# AN INDEX OF LOCAL SENSITIVITY TO NONIGNORABILITY AND A PENALIZED PSEUDOLIKELIHOOD METHOD FOR DATA WITH NONIGNORABLE NONRESPONSE

by

**Fang Zhu**

B.S in Chemistry, Tsinghua University, China, 1999

M.S in Chemistry, Ohio University, 2002

M.S in Mathematics, Ohio University, 2003

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Fang Zhu

It was defended on

June 10 2008

and approved by

Dissertation Advisor:
Gong Tang, PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Sati Mazumdar, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Bruce L. Rollman, MD, MPH
Associate Professor
Department of Medicine and Psychiatry
School of Medicine
University of Pittsburgh

Committee Member:
Chung-Chou Ho Chang, PhD
Assistant Professor
Department of Medicine
School of Medicine
University of Pittsburgh

Committee Member:
Abdus S. Wahed, PhD
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

# AN INDEX OF LOCAL SENSITIVITY TO NONIGNORABILITY AND A PENALIZED PSEUDOLIKELIHOOD METHOD FOR DATA WITH NONIGNORABLE NONRESPONSE

Fang Zhu, PhD

University of Pittsburgh, 2008

The public health significance of this study is to provide researchers and practioners more improved methods to analyze data with missing values as such data get prevalent in practice. When data are missing at random (MAR), the missing-data mechanism can be ignored. Otherwise, the mechanism needs to be modeled. Further sensitivity analyses are often necessary to evaluate the impact of alternative mechanism assumptions on the inferences. For data with nonignorable nonresponse, a pseudolikelihood method was developed, where specification of the mechanism is not necessary. A sensitivity analysis for this method and extensions to nonparametric and semi-parametric regression models were proposed in this thesis.

An index of local sensitivity to nonignorability for the maximum likelihood method ($ISNI_{ML}$) for data with missing outcome values where the missing-data mechanism was modeled by a logistic regression was developed. It is used to evaluate how a small deviation from MAR affects the maximum likelihood estimate. A new index of local sensitivity to nonignorability ($ISNI_{PL}$) was proposed for this pseudolikelihood method in this thesis. Compared with $ISNI_{ML}$, it has the advantage that functional specification of the missing-data mechanism is not required. Depending on whether or not the distribution of the covariate can be parametrically modeled, two versions of this $ISNI_{PL}$ were derived. Simulations suggested that $ISNI_{PL}$ is very close to $ISNI_{ML}$ when the likelihood is correctly specified by the latter. But it does not require assumption on the function form of the missing-data mechanism. The analysis of a real dataset was used to highlight their differences and utility.

In the second part, a penalized pseudolikelihood (PPL) method was developed for semiparametric regression models with the following form: $y = x\beta + g(t) + error$, where $g$ is an unspecified function and can be estimated by a natural cubic spline, for data with nonignorable nonresponse. Two cross-validation methods were considered to find the optimal smoothing parameter. Simulations suggested that PPL with the traditional cross-validation method yields less biased estimates of the parameter of interest and the nonparametric function. This PPL method was also illustrated in analysis of a clinical dataset.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# 1.0 INTRODUCTION

Standard statistical tools are usually designed for data with complete records. However, in practice missing values may occur for various reasons. For example, missing data may occur when study participants refuse to answer certain sensitive questions in a survey, some patients are too sick to have some outcome measures recorded in a clinical trial, some participants may miss scheduled visits or drop out in a longitudinal study. Data with nonresponse, or missing outcome values, are prevalent in survey studies and especially longitudinal studies. Simply making inference based on complete cases, or cases with complete observations, leads to inefficient usage of the data and sometimes misleading conclusions. In general, information on how missing values occurred should be taken into account in the statistical inferential procedure. Here we will focus on statistical methods for analyzing data with missing values in the outcome variables.

## 1.1 MISSING DATA AND MISSING DATA MECHANISM

Traditional statistical methods are developed for complete datasets. In order to apply these methods directly on data with missing values, incomplete cases with missing value have to be deleted before the analysis can be carried out. Such analysis is called complete case analysis and is mostly inadequate or inappropriate because the purpose of a statistical analysis is to understand the properties of the complete data, not merely those of the observed data. Research on data analysis with missing observation can be traced back to as early as 1930s (Allan *et al.*, 1930; Yates, 1933). The milestone in modern statistics analysis with missing data was in 1976, when Rubin recognized the crucial role of the missing data mechanism

(Rubin, 1976). Rubin (1976) defined the nomenclature for the missing data mechanism as the conditional distribution of the missing data indicator given the hypothetically complete data. Formally, considering a multivariate dataset where for individual $i,\ i = 1, 2, \ldots, n$, the covariates $\mathbf{x_i}$ are fully observed and the outcomes $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iK})$ are subject to missing values. The missing data indicator is denoted as $\mathbf{R_i} = (R_{i1}, R_{i2}, \ldots, R_{iK})$ with $R_{ik} = 1$ if $y_{ik}$ is observed and $R_{ik} = 0$ otherwise, $k = 1, 2, \ldots, K$. For convenience, we denote $\mathbf{y_{i,obs}}$ and $\mathbf{y_{i,mis}}$ as the observed and missing parts of $\mathbf{y_i}$. Rubin (1976) categorized general missing-data mechanisms into three classes:

(i) Missing completely at random (MCAR) if the missingness depends on neither the missing values nor the observed values:

$$Pr(\mathbf{R_i}|\mathbf{y_i}, \mathbf{x_i}; \psi) = Pr(\mathbf{R_i}; \psi)$$

.

(ii) Missing at random (MAR) if the missingness does not depend the missing values after conditioning on the observed values:

$$Pr(\mathbf{R_i}|\mathbf{y_i}, \mathbf{x_i}; \psi) = Pr(\mathbf{R_i}|\mathbf{y_{i,obs}}, \mathbf{x_i}; \psi)$$

.

(iii) Missing not at random (MNAR) if the missingness still depends on the missing data after conditioning on the observed values:

$$Pr(\mathbf{R_i}|\mathbf{y_i}, \mathbf{x_i}; \psi) = Pr(\mathbf{R_i}|\mathbf{y_{i,obs}}, \mathbf{y_{i,mis}}, \mathbf{x_i}; \psi),$$

where $\psi$ is the set of model parameters for the missing-data mechanism and $Pr(\cdot)$ will be used throughout this thesis as the probability distribution function. For example, if an individual dropped out of a longitudinal study simply because of relocation, then the missingness is most likely to be MCAR. The data would be MAR if a patient was taken off a treatment because previously observed outcome values looked worrisome to the physician. If a patient on an antidepressant quit the trial because he was not feeling well, then the missingness was more or less associated with the underlying value of psychiatric outcomes such as the Hamilton rating score for depression or PANSS (Positive and Negative Syndrome

Scale) scores. Therefore, most likely, the data are MNAR. Rubin's classification built up the foundation for statistical analysis of data with missing values.

## 1.2 METHODS FOR ANALYSIS OF DATA WITH NONRESPONSE

Many statistical methods were proposed in the past few decades for analysis of data with nonresponse. Standard methods include selection models and pattern-mixture models. Selection models require a model for the hypothetical complete data and another model for the missing-data mechanism. Pattern-mixture models stratify the data based on missing-data patterns and draw conclusion on the distribution of data within each stratum. The model parameters of the complete-data model in selection models have natural interpretation at the population level. Inference on selection models can be obtained either from likelihood-based methods or generalized estimating equation-based (GEE) methods. Multiple imputation-based methods can also be used for making inference. However, these methods often require, explicitly or implicitly, some untestable assumptions about the missing-data mechanism. For example, the functional form of the missing-data mechanism may take various forms but the dataset itself cannot tell which form is the true one. Sensitivity analysis are often recommended to check how alternative assumptions on the missing data mechanism may affect the results and subsequent conclusions. The impact of these alternative assumptions can be assessed through examining the variability of the corresponding inference.

Pattern-mixture models are useful when subpopulations are indeed different across missing data patterns and the interest is on the properties of those subpopulations. Conclusions are drawn within each subpopulation defined by the missing data pattern. Properties on the total population usually have to be formed by a mixture of the corresponding properties from the subpopulations. In general, pattern-mixture models suffer from the problem of nonidentifiability, that is, the joint distribution of variables within incomplete patterns cannot be identified because some variables are completely missing. Usually parameter restrictions across missing data patterns are used to identify model parameters. Often such parameter restrictions come from assumption on the missing-data mechanism. For example,

for a bivariate normal dataset $\{y_{i1}, y_{i2}\}_{i=1,2,\ldots,n}$ where $Y_1$ is fully observed and $Y_2$ is subject to missing values. In the incomplete pattern, only $Y_1$ is observed and the conditional distribution $[Y_2 \mid Y_1, R = 0]$, where $[\cdot]$ is used throughout this thesis to denote a generic distribution, cannot be estimated. If data are MAR, we have $[Y_2 \mid Y_1, R = 0] = [Y_2 \mid Y_1, R = 1]$ and the resulting parameter restrictions lead to identification of $[Y_1, Y_2 \mid R = 0]$. Then $E[Y_2]$ is estimated by a weighted mean of $\widehat{E}[Y_2 \mid R = 1]$ and $\widehat{E}[Y_2 \mid R = 0]$. Because of this identifiability issue and the complexity of imposing parameter restrictions, usage of pattern-mixture models to multivariate data with multiple missing-data patterns is often problematic (Tang *et al.*, 2004).

Compared to the pattern-mixture models, selection models have natural interpretation on model parameters and are more appealing to the investigators. Usually, inference on selection models is based on maximum likelihood where the missing data mechanism is modeled by a parametric form. Methods that are not likelihood-based and do not require a full specification of the missing-data mechanism have also been developed recently (Chen, 2001; Liang & Qin, 2000). A pseudolikelihood method developed by Tang *et al.* (2003) for data with outcome dependent missing is of particular interest here. It is the foundation of the two proposed methods in this thesis. Based on how much information we have on the distribution of the covariates, several variations were available.

In the following sections, we will briefly describe selection models, the subsequent sensitivity analysis, and the pseudolikelihood method by Tang *et al.* (2003). Then present a summary of a local sensitivity index for nonignorability and a penalized pseudolikelihood method for data with nonignorable nonresponse.

### 1.2.1 Standard methods

Consider a dataset $\{\mathbf{x_i}, \mathbf{y_i}\}_{i=1,\ldots n}$, where $n$ is the number of subjects. The covariates $\mathbf{x_i}$ are fully observed and the outcomes $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iK})$ are partially observed. The missing data indicator is $\mathbf{R_i} = (R_{i1}, R_{i2}, \ldots, R_{iK})$, $k = 1, 2, \ldots, K$. $R_{ik}$ is 1 if $y_{ik}$ is observed and $R_{ik} = 0$, otherwise. According to how the joint density of $f(\mathbf{x_i}, \mathbf{y_i}, \mathbf{R_i})$ is factored, standard statistical methods include selection models and pattern-mixture models. Selection models

factor this joint density into two components: one for the distribution of the underlying complete data and one for the conditional distribution of the missing data indicator given the underlying complete data:

$$p(\mathbf{x_i}, \mathbf{y_i}, \mathbf{R_i}) = p(\mathbf{x_i}, \mathbf{y_i}; \theta) p(\mathbf{R_i} | \mathbf{x_i}, \mathbf{y_i}; \psi),$$

where $\theta$, $\psi$ are model parameters and $p(\cdot)$ will be used throughout this thesis as the density function. Pattern-mixture models stratify the data by the patterns of missing values, then model distribution of data within each pattern.

$$p(\mathbf{x_i}, \mathbf{y_i}, \mathbf{R_i}) = p(\mathbf{x_i}, \mathbf{y_i} | \mathbf{R_i}; \delta) p(\mathbf{R_i}; \gamma),$$

where $\delta$ and $\gamma$ are model parameters. Usually $p(\mathbf{x_i}, \mathbf{y_i}; \theta)$, the distribution of complete data, is of interest. The inference from pattern-mixture models, on the other hand, is stratum-specific. In the following context, we will focus on statistical methods for selection models.

The maximum likelihood method (ML) maximizes the likelihood based on $(\mathbf{y_{i,obs}}, \mathbf{R_i})$, $i = 1, 2, \ldots, n$. Denote $\mathbf{X} = \{\mathbf{x_i}\}_{i=1,2,\ldots,n}$ the covariates, $\mathbf{Y} = \{\mathbf{y_i}\}_{i=1,2,\ldots,n}$ the outcome and $\mathbf{R} = \{\mathbf{R_i}\}_{i=1,2,\ldots,n}$ the missing data indicators. The likelihood function is (Diggle and Kenward, 1994; Schluchter, 1992)

$$\begin{aligned} L(\theta, \psi | \mathbf{X}, \mathbf{Y_{obs}}, \mathbf{R}) &\propto \prod_{i=1}^{n} p(\mathbf{y_{i,obs}}, \mathbf{R_i} | \mathbf{x_i}; \theta, \psi) \\ &= \prod_{i=1}^{n} \int p(\mathbf{y_{i,obs}}, \mathbf{y_{i,mis}}, \mathbf{R_i} | \mathbf{x_i}; \theta, \psi) d\mathbf{y_{i,mis}} \\ &= \prod_{i=1}^{n} \int p(\mathbf{y_{i,obs}}, \mathbf{y_{i,mis}} | \mathbf{x_i}; \theta) p(\mathbf{R_i} | \mathbf{x_i}, \mathbf{y_{i,obs}}, \mathbf{y_{i,mis}}; \psi) d\mathbf{y_{i,mis}}. \end{aligned} \quad (1.1)$$

When data are MAR, i.e., $p(\mathbf{R_i} | \mathbf{x_i}, \mathbf{y_{i,obs}}, \mathbf{y_{i,mis}}; \psi) = p(\mathbf{R_i} | \mathbf{x_i}, \mathbf{y_{i,obs}}; \psi)$ and

$$\begin{aligned} L(\theta, \psi | \mathbf{X}, \mathbf{Y_{obs}}, \mathbf{R}) &\propto \prod_{i=1}^{n} \int p(\mathbf{y_{i,obs}}, \mathbf{y_{i,mis}} | \mathbf{x_i}; \theta) p(\mathbf{R_i} | \mathbf{x_i}, \mathbf{y_{i,obs}}; \psi) d\mathbf{y_{i,mis}} \\ &= \prod_{i=1}^{n} \{ p(\mathbf{R_i} | \mathbf{x_i}, \mathbf{y_{i,obs}}; \psi) \int p(\mathbf{y_{i,obs}}, \mathbf{y_{i,mis}} | \mathbf{x_i}; \theta) d\mathbf{y_{i,mis}} \} \\ &= \prod_{i=1}^{n} p(\mathbf{R_i} | \mathbf{x_i}, \mathbf{y_{i,obs}}; \psi) \prod_{i=1}^{n} p(\mathbf{y_{i,obs}} | \mathbf{x_i}; \theta) \\ &\propto L(\psi | \mathbf{R_i}, \mathbf{x_i}, \mathbf{y_{i,obs}}) L(\theta | \mathbf{X}, \mathbf{Y_{obs}}, \mathbf{R}), \end{aligned}$$

where,

$$L(\theta|\mathbf{X}, \mathbf{Y_{obs}}, \mathbf{R}) = \prod_{i=1}^{n} p(\mathbf{y_{i,obs}}|\mathbf{x_i}; \theta)$$

is the ignorable likelihood and

$$L(\psi|\mathbf{R_i}, \mathbf{x_i}, \mathbf{y_{i,obs}}) = \prod_{i=1}^{n} p(\mathbf{R_i}|\mathbf{x_i}, \mathbf{y_{i,obs}}; \psi)$$

is only related to the missing-data mechanism. If $\theta$ and $\psi$ are also distinct, the inference on $\theta$ does not depend on the missing-data mechanism. Therefore when data are MAR, and $\theta$ and $\psi$ are distinct, the missing-data mechanism is ignorable. When data are MNAR, ignoring missing-data mechanisms could lead to biased estimates of $\theta$. In such circumstances, a parametric form has to be assumed for the missing-data mechanism in the ML method. The inference can be highly sensitive to such assumptions.

Inverse-probability weighted estimating equations (IPWEE) is an estimating equation-based method (Robins, Rotnitzky and Zhao, 1994, 1995) to make inference on selection models. A simple version of this method is to weigh each complete case by the inverse-probability of being observed while constructing the estimating equation. The motivation is that each complete case not only represent itself but also other incomplete cases with similar characteristics. It still requires specifying a model for the missing-data mechanism. Misspecification often leads to biased estimates for the model parameters.

Multiple imputation is a simulation-based approach on analysis of missing data. It imputes missing values from an explicit or implicit predictive model for the distribution of the missing values given the observed values. A total of $m > 1$ complete datasets are generated and analyzed using traditional methods as if data were complete. The analysis results from all imputed datasets are combined with between and within imputation variation considered (Rubin, 1987). Therefore multiple imputation still requires some kind of assumption on the missing-data mechanism to derive the predictive model.

These standard methods require specifying the missing-data mechanism, but the observed data do not supply such information. In practice, sensitivity analyses are used to evaluate the impact of alternative assumptions on the parameter estimates of interest.

### 1.2.2 Sensitivity analysis

Sensitivity analysis considers the estimate as a function of a parameter that related to nonignorability. By varying this nonignorability parameter in a plausible range, the impact of these parameters on the key inference is assessed (Rotnitzky *et al.*, 1998). For example, for a bivariate dataset $\{x_i, y_i\}_{i=1,\ldots,n}$, where $x_i$s are fully observed and $y_i$s are partially missing. The missing-data indicator $R_i = 1$ if $y_i$ is observed and $R_i = 0$ otherwise. The missing-data mechanism may be modeled as:

$$Pr[R_i = 1|y_i, x_i] = h(\psi_0 + \psi_1 x_i + \psi_2 y_i), \tag{1.2}$$

where $h(\cdot)$ is a known function. When $\psi_2 = 0$, data are MAR. For a fixed $\psi_2$, the MLE for the regression parameters of $Y$ given $X$, $\hat{\theta}$, is a function of $\psi_2$, $\hat{\theta} = \hat{\theta}(\psi_2)$. By varying $\psi_2$, the resulting curve $(\psi_2, \hat{\theta}(\psi_2))$ can be used to assess the impact of nonignorability on the inference. But it can be computational costly.

Local sensitivity approximations were developed on the basic idea of using an index to *measure* the dependency of the ML estimate on the nonignorability parameter at the neighborhood of MAR. If such local sensitivity is low and there is no evidence of large departure from MAR, the MAR estimate is reasonably close to the true value. Local sensitivity approximations are not as extensive as a global sensitivity test, but they require less computation and, unless there are large nonignorability, they yield reasonable results. Several methods have been proposed (Copas and Li, 1997; Copas and Eguchi, 2001; Verbeke, *et al.* 2001). But none of them can be easily adopted. Troxel *et al.* (2004) developed an index of local sensitivity to nonignorability (ISNI). It provides a more general approach to define local sensitivity with only a minor additional calculation besides MAR modeling calculation. It will be described in detail in chapter 2.

### 1.2.3 A pseudolikelihood method

A pseudolikelihood method (PL) proposed by Tang *et al.* (2003) is to make inference on data with nonresponse without modeling the missing-data mechanism for a class of nonignorable mechanisms. Consider a bivariate dataset $(X, Y)$, where $X = (x_1, x_2, \ldots, x_n)$, n is the

sample size, is fully observed and $Y = (y_1, y_2, \ldots, y_n)$ is a partially observed dependent variable. Assume that the response probability depends on the outcome variable $Y$ alone, i.e., $R$ is independent of $X$ given $Y$. This implies that the complete cases are a random sample from the conditional distribution $X$ given $Y$. Usually, the conditional distribution of $Y$ given $X$ is of interest and it is often assumed by a parametric density, $g(y|x;\theta)$. Denote $f(x;\alpha)$ the marginal distribution of $X$, Tang *et al.* (2003) proposed the following conditional likelihood method and two pseudolikelihood methods for making inference on $\theta$:
When the parametric form of $f(\cdot)$ and the true value of $\alpha$, $\alpha_0$, are known, the following conditional likelihood can be used for inference on $\theta$:

$$PL_0(\theta;\alpha_0) = \prod_{R_i=1} p(x_i|y_i,\theta,\alpha_0) = \prod_{R_i=1} \frac{g(y_i|x_i;\theta)f(x_i;\alpha_0)}{\int g(y_i|x;\theta)f(x;\alpha_0)dx} \propto \prod_{R_i=1} \frac{g(y_i|x_i;\theta)}{\int g(y_i|x;\theta)f(x;\alpha_0)dx}.$$

When $f(\cdot)$ is known but $\alpha_0$ is unknown, $\alpha$ can be estimated by maximizing the marginal likelihood of $X$: $\hat{\alpha} = \arg\max_\alpha \prod_{i=1}^n f(x_i;\alpha)$. A pseudolikelihood can be constructed as

$$PL_1(\theta;\hat{\alpha}) = \prod_{R_i=1} \frac{g(y_i|x_i;\theta)}{\int g(y_i|x;\theta)f(x;\hat{\alpha})dx},$$

and $\theta$ is estimated by maximizing $PL_1(\theta;\hat{\alpha})$ as a function of $\theta$. However, in practice, the functional form of $f(\cdot)$ is unknown and not of interest. Another pseudolikelihood method was proposed by maximizing

$$PL_2(\theta;F_n) = \prod_{R_i=1} \frac{g(y_i|x_i;\theta)}{\int g(y_i|x;\theta)dF_n(x)},$$

where $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(x \le x_i)$ is the empirical distribution of $X$.

