Journal of Statistical Planning and Inference 141 (2011) 711–724

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Distribution estimation with auxiliary information for missing data

Xu Liu [a,b,*], Peixin Liu [a], Yong Zhou [a,c]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China
[b] Department of Statistics, Yunnan University, Kunming 650091, PR China
[c] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, PR China

ARTICLE INFO

Article history:
Received 30 December 2009
Received in revised form
24 July 2010
Accepted 25 July 2010
Available online 3 August 2010

Keywords:
Auxiliary information
Empirical distribution function
Empirical likelihood
Estimating equations
Kernel regression
Missing data
Quantile estimation
Semi-parametric imputation

ABSTRACT

There is much literature on statistical inference for distribution under missing data, but surprisingly very little previous attention has been paid to missing data in the context of estimating distribution with auxiliary information. In this article, the auxiliary information with missing data is proposed. We use Zhou, Wan and Wang's method (2008) to mitigate the effects of missing data through a reformulation of the estimating equations, imputed through a semi-parametric procedure. Whence we can estimate distribution and the τth quantile of the distribution by taking auxiliary information into account. Asymptotic properties of the distribution estimator and corresponding sample quantile are derived and analyzed. The distribution estimators based on our method are found to significantly outperform the corresponding estimators without auxiliary information. Some simulation studies are conducted to illustrate the finite sample performance of the proposed estimators.

1. Introduction

Estimation of distribution is very important in order to make statistical inference for parameters or functions of parameters based on distribution. The well-know estimator of distribution is the empirical distribution that is consistent and asymptotic normal. Some stronger consequences imply the empirical distribution converges uniformly to the true distribution, and the empirical process approximates to a standard Wiener process. The empirical distribution can be used to construct a Kolmogorov–Smirnov test for null hypothesis distribution is known F_0. It is known that the empirical distribution is a non-parametric maximum estimator of unknown distribution without any auxiliary information. However, in practice, some auxiliary information can often be obtained, such as, unknown distribution is symmetric or the variance of population is a function of mean. The first motivation in this paper is to take into account auxiliary information, and then improves efficiency of distribution estimation (Qin and Lawless, 1994). An interesting question is how to improve efficiency of distribution estimator when the unknown distribution is symmetric or has a mean zero or variance is a function of mean. The question of estimation for symmetric distribution of Y can be viewed as estimating equation estimation. The unbias estimating function $\psi(Y,\mu)$ can be described as follows:

$$E[\psi(Y,\mu)] = 0,$$

* Corresponding author at: Department of Statistics, Yunnan University, Kunming 650091, PR China.
E-mail address: liuxuxm@gmail.com (X. Liu).

0378-3758/$ - see front matter Crown Copyright © 2010 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.07.015

where

$$\psi(Y,\mu)=\begin{pmatrix} Y-\mu \\ (Y-\mu)^3 \end{pmatrix} \tag{1.1}$$

or

$$\psi(Y,\mu)=\begin{pmatrix} Y-\mu \\ 1/2-I(Y\le\mu) \end{pmatrix}, \tag{1.2}$$

where $I(A)$ is an indicator function taking on the value of 1 if event A occurs and 0 otherwise. The parameter of primary interest is distribution F, but mean μ is a nuisance parameter. Note that the second function in (1.2) is discontinuous. Similarly, when variance is a function of mean, the estimating function can be

$$\psi(Y,\mu)=\begin{pmatrix} Y-\mu \\ (Y-\mu)^2-g(\mu) \end{pmatrix}, \tag{1.3}$$

where $g(\mu)=\mathrm{Var}(Y)$, and g is a known function.

In incomplete data sets, estimation of distribution function is also arguably important on statistical inference. There are many important distributions appearing in literature, for example, Kaplan and Meier (1958) gave a well-known Kaplan–Meier estimator for censored data; Lynden-Bell (1971) driven a product-limit estimator for truncated data; Tsai et al. (1987) obtained an estimator of distribution for truncated and censored data; Groeneboom and Wellner (1992) described a non-parametrical maximum likelihood estimator of distribution for interval censored data. More estimators for different incomplete data, see Turnbull (1976), Tsai and Crowley (1985), Reiss (1981) and Qin and Lawless (1994). Cheng and Chu (1996) suggested an estimator of distribution based on non-parametric kernel regression with missing data. But surprisingly very little previous attention has been paid to missing data in the context of estimating distribution with auxiliary information. In practice, it is often possible that we may have more auxiliary information which is unavailable under the assumptions of Cheng and Chu (1996). By using the auxiliary information, as we expect, we can increase the efficiency of the resulting estimator. From (1.1) and (1.3), auxiliary information can be expressed some estimating functions. Assume that we have auxiliary information that is a set of unbiased estimation functions $\psi(y,z,\theta)=(\psi_1(y,z,\theta),\ldots,\psi_q(y,z,\theta))^T$ which satisfy the moment restrictions of the form

$$E\psi(Y,Z,\theta)=0, \tag{1.4}$$

where Y is an i.i.d. response variable with unknown distribution function F and covariate Z, θ is a p-dimensional unknown parameter vector and $q\ge p$. Alternatively, we can consider the auxiliary information that is a set of unbiased estimation equations $\psi(y,z)=(\psi_1(y,z),\ldots,\psi_q(y,z))^T$ with the moment restrictions of the form $E\psi(Y,Z)=0$, in which we remove the parameter θ, that is, we have known the true value of parameter θ. For example, if we know a symmetric distribution F with mean zero, then we know that median of F is zero. Therefore, we have $E\psi(Y,0)=0$ for ψ in (1.2). In this paper, we focus mainly on estimating distribution function F of response variable Y with above auxiliary information both including the unknown parameter θ and without parameter, where Y may be missing. The methodology of analyzing data with missing is a very common issue today. We begin our analysis by considering a random sample of incomplete data:

$$(Y_i,Z_i,\delta_i),\quad i=1,2,\ldots,n,$$

where Z_i's are observed covariate of dimension d, Y_i can be observed if $\delta_i=1$, otherwise $\delta_i=0$ and Y_i is missing. Let X_i be Z_i or a sub-set of Z_i.

We focus on the case where the data are missing at random (MAR), i.e., the missing mechanism is independent of the unobserved data and is ignorable, which is the most commonly adopted baseline of analysis in the missing data literature. See for example, Cheng (1994), Chu and Cheng (1995), Cheng and Chu (1996), among others. The assumption of MAR implies that δ and Y are conditionally independent given X, i.e.,

$$P(\delta=1|X,Y)=P(\delta=1|X)=P(X),$$

or in other words, given the observed data, the missing mechanism does not depend on the unobserved data. The MAR assumption is practically justified in many situations (see Little and Rubin, 1987). All our theoretical results in fact hold also for the situation of the covariate's data being missing, see Zhou et al. (2008).

There are many methods to deal with missing data, such as EM algorithm (Dempster et al., 1977), inverse probability-weighted approach (Robins et al., 1994) and imputation scheme (Yate, 1933; Bartlett, 1937; Healy and Westmacott, 1956). Of particular relevance here is the kernel non-parametric imputation method discussed in Cheng (1994) and Cheng and Chu (1996). Wang and Rao (2002) extended this work by considering the same imputation scheme in conjunction with the empirical likelihood (EL) approach in making inference for the unknown mean. In recent decades, the method of EL has taken much of the spotlight in the statistical field since it was introduced by Owen (1988). This non-parametric method of inference has sampling properties similar to the bootstrap. It has been discussed by Owen (1990), Qin and Lawless (1994, 1995), among others. Zhou et al. (2008) combined the estimating equations (EE) and EL theory together with missing data