta})[1 + o_p(1)].$$

Following (3.6) and $\partial_\theta l(D_{o,[S_k]}|\hat{\theta}) = \partial_\theta l(D_o|\hat{\theta}) - \partial_\theta l(D_{o[S_k]}|\hat{\theta})$ where $\partial_\theta l(D_o|\hat{\theta}) = 0$ for MLE $\hat{\theta}$, we have

$$\sum_{S_k \in I_S} l(D_{o,[S_k]}|\tilde{\theta}_{S_k}) = \sum_{S_k \in I_S} l(D_{o,[S_k]}|\hat{\theta}) - \sum_{S_k \in I_S} [\partial_\theta l(D_{o[S_k]}|\hat{\theta})]^T [J_n(\hat{\theta})]^{-1} l(D_{o[S_k]}|\hat{\theta})[1 + o_p(1)],$$

where $J_n(\hat{\theta}) = -\partial_\theta^2 l(D_o|\hat{\theta})$. This yields Theorem 1 (4.5). Theorem 1 (4.6) can be obtained following similar derivation for Q-function.

Proof of Theorem 2. We consider the exhaustive splitting for the leave-m-out CV. For

any $S_k = \{\{i_1\}, \cdots, \{i_m\}\}$ that $\{i_v\}$ and $\{i_{v'}\}$ are independent when $v \neq v'$, we have

$$\begin{aligned} l(D_{o,[S_k]}|\hat{\theta}(M_l), M_l) &= \sum_{v=1}^{m} l(D_{o,i_v}|\hat{\theta}(M_l), M_l)), \\ \sum_{S_k \in I_S} l(D_{o,[S_k]}|\hat{\theta}(M_l), M_l) &= \frac{n_B m}{n} \sum_{i=1}^{n} l(D_{o,i}|\hat{\theta}(M_l), M_l), \\ &= \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \sum_{i=1}^{n} l(D_{o,i}|\hat{\theta}(M_l), M_l), \end{aligned}$$

Under Assumptions C1, C2, and C5 , we have

$$n_B^{-1} \sum_{S_k \in I_S} CD(S_k; \hat{\theta}) = \text{tr}\{[J_n(\hat{\theta})]^{-1} K_n(I_S|\hat{\theta})\} = n^{-1}\{\text{tr}[J_*^{-1} K_*(I_S)] + o_p(1)\}, \quad (4.12)$$

where $J_n(\hat{\theta}) = -\partial_\theta^2 l(D_o|\hat{\theta})$ and $J_* = \lim_{n\to\infty} n^{-1} E[J_n(\theta_*)]$, in which the expectation is taken with respect to the true data generator and θ_* denotes the true parameter. Moreover, $K_n(I_S|\hat{\theta}) = n_B^{-1} \sum_{S_k \in I_S} [\partial_\theta l(D_{o,[S_k]}|\hat{\theta})]^{\otimes 2}$ and

$$K_*(I_S) = \lim_{n\to\infty} (n_B)^{-1} \sum_{S_k \in I_S} E\{[\partial_\theta l(D_{o,[S_k]}|\theta_*)]^{\otimes 2}\},$$

where $a^{\otimes 2} = aa^T$ for any vector a.

$$\begin{aligned} E\{[\partial_\theta l(D_{o,[S_k]}|\theta_*]^{\otimes 2}\} &= \sum_{i_v, i_{v'}} E\{\partial_\theta l(D_{o,i_v}|\hat{\theta}(M_l), M_l) \partial_\theta l(D_{o,i'_v}|\hat{\theta}(M_l), M_l)^T\} \\ &= \sum_{v=1}^{m} E\{\partial_\theta l(D_{o,i_v}|\hat{\theta}(M_l), M_l)^{\otimes 2}\}, \end{aligned}$$

$$\sum_{S_k \in I_S} E\{[\partial_\theta l(D_{o,[S_k]}|\hat{\theta}(M_l), M_l)]^{\otimes 2}\} = \frac{n_B m}{n} \sum_{i=1}^{n} E\{\partial_\theta \ l(D_{o,i}|\hat{\theta}(M_l), M_l))^{\otimes 2}\}. \quad (4.13)$$

Following (4.12), (4.13) yields that $\text{CMC}(I_{LMO}) = \frac{n_B m}{n} \times \text{CMC}(I_{LOO})$.