Denote $PL_0$, $PL_1$ and $PL_2$ the estimates of the conditional and two pseudolikelihood methods respectively. Under some regularity conditions, all these estimates are consistent and asymptotically normal. $PL_1$ is more efficient than $PL_0$. Simulation studies suggested that $PL_2$, which requires no assumption about the distribution of $X$, is even more efficient than $PL_1$ (Tang *et al.* 2003).

This pseudolikelihood method can be extended to a general class of MNAR mechanisms, such as

$$Pr(R_i = 1|x_i, y_i) = \omega(\lambda y_i + x_i), \tag{1.3}$$

where $\lambda$ is a known constant *a priori*. A new dataset $(X, Y_\lambda)$, where $Y_\lambda = \lambda Y + X$ can be constructed from the original data. Inference on $\theta$ can be made by applying the pseudo-likelihood method on this generated dataset. Such extension can also be used as sensitivity analysis by looking at the parameter estimates under a range of $\lambda$ values.

## 1.3 PROPOSED METHODS FOR DATA WITH NONRESPONSE

Two methods will be proposed here for analysis of data with nonignorable nonresponse. Both of them are related to the pseudolikelihood method developed by Tang *et al.* (2003). First we would develop a local sensitivity index for this method, then extend this method to semi-parametric regression models with penalized spline for analysis of data with nonignorable nonresponse.

### 1.3.1 ISNI for a pseudolikelihood method

For bivariate data with nonresponse, as mentioned in the previous section, the standard ML method requires a model for the missing-data mechanism. A popular choice is $h = logit^{-1}$ in (1.2). The index of local sensitivity of nonignorability, ISNI, by Troxel *et al.* is defined as the first derivatives of the MLE with respect to $\psi_2$ (Troxel *et al.*, 2004, Ma, G., *et al.*, 2005). By fixing $\psi_2$, $\hat{\theta}$ or $\hat{\theta}(\psi_2)$ can be derived by maximizing the joint likelihood (1.1) and ISNI can be computed by

$$ISNI = \frac{\partial \hat{\theta}}{\partial \psi_2}\Big|_{\psi_2=0}.$$

Similarly, with $\lambda$ in (1.3) fixed, $\hat{\theta}(\lambda)$ can be estimated from the pseudolikelihood method. When $\lambda = 0$, data are MAR and the pseudolikelihood method produces the same estimate as the ignorable maximum likelihood method. A new index of local sensitivity to nonignorability for this pseudolikelihood method can be defined similarly as the first derivatives of the maximum pseudolikelihood estimate with respect to $\lambda$ at $\lambda = 0$. It does not make assumption on the parametric form of the missing-data mechanism, hence is more flexible than ISNI for

the maximum likelihood method. The details on the developement and the utility of this new local sensitivity index will be presented in Chapter 2.

### 1.3.2 A penalized pseudolikelihood method

Semi-parametric regressions with penalized spline supply a flexible and powerful regression tool when the contribution from a predictor is either nonlinear in nature or not of interest. A smoothing parameter is used to control the smoothness of the spline. It is generally chosen by cross-validation or generalized cross-validation method (Green and Silverman, 1994). The theory behind it is quite developed for complete data. However, the dependent variable may be subject to nonresponse in practice and standard semi-parametric regressions cannot be directly applied. In Chapter 3, a penalized pseudolikelihood method is proposed to incorporate data with nonignorable nonresponse. Two cross validation methods are discussed and compared via simulation studies. This penalized pseudolikelihood method is illustrated through analysis of a dataset from a psychiatric clinical study.

# 2.0 AN INDEX OF LOCAL SENSITIVITY TO NONIGNORABILITY FOR A PSEUDOLIKELIHOOD METHOD

In general, maximum likelihood inference for selection models requires specification of the missing-data mechanism unless the data are MAR. Unfortunately MAR is untestable because the dataset itself cannot tell whether or not the data are MAR. Misspecification of the missing-data mechanism often leads to biased estimates and incorrect conclusions. Sensitivity analyses are usually carried out to assess the impact that the alternative missing-data mechanism assumptions have on the parameter estimates (Rotnitzky *et al.*, 1998). Local sensitivity analyses usually check the local properties of such sensitivity analyses at the neighborhood of MAR. If parameter estimates are not sensitive to a slight deviation from the MAR assumption and there is no evidence of large departure from MAR, the parameter estimates under MAR are acceptable. ISNI, developed by Troxel *et al.* (2004), is a local sensitivity index for such purpose. The definition will be introduced in Section 2.1. ISNI supplies an intuitive measure on how fast the maximum likelihood estimates, under alternative MNAR mechanisms within a parametric family, may change when the missing-data mechanism deviates from MAR. The computation process of this index requires the assumptin on the parametric function form of the missing data mechanism (Troxel *et al.*, 2004). A popular choice is logistic regression. This assumption hampers the adaptability of ISNI. We adopted the idea of ISNI and developed a new index based on a pseudolikelihood method (Tang *et al.*, 2003). This new index can be used for analysis of a more general class of missing data.

## 2.1 ISNI FOR THE MAXIMUM LIKELIHOOD METHOD

Consider a bivariate dataset $\{x_i, y_i\}_{i=1,\ldots,n}$, where $x_i$s are fully observed and $y_i$s are missing for $i = m+1, \ldots, n$. The missing data indicator is denoted by $R_i : R_i = 1$ for $i = 1, \ldots, m$, $R_i = 0$ for $i = m+1, \ldots, n$. Assume that

$$[Y|X] \sim g(y|x, \theta),$$

where $\theta$ is the parameter of interest. A typical selection model assumes the following missing data mechanism:

$$Pr[R_i = 1|y_i, x_i] = logit^{-1}(\psi_0 + \psi_1 x_i + \psi_2 y_i). \tag{2.1}$$

Even though the logit link is used, a generalization to other mechanisms is possible if the missing data model is monotone in the outcome variable. Denote $\psi = (\psi_0, \psi_1, \psi_2)$. The log likelihood is

$$\begin{aligned} l(\theta, \psi) \quad &= \textstyle\sum_{i=1}^{n} \Big[ R_i \left\{ \log g(y_i|x_i, \theta) + \log logit^{-1}(\psi_0 + \psi_1 x_i + \psi_2 y_i) \right\} \\ &\quad + (1 - R_i) \log[\textstyle\int g(u|x_i, \theta) \left\{ 1 - logit^{-1}(\psi_0 + \psi_1 x_i + \psi_2 u) \right\} du] \Big] \end{aligned} \tag{2.2}$$

As in any sensitivity analysis, the MLE $\hat{\theta}$ can be represented as a function of $\psi_2$ for a range of possible values (Figure 1). When $\psi_2 = 0$, data are MAR. In the neighborhood of MAR, the deviation of parameter estimates under MNAR mechanisms from the estimate under MAR can be represented by the slope of the tangent line at $\psi_2 = 0$ (Figure 1). Based on this observation, a natural local index of sensitivity to nonignorability (ISNI) for the maximum likelihood method was proposed by Troxel *et al.* (2004):

$$ISNI = \frac{\partial \hat{\theta}(\psi_2)}{\partial \psi_2}\Big|_{\psi_2 = 0} \tag{2.3}$$

To differentiate it from the new local sensitivity index that would be introduced later, ISNI for the maximum likelihood method will be denoted as $ISNI_{ML}$ in the following context.

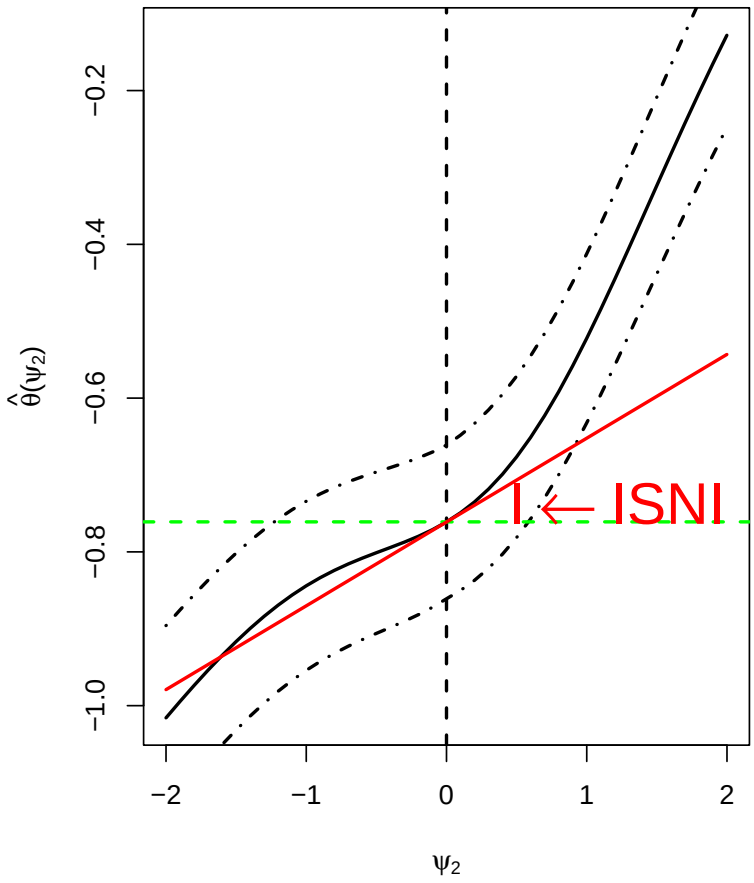


Figure 1: Definition of $ISNI$ for the maximum likelihood method

Denote the other missing data mechanism parameters by $\Psi = (\psi_0, \psi_1)$. This index can be computed by the following formula

$$ISNI_{ML} = -(\bigtriangledown^2 L_{11})^{-1} \bigtriangledown^2 L_{13}, \tag{2.4}$$

where

$$\bigtriangledown^2 L = \left( \begin{array}{ccc} \frac{\partial^2 L}{\partial\theta\partial\theta'} & \frac{\partial^2 L}{\partial\theta\partial\Psi'} & \frac{\partial^2 L}{\partial\theta\partial\psi_2} \\ \frac{\partial^2 L}{\partial\Psi\partial\theta'} & \frac{\partial^2 L}{\partial\Psi\partial\Psi'} & \frac{\partial^2 L}{\partial\Psi\partial\psi_2} \\ \frac{\partial^2 L}{\partial\psi_2\partial\theta'} & \frac{\partial^2 L}{\partial\psi_2\partial\Psi'} & \frac{\partial^2 L}{\partial\psi_2^2} \end{array} \right) \Bigg|_{\theta=\hat{\theta}_0, \Psi=\hat{\Psi}_0, \psi_2=0}$$

and $(\hat{\theta}_0, \hat{\Psi}_0)$ are the MLEs under MAR or $\psi_2 = 0$, and $\{\bigtriangledown^2 L_{ij}\}_{i,j=1,2,3}$ is subsequent $(i, j)$ element of above matrix.

$ISNI_{ML}$ depends on the scale of $Y$ when $Y$ can be re-scaled, for instance, when $Y$ is an interval variable. Under such circumstances, a sensitivity transformation below was considered by Troxel *et al.* (2004):

$$c_{ML} = |\sigma_Y SE_Y / ISNI_{ML(Y)}|,$$

where $\sigma_Y$ is the standard deviation (SD) of $Y$, $SE_Y$ is the standard error (SE) of $\hat{\theta}_0$ and $ISNI_{ML(Y)}$ is the $ISNI_{ML}$ from data with outcome $Y$ (Troxel *et al.*, 2004, Ma, G., *et al.*, 2005). In practice, the SD of $Y$ can be estimated from the observed data under the MAR assumption, $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$ and $\hat{\sigma}_Y = \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2$. This new index $c_{ML}$ is scale-independent. To interpret $c_{ML}$, notice that when $Y$ is transformed to $\frac{c_{ML}}{\sigma_Y}Y$, the missing data mechanism is

$$log\frac{Pr[R=1|y,x]}{1-Pr[R=1|y,x]} = \psi_0 + \psi_1 x + \frac{c_{ML}}{\sigma_Y} y.$$

That means a change of $\sigma_Y / c_{ML}$ in $Y$ is associated with an odds of 2.7 in response probability. At the same time, considering that $\hat{\theta}(\psi_2) \approx \hat{\theta}_0 + ISNI_{ML} \cdot \psi_2$ at the neighborhood of $\psi_2 = 0$, under this re-scaled data $\frac{c_{ML}}{\sigma_Y}Y$,

$$ISNI_{\frac{c_{ML}}{\sigma_Y}Y} / SE_{\frac{c_{ML}}{\sigma_Y}Y} = \frac{c_{ML}}{\sigma_Y} ISNI_Y / SE_Y = 1,$$

where $ISNI_{ML\left(\frac{c_{ML}}{\sigma_Y}Y\right)}$ and $SE_{\frac{c_{ML}}{\sigma_Y}Y}$ are the $ISNI_{ML}$ and standard error of $\hat{\theta}_0$ from the data with the outcome re-scaled to $\frac{c_{ML}}{\sigma_Y}Y$. The maximum likelihood estimate $\hat{\theta}$ is about one $SE_{c_{ML}Y/\sigma_Y}$ away from $\hat{\theta}_0$. A large $c_{ML}$, for example, $c_{ML} = 10$, means 0.1 SD change on $Y$

substantially changes the odds of being observed. This corresponds to a very extreme non-ignorability. Such a magnitude of nonignorability brings about a deviation of one standard error from $\hat{\theta}_0$. We would say the ML estimate is not sensitive to nonignorability assumption. If $c_{ML}$ is small, for instance, $c_{ML} = 0.1$, 10 SDs change on $Y$ is associated with a change of 2.7 in odds of being observed. However such a weak nonignorability mechanism leads to parameter estimates about one standard error from $\hat{\theta}_0$. It would suggest that the inference is quite sensitive to local deviation from MAR. More comprehensive sensitivity analyses have to be carried out in such circumstances. A cutoff point of $c_{ML} = 1$ was recommended for local sensitivity evaluation (Troxel, *et al.*, 2004).

$ISNI_{ML}$ is easy to compute. But when the missing data mechanism is not monotone in $Y$ or there is minimal information on the functional form of $Y$, it may lead to wrong conclusions about the local sensitivities. To circumvent the specification of missing data mechanism, we adopted the idea of $ISNI_{ML}$ and developed a new local sensitivity index for the pseudolikelihood method proposed by Tang *et al.* (2003).

## 2.2 ISNI FOR A PSEUDOLIKELIHOOD METHOD

For the same bivariate dataset $\{x_i, y_i\}_{i=1,\ldots,n}$ with $y_i$s subject to missing values, consider a general class of missing-data mechanisms with the following form:

$$Pr(R_i = 1|x_i, y_i) = w(x_i + \lambda y_i), \tag{2.5}$$

where $w(\cdot)$ is an arbitrary non-constant function and $\lambda$ serves as the nonignorability parameter. When $\lambda = 0$, data are MAR. Let $y_{\lambda i} = x_i + \lambda y_i$, and $\widehat{F}(x)$ be a consistent estimator of $F(x)$, the cumulative distribution function of $X$. For fixed $\lambda$, the pseudolikelihood method

maximizes

$$\begin{aligned} PL(\theta;\lambda,\widehat{F}) &= \sum_{R_i=1}\log p(x_i|y_{\lambda i};\theta,\lambda,\widehat{F}) \propto \sum_{R_i=1}\log\left\{\frac{g(y_{\lambda i}|x_i;\theta)}{p(y_{\lambda i};\theta,\lambda,\widehat{F})}\right\} \\ &\propto \sum_{R_i=1}\log\left\{\frac{g(y_i|x_i;\theta)}{p(y_{\lambda i};\theta,\lambda,\widehat{F})}\right\} \\ &= \sum_{R_i=1}\Big[\log\{g(y_i|x_i;\theta)\}-\log\{p(y_{\lambda i};\theta,\lambda,\widehat{F})\}\Big]. \end{aligned} \tag{2.6}$$

Following the same rational of $ISNI_{ML}$, a new local index for sensitivity to nonignorability is developed here for the above pseudolikelihood method (Tang, *et al.*, 2003), where the specification on the function form of the missing-data mechanism is not required. If $\hat{\theta}(\lambda)$ is the pseudolikelihood estimate of $\theta$, given a fixed $\lambda$, this index $ISNI_{PL}$ is defined as

$$ISNI_{PL} = \frac{\partial\hat{\theta}(\lambda)}{\partial\lambda}\bigg|_{\lambda=0}. \tag{2.7}$$

Consider $PL(\theta;\lambda,\widehat{F})$ as a function of $(\theta,\lambda)$, $PL(\theta,\lambda;\widehat{F})$. If $\hat{\theta}_0 = \arg\max_\theta PL(\theta,0;\widehat{F})$ is the MAR estimate, carrying out a Taylor expansion of $PL(\theta,\lambda;\widehat{F})$ at MAR point $(\theta,\lambda) = (\hat{\theta}_0,0)$ would give,

$$PL(\theta,\lambda;\widehat{F}) \approx PL(\hat{\theta}_0,0;\widehat{F}) + [(\theta-\hat{\theta}_0)',\lambda] \bigtriangledown PL + \frac{1}{2}[(\theta-\hat{\theta}_0)',\lambda] \bigtriangledown^2 PL[(\theta-\hat{\theta}_0)',\lambda]' \tag{2.8}$$

where

$$\bigtriangledown PL = \begin{pmatrix} \frac{\partial PL}{\partial\theta} \\ \frac{\partial PL}{\partial\lambda} \end{pmatrix}\bigg|_{\theta=\hat{\theta}_0,\lambda=0}$$

$$\bigtriangledown^2 PL = \begin{pmatrix} \frac{\partial^2 PL}{\partial\theta\partial\theta'} & \frac{\partial^2 PL}{\partial\theta\partial\lambda} \\ \frac{\partial^2 PL}{\partial\lambda\partial\theta'} & \frac{\partial^2 PL}{\partial\lambda^2} \end{pmatrix}\bigg|_{\theta=\hat{\theta}_0,\lambda=0}$$

and $\{\bigtriangledown^2 PL_{ij}\}_{i,j=1,2}$ is subsequent $(i,j)$ element of above matrix.

When data are MAR, $\frac{\partial PL}{\partial\theta}\Big|_{\theta=\hat{\theta}_0,\lambda=0} = 0$. Take derivatives with respect to $\theta$ from both sides of the equation (2.8) at $\hat{\theta}(\lambda)$ for any fixed $\lambda$

$$0 = \frac{\partial PL(\theta,\lambda,\widehat{F})}{\partial\theta}\bigg|_{\hat{\theta}(\lambda),\lambda} \approx (\hat{\theta}(\lambda)-\hat{\theta}_0) \bigtriangledown^2 PL_{11} + \lambda \bigtriangledown^2 PL_{12}.$$

A function form $\hat{\theta}(\lambda)$ can be derived as:

$$\hat{\theta}(\lambda) = -(\bigtriangledown^2 PL_{11})^{-1}(\hat{\theta}_0 \bigtriangledown^2 PL_{11} - \lambda \bigtriangledown^2 PL_{12}) + o(\lambda)$$

Subsequently taking the first derivative of $\hat{\theta}(\lambda)$ with respect to $\lambda$, we have

$$ISNI_{PL} = -(\bigtriangledown^2 PL_{11})^{-1} \bigtriangledown^2 PL_{12}. \tag{2.9}$$

Notice that, from Slutzky's Lemma, $Y_\lambda \to X$ in law as $\lambda \to 0$. So if $f$ is the probability distribution function of $X$, $\frac{1}{n}\sum_{R_i=1}[\log\{p(y_{\lambda i};\theta)\} - \log\{f(x_i)] \to 0$. The first and second derivatives of $\log\{p(y_{\lambda i};\theta)\}$ with respect to $\theta$ at $(\hat{\theta}_0, \lambda)$ converges to zero as $\lambda \to 0$. Thus, when $\lambda \to 0$, $\hat{\theta}_\lambda$ converges to $\hat{\theta}_0$ and $\nabla^2 PL_{11}$ depends only on $\sum_{R_i=1}\log\{g(y_i|x_i;\theta)\}$, the ignorable log-likelihood function.

