Computational Statistics and Data Analysis 103 (2016) 111–123

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Iterated imputation estimation for generalized linear models with missing response and covariate values

Fang Fang [a], Jun Shao [a,b,*]

[a] School of Statistics, East China Normal University, 500 Dongchuan Road, Shanghai, 200241, China
[b] Department of Statistics, University of Wisconsin, 1300 University Ave., Madison, WI 53706, USA

ARTICLE INFO

Article history:
Received 9 June 2015
Received in revised form 7 February 2016
Accepted 24 April 2016
Available online 13 May 2016

Keywords:
Arbitrary missing pattern
Iteration convergence
Imputation
Maximum likelihood
Missing at random
Missing covariate

ABSTRACT

A new approach named as the iterated imputation estimation is proposed for parameter estimation in generalized linear models with missing values in both response and covariates and data are missing at random. The proposed approach is much faster and easier to implement than the method of maximum likelihood or weighted estimating equation. It can be applied by directly using any existing software package for generalized linear models and treating the imputed values as observed in each iteration, which brings great convenience in programming. Theoretical results for the algorithm convergence of the iterated imputation estimation and the asymptotic distribution of the proposed estimator are obtained. Simulation studies and an illustrative example show that the iterated imputation estimation works quite well considering the trade-off between computational burden and estimation efficiency compared with the maximum likelihood estimation.

1. Introduction

Missing data is a common phenomenon in many applications in areas such as clinical trials, economics, sample surveys, and social sciences. Conventional statistical methods cannot be directly applied to incomplete data in most situations. We consider regression problems with both response and covariate variables having missing values. Most existing methods dealt with either missing response or missing covariate data (see, e.g., Little, 1992; Robins et al., 1994; Lipsitz et al., 1999; Fitzmaurice et al., 2001; Tang et al., 2003; Ibrahim et al., 2005). For linear regression models with missing response and covariate data, Chen et al. (2008) discussed theoretical properties for inference using maximum likelihood estimation via EM algorithm when the data is missing at random; Shao (2013) proposed some asymptotically unbiased and consistent estimators via direct estimation or imputation. For linear mixed models and generalized linear mixed models, Stubbendick and Ibrahim (2003, 2006) assumed full parametric models when the missing data mechanism is nonignorable and estimated the parameters by the maximum likelihood method. For longitudinal data when both responses and covariates are missing at random, Shardell and Miller (2008) and Chen et al. (2010) proposed several estimation methods based on inverse probability weighted estimating equations, in which a parametric model for the missing probability needs to be correctly specified.

In this paper, we consider a generalized linear model (GLM) involving n sampled subjects with independent and identically distributed data (y_i, x_i, z_i), $i = 1, \ldots, n$, where y_i is the outcome or response from subject i, x_i and z_i are p- and q-dimensional random vectors of covariates, y_i and some components of x_i may be missing, and z_i has no missing data. We assume missing at random (MAR), i.e., the probabilities of missing y_i and components of x_i only depend on z_i

* Corresponding author at: School of Statistics, East China Normal University, 500 Dongchuan Road, Shanghai, 200241, China.
E-mail addresses: ffang@sfs.ecnu.edu.cn (F. Fang), shao@stat.wisc.edu (J. Shao).

http://dx.doi.org/10.1016/j.csda.2016.04.010
0167-9473/

and x_i^{obs}, the observed part of x_i (Little and Rubin, 2002), but no parametric forms of the probabilities are assumed and the missing data mechanisms of y_i and x_i can be correlated each other. There are many methods for handling missing data under MAR (e.g., Little and Rubin, 2002; Kim and Shao, 2013). The simplest method is the complete case analysis that ignores the sampled subjects with incomplete data and applies the conventional statistical method. This approach is valid under the GLM but is quite inefficient if the size of incomplete data is large, especially when x_i is multivariate.

Under MAR, the maximum likelihood estimation (MLE) can be adopted in a similar way to Lipsitz et al. (1999) and Chen et al. (2008). The MLE specifies conditional distributions of y_i given x_i and z_i, $p(y_i|x_i, z_i, \beta)$, and x_i given z_i, $p(x_i|z_i, \alpha)$, where β is the unknown regression parameter vector of interest and α is an unknown vector of nuisance parameters. The parameters can be consistently estimated by maximizing the observed likelihood with a Monte Carlo EM algorithm (Lipsitz et al., 1999; Ibrahim et al., 2005). In iteration $t+1$ of the algorithm, a Monte Carlo sample of size L needs to be generated from the distribution $p(x_i^{mis}|y_i, x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ or $p(y_i, x_i^{mis}|x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ for every subject i with missing covariate values (response can be observed or missed), where x_i^{mis} is the missing components of x_i and $(\hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ is the parameter vector estimated in iteration t. The distribution $p(x_i^{mis}|y_i, x_i^{obs}, z_i, \hat{\beta}^{(t)}, \hat{\alpha}^{(t)})$ does not have an explicit form (due to the nonlinearity in GLMs) and sampling techniques such as the Gibbs sampler and the adaptive rejection algorithm of Gilks and Wild (1992) need to be used, which could be quite time-consuming especially when $p > 1$.

