Journal of Econometrics 184 (2015) 452–463

Contents lists available at ScienceDirect

Journal of Econometrics

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

Model averaging estimation of generalized linear models with imputed covariates

Valentino Dardanoni[a], Giuseppe De Luca[a], Salvatore Modica[a], Franco Peracchi[b,*]

[a] University of Palermo, Italy

[b] University of Rome Tor Vergata and Einaudi Institute for Economics and Finance (EIEF), Italy

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form
19 March 2014
Accepted 15 June 2014
Available online 24 June 2014

JEL classification:
C11
C25
C35
C81

Keywords:
Model averaging
Bayesian averaging of maximum likelihood estimators
Generalized linear models
Missing covariates
Generalized missing-indicator method
SHARE

ABSTRACT

We address the problem of estimating generalized linear models when some covariate values are missing but imputations are available to fill-in the missing values. This situation generates a bias-precision trade-off in the estimation of the model parameters. Extending the generalized missing-indicator method proposed by Dardanoni et al. (2011) for linear regression, we handle this trade-off as a problem of model uncertainty using Bayesian averaging of classical maximum likelihood estimators (BAML). We also propose a block model averaging strategy that incorporates information on the missing-data patterns and is computationally simple. An empirical application illustrates our approach.

1. Introduction

In this paper we address the problem of estimating generalized linear models (GLMs) when the outcome of interest is always observed, some covariate values are missing, and imputations are available to fill-in the missing values. This situation is becoming quite common, as public-use data files increasingly include imputations of key variables affected by item nonresponse. The focus of this paper is on how to make use of the available imputations, not on methods to impute the missing values.

Two standard approaches to the problem of missing covariate values are complete-case analysis and the fill-in approach. The first drops all the observations with missing values ignoring the imputations altogether, while the second fills-in the missing values with the available imputations without distinguishing between observed and imputed values. Under certain conditions on the missing-data mechanism and the imputation model, the choice between these two approaches generates a trade-off between bias and precision in the estimation of the parameters of interest. When the complete cases are few the loss of precision may be substantial, but just filling-in the missing values with the imputations may lead to bias when the imputation model is either incorrectly specified or uncongenial in the sense of Meng (1994), that is, the imputation model is more restrictive than the model used to analyze the filled-in data. Validity of the assumptions behind the fill-in approach is often taken for granted, so this bias-precision trade-off is usually ignored. However, when imputations are provided by an external source, the congeniality assumption may fail because the two models are based on different parametric assumptions or they condition on different sets of covariates. The estimates from the fill-in approach may therefore be inconsistent, especially in the case of nonlinear estimators.

Using the generalized missing-indicator approach originally proposed for linear regression by Dardanoni et al. (2011), we transform the bias-precision trade-off between complete-case analysis and the fill-in approach into a problem of model uncertainty regarding which covariates should be dropped from an augmented GLM, or 'grand model', which includes two subsets of regressors:

[*] Corresponding author. Tel.: +39 06 7259 5934; fax: +39 06 2040 219.
E-mail address: franco.peracchi@uniroma2.it (F. Peracchi).

http://dx.doi.org/10.1016/j.jeconom.2014.06.002
0304-4076/

the focus covariates, corresponding to the observed or imputed covariates, and a set of auxiliary regressors consisting of binary indicators for the various missing-data patterns and their interactions with the focus regressors. Our formulation of the bias-precision trade-off in terms of model uncertainty exploits the fact that complete-case analysis and the fill-in approach correspond to two extreme specifications of the grand model. Complete-case analysis corresponds to using an unrestricted specification, while the fill-in approach corresponds to using a restricted specification that includes only the focus regressors. Instead of focusing on these extreme specifications of the grand model, we consider Bayesian averaging of classical maximum likelihood estimators (BAML) that takes into account all the intermediate specifications obtained by dropping from the grand model alternative subsets of auxiliary regressors associated with the various missing-data patterns. In this way we avoid restricting attention to the complete cases but, at the same time, we exploit the available imputations in a sensible way by allowing the imputation model to be incorrectly specified or uncongenial with the GLM of interest. The extreme choices of using either the complete-case or the fill-in approach are still available, but neither is likely to emerge as the best one since all the intermediate models in the expanded model space carry information about the parameters of interest.

In addition to extending the generalized missing-indicator method to the wide class of GLMs, we depart from Dardanoni et al. (2011) in three important respects. First, we propose a new block model averaging strategy that incorporates the information on the available patterns of missing data while being computationally simple. Second, we allow the observed outcome to be multivariate, thus covering the case of seemingly unrelated regression equations models and ordered, multinomial or conditional logit and probit models. Third, we investigate the robustness of our block-BAML procedure to the choice of priors by considering two families of prior distributions: the calibrated information criteria priors introduced by Clyde (2000), which use approximations based on the Laplace method for integrals to calibrate posterior model probabilities to classical model section criteria, and the conjugate priors for GLMs introduced by Chen and Ibrahim (2003), which allow to directly estimate posterior model probabilities using a computationally simple Markov chain Monte Carlo algorithm.