To better interpret $ISNI_{PL}$, we will use the same logic as the interpretation of $ISNI_{ML}$. Because $ISNI_{PL}$ is the derivative of $\hat{\theta}$ with respect to a nonignorability parameter $\lambda$, $\hat{\theta}(\lambda) \approx \hat{\theta}_0 + ISNI_{PL} \cdot \lambda$ at the neighborhood of $\lambda = 0$. If $\lambda = 1$, the adjustment to the MAR estimate $\hat{\theta}_0$ is the corresponding $ISNI_{PL}$. We can consider the ratio of ISNI to the standard error (SE) of the parameter of interest $\theta$ when data are MAR. If this ratio is larger than one, a unit change in the nonignorability parameter would bring more than one SE deviation from the MAR estimate. A deviation of this magnitude is usually considered having substantial impact on the inference and subsequent conclusion.

Similar to $ISNI_{ML}$, $ISNI_{PL}$ is not scale free when the outcome $Y$ can be re-scaled. Denote $ISNI_{PL(Y)}$ and $ISNI_{PL(aY)}$, $a$ is any constant, the $ISNI_{PL}$s from data with outcome $Y$ and $aY$, respectively. Denote $SE_{aY}$ and $SE_Y$ the standard errors of $\hat{\theta}_0$ from data with outcome $aY$ and $Y$, respectively. Relation

$$ISNI_{PL(aY)}/SE_{(aY)} = aISNI_{PL(Y)}/SE_{(Y)}$$

holds between $ISNI_{PL}$ derived from the transformed data $aY$ and from $Y$. A parameter $c_{PL}$, that results in $ISNI_{PL(cY/\sigma_Y)}/SE_{(cY/\sigma_Y)} = 1$, is an important indicator. This transformation $c_{PL}$ can be derived as:

$$c_{PL} = \left|\sigma_Y SE_Y / ISNI_{PL(Y)}\right|.$$

So a missing data mechanism

$$Pr[R = 1|X, Y] = \omega(X + \frac{c_{PL}}{\sigma_Y} Y)$$

leads to a pseudolikelihood estimate $\hat{\theta}$ to be about $SE_Y$ from $\hat{\theta}_0$. If it is speculated that the relative impact of $Y$ in the missing data mechanism does not exceed $\frac{c_{PL}}{\sigma_Y}$, we would expect that the corresponding pseudolikelihood estimate would not differ from $\hat{\theta}_0$ by over one $SE_Y$ of $\hat{\theta}_0$

$$\hat{\theta}(\lambda) - \hat{\theta}_0 \approx ISNI_{PL(Y)} \cdot \frac{c_{PL}}{\sigma_Y} = SE_Y.$$

If $\omega$ is the logit link, when $X$ is fixed but $Y$ is changed by $\sigma_Y/c_{PL}$, this relative impact of $Y$ is associated with an odds of 2.7 of being observed. A small value of $c_{PL}$ implies a weak nonignorable mechanism may cause significant deviation from the MAR estimate. For example, if $c_{PL} = 0.1$, a change of 10 SDs on $Y$ is corresponding to an odds of 2.7 in response probability. For such a weak nonignorable mechanism, the pseudolikelihood estimate is about one $SE$ of $\hat{\theta}_0$ from the MAR estimate $\hat{\theta}_0$. Therefore the pseudolikelihood method is very sensitive to nonignorability. On the other hand, a large $c_{PL}$ would mean that the pseudolikelihood method is not sensitive to nonignorability. For example, if $c_{PL} = 10$, then a change of 0.1 SD in $Y$ is corresponding to an odds of 2.7 in response probability. For such a strong nonignorable mechanism, the pseudolikelihood estimate is about one $SE$ of $\hat{\theta}_0$ away from the MAR estimate $\hat{\theta}_0$. Therefore the pseudolikelihood estimate is not sensitive to nonignorability. For general $\omega$, it is difficult to evaluate the degree how a change in $Y$ affects the response probability but we would recommend a cutoff point of 1 in practice. When the link function is logit, a cutoff at 1 is reasonable and is consistent with the choice for ISNI under the ML approach.

A very important difference between PL and ML methods is, in (2.6), the consistent estimator of $F$, $\hat{F}$, needs to be derived from the marginal distribution of $X$. When the functional form of $F$ is known, for instance, $F(x) = F(x;\alpha)$, the estimate $\hat{F}(x)$ can be derived by replace $\alpha$ in $F(x;\alpha)$ by $\hat{\alpha} = arg\max_\alpha \prod_{i=1}^n f(x_i;\alpha)$: $\hat{F}(x) = F(x;\hat{\alpha})$. When the functional form of $F$ is unknown, ideally we would like to derive $ISNI_{PL}$ with $\hat{F} = F_n(x)$, the empirical function of $X$. However, we encountered great difficulty in deriving of the analytical form for $ISNI_{PL}$ and could not work it out at this moment. A compromise was

carried out by using an kernel estimator of $F(x)$ as $\hat{F}$ in (2.6). In the following sections, we would present the parametric version of $ISNI_{PL}$, denoted by $ISNI_{PL1}$, for bivariate normal data and the nonparametric version of $ISNI_{PL}$, denoted by $ISNI_{PL2}$. Simulation studies were carried out to evaluate their performance and analysis of a real dataset was used for illustration.

### 2.2.1 ISNI for bivariate normal data

For $ISNI_{PL1}$, consider bivariate normal data $\{x_i, y_i\}_{i=1,\ldots,n}$ with

$$[X] \sim N(\mu_x, \sigma_x^2), \quad [Y|X] \sim N(\beta_0 + \beta_1 x, \sigma^2), \tag{2.10}$$

where $\theta = (\beta_0, \beta_1, \sigma^2)$ are the parameters of interest. Assume that $y_1, y_2, \ldots, y_m$ are observed and $y_{m+1}, \ldots, y_n$ are missing. Then for a given $\lambda$, the conditional distribution of $Y_\lambda = X + \lambda Y$ given $X$ is

$$[Y_\lambda|X] \ \sim N(\lambda\beta_0 + \beta_\lambda x, \lambda^2\sigma^2), \tag{2.11}$$

and $\beta_\lambda = \lambda\beta_1 + 1$. Let $\hat{\mu}_x$ and $\hat{\sigma}_x^2$ be the consistent estimator of $\mu_x$ and $\sigma_x^2$:

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2. \tag{2.12}$$

Then a parametric estimator of $F$ is the cumulative distribution function of $N(\hat{\mu}_x, \hat{\sigma}_x^2)$. The logarithm of the pseudolikelihood function is:

$$\begin{aligned} PL(\beta_0, \beta_1, \sigma^2, \lambda; \hat{\mu}_x, \hat{\sigma}_x^2) \ &= \ -\frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \beta_0 - \beta_1 x_i)^2 \\ &\quad + \frac{m}{2}\log(\lambda^2\sigma^2 + \beta_\lambda^2\hat{\sigma}_x^2) + \frac{\sum_{i=1}^{m}(y_{\lambda i} - \lambda\beta_0 - \beta_\lambda\hat{\mu}_x)^2}{2(\lambda^2\sigma^2 + \beta_\lambda^2\hat{\sigma}_x^2)}, \end{aligned} \tag{2.13}$$

where $y_{\lambda i} = \lambda y_i + x_i$. The first derivative of $PL$ with respect to $\theta$ at MAR, $(\hat{\theta}_0 = (\hat{\beta}_{00}, \hat{\beta}_{10}, \hat{\sigma}_0^2), \lambda = 0)$ is

$$\bigtriangledown PL = \begin{pmatrix} \frac{1}{\hat{\sigma}_0^2}\sum_{i=1}^{m}(y_i - \hat{\beta}_{00} - \hat{\beta}_{10}x_i) \\ \frac{1}{\hat{\sigma}_0^2}\sum_{i=1}^{m}(y_i - \hat{\beta}_{00} - \hat{\beta}_{10}x_i)x_i \\ -\frac{m}{2\hat{\sigma}_0^2} + \frac{1}{2\hat{\sigma}_0^4}\sum_{i=1}^{m}(y_i - \hat{\beta}_{00} - \hat{\beta}_{10}x_i)^2 \\ m\hat{\beta}_{10} + \frac{1}{\hat{\sigma}_x^2}\sum_{i=1}^{m}[(y_i - \hat{\beta}_{00} - \hat{\beta}_{10}\hat{\mu}_x)^2(x_i - \hat{\mu}_x)] - \frac{1}{\hat{\sigma}_x^2}\sum_{i=1}^{m}\hat{\beta}_{10}(x_i - \hat{\mu}_x)^2 \end{pmatrix}.$$

The first three elements are exactly the same with the score matrix from ignorable likelihood. It confirmed that $\hat{\theta}(\lambda) \to \hat{\theta}_0$ as $\lambda \to 0$.

Denote $\bar{x} = \sum_{i=1}^{m} x_i$, $\bar{y} = \sum_{i=1}^{m} y_i$, $s_{11} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$, $s_{12} = \frac{1}{m-1}\sum_{i=1}^{m}\{(x_i - \bar{x})(y_i - \bar{y})\}$ and $s_{22} = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2$. Then as $\lambda \to 0$, $(\hat{\beta}_{00}, \hat{\beta}_{01}, \hat{\sigma}_0^2) = (\bar{y} - \frac{s_{12}}{s_{11}}\bar{x}, \frac{s_{12}}{s_{11}}, s_{22} - \frac{s_{12}^2}{s_{11}})$. Matrices $\nabla^2 PL_{11}$ and $\nabla^2 PL_{12}$ of (2.9) are:

$$\nabla^2 PL_{11} = -\frac{1}{\hat{\sigma}_0^2}\begin{pmatrix} m & \sum_{i=1}^{m} x & 0 \\ \sum_{i=1}^{m} x & \sum_{i=1}^{m} x_i^2 & 0 \\ 0 & 0 & \frac{m}{2\hat{\sigma}_0^2} \end{pmatrix},$$

$$\nabla^2 PL_{12} = \frac{1}{\hat{\sigma}_x^2}\begin{pmatrix} -m(\bar{x} - \hat{\mu}_x) \\ m\hat{\sigma}_x^2 - \sum_{i=1}^{m} x_i^2 + m\hat{\mu}_x\bar{x} \\ 0 \end{pmatrix} \tag{2.14}$$

$ISNI_{PL1}$ can be derived from (2.9):

$$\frac{\partial\hat{\theta}}{\partial\lambda}\Big|_{\lambda=0} = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_x^2}\begin{pmatrix} m & \sum_{i=1}^{m} x & 0 \\ \sum_{i=1}^{m} x & \sum_{i=1}^{m} x_i^2 & 0 \\ 0 & 0 & \frac{m}{2\hat{\sigma}_0^2} \end{pmatrix}^{-1}\begin{pmatrix} -m(\bar{x} - \hat{\mu}_x) \\ m\hat{\sigma}_x^2 - \sum_{i=1}^{m} x_i^2 + m\hat{\mu}_x\bar{x} \\ 0 \end{pmatrix} \tag{2.15}$$

After simplification, (2.15) can be written as

$$\begin{aligned} \frac{\partial\hat{\beta}_0}{\partial\lambda}\Big|_{\theta=\hat{\theta}_0,\lambda\to 0} &= -\frac{s_{22}s_{11} - s_{12}^2}{s_{11}^2}\bar{x} + \frac{\hat{\mu}_x}{\hat{\sigma}_x^2}\frac{s_{22}s_{11} - s_{12}^2}{s_{11}} \\ \frac{\partial\hat{\beta}_1}{\partial\lambda}\Big|_{\theta=\hat{\theta}_0,\lambda\to 0} &= \frac{s_{22}s_{11} - s_{12}^2}{s_{11}^2\hat{\sigma}_x^2}(\hat{\sigma}_x^2 - s_{11}) \\ \frac{\partial\hat{\sigma}^2}{\partial\lambda}\Big|_{\theta=\hat{\theta}_0,\lambda\to 0} &= 0 \end{aligned}$$

The detailed calculation can be found in Appendix A.1.

An alternative way of deriving $ISNI_{PL1}$ is to find the analytic forms of $\hat{\theta}_\lambda$, then their first derivatives at $\lambda = 0$. The estimator of the parameters are:

$$\begin{aligned}
\hat{\beta}_0(\lambda) &= \bar{y} - b_\lambda \bar{x} - \frac{\hat{\mu}_x}{\hat{\sigma}_x^2}(s_{12} - b_\lambda s_{11}), \\
\hat{\beta}_1(\lambda) &= \frac{1}{\hat{\sigma}_x^2}\{s_{12} + b_\lambda(\hat{\sigma}_x^2 - s_{11})\}, \\
\hat{\sigma}_y^2(\lambda) &= s_{22} + b_\lambda^2(\hat{\sigma}_x - s_{11}), \\
\hat{\sigma}^2(\lambda) &= s_{22} - \frac{b_\lambda^2}{\hat{\sigma}_x^2}(\hat{\sigma}_x^2 - s_{11})^2 + (b_\lambda^2 - \frac{2s_{12}b_\lambda}{\hat{\sigma}_x^2})(\hat{\sigma}_x^2 - s_{11}) - \frac{s_{12}^2}{\hat{\sigma}_x^2},
\end{aligned}$$

where $b_\lambda = \frac{\lambda s_{22} + s_{12}}{\lambda s_{12} + s_{11}} \overset{\lambda \to 0}{\to} \frac{s_{12}}{s_{11}}$ (Brown, 1990). Take the first derivatives with respect to $\theta$ at $\lambda = 0$, the results are exactly the same from the one we derived above. Details can be found in Appendix A.2.

Troxel *et al.* (2004) also derived $ISNI_{ML}$ for bivariate normal data with nonresponse

$$ISNI_{ML} = -\hat{\sigma}_0^2 \begin{pmatrix} m & \sum_{i=1}^m x \\ \sum_{i=1}^m x & \sum_{i=1}^m x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=m}^n h_i \\ \sum_{i=m}^n x_i h_i \end{pmatrix} \tag{2.16}$$

where $h_i = Pr[Y_i \text{ is observed} | X_i = x_i]$ and is derived by fitting a logistic regression.

The newly developed index $ISNI_{PL1}$ can be represented in matrix form as

$$ISNI_{PL1} = -\hat{\sigma}_0^2 \begin{pmatrix} m & \sum_{i=1}^m x \\ \sum_{i=1}^m x & \sum_{i=1}^m x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} m(\bar{x} - \hat{\mu}_x)/\hat{\sigma}_x^2 \\ -m + (\sum_{i=1}^m x_i^2 - m\hat{\mu}_x\bar{x})/\hat{\sigma}_x^2 \end{pmatrix}$$

Comparing $ISNI_{ML}$ and $ISNI_{PL1}$, the only difference is the last matrix. It is not clear how they are related just from these formulae. The performance of $ISNI_{ML}$ and $ISNI_{PL1}$ were compared through simulation studies in Section 2.3.

This $ISNI_{PL1}$ was developed for bivariate normal data. However, in reality, the distribution of $X$ is generally not of interest and may not be normal. In the next section, we developed the $ISNI_{PL}$ when the distribution of $X$ is unknown by using a kernel estimator for $F(x)$ in (2.6).

### 2.2.2 ISNI when the distribution of $X$ is unknown

Consider the pseudolikelihood method (2.6) when the distribution of the covariate $X$ is unknown and $\hat{F}$ is the empirical distribution of $X$, $\hat{F} = F_n(x)$. The derivation of $ISNI_{PL}$ under this scenario should still follow formula (2.9). The matrix $\nabla^2 PL_{11}$ is still the information matrix corresponding to the ignorable maximum likelihood. The second part of the equation $\nabla^2 PL_{12}$ is the partial derivatives of the logarithm of PL

$$PL(\beta_0, \beta_1, \sigma^2; \widehat{F}) = -\frac{m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \beta_0 - \beta_1 x_i)^2 + \sum_{i=1}^{m}\log p(y_{\lambda i}; \theta, \widehat{F}_n(x))$$

with respect to $\theta$ and $\lambda$. It is determined simply by the term $\log p(y_{\lambda i}; \theta, \widehat{F}_n(x))$. Although as $\lambda \to 0$, $Y_\lambda \to X$. So

$$\frac{1}{n}\sum_{i=1}^{m}\log p(y_{\lambda i}; \theta, \widehat{F}_n(x)) - \frac{1}{n}\sum_{i=1}^{m}\log f(x_i) \to 0, \tag{2.17}$$

it is not clear how to derive the analytical formula for $\nabla^2 PL_{12}$ because as $\lambda \to 0$,

$$p(y_{\lambda i}; \theta, \widehat{F}_n(x)) = \frac{1}{n\sqrt{2\pi\lambda^2\sigma^2}}\sum_{j=1}^{n}\exp\left\{-\frac{\{(y_i - \beta_0 - \beta_1 x_j) + (x_i - x_j)/\lambda\}^2}{2\sigma^2}\right\} \to \infty.$$

An alternative is to estimate the probability density function of $X$ in the PL method by a kernel estimator in the pseudolikelihood method and derive $ISNI_{PL}$ subsequently. This kernel estimator of $f(x)$ can avoid the phenomenon mentioned above.

**2.2.2.1 Kernel density estimation** A probability density function is the most fundamental concept in statistics. If $f$ is the density function of an interval random variable $X$, the probability associated with $X$ is

$$P(a < X < b) = \int_a^b f(x)dx, \quad \text{for all } a < b.$$

Density estimation is to estimate this $f(x)$ from observed data. One approach is to assume $X$ comes from a parametric family of distribution, such as a normal distribution with mean $\mu$ and variance $\sigma^2$. From the observed data, the parameters $\mu$ and $\sigma^2$ can be estimated and the distribution function can be constructed from the estimated mean and variance. There are also nonparametric methods, including the kernel density estimation, to estimate $f(x)$.

Unlike the parametric density estimation methods, the kernel density estimation method does not assume that it comes from any parametric family. It intends to retain the feature of the observed data points, while forcing certain amount of smoothing. Assume that the unknown density $f(x)$ is a smooth function of $x$. If there are $n$ data points, the kernel density estimate of $f$ with smoothing parameter $h$ is defined by (Silverman, 1986)

$$\hat{f}(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} K\left(\frac{t - x_j}{h}\right),$$

where $K(\cdot)$ is a symmetric kernel function satisfying

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \text{ and } \int t^2 K(t)dt = k \neq 0.$$

The bias and variance associated with kernel density estimation is

$$\begin{aligned} bias(f(x)) &= E\hat{f}(x) - f(x) = \frac{1}{2}h^2 f''(x)k_2 + o(h), \\ var\hat{f}(x) &\approx n^{-1}h^{-1}\int K(t)^2 dt, \end{aligned} \tag{2.18}$$

where $\lim_{h\to 0} o(h)/h = 0$ (Silverman, 1986). The smoothing parameter $h$ can be chosen by several methods. A rule of thumb is to choose $h_{opt} = 1.06\sigma_x n^{-1/5}$, where $\sigma_x^2$ is the variance of $X$ and can be estimated from $\hat{\sigma}_x^2 = \frac{1}{n}\sum_{j=1}^{n}(x_i - \bar{x})^2$. It is derived based on the assumption that $f(x)$ is $N(0, \sigma_x^2)$. In some cases when the distribution of $X$ is very skewed or it is multimodal, it tends to oversmooth the densitiy function. But it generally works well and is very easy to compute. We will use this method for the estimation of $f$.

**2.2.2.2 ISNI for the pseudolikelihood method when density of $X$ is estimated by kernel smoothing.** For computational convenience, we used Gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}).$$

Then the density estimate for the distribution of $X$ is

$$\hat{f}(x) = \frac{1}{n\sqrt{2\pi h^2}} \sum_{j=1}^{n} \exp(-\frac{(x-x_j)^2}{2h^2}).$$

If $\theta = (\beta_0, \beta_1, \sigma^2)$, the density estimation of $y_{\lambda i}$ is,

$$p(y_{\lambda i}; \theta, \widehat{F}) = \frac{1}{n} \sum_{j=1}^{n} \phi \left( \frac{\lambda y_i + x_i - \lambda\beta_0 - (\lambda\beta_1 + 1)x_j}{\sqrt{\lambda^2\sigma^2 + (\lambda\beta_1 + 1)^2 h^2}} \right) \overset{\lambda \to 0}{\to} \frac{1}{n} \sum_{j=1}^{n} \phi \left( \frac{x_i - x_j}{h} \right),$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$. The value of $\nabla^2 PL_{12}$ is then

$$\nabla^2 PL_{12} = \begin{pmatrix} \sum_{i=1}^{m} \frac{1}{h^2 \sum_{j=1}^{n} \exp\left[-\frac{(x_i - x_j)^2}{2h^2}\right]} \sum_{j=1}^{n} \exp\left[-\frac{(x_i - x_j)^2}{2h^2}\right] (x_i - x_j) \\ m - \sum_{i=1}^{m} \frac{1}{h^2 \sum_{j=1}^{n} \exp\left[-\frac{(x_i - x_j)^2}{2h^2}\right]} \sum_{j=1}^{n} \exp\left[-\frac{(x_i - x_j)}{2h^2}\right] \{(x_i - x_j)x_i\} \\ 0 \end{pmatrix} \tag{2.19}$$

The computation details can be found in Appandix B.1.

