Journal of Multivariate Analysis 129 (2014) 193–205

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Jackknife empirical likelihood inference with regression imputation and survey data

Ping-Shou Zhong [a,*], Sixia Chen [b]

[a] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA
[b] Westat, 1650 Research Blvd, Rockville, MD 20850, USA

ARTICLE INFO

Article history:
Received 10 September 2013
Available online 9 May 2014

AMS subject classifications:
62G10
62G20
62D05

Keywords:
Kernel smoothing
Missing at random
Nonignorable missing
Response mechanism
Wilks' theorem

ABSTRACT

We propose jackknife empirical likelihood (EL) methods for constructing confidence intervals of mean with regression imputation that allows ignorable or nonignorable missingness. The confidence interval is constructed based on the adjusted jackknife pseudo-values (Rao and Shao, 1992). The proposed EL ratios evaluated at the true value converge to the standard chi-square distribution under both missing mechanisms for simple random sampling. Thus the EL can be applied to construct a Wilks type confidence interval without any secondary estimation. We then extend the proposed method to accommodate Poisson sampling design in survey sampling. The proposed methods are compared with some existing methods in simulation studies. We also apply the proposed method to an Italy household income panel survey data set.

1. Introduction

Missing data appear very often in social science, survey sampling, and many other fields. A common practice is removing data with missing values and conducting statistical inference based on only the complete observations. However, simply ignoring missing values might lead to inefficient and biased inference [25,23]. Imputation is a commonly used approach to missing data by first creating plausible values for the missing observations and then conducting the statistical inference on the imputed data. Some commonly used imputation methods include linear regression imputation [44]; multiple imputation [34]; kernel regression imputation [8,41]; nearest neighborhood imputation (NNI) [29]; ratio imputation [30]; hot deck imputation [14]; and fractional imputation [19,11,21,20] among others.

Empirical likelihood (EL) [27] is a nonparametric method that has been used in statistical inference for data with missing values. Its advantages have been well documented in many papers (e.g., [41,38,7,28,37]). The EL-based confidence interval has a natural shape and orientation determined by data. Moreover, the EL method enjoys some nice properties of a parametric likelihood, for example, the Wilks theorem [26,27] and Bartlett Correctable [9,2]. See [6] for an overview. Wang and Rao [41] proposed an EL method for constructing confidence intervals for the mean functionals after kernel regression imputation under a missing at random (MAR) assumption in the sense of [33]. Wang and Chen [38] generalized the result to estimators defined by estimating equations after multiple imputation. An attraction of the standard EL is the internal studentised ability which avoids the explicit estimation of the variance. However, all the existing standard EL ratios (e.g., [41,38]) for imputed data converge to a scaled chi-square distribution instead of the standard chi-square distribution

* Correspondence to: C418 Wells Hall, 619 red cedar road, East Lansing, MI 48824, USA.
E-mail addresses: pszhong@stt.msu.edu (P.-S. Zhong), SixiaChen@westat.com (S. Chen).

http://dx.doi.org/10.1016/j.jmva.2014.04.010
0047-259X/

where the scale factor depends on the unknown variances. Hence, we need to estimate the unknown scale factor before applying their proposed methods.

This paper employs a novel jackknife empirical likelihood (JEL) method. JEL was proposed by [18] to solve nonlinear constraint problems in the standard EL formulation. JEL has been applied to the inference for the difference in ROC curves [43], case-control study [17] and testing for high dimensional means [40]. However, no existing studies of JEL could be used for inferences with imputed values and accommodating sampling designs. The paper tries to answer one question: how to construct JEL, which allows imputed values and considers sampling designs, so that the JEL ratios are still able to achieve asymptotic chi-square? We take this opportunity to study the JEL inference for the mean of data with kernel smoothing regression imputation [8] under ignorable missingness, and with semiparametric imputation under nonignorable missingness [24]. Moreover, we consider two sampling designs: simple random sample and Poisson sampling design. The proposed JEL is constructed through adjusted jackknife pseudo-values [31]. We show that the proposed JEL ratios converge to a standard chi-square distribution under ignorable or non-ignorable missingness, with simple random sampling or Poisson sampling designs. The advantages of the proposed JEL include the following: first, it maintains the good properties of the standard EL, and it is easier to implement since the proposed methods are asymptotically pivotal and no secondary estimation is needed; second, it accommodates ignorable and non-ignorable missing, together with simple random sampling or Poisson sampling. Comparing to the EL and normal approximation (NA) methods, where derivation and computation of the asymptotic variances is needed case by case, the proposed method is much easier and practical.