In our empirical illustration we analyze how cognitive functioning varies with physical health and socio-economic status using data from the fourth wave of the Survey on Health, Aging and Retirement in Europe (SHARE). Like for other household surveys, sensitive variables such as household income, household net worth, and other objective health measures are affected by substantial item nonresponse. Using the imputations contained in the public-use SHARE data, we investigate the bias-precision trade-off arising from different approaches for dealing with the problem of imputed covariates in GLMs. Further, we employ multiple imputation methods to account for the additional sampling uncertainty due to the imputation of missing covariate values.

The remainder of the paper is organized as follows. Section 2 presents our statistical framework. Section 3 discusses complete-case analysis and the fill-in approach. Section 4 describes the generalized missing-indicator method. Section 5 discusses our BAML procedure. Section 6 extends our results to the case of multivariate outcomes. Section 7 presents an empirical application. Finally, Section 8 offers some conclusions.

2. Statistical framework

We represent the available set of N observations on an outcome of interest as a realization of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$, whose components are independently distributed random variables with mean μ_n and finite nonzero variance σ_n^2.[1] We assume that the distribution of any component Y_n of $\boldsymbol{Y}$ belongs to the one-parameter linear exponential family with density function of the form

$$f(y; \gamma_n) = \exp\left[\gamma_n y - b(\gamma_n) + c(y)\right], \tag{1}$$

where γ_n is a scalar parameter called the canonical parameter, $b(\cdot)$ is a known, strictly convex and twice differentiable function, and $c(\cdot)$ is a known function.[2] By the properties of the linear exponential family, the mean and variance of Y_n are equal to $\mu_n = b'(\gamma_n)$ and $\sigma_n^2 = b''(\gamma_n)$ respectively (McCullagh and Nelder, 1989). Different choices of the functions $b(\cdot)$ and $c(\cdot)$ result in different distributions within this family. For example, letting $b(\gamma_n) = \gamma_n^2/2$ and $c(y) = -1/2[y^2 + \ln(2\pi)]$ gives the density of a normal distribution with mean γ_n and unit variance, while letting $b(\gamma_n) = \exp(\gamma_n)$ and $c(y) = -\ln(y!)$ gives the density of a Poisson distribution with intensity parameter equal to $\exp(\gamma_n)$.

In a GLM the dependence of Y_n on a vector of covariates X_n (assumed to include a constant term) is modeled by assuming that there exists a continuously differentiable and invertible function $h(\cdot)$, known as the inverse link, such that the mean of Y_n is equal to $\mu_n = h(X_n^\top \beta)$ for a unique value of the K-dimensional parameter vector β. The linear combination $\eta_n = X_n^\top \beta$ is called the linear predictor associated with the nth observation. Collecting together the linear predictors associated with the sample observations gives the N-dimensional vector $\boldsymbol{\eta} = \boldsymbol{X}\beta$, where $\boldsymbol{X}$ is the $N \times K$ matrix of observations on the covariates with nth row equal to $X_n^\top$.

In the absence of missing data, the classical approach to estimating β is maximum likelihood (ML). The sample log-likelihood for the missing-free data is

$$L(\beta) = c + \sum_{n=1}^{N} \left[\gamma_n(\beta)\, Y_n - b\left(\gamma_n(\beta)\right)\right],$$

where $\gamma_n(\beta)$ is the unique root of the equation $b'(\gamma) = h(X_n^\top \beta)$ and the missing-free data ML estimator $\widehat{\beta}$ of β is obtained by solving the system of K likelihood equations

$$0 = L'(\beta) = \sum_{n=1}^{N} \upsilon(X_n^\top \beta) \left[Y_n - h(X_n^\top \beta)\right] X_n,$$

with $\upsilon(X_n^\top \beta) = h'(X_n^\top \beta)/b''(\gamma_n(\beta))$. Provided the assumed model is correctly specified, and the mild regularity conditions in Fahrmeir and Kaufmann (1985) hold, $\widehat{\beta}$ is unique, consistent, and asymptotically normal with asymptotic variance equal to the inverse of the Fisher information matrix. The fact that β enters the likelihood equations only through the linear predictor $\eta_n = X_n^\top \beta$ is the key property of GLMs that drives our main result in Theorem 1. If $b'(\cdot) = h(\cdot)$ (the "canonical link" case), then $\gamma_n(\beta) = X_n^\top \beta$ and the likelihood equations simplify considerably because $\upsilon(X_n^\top \beta) = 1$ for all n. An example is the Gaussian model with identity link $h(X_n^\top \beta) = X_n^\top \beta$, where the likelihood equations reduce to the familiar normal equations for OLS.

In this paper we depart from the standard GLM setup by allowing some covariate values to be missing. We also assume that imputations, as provided by an external source (typically the producers of the dataset), are available to fill-in the missing covariate values. Since the constant term is always observed, the number of possible missing-data patterns is equal to 2^{K-1}. Not all the possible

[1] Vectors are always column vectors, and boldface denotes vector and matrices of sample observations or of functions of sample observations.

[2] In the original formulation of Nelder and Wedderburn (1972), the density in Eq. (1) includes an additional dispersion parameter which, without loss of generality, we set equal to one.