Journal of Multivariate Analysis 142 (2015) 117–132

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semi-parametric rank regression with missing responses

Huybrechts F. Bindele[a,*], Ash Abebe[b]

[a] 411 University Blvd. N, ILB 316, Department of Mathematics and Statistics, University of South Alabama,
Mobile AL 36688-0002, United States

[b] 221 Parker Hall, Department of Mathematics and Statistics, Auburn University, AL 36849, United States

ARTICLE INFO

Article history:
Received 4 September 2014
Available online 28 August 2015

AMS 2000 subject classifications:
primary 62J02
62G05
secondary 62F12
62G20

Keywords:
Wilcoxon estimator
Strong consistency
Asymptotic normality
Missing at random
Inverse probability
Simple imputation

ABSTRACT

We consider a semi-parametric regression model with responses missing at random and study the rank estimator of the regression coefficient. Consistency and asymptotic normality of the proposed estimator are established. Monte Carlo simulation experiments show that the proposed estimator is more efficient than the least squares estimator whenever the error distribution is heavy tailed or contaminated. When the errors follow a normal distribution, these simulation experiments show that the rank estimator can be more efficient than its least squares counterpart for cases with large proportion of missing responses.

1. Introduction

In recent decades, the problem of missing data has garnered a lot of attention within the statistical community. Responses may be missing for a number of common reasons. These include equipment malfunction, contamination of samples, manufacturing defects, drop out in clinical trials, weather conditions, incorrect data entry, etc. In this paper, we consider missing responses in the context of regression analysis. We will do so under the most commonly used missing data mechanism assumption which asserts that the responses are missing at random (MAR) as discussed in [23]. Under MAR, the probability that a response variable is observed can depend only on the values of those other variables that have been observed.

Consider the linear semi-parametric regression model

$$Y_i = \mathbf{X}_i^\tau \boldsymbol{\beta} + g(T_i) + \varepsilon_i, \quad 1 \le i \le n, \tag{1.1}$$

where $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^p$ is a vector of parameters, $\mathbf{X}_i$'s are i.i.d. p-variable random covariate vectors, T_i's are i.i.d. univariable random covariates defined on $[0, 1]$, the function $g : [0, 1] \to \mathbb{R}$ is unknown, and the model errors ε_i are independent with conditional mean zero given the covariates. Also, $E(\varepsilon_i^2|\mathbf{Z}_i) > 0$ with $\mathbf{Z}_i = (\mathbf{X}_i, T_i)$. In this paper, we are interested in inferences about the true value $\boldsymbol{\beta}_0$ of the parameter $\boldsymbol{\beta}$, when there are missing responses in the linear semi-parametric model (1.1).

[*] Corresponding author.
E-mail address: hbindele@southalabama.edu (H.F. Bindele).

http://dx.doi.org/10.1016/j.jmva.2015.08.007
0047-259X/© 2015 Elsevier Inc. All rights reserved.

For data without missingness, the partial linear semiparametric model given in (1.1) has been used to study a number of real life problems. For instance, an application of (1.1) to a mouthwash experiment was given by Speckman [24] where the model was estimated using kernel smoothing. A version of (1.1) set up as semi-parametric mixed model was used by Zeger and Diggle [35] for analyzing the CD4 cell count in HIV seroconverters where they used back-fitting along with cross-validation to estimate the model. A marketing price–volume example is studied in [6] using a penalized least squares approach. Model (1.1) has also been applied in several fields such as biometrics [5] and econometrics [14]. Other notable works include [9,20,22] among others.

Wang and Sun [28] studied the least squares estimator of the regression coefficient $\boldsymbol{\beta}$ in model (1.1) under the MAR assumption. Considering the same setting, Wang et al. [26] developed inference tools in missing response case for the mean of Y based on the least squares estimation approach and under the MAR assumption. One method for constructing confidence intervals for the true mean of Y is the empirical likelihood method introduced by Owen [18]. This approach is used to investigate a variety of statistical problems by Hall and La Scala [7], Chen and Hall [2], Kitamura [13], and Peng [19], Xue and Zhu [32], Xue and Zhu [33], Xue and Zhu [34] to mention a few. These works demonstrate that the method of empirical likelihood has a number of advantages over methods such as those based on normal approximations or the bootstrap. Studies that used this approach to study (1.1) under the MAR assumption include [27,26,25,30].

