Statistics and Probability Letters 123 (2017) 139–145

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Covariate-balancing-propensity-score-based inference for linear models with missing responses

Donglin Guo [a,b,*], Liugen Xue [a], Yuqin Hu [a,c]

[a] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China
[b] School of Mathematics and Information Science, Shangqiu Normal University, Shangqiu 476000, China
[c] School of data science, Zhejiang University of Finance and Economics, Hangzhou 310018, China

ARTICLE INFO

Article history:
Received 1 July 2016
Received in revised form 25 November 2016
Accepted 2 December 2016
Available online 16 December 2016

Keywords:
Linear model
Missing at random
Covariate balancing propensity score
GMM
Augmented inverse probability weighted
Robust estimation

ABSTRACT

Based on covariate balancing propensity score (CBPS), improved estimators for the regression coefficients and population mean of linear models are obtained, when the responses are missing at random. It is proved that the proposed estimators are asymptotically normal.

1. Introduction

Missing data is frequently encountered in statistical studies, and ignoring it could lead to biased estimation and misleading conclusions. Inverse probability weighting (Horvitz and Thompson, 1952) and imputation are two main methods for dealing with missing data. Since Scharfstein et al. (1999) noted that the augmented inverse probability weighted (AIPW) estimator in Robins and Rotnitzky (1994) was double-robust, authors have proposed many estimators with the double-robust property, see Tan (2006), Kang and Schafer (2007) and Cao et al. (2009). The estimator is doubly robust in the sense that consistent estimation can be obtained if either the outcome regression model or the propensity score model is correctly specified. In this paper, we make use of AIPW method to consider the linear model:

$$Y = X^T\beta + \nu_0(X)\varepsilon, \tag{1}$$

where Y is a scalar response variate, β is a $p \times 1$ vector of unknown regression parameter, $\nu_0(\cdot)$ is a strictly positive known function and ε is a random statistical error with $E[\varepsilon|X] = 0$. Throughout this paper, we assume that $X's$ are observed completely, Y is missing at random (Rubin, 1976). Thus, the data actually observed are independent and identically distributed $(\delta_i Y_i, \delta_i, X_i)(i = 1, \ldots, n)$, where $\delta_i = 1$ indicates that Y_i is observed and $\delta_i = 0$ indicates that Y_i is missing. The missing at random (MAR) assumption implies that δ and Y are conditionally independent given X, that is, $P(\delta = 1|X, Y) = P(\delta = 1|X) \equiv \pi(X)$. This probability is called the propensity score (Rosenbaum and Rubin, 1983).

* Corresponding author at: College of Applied Sciences, Beijing University of Technology, Beijing 100124, China.
E-mail address: gdl1105@emails.bjut.edu.cn (D. Guo).

http://dx.doi.org/10.1016/j.spl.2016.12.001
0167-7152/

The linear models with missing data have been studied in existing papers, such as Wang and Rao (2002), Xue (2009), Qin and Lei (2010) and so on. The inverse probability weighted imputation methods of Xue (2009) and other papers are based on the nonparametric estimators of the propensity score model. However, it is well known that it is difficult to obtain the nonparametric estimators because of the "curse of dimensionality". In addition, as mentioned in Kang and Schafer (2007), although the AIPW estimators are doubly robust, they can be biased when both models are misspecified.

In this paper, we construct estimators for β and μ, based on the covariate balancing propensity score (CBPS) method proposed by Imai and Ratkovic (2014), in which they made use of CBPS to estimate the average treatment in causal inference setting. Because the estimation of the average treatment effect can be treated as a two-sample missing data problem, we borrow the idea of Imai and Ratkovic (2014) to study the problems of missing data. To the best of our knowledge, in the case of missing data, there is no research based on CBPS. As mentioned in Imai and Ratkovic (2015), the weights based on CBPS are robust in the sense that they improve covariate balance even when propensity score model is misspecified. So our estimators based on CBPS are improved due to the robust weights and our method has the following merits: (1) it avoids the "curse of dimensionality" and selection of optimal bandwidth; (2) it improves performance of the usual AIPW estimators in terms of bias, standard deviation (SD) and mean-squared error (MSE), especially when both outcome regression model and propensity score model are misspecified.

The rest of this paper is organized as follows. In Section 2, based on CBPS and AIPW methods, the estimators for the parameter β and the population mean μ are proposed, and the asymptotic properties of the estimators are investigated. In Section 3, simulation studies are carried out to assess the performance of the proposed method. In Section 4, concluding remarks are made. In the Appendix, the proofs of the main results are given.

2. Construction of estimators

In this paper, we adopt the most popular choice of $\pi(X)$ and posit a logistic regression model for it:

$$\pi(X) = \frac{\exp(X^T\alpha)}{1+\exp(X^T\alpha)}, \tag{2}$$

where $\alpha \in \Theta$ is P-dimensional unknown column vector parameter.

2.1. CBPS-based estimator for the propensity score

It is necessary to estimate α of the propensity score before we construct estimators for β and μ. Based on (δ_i, X_i) $(i = 1, \ldots, n)$, the usual method estimates α by the maximum binomial likelihood estimator $\hat{\alpha}$ which maximizes the log-likelihood function:

$$L = \sum_{i=1}^{n} [\delta_i \log\{\pi(X_i, \alpha)\} + (1-\delta_i)\log\{1-\pi(X_i, \alpha)\}]. \tag{3}$$

Assuming that $\pi(X, \alpha)$ is twice continuously differentiable with respect to α, so maximizing (3) implies the first-order condition

$$\frac{1}{n}\sum_{i=1}^{n} s(\delta_i, X_i, \alpha) = 0, \quad s(\delta_i, X_i, \alpha) = \frac{(\delta_i - \pi(X_i, \alpha))\pi'(X_i, \alpha)}{\pi(X_i, \alpha)(1-\pi(X_i, \alpha))}, \tag{4}$$

where $\pi'(X_i, \alpha) = \partial\pi(X_i, \alpha)/\partial\alpha^T$. However, the main drawback of this standard method is that $\pi(X)$ may be misspecified, yielding biased estimators for the parameters β and μ. To overcome the drawback, we borrow the ideas of Imai and Ratkovic (2014). We operationalize the covariate balancing property by using inverse propensity score weighting

$$E\left\{\frac{\delta_i X_i}{\pi(X_i, \alpha)} - \frac{(1-\delta_i)X_i}{1-\pi(X_i, \alpha)}\right\} = 0. \tag{5}$$

Eq. (5) ensures that the first moment of each covariate is balanced and the weights based on CBPS are robust even when propensity score model is misspecified. The key idea behind the CBPS is that propensity score model determines the missing mechanism and covariate balancing weights, see Imai and Ratkovic (2014). The sample analogue of the covariate balancing moment condition given in Eq. (5) is

$$\frac{1}{n}\sum_{i=1}^{n} w(\delta_i, X_i, \alpha)X_i = 0, \quad w(\delta_i, X_i, \alpha) = \frac{\delta_i - \pi(X_i, \alpha)}{\pi(X_i, \alpha)(1-\pi(X_i, \alpha))}. \tag{6}$$

According to Imai and Ratkovic (2014), the CBPS is said to be *just identified* when the number of moment conditions equals that of parameters. If we use the covariate balancing conditions given in Eq. (6) alone, the CBPS is just-identified. If we combine Eq. (6) with the score condition given in Eq. (4), then the CBPS is *overidentified* because the number of moment conditions exceeds that of parameters.