Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Model averaging procedure for varying-coefficient partially linear models with missing responses

Jie Zeng [a,b], Weihu Cheng [a], Guozhi Hu [a,b,*], Yaohua Rong [a]

[a] *College of Applied Sciences, Beijing University of Technology, Beijing, 100124, China*
[b] *School of Mathematics and Statistics, Hefei Normal University, Hefei, 230601, China*

ARTICLE INFO

Article history:
Received 24 November 2017
Accepted 18 April 2018
Available online xxxx

AMS 2000 subject classifications:
primary 62G08
secondary 62G20
62G99

Keywords:
Covariate balancing propensity score
Focused information criterion
Inverse probability weighted method
Model averaging
Varying-coefficient partially linear models

ABSTRACT

This paper is concerned with model averaging procedure for varying-coefficient partially linear models with missing responses. The profile least-squares estimation process and inverse probability weighted method are employed to estimate regression coefficients of the partially restricted models, in which the propensity score is estimated by the covariate balancing propensity score method. The estimators of the linear parameters are shown to be asymptotically normal. Then we develop the focused information criterion, formulate the frequentist model averaging estimators and construct the corresponding confidence intervals. Some simulation studies are conducted to examine the finite sample performance of the proposed methods. We find that the covariate balancing propensity score improves the performance of the inverse probability weighted estimator. We also demonstrate the superiority of the proposed model averaging estimators over those of existing strategies in terms of mean squared error and coverage probability. Finally, our approach is further applied to a real data example.

1. Introduction

Varying-coefficient partially linear model (VCPLM), which is a useful extension of the popular partially linear model, offers additional flexibility by permitting interactions between a vector of covariates and a vector of unknown functions depending on another covariate, and has been a significant development in the semiparametric regression analysis. Various methods have been considered in the literature for parameter estimation in the VCPLM. See, for example, Ahmad, Leelahanon, and Li (2005), Fan and Huang (2005), Xia, Zhang, and Tong (2004) and Zhang, Lee, and Song (2002), among others. The assumption that the correctly specified model is given is the prerequisite of using these methods. Unfortunately, in practice, researchers are often confronted with a great many candidate models and are not sure which model to use. To choose the optimal model, plenty of criteria including the Akaike information criterion (AIC, Akaike, 1973), the Mallows' C_p (Mallows, 1973), the Bayesian information criterion (BIC, Schwarz, 1978) and the Focused information criterion (FIC, Claeskens & Hjort, 2003) have been proposed. Despite its long history and nice theoretical properties, such an approach may ignore uncertainty in the model selection stage, and often produces a rather unstable estimator. Model averaging, on the other hand, incorporates the uncertainty by appropriately smoothing estimators across different models rather than relying entirely on a single "winning" model.

Model averaging can be classified as Bayesian model averaging (BMA) and frequentist model averaging (FMA). BMA has long been a popular approach of incorporating model uncertainty. A comprehensive review of the Bayesian literature can be

[*] Corresponding author at: College of Applied Sciences, Beijing University of Technology, Beijing, 100124, China.
E-mail address: guozhihf@emails.bjut.edu.cn (G. Hu).

https://doi.org/10.1016/j.jkss.2018.04.004
1226-3192/© 2018 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

found in Hoeting, Madigan, Raftery, and Volinsky (1999) and Raftery, Madigan, and Hoeting (1997). A main concern for BMA is that the approach involves mixing together plenty of prior opinions regarding the parameters of interest, and it is uncertain what the results will be when some of these have clear clashes. FMA, on the other hand, eliminates the need to specify any prior distribution. Model averaging from a frequentist perspective has gained considerable attention in theoretical and applied statistics in recent years. Buckland, Burnham, and Augustin (1997) and Burnham and Anderson (2002) discussed weighting strategy based on the scores of BIC and AIC. Hansen (2007) and Wan, Zhang, and Zou (2010) developed an FMA procedure according to the Mallows' criterion. A seminal article, Hjort and Claeskens (2003), introduced FIC and proposed a smoothed FIC (S-FIC) weighting technique in a local misspecification framework. Hjort and Claeskens' (2003) investigation has been extended to several other models, including Cox's hazard regression models (Hjort & Claeskens, 2006), generalized additive partial linear models (Zhang & Liang, 2011), varying-coefficient partially linear measurement error models (Wang, Zou, & Wan, 2012), Tobit models (Zhang, Wan, & Zhou, 2012), logit models (Wan, Zhang, & Wang, 2014), linear models with missing responses (Sun, Su, & Ma, 2014) and linear quantile regression models (Peter, Gerda, & Holger, 2014). To the best of our knowledge, there are no model averaging estimators for VCPLM in missing data setting.