When $h = 0$, the density estimator will be the empirical distribution. When $h \neq 0$, this value $\bigtriangledown^2 PL$ is derived with some smoothing on the density estimation of $X$. So the bias associated with it has to be evaluated. In particular, when the true value of $\nabla^2 PL_{12}$ is close to zero, this bias can be substantial. However, this bias is related to the true form of $f(x)$ through $f''(x)$. An estimate of such bias can be difficult without knowing the true form of $f(x)$. A simple and natural approach is to assume a parametric form for $f(x)$ and derive a working estimator for $f(x)$, for example, normal distribution with mean $\mu_x$ and variance $\sigma_x^2$. Both of the parameters can be estimated from the marginal distribution of $X$ with $\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2$. Incorporate this $f$ into the bias estimation (2.18), the bias of $\hat{f}(x)$ is

$$\text{bias of } \hat{f}(x) = \frac{1}{2} h^2 \phi \left( \frac{x - \mu_x}{\sigma_x} \right) \left( \frac{(x - \mu_x)^2}{\sigma_x^4} - \frac{1}{\sigma_x^2} \right).$$

It follows that the bias of the density estimation of $y_{\lambda i}$, denoted by $\Delta$, is

$$\begin{aligned}\Delta &= \frac{1}{2}h^2\phi\left(\frac{\lambda y_i + x_i - \lambda\beta_0 - (\lambda\beta_1+1)\mu_x}{\sqrt{\sigma_\lambda^2}}\right) \\ &\quad \left[\frac{(\lambda\beta_1+1)^2}{\sigma_\lambda^4}\left[\{\lambda y_i + x_i - \lambda\beta_0 - (\lambda\beta_1+1)\mu_x\}^2 - \sigma_\lambda^2\right]\right] \\ &\overset{\lambda\to 0}{\to} \frac{1}{2}h^2\phi\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left\{(x_i-\mu_x)^2-\sigma_x^2\right\},\end{aligned}$$

where $\sigma_\lambda^2 = \lambda^2\sigma^2 + (\lambda\beta_1+1)^2\sigma_x^2$. The corresponding partial derivatives of $\Delta$ with respect to $\lambda$ and $\theta$ at $(\hat{\theta}_0, \lambda = 0)$ are:

$$\begin{aligned}\frac{\partial^2\Delta}{\partial\lambda\partial\beta_0} &\overset{\lambda\to 0}{\to} -\frac{1}{2}h^2\phi\left(\frac{x_i-\mu_x}{\sigma_x}\right)\frac{1}{\sigma^6}(x_i-\mu_x)\{(x_i-\mu_x)^2-3\sigma_x^2\}, \\ \frac{\partial^2\Delta}{\partial\lambda\partial\beta_1} &\overset{\lambda\to 0}{\to} -\frac{1}{2}h^2\phi\left(\frac{x_i-\mu_x}{\sigma_x}\right)\Big[\frac{1}{\sigma_x^4}\{(x_i-\mu_x)^2-\sigma_x^2\} \\ &\quad -\frac{1}{\sigma^6}(x_i-\mu_x)x_i\{(x_i-\mu_x)^2-3\sigma_x^2\}\Big], \\ \frac{\partial^2\Delta}{\partial\lambda\partial\sigma^2} &\overset{\lambda\to 0}{\to} 0.\end{aligned}$$

Correspondingly, if $\hat{p}(y_{\lambda i}) = \hat{p}(y_{\lambda i}; \theta, \widehat{F})$ is the density estimate of $y_{\lambda i}$ with $f(x)$ estimated from kernel density estimation, the bias corrected estimate of $\tilde{p}(y_{\lambda i})$ is $\hat{p}(y_{\lambda i}) - \Delta$. If $\theta = (\beta_0, \beta_1, \sigma^2)$, when $\lambda = 0$, all the first derivatives of $\hat{p}(y_{\lambda i})$ and $\Delta$ with respect to $\theta$ are vector of zeros. Then $\partial\tilde{p}(y_{\lambda i})/\partial\theta = \mathbf{0}$ as $\lambda \to 0$. So the correction to (2.19) for any parameter $\zeta$, where $\zeta$ can be $\beta_0$, $\beta_1$ or $\sigma^2$, is

$$\begin{aligned}\frac{\partial^2\log\tilde{p}(y_{\lambda i})}{\partial^2\zeta\partial\lambda} &= \frac{1}{\tilde{p}(y_{\lambda i})}\frac{\partial^2\tilde{p}(y_{\lambda i})}{\partial\lambda\partial\zeta} - \frac{1}{\tilde{p}(y_{\lambda i})^2}\frac{\partial\tilde{p}(y_{\lambda i})}{\partial\lambda}\frac{\partial\tilde{p}(y_{\lambda i})}{\partial\zeta} \\ &= \frac{1}{\tilde{p}(y_{\lambda i})}\frac{\partial^2\hat{p}(y_{\lambda i})}{\partial\lambda\partial\zeta} - \frac{1}{\tilde{p}(y_{\lambda i})}\frac{\partial^2\Delta}{\partial\lambda\partial\zeta} \\ &= \frac{\hat{p}(y_{\lambda i})}{\tilde{p}(y_{\lambda i})}\left\{\frac{\partial^2\log\hat{p}(y_{\lambda i})}{\partial\lambda\partial\zeta} + \frac{1}{\hat{p}(y_{\lambda i})^2}\left(\frac{\partial\hat{p}(y_{\lambda i})}{\partial\lambda}\frac{\partial\hat{p}(y_{\lambda i})}{\partial\zeta}\right)\right\} - \frac{1}{\tilde{p}(y_{\lambda i})}\frac{\partial^2\Delta}{\partial\lambda\partial\zeta} \\ &= \frac{\hat{p}(y_{\lambda i})}{\tilde{p}(y_{\lambda i})}\frac{\partial^2\log\hat{p}(y_{\lambda i})}{\partial\lambda\partial\zeta} - \frac{1}{\tilde{p}(y_{\lambda i})}\frac{\partial^2\Delta}{\partial\lambda\partial\zeta}.\end{aligned}$$

The exact bias corrected value of $\nabla^2 PL_{12}$ is:

$$\begin{aligned}
\left.\frac{\partial^2 \sum_{i=1}^m \log \tilde{p}(y_{\lambda i})}{\partial\lambda\partial\beta_0}\right|_{(\hat{\theta}_0,\lambda=0)} &= \sum_{i=1}^m \frac{\hat{p}(y_{\lambda i}) \sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)^2}{2h^2}\right](x_i-x_j)}{\tilde{p}(y_{\lambda i}) h^2 \sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)^2}{2h^2}\right]} \\
&\quad + \sum_{i=1}^m \frac{1}{2\tilde{p}(y_{\lambda i})} h^2 \phi\left(\frac{x_i-\mu_x}{\sigma_x}\right) \frac{1}{\sigma^6}(x_i-\mu_x)\{(x_i-\mu_x)^2 - 3\sigma_x^2\}, \\
\left.\frac{\partial^2 \sum_{i=1}^m \log \tilde{p}(y_{\lambda i})}{\partial\lambda\partial\beta_1}\right|_{(\hat{\theta}_0,\lambda=0)} &= \frac{\hat{p}(y_{\lambda i})}{\tilde{p}(y_{\lambda i})}\Big[m - \sum_{i=1}^m \frac{1}{h^2 \sum_{j=1}^n \exp\left\{-\frac{(x_i-x_j)^2}{2h^2}\right\}} \\
&\quad \cdot \sum_{j=1}^n \exp\left\{-\frac{(x_i-x_j)}{2h^2}\right\}(x_i-x_j)x_i\Big] \\
&\quad + \sum_{i=1}^m \frac{1}{2\tilde{p}(y_{\lambda i})} h^2 \phi\left(\frac{x_i-\mu_x}{\sigma_x}\right)\Big[\frac{1}{\sigma_x^4}\{(x_i-\mu_x)^2 - \sigma_x^2\} \\
&\quad + \frac{1}{\sigma_x^6}(x_i-\mu_x)(y_i-\beta_0-\beta_1 x_i)\{(x_i-\mu_x)^2 - 3\sigma_x^2\}\Big], \\
\left.\frac{\partial^2 \sum_{i=1}^m \log \tilde{p}(y_{\lambda i})}{\partial\lambda\partial\sigma^2}\right|_{(\hat{\theta}_0,\lambda=0)} &= 0.
\end{aligned}$$

Details can be found in Appendix B.2.

## 2.3 SIMULATION STUDIES

Two sets of simulation studies were carried out to evaluate the performance of $ISNI_{PL}$. The first set was designed to compare the performance of $ISNI_{PL}$ and $ISNI_{ML}$ under a missing-data mechanism (2.1) and the ML correctly specified the mechanism. In the second set of simulation studies, the missing data mechanism was simulated different from (2.1) and the ML method misspecified the mechanism. This set of simulations was used to demonstrate the flexibility of $ISNI_{PL}$.

### 2.3.1 Simulations with missing data mechanism correctly specified by the ML

The first set of simulations were done based on the model specified below:

$X \sim N(0,1)$,

$Y \sim \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \sim N(0,1)$,

$Pr(R = 1|X, Y) = logit^{-1}(\psi_0 + \psi_1 X + \psi_2 Y) = \frac{exp(\psi_0+\psi_1 X+\psi_2 Y)}{1+exp(\psi_0+\psi_1 X+\psi_2 Y)}$,

where $\beta_0 = 0$ and $\beta_1 = 1$; parameters $(\psi_1, \psi_2)$ controlled the nature and magnitude of the nonignorable mechanism with the following choices: (0,0), (1,0), (1,1,), (1,-1), (1,3), (0,1), and (0,3); parameter $\psi_0$ was used to maintain 50% overall completed cases. Under each setting, 500 datasets were simulated and there were 1000 observations within each simulated dataset. In these simulation studies, the primary of interest is $\beta_1$, the regression slope. Denote the sensitivity transformations for the maximum likelihood method, pseudolikelihood with bivariate normal distribution and pseudolikelihood with the kernel density method as $c_{ML}$, $c_{PL1}$ and $c_{PL2}$, respectively. For each simulated dataset, the estimator of $\beta_1$ under the ignorable ML method and the three sensitivity transformations were obtained. These results were summarized and compared in Table 1 and Table 2.

In both tables, the true values of $(\psi_1, \psi_2)$ are listed in the first two columns. Under each parameter setting for the missing-data mechanism, Table 1 presents the empirical median, the empirical 90% confidence interval that consist of the 5th percentile and 95th percentile of the sensitivity transformations from 500 simulated datasets for each of those three methods. Histograms for $c_{ML}$s, $c_{PL1}$s and $c_{PL2}$s from 500 simulated datasets under $(\psi_1, \psi_2) = (0,0), (1,0)$ are also presented in Figure 2. Under each parameter setting for the missing-data mechanism, Table 2 presents the empirical bias, empirical standard deviation of the estimators under the ignorable ML method from 500 simulated datasets and the proportions of those 500 simulated datasets with $c_{ML} < 1$, $c_{PL1} < 1$, and $c_{PL2} < 1$, respectively.

Table 1 shows that the $c_{PL}$s and the $c_{ML}$ were in general consistent with each other on the local sensitivity of the data to nonignorability. The local sensitivity indices for the pseudolikelihood methods tended to be more conservative compared to $ISNI_{ML}$ and had a larger variability in most cases. This is probably a reflection of the loss of information on the pseudolikelihood method from the ML method. Table 2 shows that except for the case when $(\psi_1, \psi_2) = (1,0)$, for those mechanisms when the ignorable MLEs were biased, the $c_{PL}$s and the $c_{ML}$ concluded that the datasets were sensitive to local deviation from MAR and more extensive sensitivity analyses would be warranted; for those mechanisms when the ignorable MLEs were not substantially biased, most of the $c_{PL}$s and the $c_{ML}$s concluded that

Table 1: Distribution characteristics of three sensitivity transformations for bivariate normal data when the mechanism is correctly specified by ML.

| $\psi_1$ | $\psi_2$ | Median (90% CI) of $c_{ML}$ | Median (90% CI) of $c_{PL1}$ | Median (90% CI) of $c_{PL2}$ |
|---|---|---|---|---|
| 0 | 0 | 6.215 ( 2.103 - 74.396 ) | 2.139 ( 0.785 - 23.178 ) | 1.602 ( 0.62 - 15.303 ) |
| 1 | 0 | 0.322 ( 0.295 - 0.356 ) | 0.316 ( 0.221 - 0.587 ) | 0.279 ( 0.206 - 0.418 ) |
| 1 | 1 | 0.238 ( 0.226 - 0.252 ) | 0.139 ( 0.110 - 0.183 ) | 0.130 ( 0.109 - 0.157 ) |
| 1 | -1 | 6.316 ( 2.119 - 75.873 ) | 2.133 ( 0.730 - 19.86 ) | 1.858 ( 0.608 - 19.60 ) |
| 1 | 3 | 0.222 ( 0.212 - 0.234 ) | 0.113 ( 0.093 - 0.142 ) | 0.106 ( 0.093 - 0.124 ) |
| 0 | 1 | 0.349 ( 0.315 - 0.391 ) | 0.406 ( 0.251 - 1.034 ) | 0.349 ( 0.236 - 0.555 ) |
| 0 | 3 | 0.247 ( 0.232 - 0.264 ) | 0.169 ( 0.128 - 0.233 ) | 0.155 ( 0.128 - 0.197 ) |

the datasets were not sensitive to local deviation from MAR and the ignorable MLE would probably be fine for the parameter estimation and subsequent conclusion. Similar to Table 1, the $c_{PL}$s were more conservative than the $c_{ML}$ although they were mostly consistent.

### 2.3.2 Simulation results when the missing data mechanism was misspecified by the ML

For the next set of simulations, we would like to study the adaptability of $ISNI_{PL}$ in detecting the sensitivity of the estimate under other nonresponse mechanisms. For this simulation study, the complete data were simulated from the same distribution as the previous simulation study and a quadratic function was used to simulate the missing-data mechanism:

$$X \sim N(0,1)$$

$$Y \sim \beta_0 + \beta_1 X + \epsilon, \text{ where } \epsilon \sim N(0,1)$$

$$Pr(R = 1|X, Y) = logit^{-1}\{\psi_0 + (\psi_1 X + \psi_2 Y)^2\}$$

As in the previous simulation study, $(\beta_0, \beta_1) = (0, 1)$. Similarly parameters $(\psi_1, \psi_2)$ controlled the nature and magnitude of the nonignorable mechanism with the following choices: (0,0), (1,0), (1,1,), (1,-1), (1,3), (0,1), and (0,3); parameter $\psi_0$ was used to maintain 50% overall completed cases. Under each parameter setting for the complete data model and nonresponse mechanism, 500 datasets with 1000 observations were simulated. The results

Table 2: Empirical bias for the ignorable MLE and proportions with $c_{ML} < 1$, $c_{PL1} < 1$, and $c_{PL2} < 1$ when mechanism is correctly specified by the ML.

| $\psi_1$ | $\psi_2$ | Bias (standard error) of $\hat{\beta}_{10}$ | % $[c_{ML} < 1]$ | % $[c_{PL1} < 1]$ | % $[c_{PL2} < 1]$ |
|---|---|---|---|---|---|
| 0 | 0 | -0.004(0.044) | 0 | 14.2 | 23.8 |
| 1 | 0 | -0.004(0.051) | 100 | 99.8 | 100 |
| 1 | 1 | -0.235(0.056) | 100 | 100 | 100 |
| 1 | -1 | -0.002(0.041) | 0 | 13.2 | 24.4 |
| 1 | 3 | -0.426(0.051) | 100 | 100 | 100 |
| 0 | 1 | -0.154(0.046) | 100 | 94.8 | 99.2 |
| 0 | 3 | -0.373(0.045) | 100 | 100 | 100 |

are averaged over 500 datasets for each parameter setting. The true selection model for these datasets is non-monotone in the outcome. The $ISNI_{ML}$ assumed a wrong missing-data mechanism and would be misleading for datasets simulated under such mechanisms. Results on $c_{ML}$ were not collected. The empirical medians and 90% empirical confidence intervals for both $c_{PL1}$ and $c_{PL2}$, proportions of $c_{PL1} < 1$ and $c_{PL2} < 1$ were obtained and compared in Table 3.

In general, these two sensitivity indices agreed with each other on whether datasets are sensitive to local deviation from MAR. With less assumption on the distribution of the covariate, $c_{PL2}$ was more conservative than $c_{PL1}$.

## 2.4 EXAMPLE: SMOKING AND MORTALITY DATA

A dataset from Troxel *et al.* was used for illustrating the proposed index. That dataset contains the smoking and the mortality information for 25 occupational groups in England and Wales in early 1970s (Troxel, *et al.*, 2004). The variables include the smoking index, the ratio of the average number of cigarettes smoked per day by men in the occupational group to the average number of cigarettes smoked per day by all men, and the mortality index, the ratio of the rate of deaths from lung cancer among men in the occupational group to

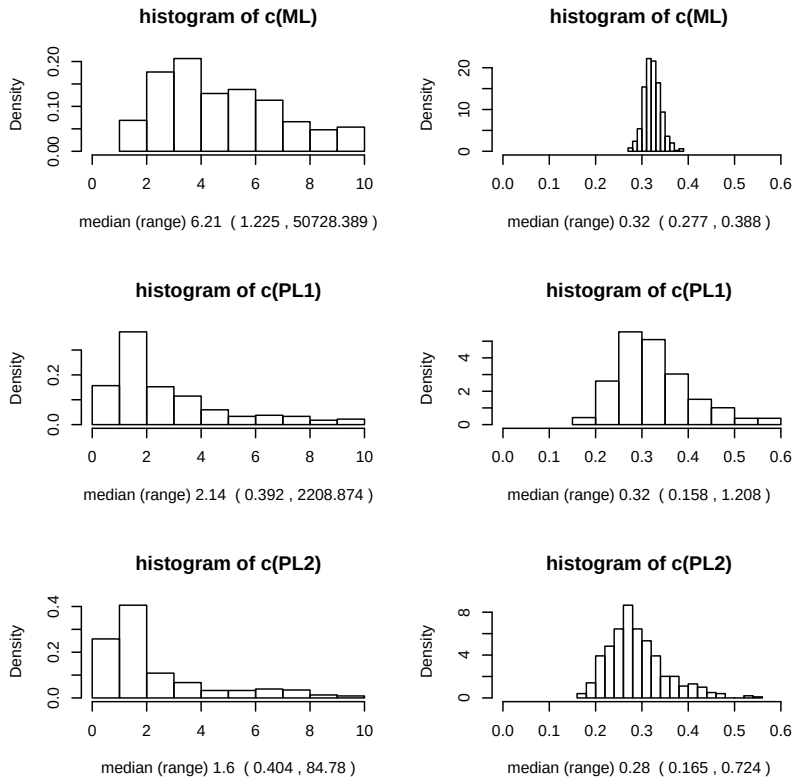


Figure 2: Histograms of sensitivity transformations when the missing data mechanism is correctly specified by maximum likelihood method. The three plots in the left panel are corresponding to $(\psi_1, \psi_2) = (0, 0)$. The three plots in the right panel are corresponding to $(\psi_1, \psi_2) = (1, 0)$.

Table 3: Simulation results on sensitivity transformations $c_{PL1}$ and $c_{PL2}$ when mechanism is misspecified by ML.

| $\psi_1$ | $\psi_2$ | $Bias(SE)$ of $\hat{\beta}_{10}$ | Median (90% CI) of $c_{PL1}$ | Median(90% CI) $c_{PL2}$ | % $c_{PL1} < 1$ | % $c_{PL2} < 1$ |
|---|---|---|---|---|---|---|
| 0 | 0 | -0.004(0.044) | 2.139 ( 0.785 - 23.178 ) | 1.602 ( 0.62 - 15.303 ) | 15 | 23.8 |
| 1 | 0 | 0.004(0.043) | 0.168 ( 0.153 - 0.182 ) | 0.191 ( 0.172 - 0.216 ) | 100 | 100 |
| 1 | 1 | 0.194(0.033) | 0.158 ( 0.145 - 0.173 ) | 0.173 ( 0.158 - 0.192 ) | 100 | 100 |
| 1 | -1 | -0.007(0.058) | 1.994 ( 0.667 - 28.914 ) | 1.663 ( 0.53 - 13.661 ) | 17.6 | 30.2 |
| 1 | 3 | 0.257(0.033) | 0.181 ( 0.165 - 0.200 ) | 0.202 ( 0.182 - 0.226 ) | 100 | 100 |
| 0 | 1 | 0.252(0.037) | 0.252 ( 0.216 - 0.299 ) | 0.295 ( 0.25 - 0.369 ) | 100 | 100 |
| 0 | 3 | 0.299(0.036) | 0.212 ( 0.19 - 0.242 ) | 0.243 ( 0.212 - 0.287 ) | 100 | 100 |

the rate of deaths from lung cancer among all men. The dependent variable is the mortality index. The purpose was to find how smoking, as represented by the smoking index, affects the lung cancer mortality. The original dataset is complete. The estimates of the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$ from the complete data are $\hat{\beta}_0 = -2.885 \pm 23.034$ and $\hat{\beta}_1 = 1.088 \pm 0.221$ (Estimate±se).

