Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–13

DOI: 10.1177/0962280219876959
journals.sagepub.com/home/smm
SAGE

Cluster-specific nonignorably missing, endogenous, and continuous regressors in multilevel model for binary outcome

Gi-Soo Kim,[1] Youngjo Lee,[1] Hongsoo Kim[2,3,4] and Myunghee Cho Paik[1]

Abstract

In multilevel regression models for observational clustered data, regressors can be correlated with cluster-level error components, namely endogenous, due to omitted cluster-level covariates, measurement error, and simultaneity. When endogeneity is ignored, regression coefficient estimators can be severely biased. To deal with endogeneity, instrument variable methods have been widely used. However, the instrument variable method often requires external instrument variables with certain conditions that cannot be verified empirically. Methods that use the within-cluster variations of the endogenous variable work under the restriction that either the outcome or the endogenous variable has a linear relationship with the cluster-level random effect. We propose a new method for binary outcome when it follows a logistic mixed-effects model and the endogenous variable is normally distributed but not linear in the random effect. The proposed estimator capitalizes on the nested data structure without requiring external instrument variables. We show that the proposed estimator is consistent and asymptotically normal. Furthermore, our method can be applied when the endogenous variable is missing in a cluster-specific nonignorable mechanism, without requiring that the missing mechanism be correctly specified. We evaluate the finite sample performance of the proposed approach via simulation and apply the method to a health care study using a San Diego inpatient dataset. Our study demonstrates that the clustered structure can be exploited to draw valid analysis of multilevel data with correlated effects.

Keywords

Cluster-specific nonignorable missingness, correlated effects, endogeneity, instrumental variable

1 Introduction

Consider observational health care data collected in San Diego, California,[1] where the subjects are 41,179 elderly inpatients who were admitted and discharged alive from 20 acute care hospitals from April to September 2006. Among the patients, about 11.03% experienced unscheduled rehospitalizations (UR) within 30 days. Thirty-day UR are likely to be due to a lack of discharge planning and care coordination from the index hospital.[2] Moreover, the cost of UR of Medicare patients has been estimated to be up to $17.4 billion per year.[3] In October 2012, the Centers for Medicare & Medicaid Services adopted rehospitalizations for certain conditions within 30 days as a quality measure and implemented a pay-for-performance scheme related to the indicator for Medicare beneficiaries.[4] This policy change has increased the interest in identifying the factors affecting UR.

Our specific interest is the effect of health care cost on UR. We can fit a generalized linear mixed model (GLMM) with hospital-level random effects using the indicator of UR as a binary outcome and the patient-level and hospital-level observed characteristics as regressors. In a naive GLMM, however, the true effect of the cost can be confounded if the cost variable is correlated with unmeasured hospital conditions represented by the random effects. When correlation exists, the cost variable is called endogenous and the unmeasured hospital

[1]Department of Statistics, Seoul National University, Seoul, South Korea
[2]Graduate School of Public Health, Dept. of Public Health Sciences, Seoul National University, Seoul, South Korea
[3]Institute of Aging, Seoul National University, Seoul, South Korea
[4]Institute of Health and Environment, Seoul National University, Seoul, South Korea

Corresponding author:
Myunghee Cho Paik, Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea.
Email: myungheechopaik@snu.ac.kr

variables are called confounders. Potential hospital-level confounders may include general service quality, competence of the hospital crew, and policies for discharged patients such as education programs or connections to other health facilities. Our goal is to remove such endogeneity and estimate the effect of the cost associated with illness or patient condition but not with hospital condition.

In this paper, we propose a new method for estimating the regression coefficients in a logistic GLMM when a normally distributed regressor is correlated with cluster-level random effects due to omitted cluster-level covariates, measurement error, and simultaneity. As will be shown in section 6, the variance component of the random effects is not significant using the standard GLMM, while the proposed method shows otherwise. Also, by ignoring endogeneity the effect estimate of health care cost can be misleading.