This paper extends the FIC and FMA procedures to a VCPLM when the response variable is assumed to be missing at random (MAR). We make use of the inverse probability weighted (IPW) method to handle missing data given that its resulting estimator has a credible "double robustness" property. To avoid the "curse of dimensionality", a fully parametric specification of the propensity score function is considered, a logistic regression model, for instance. Then we borrow the covariate balancing propensity score (CBPS, Imai & Ratkovic, 2014) method from causal inference setting to construct its parametric estimators. As a method robust to mild misspecification of parametric propensity score, CBPS can improve the performance of the usual IPW estimators by optimizing the covariate balance. Guo, Xue, and Hu (2017) considered CBPS-based inference for linear models with missing responses. To our best knowledge, there is no existing work considering FMA depending on CBPS-based IPW procedures for incomplete data setting. The main purpose of this paper is to fill this gap. However, we emphasize that such an extension is by no means straightforward and routine, for the following reasons: (i) compared with the existing imputation-based FMA procedure (Sun et al., 2014), our proposed IPW-based FMA process is much more complex for it involves the estimation of the completely unknown propensity score function; (ii) because the CBPS method is adopted in the estimation process, the derivation of the asymptotic properties of the parametric estimators in each partially restricted model should take into account the CBPS estimation's property.

The remainder of the paper is organized as follows. Section 2 outlines the model framework and provides estimation method in each plausible model. Section 3 specifies the FIC and the corresponding FMA process, along with the confidence intervals for the focus parameters. Section 4 conducts simulation studies to detect the finite sample properties of the proposed estimators. An application to a real dataset is presented in Section 5. Regularity conditions and the proofs of the main results are represented in the Appendix.

2. Model framework and estimation

Consider the following VCPLM:

$$Y = Z^{\top}\beta + X^{\top}\alpha(T) + \epsilon, \tag{1}$$

where Y is a response variable, (Z, X, T) are covariates, β is a $q \times 1$ column vector of unknown regression parameters, $\alpha(\cdot) = (\alpha_1(\cdot), \dots, \alpha_p(\cdot))^{\top}$ is a p-dimensional unknown coefficient functions, and ϵ is a random error with mean 0. For simplicity, we assume T is a one-dimensional random variable. Throughout this paper, we assume that some of the Y values in a sample size of n may be MAR whereas the Z, X and T values are completely observed. The response probability, also called the propensity score in the causal inference literature, is given by

$$P(\delta = 1|Y, Z, X, T) = P(\delta = 1|Z, X, T) = \Delta(Z, X, T), \tag{2}$$

where δ is a binary observation indicator. Consider a random sample of incomplete data $\{(Y_i, \delta_i, Z_i, X_i, T_i), i = 1, \dots, n\}$ from model (1), where $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$.

2.1. CBPS-based estimator for the propensity score

It is essential to estimate the completely unknown propensity score function before we construct estimators for β and $\alpha(\cdot)$. The current paper posits the following logistic model for it:

$$\Delta_i(d) = \Delta(L_i, d) = \frac{\exp(L_i^{\top} d)}{1 + \exp(L_i^{\top} d)}, \tag{3}$$

where $d \in \Theta$ is an unknown parameter vector, $L_i = (Z_i^{\top}, X_i^{\top}, T_i)^{\top}$. Naturally, d can be estimated by maximizing the log-likelihood function:

$$\sum_{i=1}^{n}\Big\{\delta_i \log[\Delta(L_i, d)] + (1 - \delta_i)\log[1 - \Delta(L_i, d)]\Big\}. \tag{4}$$