We ordered observations according to the predictor, smoking index. Missing data were artificially created by orderly deleting one or five consecutive values of the mortality index. The impact of any particular point/points and the proportion of missing on the sensitivity are of particular interest here. Sensitivity transformations, $c_{ML}$, $c_{PL1}$ and $c_{PL2}$, were computed from the observed data after artificial deletions. Due to the large variance of smoking index, the smoothing parameter $h$, which is proportional to the standard deviation of the covariate, is very large. It causes the bias to soar beyond the estimated density of the outcome variable. To resolve this problem, the smoking index were scaled with its standard deviation. The subsequent $ISNI_{PL2}$ was rescaled back to the one corresponding to the original data.

For each artificially created dataset with missing values, the estimated ignorable ML estimator and its standard error, the $c_{ML}$, $c_{PL1}$ and $c_{PL2}$ corresponding to the regression slope are listed in Table 4. Overall, there are some discrepancies between these indices. Except for the scenario where the first 5 points are missing, the complete data estimate (1.088) is within all one SE of the ignorable MLE. Despite some occasional jump, the sensitivity

Table 4: Smoking and mortality data with missing points

| Missing Point(s) | $\hat{\beta}_{10}(SE)$ | $\hat{\sigma}^2$ | $c_{ML}$ | $c_{PL1}$ | $c_{PL2}$ |
|---|---|---|---|---|---|
| 1 | 0.967(0.235) | 315 | $\infty$ | 25.23 | 25.45 |
| 2 | 1.000(0.222) | 313 | 4.77 | 67.00 | 52.08 |
| 12 | 1.085(0.21) | 315 | 141.94 | 125.31 | 12.22 |
| 1 - 5 | 0.655(0.317) | 315 | $\infty$ | 7.11 | 2.85 |
| 3 - 7 | 1.088(0.265) | 374 | 1.73 | 88.0 | 2.62 |
| 10 - 14 | 1.086(0.222) | 347 | 91.70 | 25.91 | 48.42 |
| 20 - 24 | 1.036(0.261) | 324 | 1.92 | 29.41 | 1.38 |

transformations were generally larger when only one point was missing. Considering that the proportion of missing is much larger when 5 observations are missing at a time than when only one observation is missing at a time, this phenomenon is consistent with past reports that the proportion of missingness is relevant to the sensitivity of the estimate. A larger proportion of missing is usually related to higher sensitivity (Troxel, *et al.*, 2006). The point 1 seems to have a large impact on the estimates. The $\hat{\beta}_{10}$s are the worst when data points 1 or 1-5 are missing. Somehow, $c_{ML}$s, in both cases, suggest complete insensitivity towards nonignorability assumption. At the same time, $c_{PL1}$ is among the lowest in both situations and $c_{PL2}$ confirms that with points 1-5 missing, these methods are not sensitive to deviation from MAR. With points 20-24 missing, $c_{PL2}$ agrees with $c_{ML}$, confirming that the estimate is somewhat sensitive, while $c_{PL1}$ rejected such indication. This discrepancy between $c_{PL1}$ and $c_{PL2}$ may be heavily affected by the distribution of $X$. With points 3-7 missing, the MAR estimate is almost the same with the complete data analysis. Only $c_{PL1}$ suggested strongly that the method PL1 is not sensitive to deviation from MAR. With points 10-14 missing, all of them agree that they are not sensitive to the missing data mechanism assumption.

We cannot tell from this data analysis, which one of the methods is more close to the truth. With only a small sample of data, the complete data analysis may not be trustworthy

either. But there are still noticeable difference between these indices. Each index has its advantage and weakness. $ISNI_{ML}$ depends on the missing data mechanism and $ISNI_{PL}$s do not. Between the two $ISNI_{PL}$s, $ISNI_{PL1}$ need the parametric distribution of $X$, while $ISNI_{PL2}$ does not. But its values tend to be lower. Precautions are required when making any judgment based on only one of them. If possible, all of them can be computed and if any of them indicate important sensitivity, more comprehensive sensitivity analyses should be carried out.

# 3.0 A PENALIZED PSEUDOLIKELIHOOD METHOD

Parametric regressions generally require the assumption that the contributions of the predictors to the mean of the outcome or its transformation are either linear or polynomial. For example, in multiple linear regressions, the mean of the outcome is linear in predictors. In logistic regression model, the logit of the mean of the outcome is linear in predictors. With such assumptions, the effect of the predictors on the outcome is easy to interpret. In most situations, a parametric model is good enough for a regression analysis. However, sometimes the effect of some predictors on the outcome cannot be simply characterized by parametric functions. Instead the effect of such variables are often modeled nonparametrically. For example, for a bivariate dataset $\{t_i, y_i\}_{i=1,\ldots,n}$, the contribution of $T$ on $Y$ can be described by an unspecified function $g$. The subsequent model is:

$$y_i = g(t_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2), \tag{3.1}$$

where $\epsilon_i$s are iid. Some restrictions on $g$ are then required in order to estimate $g$. In general, these nonparametric models require fewer assumptions and provide more freedom in model fitting. In some semi-parametric regression models, the mean structure is a combination of parametric functions of some predictors and nonparametric functions of other predictors. These semi-parametric regression models are especially useful when the contribution from some predictors is not well understood or not of interest but needs to be adjusted. For example, for a dataset $(x_i, t_i, y_i),\ i = 1, \ldots, n$, where $n$ is the number of subjects, $x_i$ is the predictor of interest, $t_i$ is a confounder whose effect is not of interest but needs to be adjusted, and $y_i$ is the outcome or dependent variable. A typical semi-parametric regression model for such dataset is

$$y_i = x_i\beta + g(t_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2), \tag{3.2}$$

where $g(t)$ is the contribution of $T$ on $Y$ and its functional form is not well understood and unspecified. Let $\theta = (\beta, g, \sigma^2)$, the likelihood function is

$$L(\theta) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i - x_i\beta - g(t_i)\}^2.$$

Without further information, $\hat{\theta}$ cannot be identified. For such semi-parametric regression models, a useful method for estimation is the penalized likelihood as the following(Good & Gaskins, 1971)

$$L(\theta) - \frac{1}{2}\lambda J(g),$$

where $J(g)$ is a pre-specified roughness penalty, which increases as $g$ becomes less smooth. For example, it can be $J(g) = \int |g''(x)|dx$. When $g$ is linear, the regression curve is smooth and this quantity $J(g) = 0$. When $g(x) = \sin(x)$ with $x$ bounded by $[0, \pi]$, the regression curve is not as smooth and $J(g) = 2$. Parameter $\lambda$ is a nonnegative smoothing parameter that is used to control the influence of the smoothness on the model fitting. Semi-parametric models have been successfully adopted to solve many complex problems. However, few studies have been done when there are nonignorable nonresponse in $Y$. We extended the pseudolikelihood method for multivariate monotone data with nonignorable nonresponse (Tang *et al.*, 2003) by incorporating a roughness penalty term in the logarithm of the pseudolikelihood function. Two cross-validation ($CV$) methods were explored to choose the optimal $\lambda$. The properties of the proposed penalized pseudolikelihood method and two $CV$ methods were evaluated through simulation studies and illustrated by analysis of a psychiatric clinical study dataset.

## 3.1 NONPARAMETRIC REGRESSIONS AND THE STATISTICAL METHODS FOR NONPARAMETRIC REGRESSIONS

Consider a bivariate dataset $\{t_i, y_i\}_{i=1,\ldots,n}$, where $n$ is the sample size. For simplicity we assume that the predictor $T$ has no tie. When there is no prior knowledge on the functional form of effect of $T$ on $Y$, their relationship can be described by model (3.1). The mean

structure function $g$ is square integrable and has $m$ continuous derivatives. Denote $L_2[a,b]$ as the function space of all square integrable functions on a pre-specified interval [a, b] that covers the observed $t_i$s. This function $g(\cdot)$ is assumed to be a member of the following space(Eubank, 1999)

$$W_2^m[a,b] = \left\{ g : \begin{array}{c} g^{(j)} \text{ is absolutely continuous,} \\ j = 0, \ldots, m-1, \text{ and } g^{(m)} \in L_2[a,b] \end{array} \right\}.$$

The exact function form of $g$ is not well understood and cannot be modeled by linear or polynomial regression. This function $g(T)$ will be estimated using a nonparametric model.

### 3.1.1 The penalized likelihood method for nonparametric regression

The motivation of using a nonparametric model for $g(T)$ is to preserve the key features of the real function $g$, while control for the overall smoothness. Let $\theta = (g, \sigma^2)$. Given a smoothing parameter $\lambda > 0$ and a penalty function $J(g) = \int_a^b \{g''(x)\}^2 dx$, the penalized likelihood is:

$$L(\theta) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i - g(t_i)\}^2 + \lambda\int_a^b \{g''(x)\}^2 dx. \tag{3.3}$$

The estimates are obtained by maximizing this penalized likelihood: $\hat{\theta} = arg\max_\theta L(\theta)$. With $\lambda$ fixed, there is a unique solution to this optimization problem, in which $\hat{g}$ is a natural cubic spline (NCS) with knots at all unique points $t_i$ (Green and Silverman, 1994). That is, if $t_1, t_2, \ldots, t_n$ satisfying $a < t_1 < t_2 < \ldots < t_n < b$ are the unique ordered values of $T$, $\hat{g}$ is cubic polynomial on each of the intervals $(a, t_1), (t_1, t_2), \ldots, (t_n, b)$. Function $\hat{g}$ itself and its first and second derivatives are continuous on $[a, b]$. The second and third derivatives are zero at $a$ and $b$. The estimating procedure for $\hat{g}$ is straightforward. However, in reality, if all the unique data points are chosen as knots, the complexity of computation increases quickly with the sample size. It has been suggested that when the total number of knots $k$ is sufficiently large, increasing $k$ has little influence on the fit from the penalized likelihood function. It was discussed in Ruppert (2002) that a default of at most 35 knots can provide a good fit for almost all sample size. In most data analysis, the knots $\tau_1, \tau_2, \ldots, \tau_k$ on [a, b], where $a < \tau_1 < \tau_2 < \ldots < \tau_k < b,\ k \leq n$, were selected beforehand.

After $k$ knots have been selected, the value of the roughness penalty $J(g)$ and $\hat{g}(t)$, for any $t \in [a, b]$, can be computed from $\hat{g}(\tau_i),\ i = 1, \ldots, k$. Denote $g_i = g(\tau_i)$ and $\gamma_i = g"(\tau_i)$ for $i = 1, \ldots, k$ and $\gamma_1 = \gamma_k = 0$. Denote $\mathbf{g} = (g_1, \ldots, g_k)^T$ and $\gamma = (\gamma_2, \ldots, \gamma_{k-1})^T$, $h_i = \tau_{i+1} - \tau_i$ for $i = 1, \ldots, k-1$. Matrices $Q_{k\times(k-2)}$ and $R_{(k-2)\times(k-2)}$ are two band matrices. The entries $q_{ij}$, $i = 1, \ldots, k$ and $j = 2, \ldots, k-1$, of Q is given by

$$q_{j-1,j} = h_{j-1}^{-1},\ q_{jj} = -h_{j-1}^{-1} - h_j^{-1},\ q_{j+1,j} = h_j^{-1}, \text{ and } q_{ij} = 0 \text{ for } |i-j| \geq 2.$$

R is a symmetric matrix with elements $r_{ij}$, for $i$ and $j$ from 2 to $(k-1)$, given by

$$r_{ii} = \frac{1}{3}(h_{i-1} + h_i) \text{ for } i = 2, \ldots, k-1,\ r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i \text{ for } i = 2, \ldots, k-2,$$

and $r_{ij} = 0$ for $|i-j| \geq 2$. Matrix $K = QR^{-1}Q^T$.

For the band matrices defined as above, $Q^T\mathbf{g} = R\gamma$. The second derivatives vector $\gamma$ can be computed from $\mathbf{g}$ and the band matrices. The roughness penalty can be derived by

$$\int_a^b g''(t)^2 dt = \gamma^T R\gamma = \mathbf{g}^T K\mathbf{g} \tag{3.4}$$

(Green and Silverman, 1994). For any $\tau_{j-1} < t < \tau_j,\ j = 1, \ldots, k$, $\hat{g}(t)$ can be computed by

$$\hat{g}(t) = \frac{t_l}{h_j}g_{j+1} + \frac{t_r}{h_j}g_{j+1} - t_l t_r\{\gamma_{j+1}(1 + \frac{t_l}{h_j}) + \gamma_k(1 + \frac{t_r}{h_j})\}/6, \tag{3.5}$$

where $t_l = t - \tau_{j-1}$ and $t_r = \tau_j - t$.

While fitting a smoothing spline, choosing a optimal smoothing parameter $\lambda$ is essential. When a very large value of $\lambda$ is used, the penalty term would dominate the penalized likelihood function and force the spline close to a straight line. Such a smooth fit often leads to substantial bias. When $\lambda$ is too small, the regression fit of the data will dominate the penalized likelihood and lead to a volatile fit with small bias. Data-driven methods such as cross validation (CV) and generalized cross-validation (GCV) are the most common methods to find the optimal $\lambda$ that balance the bias and variation in practice. The leave-one-out $CV$ is to find a $\lambda$ that minimizes the following function:

$$CV(\lambda) = n^{-1}\sum_{i=1}^{n}\{y_i - \hat{g}^{(-i)}(t_i; \lambda)\}^2,$$

where $\hat{g}^{(-i)}(t;\lambda)$ is estimated from the dataset after deleting an observation $(t_i, y_i)$, given the smoothing parameter $\lambda$. If we consider observation $(t_i, y_i)$ a new observation, the estimated mean for $y_i$ is $\hat{g}^{(-i)}(t_i)$ and the corresponding prediction error would be $\{y_i - \hat{g}^{(-i)}(t_i;\lambda)\}$. The $CV$ score can be roughly considered as the estimate of mean squared prediction error. The optimal $\lambda$ is obtained by minimizing $CV$. Generalized cross-validation (GCV) is a modified form of the simple leave-one-out cross-validation. It adaptively chooses $\lambda$ that minimizes a $GCV$ score

$$GCV(\lambda) = n^{-1}\frac{\sum_{i=1}^{n}\{y_i - \hat{g}(t_i)\}^2}{\{1 - n^{-1}\mathrm{tr}A(\lambda)\}},$$

where $A(\lambda) = (I_n + \lambda Q R^{-1} Q^T)^{-1}$ and $I_n$ is $n \times n$ identity matrix.

These traditional methods for nonparametric models have been successfully employed in many statistical problems. But they only deal with data with complete records and may yield biased estimates when the data are not complete. The proposed statistical methods for nonparametric regression with incomplete data mainly include imputation methods (Cheng, 1994), the propensity score method (Hahn, 1998) and imputed empirical likelihood methods (Wang and Rao, 2002). However, all of them require the data to be MAR. Few studies has been done to nonparametric regression of data with nonignorable nonresponse. To address this issue, we expanded the pseudolikelihood method (Tang, *et al.*, 2003) to the analysis of nonparametric regression models for data with nonignorable nonresponse.

### 3.1.2 A penalized pseudolikelihood method (PPL)

Consider bivariate data $\{t_i, y_i\}_{i=1,\ldots,n}$, $t_i$s are fully observed and $y_i$s are observed for $i = 1, \ldots, m$, missing for $i = m+1, \ldots, n$. The missing data indicator $R_i = 1$ for $i = 1, \ldots, m$ and $R_i = 0$ for $i > m$. Assume $R_i$ depends completely on $y_i$

$$Pr[R_i = 1|t_i, y_i] = Pr[R_i = 1|y_i] = \omega(y_i;\psi),$$

for some unknown function $\omega(\cdot)$ and parameter $\psi$. Under such circumstances, $R$ is independent of $T$, given $Y$. Then observed data are a random sample of $T$, given $Y$.

The corresponding pseudolikelihood function for $\theta = (\mathbf{g}, \sigma^2)$ is

$$L(\theta; \widehat{F}, \lambda) = -\sum_{i=1}^{m}\frac{\{y_i - g(t_i)\}^2}{2\sigma^2} - \sum_{i=1}^{m}\log\int \exp\left[-\frac{\{y_i - g(t)\}^2}{2\sigma^2}\right] d\widehat{F}(x),$$

where $\widehat{F}(x)$ is a consistent estimator of the cumulative distribution function of $X$. In order to obtain a smooth fit of $g$, the following penalized pseudolikelihood was proposed with knots $\tau_1, \tau_2, \ldots, \tau_k$ predetermined:

$$\begin{aligned} L(\theta; \widehat{F}, \lambda) &= -\sum_{i=1}^{m} \frac{\{y_i - g(t_i)\}^2}{2\sigma^2} - \sum_{i=1}^{m} \log \int \exp\left[-\frac{\{y_i - g(t)\}^2}{2\sigma^2}\right] d\widehat{F}(x) \\ &\quad -\lambda \int_a^b g''(t)^2 dt, \end{aligned} \tag{3.6}$$

In the following context, the empirical distribution of $T$, $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i \le t)$ is used as the consistent estimator of $F$. Parameter $\theta$ can be estimated by

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} L(\theta; F_n) \\ &= \arg\max_{\theta} \left[ -\sum_{i=1}^{m} \frac{\{y_i - g(t_i)\}^2}{2\sigma^2} - \sum_{i=1}^{m} \log \int \exp[-\frac{\{y_i - g(t)\}^2}{2\sigma^2} dF_n(t)] \right. \\ &\quad \left. -\lambda \int_a^b g''(t)^2 dt \right] \\ &= \arg\max_{\theta} \left[ -\sum_{i=1}^{m} \frac{\{y_i - g(t_i)\}^2}{2\sigma^2} - \sum_{i=1}^{m} \log \left[ \frac{1}{n} \sum_{j=1}^{n} \exp\left\{ -\frac{\{y_i - g(t_j)\}^2}{2\sigma^2} \right\} \right] \right. \\ &\quad \left. -\lambda \mathbf{g}^T K \mathbf{g} \right] \end{aligned} \tag{3.7}$$

### 3.1.3 Cross validation method for the penalized pseudolikelihood method

Leave-one-out cross validation is rooted on the assumption that any data point is a random sample from the study population. If the data are complete, $CV$ is a slightly biased estimate of the mean squared prediction error. However, for data with nonignorable nonresponse, the missing data indicator needs to be taken into account while comparing observed outcome and the predicted value. For example, for the nonparametric regression model (3.1),

$$[Y|T, R = 1] \neq Pr[Y|T].$$

The formula

$$\frac{1}{m}\sum_{i=1}^{m} [y_i - \hat{E}\{y|\hat{g}^{(-i)}(t_i; \lambda), \hat{\sigma}^{2(-i)}\}]^2,$$

where $i = 1, 2, \ldots, m$, is only a rough estimate of the squared prediction error of $y_i$ given $R_i = 1$ and $t_i$. This $CV$ is not an ideal candidate for data with outcome-dependent missingness. But this $CV$ method is easy to implement and will be considered as an option for choosing the optimal smoothing parameter $\lambda$.

An alternative cross-validation method for the penalized pseudolikelihood method was also considered. Its rationale was to trace back to the assumption on the missing-data mechanism. Under the outcome-dependent assumption,

$$[T|Y, R = 1] = [T|Y],$$

the difference

$$t_i - \hat{E}\{t|y_i, \hat{g}^{(-i)}(t_i; \lambda), \hat{\sigma}^{2(-i)}\},$$

offers a legitimate evaluation of the prediction error when $y_i$ is used to predict $t_i$. Let $\theta = (g, \sigma^2)$, we can construct a reversed cross validation (RCV) in terms of the mean squared prediction error of $T$,

$$RCV(\lambda) = \frac{1}{m}\sum_{i=1}^{m}[t_i - \widehat{E}\{t|y_i; \hat{g}^{(-i)}(t_i; \lambda), \hat{\sigma}^{2(-1)}\}]^2 \tag{3.8}$$

where,

$$\widehat{E}\{t|y_i; \theta^{(-i)}, \widehat{F}\} = \int tp(t|y_i; \hat{\theta}^{(-i)})dt = \int t\frac{p(y_i|t; \hat{\theta}^{(-i)})}{\int p(y_i|t; \hat{\theta}^{(-i)})d\widehat{F}(t)}d\widehat{F}(t),$$

and $\widehat{F}(t)$ is a consistent estimate of the cumulative distribution function of $T$.

