Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian local influence analysis of general estimating equations with nonignorable missing data

Yan-Qing Zhang, Nian-Sheng Tang *

Department of Statistics, Yunnan University, Kunming 650091, PR China

HIGHLIGHTS

- Propose a Bayesian local influence to assess the effect of various perturbations.
- Introduce a perturbation model to simultaneously characterize various perturbations.
- Construct a Bayesian perturbation manifold to characterize the structure of various perturbations.
- Develop the adjusted local influence measures to quantify the effect of various perturbations.
- Present goodness-of-fit statistics to assess the plausibility of the posited EEs.

ARTICLE INFO

Article history:
Received 31 December 2015
Received in revised form 5 August 2016
Accepted 12 August 2016
Available online xxxx

Keywords:
Bayesian empirical likelihood
Bayesian local influence
Estimating equations
Goodness-of-fit
Nonresponse instrumental variable
Nonignorable missing data

ABSTRACT

Bayesian empirical likelihood (BEL) method with missing data depends heavily on the prior specification and missing data mechanism assumptions. It is well known that the resulting Bayesian estimations and tests may be sensitive to these assumptions and observations. To this end, a Bayesian local influence procedure is proposed to assess the effect of various perturbations to the individual observations, priors, estimating equations (EEs) and missing data mechanism in general EEs with nonignorable missing data. A perturbation model is introduced to simultaneously characterize various perturbations, and a Bayesian perturbation manifold is constructed to characterize the intrinsic structure of these perturbations. The first- and second-order adjusted local influence measures are developed to quantify the effect of various perturbations. The proposed methods are adopted to systematically investigate the tenability of nonignorable missing mechanism assumption, the sensitivity of the choice of the nonresponse instrumental variable and the sensitivity of EEs assumption, and goodness-of-fit statistics are presented to assess the plausibility of the posited EEs. Simulation studies are conducted to investigate the performance of the proposed methodologies. An example is analyzed.

1. Introduction

Missing data are commonly encountered in various fields including public health, medicine, economics and social sciences. In missing data analysis, although missing at random (MAR) mechanism that the probability of missing data only depends upon the observed data but not the unobserved data (Little and Rubin, 2002) is a widely used assumption, this assumption may be questionable in some situations. For example, in the surveys about income or the history of committing, the nonresponse rates tend to be related to the values of nonresponses. In this case, missing not at random (MNAR, also

* Corresponding author. Fax: +86 871 65033700.
E-mail address: nstang@ynu.edu.cn (N.-S. Tang).

http://dx.doi.org/10.1016/j.csda.2016.08.010
0167-9473/

Please cite this article in press as: Zhang, Y.-Q., Tang, N.-S., Bayesian local influence analysis of general estimating equations with nonignorable missing data. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.08.010

referred to as nonignorable missingness) assumption may be more reasonable than the classical MAR assumption. There has been many literature on MNAR data analysis in recent years. For example, Lee and Tang (2006) proposed a Bayesian approach to analyze nonlinear structural equation models with MNAR data; (Kim and Yu, 2011) proposed an exponential tilting model and developed a semiparametric estimation method of mean functionals with nonignorable missing data; Zhao et al. (2013b) developed an empirical likelihood (EL) approach to inference on mean functionals with nonignorable missing response data; Zhao et al. (2013a) proposed a nonparametric/semiparametric estimation method and an augmented inverse probability weighted imputation approach to estimate distribution function and quantiles of a response variable with nonignorable missing data; Tang et al. (2014) developed an EL approach to make inference on generalized estimating equations (EEs) based on the exponential tilting model for the response probability; Linero and Daniels (2015) developed a Bayesian nonparametric model for a longitudinal response in the presence of MNAR data; Jiang et al. (2016) proposed a composite quantile regression procedure for parameter estimation of linear regression model with nonignorable missing covariates. The above mentioned works are mainly developed on the basis of only Bayesian method or only EL method.

There is considerable interest in the development of EL approach under Bayesian paradigm with MNAR nonresponse, that is, Bayesian empirical likelihood (BEL) inference (Lazar, 2003). Recently, many works have been done on BEL (e.g. Fang and Mukerjee, 2006; Chang and Mukerjee, 2008; Grendár and Judge, 2009). It has been shown that the BEL method may improve the efficiency of parameter estimation using informative priors and help the search of the maximum EL estimates (Yang and He, 2012). However, because the BEL method heavily depends on the prior, the resulting Bayesian estimations and tests may be sensitive to the prior assumption and observations. Hence, it is interesting to develop an appropriate approach to conduct the sensitivity analysis for the BEL method.