When dealing with missing data, the main approach is to impute a plausible value for each missing datum and then analyze the results as if they were complete. In most of the regression problems, the commonly used approaches include linear regression imputation [8], nonparametric kernel regression imputation [3,27], and semi-parametric regression imputation [28]. An alternative approach for handling missing data is the inverse probability weighting. This approach has gained considerable attention as a way to deal with missing data problems. For a discussion of this approach, see [21,36,29,26] and references therein. As pointed out by Wang and Sun [28], for missing problems, the inverse probability weighting approach usually depends on high dimensional smoothing for estimating the completely unknown propensity score function. This suffers from the *curse of dimensionality* that may restrict the use of the resulting estimator. One way to avoid such a problem is to use the inverse marginal probability weighted method proposed by Wang et al. [26].

In this paper, we study the rank estimator of $\boldsymbol{\beta}$ in model (1.1) with MAR responses in an effort to mitigate the adverse effects of heavy tails and gross outliers on the least squares estimator of $\boldsymbol{\beta}$. To that end, $\boldsymbol{\beta}$ will be defined as the minimizer of the general rank dispersion function proposed by Jaeckel [11], where the missing responses are imputed either using simple imputation or inverse marginal probability weighting and the function g is estimated using kernel smoothing.

The paper is organized as follows. We define our estimator and give some preliminary results in Section 2. The asymptotic normality of the proposed estimator is established in Section 3. Section 4 addresses some practical issues encountered in the estimation process including the estimation of g and standard errors of estimated coefficients. Section 5 gives a simulation study and real data examples to illustrate the use of the proposed estimator. Assumptions used in our development as well as sketch of proofs of our results are given in the Appendix.

2. Rank estimator

In model (1.1), consider the case where some values of Y in the sample of size n may be missing, but $\mathbf{X}$ and T are fully observed. That is, we obtain the following incomplete observations

$$(Y_i, \delta_i, \mathbf{X}_i, T_i), \quad i = 1, 2, \ldots, n$$

from (1.1), where $\mathbf{X}_i$'s and T_i's are observed, and

$$\delta_i = \begin{cases} 0, & \text{if } Y_i \text{ is missing,} \\ 1, & \text{otherwise.} \end{cases}$$

As discussed above, we assume that Y is missing at random (MAR). The MAR assumption implies that δ and Y are conditionally independent given $\mathbf{X}$ and T, that is, $P(\delta = 1|Y, \mathbf{Z}) = P(\delta = 1|\mathbf{Z})$, where $\mathbf{Z} = (\mathbf{X}, T)$ as defined above. Please see [16] for an in-depth discussion regarding the MAR assumption.

Denote $\Delta(\mathbf{z}) = P(\delta = 1|\mathbf{Z} = \mathbf{z})$, $\sigma^2(\mathbf{Z}) = E(\varepsilon^2|\mathbf{Z})$, and $\Gamma(t) = P(\delta = 1|T = t)$. Now consider the rank objective function proposed by Jaeckel [11]

$$D_n^C(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \varphi\Big(\frac{R(e_i(\boldsymbol{\beta}))}{n+1}\Big) e_i(\boldsymbol{\beta}), \tag{2.1}$$

where $e_i(\boldsymbol{\beta}) = \delta_i\varepsilon_i$ and $R(e_i(\boldsymbol{\beta}))$ is the rank of $e_i(\boldsymbol{\beta})$ among $e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})$. Due to the MAR assumption $E[e_i(\boldsymbol{\beta})|\mathbf{Z}_i] = 0$.

Note that in the expression of $D_n^C(\boldsymbol{\beta})$ in (2.1), $\boldsymbol{\beta}$ and g are unknown. So, before dealing with the estimation of $\boldsymbol{\beta}$, let us first consider the estimation of g based on the completely observed data; that is, estimating g as a known function of t but unknown with respect to $\boldsymbol{\beta}$. As discussed in [26], pre-multiplying (1.1) by the observation indicator and taking conditional expectation given $T = t$, we have

$$E[\delta_i Y_i|T_i = t] = E[\delta_i \mathbf{X}_i|T_i = t]\beta + E[\delta_i|T_i = t]g(t).$$