The estimated value $\hat{E}[t|y_i, \theta^{(-i)}]$ can be obtained by using empirical estimator $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(t_i \le t)$:

$$\begin{aligned} H(\lambda) &= \sum_{j=1, j\neq i}^{n} t_j \frac{\frac{1}{n-1}p(y_i|t_j\hat{\theta}^{(-i)}(\lambda))}{\frac{1}{n-1}\sum_{k=1,k\neq i}^{n} p(y_i|t_k; \hat{\theta}^{(-i)}(\lambda))} \\ &= \sum_{j=1, j\neq i}^{n} t_j \frac{p(y_j|t_j; \hat{\theta}^{(-i)}(\lambda))}{\sum_{k=1,k\neq i}^{n} p(y_j|t_k; \hat{\theta}^{(-i)}(\lambda))} \end{aligned} \tag{3.9}$$

The corresponding version of RCV for semi-parametric model (3.2) and $\theta = (\beta, g, \sigma^2)$, when $X$ and $T$ are independent, is

$$RCV(\lambda) = \frac{1}{m}\sum_{i=1}^{m}[t_i - \widehat{E}\{t|y_i, x_i; \hat{\beta}^{(-i)}(\lambda), \hat{g}^{(-i)}(t_i; \lambda), \hat{\sigma}^{2(-1)}\}]^2 \tag{3.10}$$

where,

$$\widehat{E}\{t|y_i,x_i;\theta^{(-i)},\widehat{F}\} = \int tp(t|y_i,x_i;\hat{\theta}^{(-i)})dt = \int t\frac{p(y_i|x_i,t;\hat{\theta}^{(-i)})p(x_i|t)}{\int p(y|x_i,t;\hat{\theta}^{(-i)})p(x_i|t)d\widehat{F}(t)}d\widehat{F}(t)$$
$$\overset{X\perp T}{=} \int t\frac{p(y_i|x_i,t;\hat{\theta}^{(-i)})}{\int p(y_i|x_i,t;\hat{\theta}^{(-i)})d\widehat{F}(t)}d\widehat{F}(t).$$

The above term can be estimated by an empirical estimator $F_n(t)=\frac{1}{n}\sum_{i=1}^n I(t_i\le t)$:

$$H(\lambda) = \sum_{j=1,j\neq i}^{n} t_j\frac{\frac{1}{n-1}p(y_i|x_i,t_j;\hat{\theta}^{(-i)}(\lambda))}{\frac{1}{n-1}\sum_{k=1,k\neq i}^{n}p(y_i|x_i,t_k;\hat{\theta}^{(-i)}(\lambda))} \tag{3.11}$$
$$= \sum_{j=1,j\neq i}^{n} t_j\frac{p(y_i|x_i,t_j;\hat{\theta}^{(-i)}(\lambda))}{\sum_{k=1,k\neq i}^{n}p(y_i|x_i,t_k;\hat{\theta}^{(-i)}(\lambda))}$$

By minimizing (3.8) and (3.10), the optimal $\lambda$ can be obtained. This RCV method, along with CV method, will be considered as the candidate for choosing an optimal $\lambda$.

### 3.1.4 Simulation studies

We conduced simulation studies to examine the performance of PPL and to select the best cross-validation method. Several issues were considered and studied in the simulation studies: i) The fitted curves derived from PPL methods are affected by the smoothing parameter $\lambda$. They tend to be rough for small $\lambda$ and smooth for large $\lambda$. ii) The cross validation methods have to pick a better fit for the nonparametric curve. These issues were investigated through one set of simulation studies to evaluate the performance of the PPL method for nonparametric regression models here.

In this simulation study, a bivariate dataset $\{t_i,y_i\}_{i=1,\ldots,n}$ was simulated from

$$T\sim Unif(0,1) \text{ and } Y=10T^2+N(0,1). \tag{3.12}$$

The nonresponse in $Y$ was created based on the following mechanism:

$$Pr[R=1|T,Y]=\Phi(-1.7+0.5Y). \tag{3.13}$$

Each simulated dataset had 300 observations. Due to high computation complexity, the global minimization of either $CV$ or $RCV$ was not carried out. The optimal $\lambda$ among a

grid of ten values: $(0.0, 10^{-6}, 10^{-5}, 10^{-4}, 0.001, 0.005, 0.01, 0.1, 1, 10)$ was used instead in the simulation studies. For each simulated dataset, the corresponding values of $CV(\lambda)$ and $RCV(\lambda)$ over these ten values were computed and the smallest one was used to determine the corresponding optimal $\lambda$. The loss function

$$LOSS(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_\lambda(t_i) - g(t_i))^2$$

was used to evaluate how the fitted curve, from the $CV$ or $RCV$ method, differed from the true curve. A large value of $LOSS(\lambda)$ would indicate a poor fit with large squared bias. The smaller $LOSS(\lambda)$ indicates a better fit to the curve. For each $\lambda$, the penalized pseudolikelihood was fitted with 10 evenly distributed knots within (0, 1).

The results from one simulated dataset is shown in Figure 3. In the figure, $CC$ denotes results from the complete case analysis by minimizing the penalized likelihood function (3.3). It was based solely on the complete cases. This was fitted by "smooth.spline" function in R with similar degree of freedom and CV method (The R Development Core Team, 2008). The figure gives the PPL fitted lines under eight values of $\lambda$ and the complete case estimator. The influence of $\lambda$ is visible. When the value of $\lambda$ is small, the part with less observed data is very rough. When the value of $\lambda$ is large, the curve is close to a straight line. The complete case analysis is obviously bias toward the low end of $T$. In Table 5, the values of $RCV(\lambda)$, $CV(\lambda)$ scores and $LOSS(\lambda)$ corresponding to each $\lambda$ are listed. The last row gives the optimal $\lambda$, the corresponding $CV(\lambda)$ and $LOSS(\lambda)$ from the complete case analysis. The $RCV(\lambda)$ and $CV(\lambda)$ were minimized at $\lambda = 0.0001$ and $\lambda = 0.005$ over the grid of ten values, respectively. $LOSS(\lambda)$ is the smallest at $\lambda = 0.005$, which is exactly the same one that was picked by CV method. At $\lambda = 0.0001$, however, the $LOSS$ is relatively large. Therefore the CV method performed better than the RCV method on this simulated dataset. The optimal $\lambda$ from the complete case analysis was larger than that from the penalized likelihood methods and $LOSS$ was also much larger. However, $CV$ score from the complete case analysis was smaller than the $CV$ score from the PPL methods because the complete case analysis was aimed to fit the complete cases without incorporating information from the incomplete cases.

In the subsequent simulation study, 1000 datasets with the above parameter setting were simulated. For each simulated dataset, the optimal $\lambda$s chosen by the CV and RCV

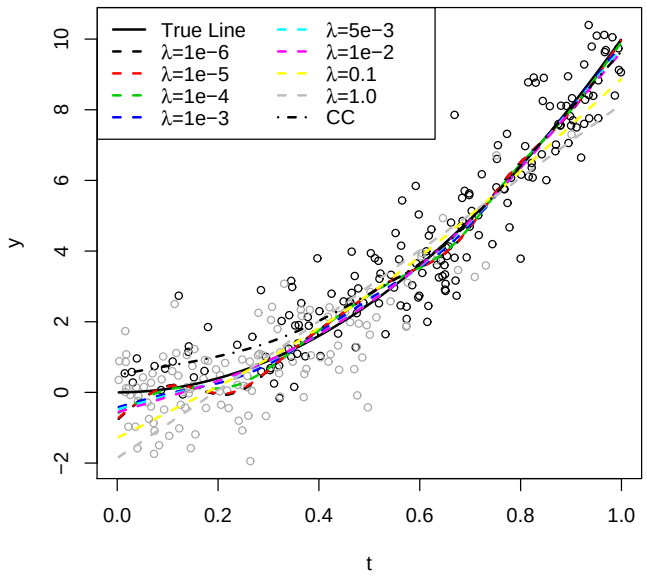


Figure 3: Regression lines from one set of simulation

Table 5: $RCV$, $CV$ and $LOSS$ for the first simulation study

| method | $\lambda$ | $RCV(\lambda)$ | $CV(\lambda)$ | $LOSS(\lambda)$ |
|---|---|---|---|---|
| PL | 0 | 0.00953561 | 1.112577 | 0.06188 |
| PL | 0.000001 | 0.00953559 | 1.112018 | 0.06141 |
| PL | 0.00001 | 0.00953541 | 1.107453 | 0.05768 |
| PL | 0.0001 | 0.00953441 | 1.083862 | 0.04026 |
| PL | 0.001 | 0.00954016 | 1.054673 | 0.01904 |
| PL | 0.005 | 0.00957467 | 1.046275 | 0.01884 |
| PL | 0.01 | 0.00960019 | 1.047713 | 0.02655 |
| PL | 0.1 | 0.00994471 | 1.196050 | 0.22773 |
| PL | 1 | 0.01029818 | 1.504486 | 0.57325 |
| CC | 0.008 | | 0.922399 | 0.16393 |

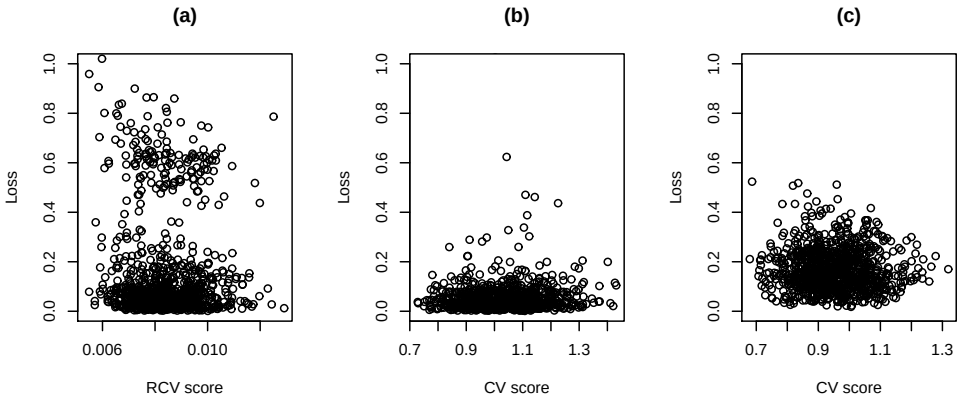


Figure 4: Relations of $RCV$, $CV$ scores from PPL, $CV$ from complete case analysis and LOSS

Table 6: Summary of $RCV$,$CV$ score and $LOSS$

| method | $\lambda$ | RCV($CV$) score | LOSS |
|---|---|---|---|
| | Median(Range) | Mean(std) | Mean(std) |
| $RCV$ | 0.005(0-10) | 0.008(0.001) | 0.156(0.214) |
| $CV$ | 0.005(0-0.1) | 1.031(0.119) | 0.050(0.062) |
| $CC$ | 0.018(0-0.046) | 0.952(0.102) | 0.167(0.083) |

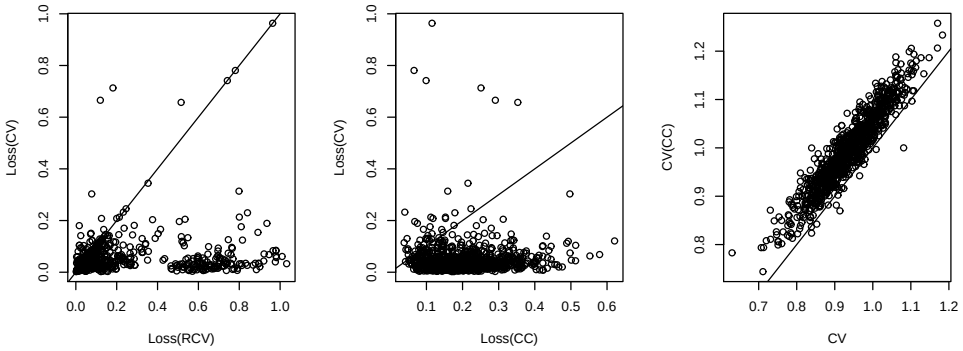


Figure 5: Comparison of $LOSS$ and $CV$ score between different methods

methods were recorded and the corresponding loss functions were computed. The results were summarized in Table 6. The numbers of complete cases were between 152 and 215 with an average of 180. The optimal $\lambda$s based on the loss function had a median of 0.005 and range (0.0001, 0.1). The optimal $\lambda$s chosen by the RCV and CV methods were both centered at 0.005. However, the optimal $\lambda$s chosen by the RCV method had a much larger range. Overall the $LOSS$ from the RCV method were larger with much larger variance than that from the CV method. The complete case analysis still had the largest $LOSS$. The scatter plots of $LOSS$ versus the cross-validation scores, corresponding to the $RCV$ for the PPL method, $CV$ for the PPL method, and complete case analysis over 1000 simulated datasets are presented in Figure 4. The CV method with PPL had the lowest $LOSS$ function and the corresponding $LOSS$ was generally stable. The $LOSS$ associated with the $RCV$ method had much wider spread and suggested a worse fit compared to the fit under the $CV$ method for the PPL approach. Scatter plots of those three $LOSS$ functions over 1000 simulated datasets are also presented in Figure 5. In the first plot, $LOSS$ from PPL with CV method ($LOSS(CV)$) is compared with $LOSS$ from the PPL with RCV method ($LOSS(RCV)$). This plot suggests that mostly the CV method can find a better fitted spline with smaller $LOSS$ than the RCV method when PPL is used. In the second plot, similar comparison was done with $LOSS$ from the PPL with CV method and the complete case analysis ($LOSS(CC)$). In the last plot, the $CV$ score from PPL method ($CV$) and $CV$ score from complete case analysis ($CV(CC)$) is compared. These plots showed that the PPL method with CV method was associated with smaller $LOSS$ function and better fit of the true association between the predictor and the outcome, even though the complete case analysis fit the complete cases better. So CV method still has its flaw because it only respond to complete cases. But overall, CV method in combination with PPL consistently chose a better fitted curve than the RCV method.

These simulation studies suggest that the RCV method is not stable and performs worse than the CV method in cross validation. It is probably because that the RCV method chooses the optimal smoothing parameter $\lambda$ based on the predictive error of $E[T|Y]$ from $T$. However, whether a curve fits the data well or not should still be evaluated based on the predictive error of $E[Y|T]$ from $Y$. Although the traditional cross validation method is not ideal because the complete cases are not a random sample of $[Y|T]$, it is still better.

## 3.2 SEMI-PARAMETRIC REGRESSIONS AND THE STATISTICAL METHODS FOR SEMI-PARAMETRIC REGRESSIONS

In Section 3.1, the PPL method for fitting a nonparametric regression, a NCS in specific, was explored. However, very often, we would have a dataset where some variables can be modeled by a parametric model and other variables cannot. The prime interest is generally to understand the variables that can be modeled parametrically after adjusted by the other variables. A semi-parametric model that consists of both parametric and nonparametric component can be used. Traditionally, a penalized likelihood is maximized to obtain the estimate of both parametric parameter and the nonparametric regression curve when data are complete. When data are MNAR, it may yield invalid inferences. In this section, we expanded our PPL method to incorporate the parametric component for data with nonignorable nonresponse.

### 3.2.1 A penalized likelihood method for semi-parametric regressions

For semi-parametric regression model (3.2), NCS can also be used to describe the effect of a predictor on the outcome with a similar penalty term in the penalized likelihood function. Consider a trivariate dataset $\{x_i, t_i, y_i\}_{i=1,\ldots,n}$ from model (3.2). The effect of $X$ on the outcome $Y$ can be modeled in a linear form and it is of primary interest. The effect of $T$ on $Y$ cannot be modeled by a parametric model and is not of primary interest. If $\theta = (\beta, g, \sigma^2)$, the penalized likelihood for a semi-parametric model can be defined similarly as a nonparametric model with the same roughness penalty as:

$$L(\theta) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{\sigma^2}\sum_{i=1}^{n}\{y_i - x_i\beta - g(t_i)\}^2 + \lambda\int_a^b \{g''(x)\}^2 dx. \tag{3.14}$$

In this model, the roughness penalty term is only a function of the nonparametric component. But it also impacts the parametric component. Variations of the $CV$ or $GCV$ to include the $X$ term were used to choose the optimal $\lambda$:

$$\begin{aligned} CV(\lambda) &= n^{-1}\sum_{i=1}^{n}\{y_i - x_i\hat{\beta}^{(-i)} - \hat{g}^{(-i)}(t_i;\lambda)\}^2 \\ GCV(\lambda) &= n^{-1}\frac{\sum_{i=1}^{n}\{y_i - x_i\hat{\beta}^{(-i)} - \hat{g}(t_i)\}^2}{\{1 - n^{-1}\mathrm{tr}A(\lambda)\}}. \end{aligned}$$

In both formula, $\hat{\beta}^{(-i)}$ is the estimate of $\beta$ after deleting observation $i$. The optimal $\lambda$ is derived by maximizing either $CV$, or GCV. The final estimates of $\hat{\beta}$ and $\hat{g}(\cdot)$ are then estimated from (3.14) with the optimal $\lambda$.

### 3.2.2 The penalized pseudolikelihood method for semi-parametric regression

Similarly, for the semi-parametric regression model (3.2) with dataset $\{x_i, t_i, y_i\}_{i=1,\ldots,n}$, where $y_i,\ i = 1, \ldots, m$ are observed and $y_i,\ i = m+1, \ldots, n$ are missing, to obtain the estimate of $\theta = (\beta, g, \sigma^2)$, the penalized pseudolikelihood is maximized:

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} L(\theta; F_n, \lambda) \\
&= \arg\max_{\theta} \Bigg[ -\sum_{i=1}^{m} \frac{\{y_i - x_i\beta - g(t_i)\}^2}{2\sigma^2} - \sum_{i=1}^{m} \log \int\int \exp[-\frac{\{y_i - x\beta - g(t)\}^2}{2\sigma^2} dF_n(x,t)] \\
&\quad -\lambda \int_a^b g''(t)^2 dt \Bigg] \\
&= \arg\max_{\theta} \Bigg[ -\sum_{i=1}^{m} \frac{\{y_i - x_i\beta - g(t_i)\}^2}{2\sigma^2} - \sum_{i=1}^{m} \log \Big[ \frac{1}{n} \sum_{j=1}^{n} \exp\Big\{ -\frac{\{y_i - x_j\beta - g(t_j)\}^2}{2\sigma^2} \Big\} \Big] \\
&\quad -\lambda \mathbf{g}^T K \mathbf{g} \Bigg], \qquad (3.15)
\end{aligned}$$

where $F_n(x,t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x,\ t_i \le t)$.

An analytical form of the standard error of $\hat{\beta}$ is difficult to derive. Resampling based methods such as Bootstrap and Jackknife can be used for standard error estimation.

### 3.2.3 A simulation study

A simulation study was carried out to evaluate the performance of this penalized pseudolikelihood method for semi-parametric regression model on data with nonignorable nonresponse. In particular, the PPL method should yield an unbiased estimate of the parameter of interest, $\beta$. The complete data $\{t_i, x_i, y_i\}_{i=1,\ldots,n}$ were generated from

$$X \sim N(0,1), \quad T \sim Unif(0,1), \text{ and } Y = X + 10T^2 + N(0,1). \qquad (3.16)$$

Table 7: Empirical bias and standard deviation of $\hat{\beta}$ when under three methods

| Method | Empirical Bias(se)of $\hat{\beta}$ | $LOSS$ |
|---|---|---|
| $RCV$ | 0.013(0.084) | 0.161(0.235) |
| $CV$ | 0.006(0.078) | 0.048(0.069) |
| $CV(CC)$ | 0.065(0.078) | 0.204(0.091) |

The nonresponse in $Y$ was created based on the following mechanism:

$$Pr[R = 1|T, X, Y] = \Phi(-1.7 + 0.5Y). \tag{3.17}$$

One thousand datasets were simulated and the number of complete cases ranged from 153 to 205, with an average of 180. For each dataset, the values of $\hat{\beta}$ when the $RCV$ score and the $CV$ score are minimal were recorded. The complete case analysis based on a semi-parametric model was also conducted to each simulated data for comparison purpose. The *gam(mgcv)* function in R was used (The R Development Core Team, 2008). $LOSS$ is calculated by

$$LOSS(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[\hat{(}\hat{\beta}_\lambda - \beta)x_i + \hat{g}_\lambda(t_i) - 10t^2]^2.$$

Table 7 presents the empirical bias and empirical standard deviation of $\hat{\beta}$ under the PPL method with either $RCV$ and $CV$ for cross validation and the complete case analysis. This simulation study suggests that the PPL method has negligible bias and the complete case analysis is apparently biased. Among the two cross validation methods for the PPL method, the $CV$ method is associated with smaller bias and $LOSS$ function.

