Computational Statistics and Data Analysis 67 (2013) 15–24

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Moment adjusted imputation for multivariate measurement error data with applications to logistic regression

Laine Thomas[a,*], Leonard A. Stefanski[b], Marie Davidian[b]

[a] *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA*

[b] *Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA*

ARTICLE INFO

Article history:
Received 10 December 2012
Received in revised form 26 April 2013
Accepted 27 April 2013
Available online 11 May 2013

Keywords:
Moment adjusted imputation
Multivariate measurement error
Logistic regression
Regression calibration

ABSTRACT

In clinical studies, covariates are often measured with error due to biological fluctuations, device error and other sources. Summary statistics and regression models that are based on mis-measured data will differ from the corresponding analysis based on the "true" covariate. Statistical analysis can be adjusted for measurement error, however various methods exhibit a tradeoff between convenience and performance. Moment Adjusted Imputation (MAI) is a measurement error in a scalar latent variable that is easy to implement and performs well in a variety of settings. In practice, multiple covariates may be similarly influenced by biological fluctuations, inducing correlated, multivariate measurement error. The extension of MAI to the setting of multivariate latent variables involves unique challenges. Alternative strategies are described, including a computationally feasible option that is shown to perform well.

1. Introduction

The problem of measurement error arises whenever data are measured with greater variability than the true quantities of interest, $\boldsymbol{X}$. It can be attributed to sources like device error, assay error, and biological fluctuations. Summary statistics and regression models based on mis-measured data, $\boldsymbol{W}$, may have biased parameter estimators, reduced power and insufficient confidence interval coverage (Fuller, 1987; Armstrong, 2003; Carroll et al., 2006).

Many strategies to adjust for measurement error have been proposed. Correction for measurement error in covariates, $\boldsymbol{X}$, in linear and generalized linear models is commonly achieved by regression calibration (RC), which substitutes an estimate of the conditional mean $E(\boldsymbol{X}|\boldsymbol{W})$ for the unknown $\boldsymbol{X}$ (Carroll and Stefanski, 1990; Gleser, 1990). When estimation of the density of $\boldsymbol{X}$ is of interest, this is not particularly useful, as $E(\boldsymbol{X}|\boldsymbol{W})$ is over-corrected in terms of having reduced spread (Eddington, 1940; Tukey, 1974). However, the linear regression based on this substitution estimates the underlying parameters of interest. RC is also implemented in non-linear models because of its simplicity, but is typically most effective for general linear models when the measurement error is not large (Rosner et al., 1989; Carroll et al., 2006). Alternatives for non-linear models are thoroughly reviewed by Carroll et al. (2006) and include maximum likelihood, conditional score, SIMEX and Bayesian methods. Density estimation is addressed in a separate literature, where deconvolution methods are a prominent approach for scalar variables (Carroll and Hall, 1988; Stefanski and Carroll, 1990). Recent papers add to the theoretical basis for deconvolution estimators and offer an extension to heteroscedastic measurement error (Carroll and Hall, 2004; Delaigle, 2008; Delaigle et al., 2008; Delaigle and Meister, 2008).

Despite the variety of available methodology, measurement error correction is rarely implemented (Jurek et al., 2004), and when it is, RC remains popular. Convenience may be a priority for a method to achieve widespread use. The preceding

[*] Corresponding author. Tel.: +1 9196841289.
E-mail addresses: laine.thomas@duke.edu, leellio2@gmail.com (L. Thomas).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.04.017