Sensitivity analysis based on a general Bayesian approach has been developed in recent years. For example, Zhu et al. (2011) developed a Bayesian local influence approach in a class of parametric models from a viewpoint of differential geometry; Kaciroti and Raghunathan (2014) proposed a Bayesian sensitivity analysis of incomplete data in pattern-mixture and selection models; Zhu et al. (2014) proposed a geometric framework of Bayesian influence analysis to assess various perturbation schemes with missing data. However, there is little work done on developing a novel Bayesian local influence analysis in general EEs with nonignorable missing data because general EEs have been widely used to deal with missing data due to their robustness to model misspecification (Robins et al., 1994; Lipsitz et al., 1999).

We propose a BEL function in general EEs by incorporating prior distributions of parameters and missing data mechanism models, and explore Bayesian estimations of parameters via Markov chain Monte Carlo (MCMC) algorithm, which avoids the difficulties for directly maximizing EL function. We develop a Bayesian diagnostic procedure in a differential geometric framework, called Bayesian perturbation manifold, and establish the first- and second-order adjusted influence measures to assess the effect of various perturbations, including perturbations to EEs, priors of parameters, individuals and missing data mechanism models. In addition, some goodness-of-fit statistics are presented to assess the plausibility of the posited EEs.

The rest of this paper is organized as follows. In Section 2, we develop a BEL estimation approach under MNAR assumption. Section 3 proposes several Bayesian local influence measures to assess various perturbations. Section 4 constructs goodness-of-fit statistics to measure the plausibility of the posited EEs. Section 5 presents simulation studies and an example to illustrate the proposed methods. Some concluding remarks are given in Section 6. The technical details are presented in the Appendix A and Supplementary materials (see Appendix B).

2. Bayesian empirical likelihood with MNAR data

Consider a data set $\{(Z_i^\top, Y_i^\top)^\top, i = 1, \ldots, n\}$ with n individuals, where Z_i is a $d_z \times 1$ observable vector and Y_i is a $d_y \times 1$ vector subject to missingness. Suppose that $(Z_i^\top, Y_i^\top)^\top$'s are independent and identically distributed (i.i.d.) as an unknown distribution $F(z, y)$. Without assuming a specific form of $F(z, y)$, our main interest is to make inference on unknown parameters $\theta \in \mathcal{R}^p$ based on q ($q \geqslant p$) functionally independent EEs: $\psi(Y_i, Z_i; \theta) = (\psi_1(Y_i, Z_i; \theta), \ldots, \psi_q(Y_i, Z_i; \theta))^\top$. The EEs satisfy the following unconditional moment restrictions $E_F\{\psi(Y_i, Z_i; \theta_0)\} = 0$, where θ_0 is the true value of θ and E_F is the expectation taken with respect to the unknown distribution $F(z, y)$. For missing data, we define $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ otherwise. For simplicity, we decompose $Z_i = (X_i^\top, U_i^\top)^\top$ in which the dimension of X_i is denoted as d_x, the dimension of U_i is greater than or equal to one, and U_i is related with Y_i. We assume that δ_i is independent of δ_j for any $i \neq j$, and δ_i depends on Y_i and X_i such that $\Pr(\delta_i = 1|Z_i, Y_i) = \Pr(\delta_i = 1|X_i, Y_i)$. That is, the missing data mechanism is MNAR (Little and Rubin, 2002). We consider the following parametric model for the respondent indicator:

$$\Pr(\delta_i = 1|X_i, Y_i; \gamma) = \pi(g_1(X_i; \gamma^{(1)}) + g_2(X_i, Y_i; \gamma^{(2)})) := \pi_i(\gamma), \quad (1)$$

where $\gamma = (\gamma^{(1)\top}, \gamma^{(2)\top})^\top$ is a v-dimensional unknown parameter vector, $\pi(\cdot)$, $g_1(\cdot)$ and $g_2(\cdot)$ are arbitrary user-specified known functions, and $g_2(\cdot)$ satisfies $g_2(X, Y; 0) = 0$. The model (1) implies that the indicator δ does not depend on variable U conditional on X and Y. That is, variable U is regarded as a nonrespondent instrumental variable. Thus, the above considered parametric missing data mechanism model is identifiable (Wang et al., 2014). If $\gamma^{(2)} = 0$, the model (1) reduces to a MAR mechanism model, that is, $\gamma^{(2)}$ measures the amount of departure from MAR assumption. If $\pi_i(\gamma)$ are known (i.e., γ is