Alternatively, one may use the weighted estimating equation (WEE) that was mainly developed by Robins et al. (1994), Zhao et al. (1996), and Lipsitz et al. (1999). The WEE with an augmented term has a double robustness property in the sense that the estimation is consistent when either $p(x_i|z_i, \alpha)$ or $\pi_i = p(\delta_i^y = 1, \delta_i^x = 1|x_i^{obs}, z_i)$ is correctly specified, where $\delta_i^y = 1$ if y_i is observed, $\delta_i^y = 0$ if y_i is missing, $\delta_i^x = 1$ if x_i is fully observed and $\delta_i^x = 0$ if some components of x_i are missing. However, the efficiency of WEE relies on the correct specifications of both $p(x_i|z_i, \alpha)$ and π_i; and modeling π_i is not easy especially when $p > 1$. Furthermore, the WEE has even heavier computational burden than the MLE since a EM-Type algorithm is also needed and the Monte Carlo sampling procedure needs to be done for all the subjects (not just the subjects with $\delta_i^x = 0$ as in the MLE).

Kim (2011) proposed a parametric fraction imputation (PFI) method using the idea of importance sampling and calibration weighting to reduce the computational burden of the MLE and WEE. However, the PFI method needs to carefully arrange the imputed data and apply a weighted GLM, where the weights have to be updated in each iteration. Moreover, the PFI needs a large number of multiple imputations to achieve the consistency and efficiency of the resulting estimators. Although a calibrated PFI was proposed for a moderate number of multiple imputations, it brings more complexity in computation.

In this paper, we propose an iterated imputation estimation (IIE) approach. It has much less computational burden than the MLE and WEE and hence is more applicable for large samples with multivariate x_i having missing values. The IIE can directly use any existing software package for GLMs by treating the imputed values as observed in each iteration, which brings great convenience in programming. The methodology is described in Section 2. Some theoretical results concerning the convergence of the iterative algorithm and asymptotic distribution of the proposed estimator are presented in Section 3. Some simulations in Section 4 show that the IIE is much faster than the MLE and WEE but its efficiency loss is minor. The capability of the IIE to handle multivariate x_i and a relatively large sample size is also checked. An illustrative example is given in Section 5. Some concluding remarks are given in Section 6. All the proofs are in the Appendix.

2. Iterated imputation estimation

2.1. Notation and model

Under a generalized linear model, the conditional distribution of y_i given x_i and z_i has a density $p(y_i|x_i, z_i, \beta, \tau) = \exp\{\tau^{-1}(y_i\eta_i - b(\eta_i)) + c(y_i, \tau)\}$, where b and c are known functions, $\tau > 0$ is an unknown dispersion parameter, $\eta_i = \eta(\beta_x^{\mathrm{T}}x_i + \beta_z^{\mathrm{T}}z_i)$, β_x and β_z are p- and q-dimensional subvectors of β, a^{T} denotes the transpose of a, and η is a known one-to-one, continuously differentiable function. This includes many useful regression models as special cases, such as normal linear regression, logistic regression, probit regression, Poisson regression, gamma regression, etc. The covariate z_i may contain a constant component so that the corresponding component of β_z is the intercept effect. Since our main interest is to estimate β with missing values, without loss of generality, we assume throughout that $\tau = 1$ and write $p(y_i|x_i, z_i, \beta) = p(y_i|x_i, z_i, \beta, \tau = 1)$. If there is no missing data, β is estimated by maximizing the full data likelihood function, or equivalently, by resolving the full data score equation

$$S(\beta) = \sum_{i=1}^{n} \frac{\partial \log\{p(y_i|x_i, z_i, \beta)\}}{\partial \beta} = \sum_{i=1}^{n} g(x_i, z_i, \beta)\left\{y_i - h(\beta_x^{\mathrm{T}}x_i + \beta_z^{\mathrm{T}}z_i)\right\} = 0 \tag{1}$$

with $h(\beta_x^{\mathrm{T}}x_i + \beta_z^{\mathrm{T}}z_i) = b'(\eta_i) = E(y_i|x_i, z_i)$, $g(x_i, z_i, \beta) = \{\partial h(\beta_x^{\mathrm{T}}x_i + \beta_z^{\mathrm{T}}z_i)/\partial\beta\}/v_i$, and $v_i = b''(\eta_i) = \mathrm{Var}(y_i|x_i, z_i)$.

When y_i and some components of x_i are missing for some subjects, the equation in (1) cannot be solved. The complete case analysis replaces (1) with

$$S_{cc}(\beta) = \sum_{i=1}^{n} \delta_i g(x_i, z_i, \beta)\left\{y_i - h(\beta_x^{\mathrm{T}}x_i + \beta_z^{\mathrm{T}}z_i)\right\} = 0, \tag{2}$$