methods target estimation of parameters in a specific regression context; a different method must be implemented for every type of regression model in which $\boldsymbol{X}$ is used, and density estimation is treated separately. An alternative approach is to focus on re-creating the true $\boldsymbol{X}$ from the observed $\boldsymbol{W}$, at least approximately, as the primary quantity of interest or as a means to improving parameter estimation (Louis, 1984; Bay, 1997; Shen and Louis, 1998; Freedman et al., 2004). Most recently, Thomas et al. (2011) introduce a Moment Adjusted Imputation (MAI) method that aims to replace scalar, mis-measured data, W, with estimators, $\widehat{X}$, that have asymptotically the same joint distribution with a response, Y, and potentially error-free covariates $\boldsymbol{Z}$, as does the latent variable, X, up to some number of moments. Originally developed in an unpublished dissertation, (Bay, 1997), MAI extends the idea of moment reconstruction (MR), which focuses on the first two moments of the joint distribution (Freedman et al., 2004, 2008). Thomas et al. (2011) investigate the performance of MAI in logistic regression and demonstrate superior results to RC and MR when the distribution of X is non-normal. Moreover, the $\widehat{X}$ can be used for density estimation, with a guarantee of matching the latent variable mean, variance and potentially higher moments.

Thomas et al. (2011) discuss an important application where measurement error is likely present in multiple covariates and a scalar adjustment method would not be adequate. In this example, Gheorghiade et al. (2006) studied blood pressure at admission in patients hospitalized with acute heart failure using data from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF) registry. Logistic regression was used to describe the relationship in-hospital mortality and both systolic and diastolic blood pressure. These variables, when measured at the same time, are likely to have correlated measurement error due to common biological fluctuations and measurement facilities. The original analysis by Gheorghiade et al. (2006) regards both variables as error free. Thomas et al. (2011) revise the analysis to account for measurement error in systolic blood pressure, which is of primary interest, but not diastolic blood pressure. Adjustment of both variables could be important. When multiple covariates are measured with error, one could apply a univariate adjustment separately. However, this would not account for correlation between the latent variable measurement errors.

Here, we introduce the extension of MAI to multivariate mis-measured data and provide a computationally convenient approach to implementation. The result is quite similar to univariate MAI, but unique challenges are addressed. In Section 2, we define a set of moments that are feasible for matching with multivariate measurement error and introduce the natural extension of the MAI algorithm, used to obtain adjusted data with appropriate moments. In Section 3, a numerically convenient method for the implementation of multivariate MAI is proposed. In Section 4, the alternative implementations of multivariate MAI are evaluated and MAI is compared to other imputation methods via simulation in applications to density estimation and logistic regression. In Section 5, we revise the OPTIMIZE-HF analysis to account for measurement error and obtain estimates that describe the features of "true" diastolic and systolic blood pressure.

2. The method

Here we introduce notation for the current problem that is similar to Thomas et al. (2011) but not identical. Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iG})^T$ be a $(G \times 1)$ vector of latent variables for $i = 1, \ldots, n$. The observed data are $\boldsymbol{W}_i = \boldsymbol{X}_i + \boldsymbol{U}_i$, where $\boldsymbol{U}_i \sim MVN(\boldsymbol{0}, \Sigma_{ui})$, $MVN(\boldsymbol{\mu}, \Sigma)$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, $\boldsymbol{0}$ is a $G \times 1$ vector of zeros, $\boldsymbol{U}_i$ is independent of $\boldsymbol{X}_i$, and $\boldsymbol{U}_i$ are mutually independent. We assume that Σ_{ui} is known. The latent variables $\boldsymbol{X}_i$ may be of particular interest, as in density estimation or as predictors in a regression model. In the latter case, we also have a response Y_i and potentially a vector of $(K - 1)$ error-free covariates $\boldsymbol{Z}_i$. These additional variables are collected to create $\boldsymbol{V}_i = (Y_i, \boldsymbol{Z}_i^T)$, with components V_{ik} for $k = 1, \ldots, K$. We make the usual *surrogacy* assumption that Y_i and $\boldsymbol{Z}_i$ are not related to the measurement error in $\boldsymbol{X}_i$, so that $\boldsymbol{V}_i$ is conditionally independent of $\boldsymbol{W}_i$ given $\boldsymbol{X}_i$ (Carroll et al., 2006).