Therefore, we have

$$\mathrm{CIC}(I_{LMO}, M_l) = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \mathrm{CIC}(I_{LOO}, M_l),$$

which yields Theorem 2 (i). Theorem 2 (ii) directly follows from Assumption C5.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Ali, M. W. and Siddiqui, O. (2000). Multiple imputation compared with some informative dropout procedures in the estimation and comparison of rates of change in longitudinal clinical trials with dropouts. *Journal of biopharmaceutical statistics*, 10(2):165–181.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical bayesian and empirical bayes models. *Biometrika*, 94(2):443–458.

Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Banerjee, M. and Frees, E. W. (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*, 92(439):999–1005.

Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73.

Blyth, S. (1994). Local divergence and association. *Biometrika*, 81(3):579–584.

Bradlow, E. T. and Zaslavsky, A. M. (1997). Case influence analysis in bayesian inference. *Journal of Computational and Graphical Statistics*, 6(3):314–331.

Carlin, B. P. and Carlin, B. P. (1991). An expected utility approach to influence diagnostics. *Journal of the American Statistical Association*, 86(416):1013–1021.

Chatterjee, S. and Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, pages 379–393.

Christensen, R., Pearson, L. M., and Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34(1):38–45.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, pages 15–18.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–169.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete data bias (with discussion). *Journal of the Royal Statistical Society Series B.*, 67:459–512.

Daniels, M. J. and Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press.

Davison, A. and Tsai, C.-L. (1992). Regression model diagnostics. *International Statistical Review/Revue Internationale de Statistique*, pages 337–353.

EMA (2010). *Guideline on Missing Data in Confirmatory Clinical Trials*. European Medicines Agency.

Fairclough, D. L. (2010). *Design and analysis of quality of life studies in clinical trials*. CRC press.

FDA et al. (2009). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register*, 74(235):65132–65133.

Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010). Variable selection for regression models with missing data. *Statistica Sinica*, 20(1):149.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.

Haslett, J. and Dillane, D. (2004). Application of delete= replaceto deletion diagnostics for variance component estimation in the linear mixed model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):131–143.

Hedeker, D. and Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological methods*, 2(1):64.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88(2):551–564.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):173–190.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the em algorithm. *Journal of the American Statistical Association*, 103(484).

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. G. (2006). The nature of sensitivity in monotone missing not at random models. *Computational Statistics and Data Analysis*, 50:830–858.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A local in uence approach to binary data from a psychiatric study. *Biometrics*, 59:410–419.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association*, 78(381):137–144.

Johnson, W. and Geisser, S. (1985). Estimative influence measures for the multivariate general linear model. *Journal of Statistical Planning and Inference*, 11(1):33–56.

Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91(1):27–43.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83(4):875–890.

Lin, D. Y., Wei, L.-J., and Ying, Z. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572.

Lipsitz, S. R. and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83(4):916–922.

Little, R. et al. (2011). Calibrated bayes, for statistics in general, and missing data in particular. *Statistical science*, 26(2):162–174.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. New York: Wiley.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):141–167.

Mangel, A. W., Bornstein, J. D., Hamm, L. R., Buda, J., Wang, J., Irish, W., and Urso, D. (2008). Clinical trial: asimadoline in the treatment of patients with irritable bowel syndrome. *Alimentary pharmacology & therapeutics*, 28(2):239–249.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, 21(8):1023–1041.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., and Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3):445–464.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001). Influence analysis to assess sensitivity of the dropout process. *Computational statistics & data analysis*, 37(1):93–113.

Murata, N., Yoshizawa, S., and Amari, S.-I. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *Neural Networks, IEEE Transactions on*, 5(6):865–872.

on Handling Missing Data in Clinical Trials; National Research Council, P. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press.