When endogeneity is present, naive methods that ignore endogeneity yield substantial bias for parameter estimates. Ebbes et al.[5] showed through simulation studies that even a moderate correlation between a regressor and random component can result in severe bias in the parameter estimates. Most of the existing methods that deal with endogeneity are instrumental variable (IV)-based methods.[6] An instrument is an external variable that is required to be uncorrelated with the random component while partially explaining variability in the endogenous regressor. In the simple linear models, the IV method replaces the endogenous variable with its conditional expectation given the instrument and exogenous regressors, assuming a linear relationship. As long as the conditions of the IV are satisfied, the resulting coefficient estimates are consistent. The method requires at least as many linearly independent instruments as the number of endogenous regressors to avoid identifiability problem. This method was later extended to the case of nonlinear models.[7–9]

A drawback of the IV method is that the necessary conditions of an instrument cannot be verified empirically and its selection is mostly based on a priori subject-matter knowledge. Such external instrument, if available, yields inefficient estimates when the correlation with the endogenous regressor is weak. Moreover, Bound et al.[10] showed that the IV estimates can be more biased than the already inconsistent naive estimators if a weak relationship exists between the instrument and the random component. Mundlak,[11] Hausman and Taylor,[12] and Kim and Frees[13] proposed an alternative approach applicable to linear mixed-effects models (LMM) where regressors are correlated with cluster-level random effects. Instead of using an external IV, their method exploited the clustered structure of the data. However, this method does not work in the case of nonlinear models. For GLMMs, Neuhaus and McCulloch[14] proposed to decompose the endogenous variable into between- and within-cluster components using either conditional or partitioning methods. The conditional likelihood method, however, does not provide estimates of the cluster-level covariate effects, while the partitioning method requires that the endogenous variable has a linear relationship with the cluster-level random effects.

In this paper, we proposed a new method that capitalizes on the clustered structure of the data to draw valid analysis in logistic GLMMs when a normally distributed regressor is endogenous. A caveat of the new method is to utilize a link-preserving imputation introduced by Paik and Sacco[15] and Chen et al.[16] This strategy was initially proposed to deal with missing regressors in logistic regression models. The proposed estimator is consistent and asymptotically normal, and does not require a linear relationship between the endogenous variable and random effect. Our method can also be applied when the endogenous regressor is missing under a cluster-specific non-ignorable (CSNI) mechanism,[17] where missingness depends on the observed variables and cluster-level random component, but not on the missing regressor. It takes a nonparametric approach to the missing mechanism in the sense that the functional form of the missing probability needs not to be specified. To the best of our knowledge, our method is the first to provide a consistent estimator for parameters in the logistic mixed-effects model when the endogenous variable is not missing at random.

Section 2 describes various forms of endogeneity in regression models and existing methods established in the literature. Section 3 presents the proposed estimator using link-preserving imputation. Section 4 extends the proposed method to allow missingness in the endogenous variable. Section 5 evaluates finite sample properties of the proposed estimator through simulation. We then apply the existing methods and the proposed method to the 2006 San Diego inpatient dataset in section 6. The conclusion follows in section 7.

2 Literature on endogeneity

In this section, we describe various forms of endogeneity in regression models and the existing estimation methods. Suppose that

$$\mathbb{E}(Y_i|X_i, \mathbf{Z}_i, \alpha_i) = h(\beta_0 + X_i\beta_1 + \mathbf{Z}_i^T\boldsymbol{\beta_2} + \alpha_i), \quad i = 1, \ldots, N \tag{1}$$

where Y_i is an outcome, $h(\cdot)$ is a known twice differentiable function, α_i is an unobserved random component with mean 0, $\mathbf{Z}_i$ is a vector of exogenous regressors, X_i is an endogenous variable, correlated with α_i, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}_2^T)$ is a vector of unknown parameters. In equation (1), α_i can represent an omitted variable term, $\alpha_i = Z_{ui}\beta_u$, where Z_{ui} is the omitted variable correlated with X_i and β_u is the coefficient. If we fit equation (1) assuming α_i is not correlated with X_i, estimates of $\boldsymbol{\beta}$ can be severely biased. Let

$$\varepsilon_i = Y_i - \mathbb{E}(Y_i|X_i, \mathbf{Z}_i, \alpha_i)$$

so that $\mathbb{E}(\varepsilon_i|X_i, \mathbf{Z}_i, \alpha_i) = 0$. For simplicity, we assume that X_i is a scalar.