The goal is to obtain adjusted versions of the $\boldsymbol{W}_i$, $\widehat{\boldsymbol{X}}_i$, whose distribution closely resembles that of $\boldsymbol{X}_i$ and possibly the joint distribution of $\boldsymbol{X}_i$ and other variables. In terms of moments, we require that $E(n^{-1} \sum_{i=1}^n \widehat{\boldsymbol{X}}_i) = E(\boldsymbol{X}_i)$, $E(n^{-1} \sum_{i=1}^n \widehat{\boldsymbol{X}}_i \widehat{\boldsymbol{X}}_i^T) = E(\boldsymbol{X}_i \boldsymbol{X}_i^T)$, $E(n^{-1} \sum_{i=1}^n \widehat{\boldsymbol{X}}_i V_{ik}) = E(\boldsymbol{X}_i V_{ik})$ for $k = 1, \ldots, K$, and $E(n^{-1} \sum_{i=1}^n \widehat{\boldsymbol{X}}_i^r) = E(\boldsymbol{X}_i^r)$ for $r = 3, \ldots, M$, and the rth power is applied component-wise. This differs from Thomas et al. (2011) in that only first order cross products between $\boldsymbol{X}_i$ and error free covariates are matched. Additional moment constraints can be added, but unlike the case of univariate MAI, estimators for higher order cross products are not straightforward. They are complex functions of many parameters that each have to be estimated. We therefore favor a reduced set cross products for parsimony and good performance in simulations.

2.1. Implementation

The first step is to define unbiased estimators for the unknown quantities $\boldsymbol{m}_r = E(\boldsymbol{X}_i^r)$, $(G \times 1)$ and $r = 1, \ldots, M$, $\boldsymbol{m}^* = E(\boldsymbol{X}_i \boldsymbol{X}_i^T)$, $(G \times G)$, $\boldsymbol{m}_k^V = E(\boldsymbol{X}_i V_{ik})$, $(G \times 1)$. Because $\boldsymbol{W}_i | \boldsymbol{X}_i \sim MVN(\boldsymbol{X}_i, \Sigma_{ui})$, we know that $E(\boldsymbol{W}_i) = E\{E(\boldsymbol{W}_i | \boldsymbol{X}_i)\} = E(\boldsymbol{X}_i)$ so $\widehat{\boldsymbol{m}}_1 = n^{-1} \sum_{i=1}^n \boldsymbol{W}_i$. Unbiased estimators for the higher-order moments are defined using the recursion formula $H_0(z) = 1$, $H_1(z) = z$, $H_r(z) = zH_{r-1}(z) - (r-1)H_{r-2}(z)$ for $r = 2, 3, \ldots$ (Cramer, 1957). Stulajter (1978) proved that if $W \sim N(\mu, \sigma^2)$, then $E\{\sigma^r H_r(W/\sigma)\} = \mu^r$ (Stefanski, 1989; Cheng and Van Ness, 1999). Let W_{ig} and X_{ig} denote the gth component of $\boldsymbol{W}_i$ and $\boldsymbol{X}_i$, respectively, and let $\Sigma_{ui,gg'}$ denote the element of Σ_{ui} in the gth row and g'th column. Marginally, $W_{ig} | X_{ig} \sim N(X_{ig}, \Sigma_{ui,gg})$. Letting $P_r(w, \sigma) = \sigma^r H_r(w/\sigma)$, we have $E\{P_r(W_{ig}, \Sigma_{ui,gg})\} = E[E\{P_r(W_{ig}, \Sigma_{ui,gg}) | X_{ig}\}] = E(X_{ig}^r)$. The gth component of $\widehat{\boldsymbol{m}}_r$ is $\widehat{m}_{rg} = n^{-1} \sum_{i=1}^n P_r(W_{ig}, \Sigma_{ui,gg})$ for $r = 1, \ldots, M$. In addition, $E(\boldsymbol{W}_i \boldsymbol{W}_i^T | \boldsymbol{X}_i) = \boldsymbol{X}_i \boldsymbol{X}_i^T + \Sigma_{ui}$, so