## 3.3 EXAMPLE: DATA FROM CLINICAL TRIAL TO TREAT PANIC AND GENERALIZED ANXIETY DISORDERS (PD/GAD)

This penalized pseudolikelihood method was further illustrated by using data collected as part of an NIH-funded clinical trial to treat panic and/or generalized anxiety disorders (PD/GAD) (Rollman, *et al.*, 2005). Over a 22-month period (7/00-4/02), 191 primary care patients with PD/GAD were randomized into two groups: a telephone-based collaborative care intervention or a "usual care" control condition with a ratio of 3 versus 2. Afterwards blinded telephone follow-up assessments were conducted at 2-, 4-, 8-, and 12-months. By 12-months, 15% of patients dropped-out, 65%-75% completed a follow-up assessment at the appropriate time-point, and 95% completed $\geq 1$ follow-up assessments. Using the continuous outcome of decline in Rating Scale for Anxiety (HRS-A) score from baseline at 12 month (change), we compared outcomes using three methods: (a) Penalized pseudolikelihood with RCV method. (b) Penalized pseudolikelihood with $CV$ method. (c) Function *gam(mgcv)* in R. We used spline to fit the baseline HRS-A score and the treatment effect is of interest.

In this dataset, 143 patients had HRS-A score recorded at 12 months and the other 48 had missing 12-month HRS-A score. Among these 48 patients, 22 withdrew from the study due to: time constraint, no longer interested, or simply refused. Two patients were deceased before 12 months and another 14 patients could not be reached. The other 10 were excluded because they were later found out to be ineligible by protocol. The treatment effect (trt) is of interest. Normally, patients with higher baseline score (base) may have more reduction, but they are also found to be more treatment resistant. There the baseline score needs to be adjusted in regression analysis. Figure 6 shows the relation between these three variables. It seems that patients in intervention (trt=1) had more reduction on HRS-A at 12 month than patients in UC (trt=0). Those who have higher baseline score was associated with more absolute improvement. The following semi-parametric regression model was considered:

$$change = \beta \cdot trt + g(base) + N(0, \sigma^2).$$

In this analysis, the standard error of the estimate is derived from 500 bootstrap samples of size 100. The estimate of the treatment effect is listed in Table 8. In the original report,

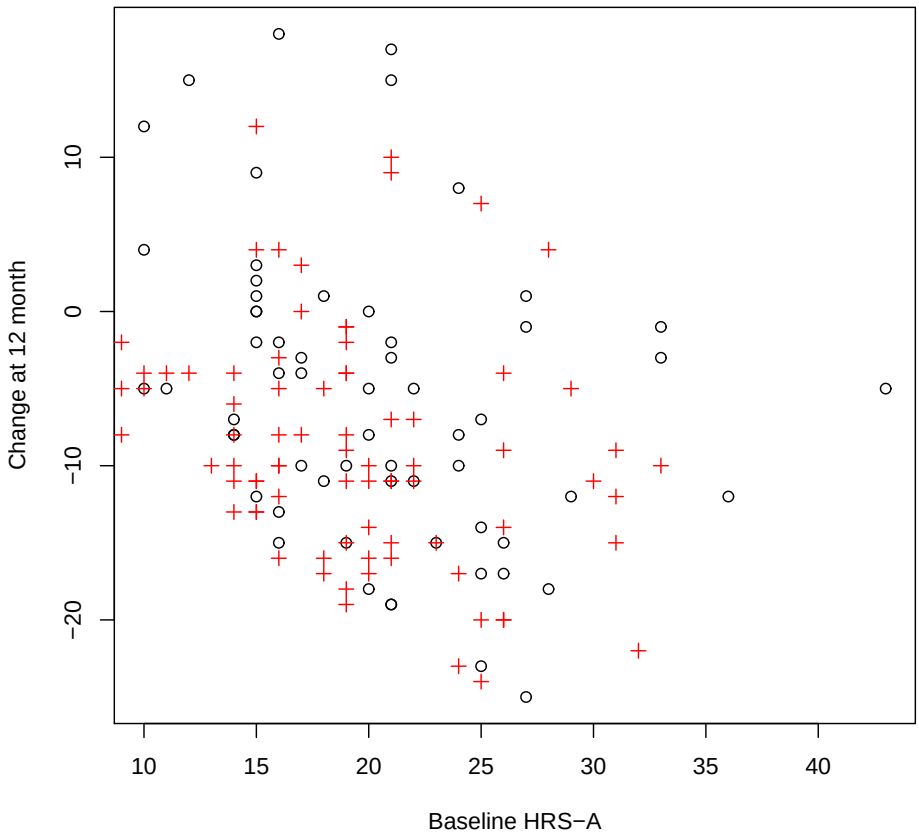


Figure 6: Baseline HRS-A, treatment group and changes on HRS-A at 12 month

Table 8: Estimate of treatment effect

| Method | $\lambda$, Median(range) | $\hat{\beta}$, estimate(bootstrap se) |
|---|---|---|
| *RCV* | 0.1(0-10) | -3.477(3.003) |
| *CV* | 1.0(0-10) | -3.205(2.390) |
| *CC* | | -3.694 (1.385) |

random regression models were used on all data from 5 time points to account for between subject variation (Rollman *et al.*, 2005). The difference in change score between UC and intervention estimated at 12 month was -3.6 points with 95% CI of $(-6.4, -0.8)$. It is almost the same from our complete case analysis. Estimates from PPL are very similar but with larger standard errors. The splines from PPL have a very large variation. The optimal $\lambda$ covers from 0 to 10. It may have caused the large variance of the estimate of the treatment effect.

In complete case analysis, an ignorable penalized likelihood method was used. It assumes that the missingness depend only on the covariates. The PPL, on the other hand, assumes that the missingness depends only on the outcome. Neither of them are testable. The results from PPL are reasonably close to that using ignorable penalized likelihood method. Even though the larger standard error will make the treatment effect insignificant, we can still conclude that the results supported the ignorable ML method.

# 4.0 DISCUSSION AND CONCLUSION

## 4.1 AN INDEX OF LOCAL SENSITIVITY TO NONIGNORABILITY FOR A PSEUDOLIKELIHOOD METHOD

Selection models supply a standard framework for analysis of data with missing values. Specification of a parametric model is usually required for making the maximum likelihood inference. When data are MAR and complete-data model parameters are distinct from the parameters for the missing-data mechanism, the mechanism is ignorable and the inference can be made on the likelihood based on observed values. Without prior knowledge on the missing-data mechanism, sensitivity analysis is necessary in order to evaluate the impact from alternative assumptions on the missng-data mechanism. $ISNI_{ML}$ is a local sensitivity index to nonignorability developed for the maximum likelihood method (Troxel *et al.*, 2004). It is used to evaluate how a small deviation from MAR may affect the ML estimate. If the estimate is not sensitive to deviation from MAR, then the ignorable ML estimate may be used for subsequent conclusion. The developed $ISNI_{ML}$ requires a parametric model, often a logistic regression, for the missing-data mechanism. Here we developed a new local sensitivity index to nonignorability based on a pseudolikelihood method (Tang *et al.*, 2003) that does not require parametric specification fo the missing-data mechanism. Through simulation studies and analysis of a dataset, it was demonstrated that $ISNI_{PL}$ had similar performance with $ISNI_{ML}$ when the ML method used a correctly specified model for the missing-data mechanism. In some cases $ISNI_{PL}$s were a little more conservative than $ISNI_{ML}$ because fewer assumptions were used. When the missing-data mechanism is quite differenct from what was required by $ISNI_{ML}$, $ISNI_{PL}$ still supplied reationable guidance on the local sensitivity, but $ISNI_{ML}$ would be misleading. Since both $ISNI_{ML}$ and $ISNI_{PL}$ are not

difficult to compute, it is recommended that all of them are computed and compared. If either of them indicate potential local sensitivity to nonignorability, more comprehensive sensitivity analyses on how these estimates, such as the estimate of the prime interest, change on a range of nonignorability parameters.

Based on whether the distribution of the covariate can be assumed parametric or not, two versions of $ISNI_{PL}$ were developed here. If the distribution of $X$ is not normal, the performance of the bias correction is essential. As a matter of fact, the bias from the kernel smoothing can take over the entire computation of $ISNI_{PL}$. This is not desired. A rescaling of the covariates are needed. In the smoking data example, the standard deviation of the covariate is very large. The optimal $h$ is proportional to it. A large $h$ also leads to large bias and re-scale of the covariates is necessary.

## 4.2 SEMI-PARAMETRIC REGRESSIONS AND THE STATISTICAL METHODS FOR SEMI-PARAMETRIC REGRESSIONS

In this thesis, we developed a penalized pseudolikelihood method for a nonparametric/semi-parametric regression. A new cross validation method was proposed based on the predication error of the nonparametric variable. This RCV method was compared to traditional CV method through series of simulations.

In the simple nonparametric regression simulations, the lines fitted with PPL has less bias than that from ignorable penalized likelihood method. It does not matter which cross validation was used. The CV method with PPL out performed the RCV method, judging by the $LOSS$ of the fitted lines chosen by these two methods. The reason may be that the relation between $Y$ and $E[Y|T]$ is not reflected in that between $T$ and $[T|Y]$. In the semi-parametric regression simulations, judging by $LOSS$ or the bias of $\hat{\beta}$, the CV method with PPL method also out performed the RCV method. Despite its advantage over RCV, the CV scores in nonparametric complete case analysis is still lower than the CV scores obtained from PPL. It suggested that CV method is only controlled by the complete case and the missing data are completely ignored. It is not ideal and has to be used in combination with

PPL. The PPL method compensate some of the bias caused by the nonignorable missingness.

The limitation of this study would be that due to computation difficulty, a bootstrap for each simulated dataset was not done. Only empirical standard errors were obtained. We are unable to acquire information on the coverage probability. Several studies on complete data have come to the conclusion that the estimating process of $\beta$ in semi-parametric regression is confounded by the smoothing process of the NCS (Green, 1987). In fact, for semi-parametric regression on trivariate dataset, the bias of $\hat{\beta}$ consists of two parts. The first part is in the order of $o(n^{-1/2})$, while the second part is bounded by the square root of integrated squared bias of $\hat{g}$. The variance of $\hat{\beta}$ is in the order of $o(n^{-1/2})$. If the bias can achieve the order of $o(n^{-1/2})$, this bias reduces at least as fast as the variance. However, it is almost always at the expense of undersmoothing the nonparametric components (Rice, 1986). Given the computation power, a study on the coverage probabilities may provide some insights on how the bias and variance of $\hat{\beta}$ from PPL evolve with different sample sizes.

# APPENDIX A

# DERIVING ISNI FOR A PSEUDOLIKELIHOOD OF BIVARIATE NORMAL DATA

In Section 2.2.1, $ISNI_{PL1}$ was derived for bivariate normal data. In this part, the details of how it was derived from (a) Pseudolikelihood method and (b) Brown's estimators were presented in Section A.1 and Section A.2, respectively.

## A.1 FROM PSEUDOLIKELIHOOD

The details of deriving $\bigtriangledown PL$:

$$
\begin{aligned}
\frac{\partial PL}{\partial \beta_0} &= \frac{1}{\sigma^2}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1 x_i) - \frac{\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)\lambda}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2} \overset{\lambda\to 0}{\to} \frac{1}{\sigma^2}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1 x_i) \\
\frac{\partial PL}{\partial \beta_1} &= \frac{1}{\sigma^2}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1 x_i)x_i + \frac{m\lambda\beta_\lambda\hat{\sigma}_x^2}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2} - \frac{\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)\lambda\hat{\mu}_x}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2} \\
&\quad -\frac{\lambda\beta_\lambda\hat{\sigma}_x^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2}\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)^2 \overset{\lambda\to 0}{\to} \frac{1}{\sigma^2}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1 x_i)x_i \\
\frac{\partial PL}{\partial \sigma^2} &= -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1 x_i)^2 + \frac{m\lambda^2}{2(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)} \\
&\quad -\frac{\lambda^2}{2(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2}\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)^2 \overset{\lambda\to 0}{\to} -\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1 x_i)^2
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial PL}{\partial \lambda} &= \frac{m(\lambda\sigma^2+\beta_\lambda\beta_1\hat{\sigma}_x^2)}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2}+\frac{1}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2}\sum_{i=1}^{m}[(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)(y_i-\beta_0-\beta_1\hat{\mu}_x)] \\
&\quad -\frac{1}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2}\sum_{i=1}^{m}[(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)^2(\lambda\sigma^2+\beta_\lambda\beta_1\hat{\sigma}_x^2)] \\
&\quad \overset{\lambda\to 0}{\to} m\beta_1+\frac{1}{\hat{\sigma}_x^2}\sum_{i=1}^{m}[(y_i-\beta_0-\beta_1\hat{\mu}_x)(x_i-\hat{\mu}_x)]-\frac{1}{\hat{\sigma}_x^2}\sum_{i=1}^{m}\beta_1(x_i-\hat{\mu}_x)^2
\end{aligned}
$$

Detail of deriving $\bigtriangledown^2 PL$:

$$
\begin{aligned}
\frac{\partial^2 PL}{\partial \beta_0^2} &= -\frac{m}{\sigma^2}+\frac{m\lambda}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2}\overset{\lambda\to 0}{\to}-\frac{m}{\sigma^2} \\
\frac{\partial^2 PL}{\partial \beta_0\partial \beta_1} &= -\frac{m\bar{x}}{\sigma^2}+\frac{m\lambda^2\hat{\mu}_x}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2}+\frac{2\lambda^2\beta_\lambda\hat{\sigma}_x^2\sum_{i=1}^{m}(y_{\lambda i}-\beta_0-\beta_\lambda\hat{\mu}_x)}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2} \\
&\overset{\lambda\to 0}{\to} -\frac{m\bar{x}}{\sigma^2} \\
\frac{\partial^2 PL}{\partial \beta_0\partial \sigma^2} &= -\frac{1}{\sigma^4}\sum_{i=1}^{m}(y_{\lambda i}-\beta_0-\beta_\lambda x_i)+\frac{\sum_{i=1}^{m}\lambda^3(y_{\lambda i}-\beta_0-\beta_\lambda\hat{\mu}_x)}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2} \\
&\overset{\lambda\to 0}{\to} -\frac{1}{\sigma^4}\sum_{i=1}^{m}(y_{\lambda i}-\beta_0-\beta_\lambda x_i) \\
\frac{\partial^2 PL}{\partial \beta_0\partial \lambda} &= -\frac{1}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2}\sum_{i=1}^{m}[2\lambda(y_i-\beta_0-\beta_1\hat{\mu}_x)+(x_i-\hat{\mu}_x)] \\
&\quad -\frac{2\lambda\sum_{i=1}^{m}[(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)(\lambda\sigma^2+\beta_\lambda\beta_1\hat{\sigma}_x^2)]}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2} \\
&\overset{\lambda\to 0}{\to} -\frac{\sum_{i=1}^{m}(x_i-\mu_x)}{\hat{\sigma}_x} \\
\frac{\partial^2 PL}{\partial \beta_1^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{m}x_i^2+\frac{m\lambda^2\hat{\sigma}_x^2}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2}-\frac{2m\lambda^2\beta_\lambda^2\hat{\sigma}_x^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2} \\
&\quad +\frac{\sum_{i=1}^{m}\lambda^2(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)\hat{\mu}_x^2}{\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2}-\frac{\lambda^2\hat{\sigma}_x^2\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^2} \\
&\quad +\frac{4\lambda^2\beta_\lambda^2\hat{\sigma}_x^4\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat{\mu}_x)^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat{\sigma}_x^2)^3} \\
&\overset{\lambda\to 0}{\to} \frac{1}{\sigma^2}\sum_{i=1}^{m}x_i^2
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 PL}{\partial\beta_1\partial\sigma^2} &= -\frac{1}{\sigma^4}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1x_i)x_i - \frac{m\lambda^3\beta_\lambda\hat\sigma_x^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} \\
&\quad + \frac{\lambda^3\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)\hat\mu_x}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} + \frac{2\lambda^3\beta_\lambda\hat\sigma_x^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^3}\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)^2 \\
&\overset{\lambda\to0}{\to} -\frac{1}{\sigma^4}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1x_i)x_i \\
\frac{\partial^2 PL}{\partial\beta_1\partial\lambda} &= \frac{m(2\beta_\lambda-1)\hat\sigma_x^2}{\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2} - \frac{2m\lambda\beta_\lambda\hat\sigma_x^2(\lambda\sigma^2+\beta_\lambda\beta_1\hat\sigma_x^2)}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} \\
&\quad - \frac{\hat\mu_x}{\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2}\sum_{i=1}^{m}[2\lambda(y_i-\beta_0-\beta_1\hat\mu_x)+(x_i-\hat\mu_x)] \\
&\quad + \frac{2\lambda\hat\mu_x(\lambda\sigma^2+\beta_\lambda\beta_1\hat\sigma_x^2)\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} \\
&\quad - \frac{(2\beta_\lambda-1)\hat\sigma_x^2\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} \\
&\quad - \frac{2\lambda\beta_\lambda\hat\sigma_x^2\sum_{i=1}^{m}\{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)(y_i-\beta_0-\beta_1\hat\mu_x)\}}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} \\
&\quad + \frac{4\lambda\beta_\lambda\hat\sigma_x^2(\lambda\sigma^2+\beta_\lambda\beta_1\hat\sigma_x^2)\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^3} \\
&\overset{\lambda\to0}{\to} m - \frac{m\hat\mu_x(\bar x-\hat\mu_x)}{\hat\sigma_x^2} - \frac{\sum_{i=1}^{m}(x_i-\hat\mu_x)^2}{\hat\sigma_x^2} \\
\frac{\partial^2 PL}{\partial\sigma^4} &= \frac{m}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1x_i)^2 \\
&\quad - \frac{m\lambda^4}{2(\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} + \frac{\lambda^4}{(\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^3}\sum_{i=1}^{m}(y_{\lambda i}-\beta_0-\beta_\lambda\hat\mu_x)^2 \\
&\overset{\lambda\to0}{\to} \frac{m}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{m}(y_i-\beta_0-\beta_1x_i)^2 \\
\frac{\partial^2 PL}{\partial\sigma^2\partial\lambda} &= \frac{m\lambda}{\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2} - \frac{m\lambda^2(\lambda\sigma^2+\beta_\lambda\beta_1\hat\sigma_x^2)}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} - \frac{\lambda\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)^2}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} \\
&\quad - \frac{\lambda^2\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)(y_i-\beta_0-\beta_1\hat\mu_x)}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^2} \\
&\quad + \frac{2\lambda^2(\lambda\sigma^2+\beta_\lambda\beta_1\hat\sigma_x^2)}{(\lambda^2\sigma^2+\beta_\lambda^2\hat\sigma_x^2)^3}\sum_{i=1}^{m}(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\hat\mu_x)^2 \\
&\overset{\lambda\to0}{\to} 0
\end{aligned}
$$

Let $(\hat\beta_{00},\ \hat\beta_{10},\ \hat\sigma_0^2)$ and $\hat\beta_{\lambda0}=\lambda\hat\beta_{10}+1$ be the maximum likelihood estimate at $\lambda\to0$. Using

the notations in Section 2.2.1, at $(\hat{\beta}_{00},\ \hat{\beta}_{10},\ \hat{\sigma}_0^2,\ \lambda=0)$,

$$
\begin{aligned}
\frac{\partial^2 PL}{\partial \beta_0^2}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= -\frac{m}{\hat{\sigma}_0^2} \\
\frac{\partial^2 PL}{\partial \beta_0 \partial \beta_1}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= -\frac{m\bar{x}}{\hat{\sigma}_0^2} \\
\frac{\partial^2 PL}{\partial \beta_0 \partial \sigma^2}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= -\frac{\sum_{i=1}^m (y_i - \hat{\beta}_{00} - \hat{\beta}_{01}x_i)}{\hat{\sigma}_0^4} \\
\frac{\partial^2 PL}{\partial \hat{\beta}_0 \partial \lambda}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= -\frac{m(\bar{x} - \hat{\mu}_x)}{\hat{\sigma}_x^2} \\
\frac{\partial^2 PL}{\partial \beta_1^2}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= -\frac{\sum_{i=1}^m x_i^2}{\hat{\sigma}_0^2} \\
\frac{\partial^2 PL}{\partial \beta_1 \partial \sigma^2}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= -\frac{\sum_{i=1}^m (y_i - \hat{\beta}_{00} - \hat{\beta}_{01}x_i)x_i}{\hat{\sigma}_0^4} \\
\frac{\partial^2 PL}{\partial \hat{\beta}_1 \partial \lambda}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= m - \frac{m\hat{\mu}_x(\bar{x} - \hat{\mu}_x)}{\hat{\sigma}_x^2} - \frac{\sum_{i=1}^m (x_i - \hat{\mu}_x)^2}{\hat{\sigma}_x^2} \\
\frac{\partial^2 PL}{\partial \sigma^4}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= \frac{m}{2\hat{\sigma}_0^4} - \frac{\sum_{i=1}^m (y_i - \hat{\beta}_{00} - \hat{\beta}_{01}x_i)^2}{\hat{\sigma}_0^6} \\
\frac{\partial^2 PL}{\partial \sigma^2 \partial \lambda}\Big|_{\hat{\beta}_{00},\hat{\beta}_{10},\hat{\sigma}_0^2,\lambda=0} &= 0
\end{aligned}
$$

MLE as $\lambda \to 0$ is $(\hat{\beta}_{00}, \hat{\beta}_{01}, \hat{\sigma}_0^2) = (\bar{y} - \frac{s_{12}}{s_{11}}\bar{x}, \frac{s_{12}}{s_{11}}, s_{22} - \frac{s_{12}^2}{s_{11}})$. Then

$$
ISNI_{PL1} = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_x^2}
\begin{pmatrix}
m & \sum_{i=1}^m x & 0 \\
\sum_{i=1}^m x & \sum_{i=1}^m x_i^2 & 0 \\
0 & 0 & \frac{m}{2\hat{\sigma}_0^2}
\end{pmatrix}^{-1}
\begin{pmatrix}
-m(\bar{x} - \hat{\mu}_x) \\
m\hat{\sigma}_x^2 - \sum_{i=1}^m x_i^2 + m\hat{\mu}_x\bar{x} \\
0
\end{pmatrix}
$$

## A.2 FROM BROWN'S ESTIMATOR

In this part of appendix, Brown's estimators for dataset $(X, Y_\lambda)$ are presented. The derivatives of them with respect to $\lambda$ were calculated. The results were compared with what we had from appoximate method.