The paper is organized as follows. In Section 2, we introduce the basic concept of missing mechanisms. Asymptotic results on JEL estimators and JEL ratios for mean estimators with regression imputation under ignorable or nonignorable missing mechanisms are presented in Section 3. The extension of the proposed method to Poisson sampling is given in Section 4. In Section 5, we demonstrate the proposed method by simulation studies. We also apply the proposed methods to Italy Household Income Panel Survey (IHIPS) data in Section 6. All the technical proofs are relegated to a supplemental paper [45] (see Appendix B).

2. Basic setup

Let (X_i, Y_i) $(i = 1, 2, \ldots, n)$ be a set of independent and identically distributed (IID) random vectors from an infinite population $\mathcal{F}$ where Y_i is a scalar response and X_i is a d-dimensional random vector. We assume that Y_i may be subject to missingness, but X_i is always observed. Let r_i be nonmissing response indicator such that $r_i = 1$ if Y_i is observed, and $r_i = 0$ if Y_i is missing.

The missing data mechanism is ignorable or missing at random (MAR) [33] if

$$\mathrm{pr}(r_i = 1|X_i, Y_i) = \mathrm{pr}(r_i = 1|X_i) = \pi(X_i). \tag{1}$$

That is, the conditional missing probability depends only on the observed data. The missingness is called nonignorable or not MAR (NMAR) if (1) does not hold. That is, the conditional missing probability may also depend on the unobserved data Y_i. In this paper, we assume a semi-parametric response probability model [24] when missingness is NMAR. Namely,

$$\mathrm{pr}(r_i = 1|X_i, Y_i) = \pi(X_i, Y_i) = \frac{\exp\{g(X_i) - \gamma Y_i\}}{1 + \exp\{g(X_i) - \gamma Y_i\}}, \tag{2}$$

for a completely unspecified function $g(\cdot)$ and parameter γ. Note that, when $\gamma = 0$, the NMAR assumption reduces to MAR.

The parameter of interest is $\theta_0 = E(Y)$. A consistent estimator of θ_0 under either ignorable or nonignorable missing is

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i Y_i + (1 - r_i)\hat{m}_0(X_i; \hat{\gamma})\right\}, \tag{3}$$

where $\hat{m}_0(X_i; \hat{\gamma})$ is a consistent estimator of $m_0(X_i; \gamma) = E(Y_i|X_i, r_i = 0)$. The estimation of $m_0(X_i; \gamma)$ depends on a missing mechanism, which is introduced below separately under both missing mechanisms.

When the missing data are MAR, $m_0(X_i; \gamma)$ is independent of γ and $m_0(X_i; \gamma) = m(X_i)$ where $m(X_i) = E(Y_i|X_i)$. Cheng [8] proposed a Nadaraya–Watson (NW) estimator $\hat{m}(X_i)$ to estimate $m(X_i)$ where $\hat{m}(X_i)$ is

$$\hat{m}(X_i) = \frac{\sum_{j\neq i}^{n} r_j Y_j K_h(X_i, X_j)}{\sum_{j\neq i}^{n} r_j K_h(X_i, X_j)}, \tag{4}$$

with $K_h(x, y) = h^{-d}K((x - y)/h)$ where K is a kernel function and h is the bandwidth.

When the missing data are NMAR, by using the following relationship between conditional densities $f(y_i|x_i, r_i = 0)$ and $f(y_i|x_i, r_i = 1)$ [24],

$$f(y_i|x_i, r_i = 0) = f(y_i|x_i, r_i = 1)\frac{O(x_i, y_i)}{E\{O(x_i, Y_i)|x_i, r_i = 1\}},$$