O'Neill, R. and Temple, R. (2012). The prevention and treatment of missing data in clinical trials: an fda perspective on the importance of dealing with it. *Clinical Pharmacology & Therapeutics*, 91(3):550–554.

Pauler, D. K., McCoy, S., and Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in medicine*, 22(5):795–809.

Peng, F. and Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, 23(2):199–213.

Pettit, L. (1986). Diagnostics in bayesian model choice. *The Statistician*, pages 183–190.

Post, W. J., Buijs, C., Stolk, R. P., de Vries, E. G., and Le Cessie, S. (2010). The analysis of longitudinal quality of life measures with informative drop-out: a pattern mixture approach. *Quality of Life Research*, 19(1):137–148.

Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika*, 83(3):551–562.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Sherrill, B., Di Leo, A., Amonkar, M. M., Wu, Y., Zvirbule, Z., Aziz, Z., Bines, J., and Gomez, H. L. (2010). Quality-of-life and quality-adjusted survival (q-twist) in patients receiving lapatinib in combination with paclitaxel as first-line treatment for metastatic breast cancer. *Current Medical Research & Opinion*, 26(4):767–775.

Shi, X. Y., Zhu, H., and Ibrahim, J. G. (2009). Local influence for generalized linear models with missing covariates. *Biometrics*, 65:1164–1174.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Math. Sci*, 153:12–18.

Troxel, A. B. (1998). A comparative analysis of quality of life data from a southwest oncology group randomized trial of advanced colorectal cancer. *Statistics in Medicine*, 17:767–779.

van Steen, K., Molenberghs, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal*, 1:125–142.

Verbeke, G., Molenberghs, G., Thijs, H., Lasaffre, E., and Kenward, M. G. (2001a). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, 57:43–50.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001b). Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics*, 57(1):7–14.

Weiss, R. (1996). An approach to bayesian sensitivity analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 739–750.

Weiss, R. E. and Cook, R. D. (1992). A graphical case statistic for assessing posterior influenceo. *Biometrika*, 79(1):51–55.

Weissfeld, L. A. (1990). Influence diagnostics for the proportional hazards model. *Statistics & probability letters*, 10(5):411–417.

Zhou, X., Cella, D., Cameron, D., Amonkar, M. M., Segreti, A., Stein, S., Walker, M., and Geyer, C. E. (2009). Lapatinib plus capecitabine versus capecitabine alone for her2+ (erbb2+) metastatic breast cancer: quality-of-life assessment. *Breast cancer research and treatment*, 117(3):577–589.

Zhu, H., Ibahim, J., Cho, N., and N, T. (2010). *Bayesian Influence Methods, in Frontiers of Statistical Decision Making and Bayesian Analysis*. eds. M. H. Chen, D. K. Dey, P. Muller, D. Sun and K. Ye, New York: Springer-Verlag, pp. 219-237.

Zhu, H., Ibrahim, J. G., and Chen, Q. (2014a). Bayesian case-deletion model complexity and information criterion. *Statistics and Its Interface*, in press.

Zhu, H., Ibrahim, J. G., and Cho, H. (2012a). Perturbation and scaled cook's distance. *Annals of Statistics*, 40(2):785.

Zhu, H., Ibrahim, J. G., Cho, H., and Tang, N. (2012b). Bayesian case influence measures for statistical models with missing data. *J Comput Graph Stat*, 21(1):253–271.

Zhu, H., Ibrahim, J. G., and Shi, X. (2009). Diagnostic measures for generalized linear models with missing covariates. *Scandinavian Journal of Statistics*, 36(4):686–712.

Zhu, H., Ibrahim, J. G., and Tang, N. (2014b). Bayesian sensitivity analysis of statistical models with missing data. *Statistica Sinica*, 24(2):871–896.

Zhu, H., Lee, S.-Y., Wei, B.-C., and Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, 88(3):727–737.

Zhu, H. T. and Lee, S.-Y. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 63(1):111–126.