Using the notations of section ( 2.2.1), Brown' estimators for $(X, Y_\lambda)$ are:

$$\begin{aligned}
\hat{\beta}_0(\lambda) &= \bar{y} + \frac{\bar{x} - \hat{\mu}_x}{\lambda} + (\hat{\mu}_x - \bar{x})\frac{\lambda s_{22} + s_{11}/\lambda + 2s_{12}}{\lambda s_{12} + s_{11}} \\
&\quad -\frac{\hat{\mu}_x}{\hat{\sigma}_x^2}\{s_{12} + b_\lambda(\sigma_x^2 - s_{11})\} \\
&= \bar{y} - b_\lambda \bar{x} - \frac{\hat{\mu}_x}{\hat{\sigma}_x^2}(s_{12} - b_\lambda s_{11}) \\
\hat{\beta}_1(\lambda) &= \frac{1}{\hat{\sigma}_x^2}\{s_{12} + b_\lambda(\hat{\sigma}_x^2 - s_{11})\} \\
\hat{\sigma}_y^2(\lambda) &= s_{22} + b_\lambda^2(\hat{\sigma}_x - s_{11}) \\
\sigma^2(\lambda) &= \hat{\sigma}_y^2(1 - \frac{\beta_1^2\hat{\sigma}_x}{\hat{\sigma}_y^2}) \\
&= s_{22} - \frac{b_\lambda^2}{\hat{\sigma}_x^2}(\hat{\sigma}_x^2 - s_{11})^2 + (b_\lambda^2 - \frac{2s_{12}b_\lambda}{\hat{\sigma}_x^2})(\hat{\sigma}_x^2 - s_{11}) - \frac{s_{12}^2}{\hat{\sigma}_x^2}
\end{aligned}$$

where $b_\lambda = \frac{\lambda s_{22} + s_{12}}{\lambda s_{12} + s_{11}} \overset{\lambda \to 0}{\to} \frac{s_{12}}{s_{11}}$ and $b'_\lambda = \frac{s_{22}s_{11} - s_{12}^2}{(\lambda s_{12} + s_{11})^2} \overset{\lambda \to 0}{\to} \frac{s_{22}s_{11} - s_{12}^2}{s_{11}^2}$.
$ISNI_{PL1}$ derived from these estimators would be:

$$\begin{aligned}
\frac{\partial \beta_0}{\partial \lambda} &= -b'_\lambda \bar{x} + \frac{\hat{\mu}_x}{\hat{\sigma}_x^2} b'_\lambda s_{11} \\
&\overset{\lambda \to 0}{\to} -\frac{s_{22}s_{11} - s_{12}^2}{s_{11}^2}\bar{x} + \frac{\hat{\mu}_x}{\hat{\sigma}_x^2}\frac{s_{22}s_{11} - s_{12}^2}{s_{11}} \\
\frac{\partial \beta_1}{\partial \lambda} &= \frac{1}{\hat{\sigma}_x^2} b'_\lambda(\hat{\sigma}_x^2 - s_{11}) \\
&\overset{\lambda \to 0}{\to} \frac{s_{22}s_{11} - s_{12}^2}{s_{11}^2 \hat{\sigma}_x^2}(\hat{\sigma}_x^2 - s_{11}) \\
\frac{\partial \sigma^2}{\partial \lambda} &= -\frac{2b_\lambda b'_\lambda}{\hat{\sigma}_x^2}(\hat{\sigma}_x^2 - s_{11})^2 + (b_\lambda - \frac{2s_{12}}{\hat{\sigma}_x^2})(\hat{\sigma}_x^2 - s_{11})b'_\lambda \\
&\overset{\lambda \to 0}{\to} 0
\end{aligned}$$

If substituting $\sum_{i=1}^m x_i^2 = s_{11} + m\bar{x}^2$, the results derived from pseudolikelihood is the same with $ISNI_{PL1}$ derived from Brown's protective estimators.

# APPENDIX B

# DERIVING ISNI FOR A PSEUDOLIKELIHOOD WITH KERNEL SMOOTHING

In Section 2.2.2.2, $ISNI_{PL2}$ for bivariate dataset $\{x_i, y_i\}_{i=1,\ldots,n}$ with the distribution of $X$ estimated from kernel smoothing is presented. In B.1, the computation details of how the index was derived is presented and in B.2, the details of how the bias corrections were computed is presented.

## B.1 DERIVING THE INDEX

Using Gaussian Kernel $\hat{f}(x) = \frac{1}{n}\sum_{i:1,n}\psi(\frac{x-x_i}{h})$

$$\begin{aligned} P(y_{\lambda i}) &= \int \phi(\frac{\lambda y_i + x_i - \lambda\beta_0 - (\lambda\beta_1+1)x}{\lambda^\sigma}) \cdot \frac{1}{n}\sum_{j=1}^{n}\psi(\frac{x-x_j}{h})dx \\ &= \frac{1}{n}\sum_{j=1}^{n}\int \phi(\frac{\lambda y_i + x_i - \lambda\beta_0 - (\lambda\beta_1+1)x}{\lambda\sigma}) \cdot \psi(\frac{x-x_j}{h})dx \\ &= \frac{1}{n}\sum_{j=1}^{n}\phi\left(\frac{\lambda y_i + x_i - \lambda\beta_0 - (\lambda\beta_1+1)x_j}{\sqrt{\lambda^2\sigma^+(\lambda\beta_1+1)^2h^2}}\right) \end{aligned}$$

Let $V_y = \lambda^2\sigma^2 + (\lambda\beta_1+1)^2h^2$ and $A = -\frac{\{\lambda(y_i-\beta_0-\beta_1x_j)+(x_i-x_j)\}^2}{2V_y}$ then

$$\log P(y_{\lambda i}) = -\frac{1}{2}\log(V_y) + \log\{\frac{1}{n}\sum_{j=1}^{n}\exp(A)\}$$

$$
\begin{aligned}
\frac{\partial \log P(y_{\lambda i})}{\partial \lambda} &= -\frac{\lambda\sigma^2 + (\lambda\beta_1+1)\beta_1 h^2}{V_y} - \frac{1}{\sum_{j=1}^n \exp(A)} \sum_{j=1}^n \exp(A) \\
&\quad \Big[ \frac{\{\lambda(y_i - \beta_0 - \beta_1 x_j) + (x_i - x_j)\}(y_i - \beta_0 - \beta_1 x_j)}{V_y} \\
&\quad - \frac{\{\lambda(y_i - \beta_0 - \beta_1 x_j) + (x_i - x_j)\}^2 \{\lambda\sigma^2 + (\lambda\beta_1+1)\beta_1 h^2\}}{V_y^2} \Big] \\
&\overset{\lambda\to 0}{\to} -\beta_1 - \frac{1}{\sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)^2}{2h^2}\right]} \cdot \\
&\quad \frac{1}{h^2} \sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)^2}{2h^2}\right] \Big[(x_i - x_j)(y_i - \beta_0 - \beta_1 x_j) - (x_i - x_j)^2 \beta_1\Big]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \log P(y_{\lambda i})}{\partial \lambda \partial \beta_0} &\overset{\lambda\to 0}{\to} -\frac{1}{h^2 \sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)^2}{2h^2}\right]} \sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)^2}{2h^2}\right](x_i - x_j) \\
\frac{\partial \log P(y_{\lambda i})}{\partial \lambda \partial \beta_1} &\overset{\lambda\to 0}{\to} -1 + \frac{1}{h^2 \sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)^2}{2h^2}\right]} \sum_{j=1}^n \exp\left[-\frac{(x_i-x_j)}{2h^2}\right](x_i - x_j)x_i
\end{aligned}
$$

## B.2 BIAS CORRECTION

If $f(x)$ is a normal distribution with mean $\mu_x$ and variance $\sigma^2$, the bias of $\hat{f}(x)$ is

$$
\text{bias of } \hat{f}(x) = \frac{1}{2}h^2 f''(x) = \frac{1}{2\sqrt{2\pi\sigma_x^2}} h^2 \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^4} - \frac{1}{\sigma_x^2}\right\}.
$$

The bias of $P(y_{\lambda i})$ $(\Delta)$ is

$$\begin{aligned}
\Delta &= \frac{1}{2}h^2 \int \frac{1}{\sqrt{2\pi\sigma^2\lambda^2}} \exp\left\{-\frac{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda x)^2}{2\lambda^2\sigma^2}\right\} \\
&\quad \cdot\frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^4}-\frac{1}{\sigma_x^2}\right\}dx \\
&= -\frac{1}{2\sigma_x^2}h^2\phi\left(\frac{y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x}{\sqrt{\sigma_\lambda^2}}\right)+\frac{1}{2}h^2\frac{1}{2\pi\sqrt{\sigma_x^2\sigma^2\lambda^2}}\exp\left\{\frac{(y_{\lambda i}-\lambda\beta_0)^2}{2\lambda^2\sigma^2}\right. \\
&\quad \left.-\frac{\mu_x^2}{2\sigma_x^2}+\frac{1}{2\sigma_\lambda^2\lambda^2\sigma^2\sigma_x^2}(\beta_\lambda(y_{\lambda i}-\lambda\beta_0)\sigma_x^2+\mu_x\lambda^2\sigma^2)^2\right\} \\
&\quad \int\frac{(x-\mu_x)^2}{2\sigma_x^4}\exp\left[-\frac{\sigma_\lambda^2}{2\lambda^2\sigma^2\sigma_x^2}[x-\frac{1}{\sigma_\lambda^2}\{\beta_\lambda(y_{\lambda i}-\lambda\beta_0)\sigma_x^2+\mu_x\lambda^2\sigma^2\}]\right]dx \\
&= -\frac{1}{2\sigma_x^2}h^2\phi\left(\frac{y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x}{\sqrt{\sigma_\lambda^2}}\right)+\frac{1}{2}h^2\phi\left(\frac{y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x}{\sqrt{\sigma_\lambda^2}}\right) \\
&\quad \left[\frac{1}{\sigma_x^4}[\frac{1}{\sigma_\lambda^2}(\beta_\lambda(y_{\lambda i}-\lambda\beta_0)\sigma_x^2+\mu_x\lambda^2\sigma^2)-\mu_x]^2+\frac{\lambda^2\sigma^2}{\sigma_x^2\sigma_\lambda^2}\right] \\
&= \frac{1}{2}h^2\phi\left(\frac{y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x}{\sqrt{\sigma_\lambda^2}}\right)\left[\frac{\beta_\lambda^2}{\sigma_\lambda^4}\{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x)^2-\sigma_\lambda^2\}\right] \\
&\overset{\lambda\to 0}{\to} \frac{1}{2\sigma_x^4}h^2\phi\left(\frac{x_i-\mu_x}{\sigma_x}\right)\{(x_i-\mu_x)^2-\sigma_x^2\}
\end{aligned}$$

Let $B=\frac{y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x}{\sqrt{\sigma_\lambda^2}}$. The first derivatives of $\Delta$ with respect to $\lambda$ is

$$\begin{aligned}
\frac{\partial\Delta}{\partial\lambda} &= \frac{1}{2}h^2\phi(B)\Big[-\frac{\lambda\sigma^2+\beta_\lambda\beta_1\sigma_x^2}{\sigma_\lambda^2}\frac{\beta_\lambda^2}{\sigma_\lambda^4}\{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x)^2-\sigma_x^2\} \\
&\quad +\{-\frac{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x)(y_i-\beta_0-\beta_1\mu_x)}{\sigma_\lambda^2}+ \\
&\quad \frac{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x)^2(\lambda\sigma^2+\beta_\lambda\beta_1\sigma_x^2)}{\sigma_\lambda^4}\}\cdot[\frac{\beta_\lambda^2}{\sigma_\lambda^4}\{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x)^2-\sigma_x^2\}] \\
&\quad +\left\{\frac{2\beta_\lambda\beta_1}{\sigma_\lambda^4}-\frac{4\beta_\lambda^2(\lambda\sigma^2+\beta_\lambda\beta_1\sigma_x^2)}{\sigma_\lambda^6}\right\}\{(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x)^2-\sigma_x^2\} \\
&\quad +\frac{\beta_\lambda^2}{\sigma_\lambda^4}\{2(y_{\lambda i}-\lambda\beta_0-\beta_\lambda\mu_x)(y_i-\beta_0-\beta_1\mu_x)-2(\lambda\sigma^2\beta_\lambda\beta_1\sigma_x^2)\}\Big] \\
&\overset{\lambda\to 0}{\to} -\frac{1}{2}h^2\phi\left(\frac{x_i-\mu_x}{\sigma_x}\right)\Big[\frac{\beta_1}{\sigma_x^4}\{(x_i-\mu_x)^2-\sigma_x^2\} \\
&\quad +\frac{1}{\sigma_x^6}(x_i-\mu_x)(y_i-\beta_0-\beta_1x_i)\{(x_i-\mu_x)^2-3\sigma_x^2\}\Big]
\end{aligned}$$

Then the partial derivatives of $\Delta$ with respect to $\lambda$ and $\theta = (\beta_0, \beta_1, \sigma^2)$ is:

$$\begin{aligned}
\frac{\partial^2 \Delta}{\partial \lambda \partial \beta_0} &\overset{\lambda \to 0}{\to} -\frac{1}{2} h^2 \phi \left( \frac{x_i - \mu_x}{\sigma_x} \right) \frac{1}{\sigma^6} (x_i - \mu_x) \{ (x_i - \mu_x)^2 - 3\sigma_x^2 \}. \\
\frac{\partial^2 \Delta}{\partial \lambda \partial \beta_1} &\overset{\lambda \to 0}{\to} -\frac{1}{2} h^2 \phi \left( \frac{x_i - \mu_x}{\sigma_x} \right) \Big[ \frac{1}{\sigma_x^4} \{ (x_i - \mu_x)^2 - \sigma_x^2 \} \\
&\quad -\frac{1}{\sigma^6} (x_i - \mu_x) x_i \{ (x_i - \mu_x)^2 - 3\sigma_x^2 \} \Big]. \\
\frac{\partial^2 \Delta}{\partial \lambda \partial \sigma^2} &\overset{\lambda \to 0}{\to} 0.
\end{aligned}$$

## BIBLIOGRAPHY


[1] Allan FG and Wishart, J (1930). A method of estimating the yield of a missing plot in field experiments. *J. Agric. Sci.*, **20**, 399–406.

[2] Brown CH (1990). Protecting against nonrandomly missing data in longitudinal studies, *Biometrics*, **46**, 143–155.

[3] Chen C (2001). Parametric models for response-biased sampling, *J. Roy. Statist. Soc. Ser. B*, **63**, 775–789.

[4] Cheng, PE (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, **89**, 81–87.

[5] Copas JB and Li HG (1997). Inference for Non-Random Sampling, *J. Roy. Statist. Soc. Ser. B*, **59**, 55–95.

[6] Copas JB and Eguchi S (2001). Local sensitivity approximations for selectivity bias, *J. Roy. Statist. Soc. Ser. B*, **63**, 871–895.

[7] Craven P and Wahba G (1979). Smoothing Noisy Data With Spline Functions, *Numerische Mathematik*, **31**, 377-403.

[8] Diggle P and Kenward MG(1994). Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.*, **43**, 49–73.

[9] Eubank RL (1988). *Nonparametric regression and spline smoothing*, 2nd ed, Marcel Dekker, Inc.

[10] Gasser T, Sroka L, and Jennen Steinmetz C (1986). Residual Variance and Residual Pattern in Nonlinear Regression, *Biometrika*, **73**, 625–633.

[11] Green PJ (1987). Penalized Likelihood for general semi-parametric regression models, *International statistical Review*, **55**, 245–259.

[12] Green PJ and Silverman BW (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall.

[13] Good IJ and R.A. Gaskins RA (1971). Nonparametric roughness penalties for probability densities, *Biometrika*, **58**, 255-277.

[14] Hahn J (1998). On the role of propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66**, 315-331.

[15] Huber PJ (1967). The behavior of maximum likelihood estimates under non-standard conditions. *In Fifth Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley: University of California Press, 221–233.

[16] Kenward MG and Molenberghs G (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, **13**, 236–47.

[17] Liang KY and Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

[18] Liang KY and Qin J (2000). regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *J. Roy. Statist. Soc. Ser. B*, **62**, 773–786.

[19] Little RJA (1993). Pattern-Mixture Models for Multivariate Incomplete Data, *J. Amer. Statist. Assoc.*, **88**, 125-134.

[20] Little RJA (1994). A class of pattern-mixture models for multivariate incomplete data. *Biometrika*, **81**, 471–483.

[21] Little, RJA and Rubin DB (2002). *Statistical Analysis with Missing Data*, Second edition. New York: Wiley.

[22] Ma G, Troxel AB and Heitjan DF (2005). An index of local sensitivity to nonignorable drop-out in longitudinal modelling, *Statist. Med.*, **24**, 2129–2150.

[23] Ma G, Troxel AB and Heitjan DF (2006). LETTER TO THE EDITOR: An index of local sensitivity to nonignorable drop-out in longitudinal modelling, *Statist. Med.*, **25**, 3217–3223.

[24] Molenberghs G, Kenward MG and Lesaffire E (1997). The Analysis of Longitudinal ordinal data with informative dropout. *Biometrika*, **84**, 33–44. bibitem Rice J (1986). Convergence rates for partially splined models, *Statistics & Probability Letters*, **4**, 203–208.

[25] Rollman BL, Belnap BH, Mazumdar S, Houck PR, Zhu F, Gardner W, Reynolds CF, Schulberg HC and Shear KM (2005). A randomized trial to improve the quality of treatment for panic and generalied anxiety disorders in primary care, *Archive of General Psychiatry*, **62**, 1332–1341.

[26] Rotinitzky A, Robins JM and Scharfstin DO (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse, *J. Amer. Statist. Assoc.*, **93**, 1321–1339.

[27] Rubin DB (1976). Inference and Missing Data, *Biometrika*, **63**, 581–592.

[28] Rubin DB (1987). Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

[29] Schluchter MD (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, **11**, 1861–1870.

[30] Silverman BW (1994). *Density Estimation for statistics and data analysis*, Chapman & Hall.

[31] Tang G, Little RJA and Raghunathan TE (2003). Analysis of multivariate missing data with nonignorable nonresponse, *Biometrika*, **90**, 747–764.

[32] Tang G, Little RJA and Raghunathan TE (2004). Analysis of Multivariate Monotone Missing Data by a Pseudolikelihood Method. *Proceedings of the 2nd. Seattle Symposium in Biostatistics: Analysis of Correlated Data. Lecture Notes in Statistics*. **179**, Ed.D Lin and P.J.Heagerty. New York: Springer Verlag.

[33] The R Development Core Team (2008). *R: A Language and Environment for Statistical Computing Reference Index Version 2.7.0*, R Foundation for Statistical Computing.

[34] Troxel AB, Ma G and Heitjan DF(2004). An index of local sensitivity to nonignorability, *Statistica Sinica*, **14**, 1221–1237.

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

[36] Wang QH and Rao JNK (2002). Empirical likelihood-based inference under imputation with missing response, *The Annuals of Statistics*, **30**, 896–924.

[37] Yates F (1933). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *J. Amer. Statist. Assoc.*, **57**, 348–368.