In linear models, i.e. when $h(\cdot)$ is the identity function, $(\alpha_i + \varepsilon_i)$ can be considered as a single error term and the IV method by Bowden and Turkington[6] can be used. We denote an external IV for X_i by X_i^*, which is required to be uncorrelated with both α_i and ε_i while explaining some variability in X_i. When $\mathbf{Z}$ is null, the method obtains ordinary least squares estimator after replacing $\mathbf{X} = \{\mathbf{1}, (X_1, \ldots, X_N)^T\}$ with $\widehat{\mathbf{X}} = \mathbf{H}^*\mathbf{X}$ where $\mathbf{H}^* = \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}$ and $\mathbf{X}^* = \{\mathbf{1}, (X_1^*, \ldots, X_N^*)^T\}$. Denoting $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$, and $\mathbf{E} = (\alpha_1 + \varepsilon_1, \ldots, \alpha_N + \varepsilon_N)^T$, the resulting estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, say $\widehat{\boldsymbol{\beta}}^{IV}$ is consistent since $\widehat{\boldsymbol{\beta}}^{IV} = (\widehat{\mathbf{X}}^T\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{H}^*\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}^*[\mathbf{X}\boldsymbol{\beta} + \mathbf{E}] = \boldsymbol{\beta} + o_p(1)$, regardless of the relationship between $\mathbf{X}$ and $\mathbf{X}^*$.

When $h(\cdot)$ is nonlinear, α_i and ε_i may not be combined as a single error term. Terza et al.[9] proposed to use the two-stage residual inclusion method, which is analogous to the IV method by Bowden and Turkington.[6] Instead of replacing X_i with $\hat{X}_i$, it uses the fitted residual $(X_i - \hat{X}_i)$ as an additional regressor, which has an effect of accounting for α_i and thus removing the endogeneity issue. Unlike in the linear $h(\cdot)$ case, the regression model for the endogenous variable should be correctly specified to achieve consistency. When $h(\cdot)$ is linear, the two-stage residual inclusion method is shown to be numerically equivalent to the IV method by Bowden and Turkington.[6]

In this paper, we consider clustered data Y_{ij} for the ith cluster, $i = 1, \ldots, K$ and the jth unit, $j = 1, \ldots, n_i$. Suppose that

$$\mathbb{E}(Y_{ij}|X_{ij}, \mathbf{Z}_{ij}, \alpha_i) = h(\beta_0 + X_{ij}\beta_1 + \mathbf{Z}_{ij}^T\boldsymbol{\beta}_2 + \alpha_i)$$

where α_i represents a cluster-level unobserved random component correlated with X_{ij}. When data are clustered, new methods do not require external IV X^*. A simple approach is to treat α_i as fixed effects and introduce cluster-level dummy variables, thereby removing endogeneity. However, even without endogeneity, such analysis is shown to yield severely biased estimates when the response is not normal and clusters are not large, due to a large number of incidental parameters.[18] Also, the parameter estimates are available only for unit-level regressors.[19]

Kim and Frees[13] and Hausman and Taylor[12] derived an internal IV method for LMMs. They eliminate the source of endogeneity, α_i, by subtracting the clusterwise means from both sides of the model equation and use the within-cluster variations of the endogenous variable as instruments. Mundlak[11] called the resulting estimator the ''within'' estimator. Details are presented in Appendix 1. Though the method is novel, it is not easily extendable to the case of nonlinear models as it capitalizes on the linearity of the outcome as well as the clustered structure.

When Y_{ij} follows a distribution from the exponential family and $h(\cdot)$ is the canonical link function, Neuhaus and McCulloch,[14] Brumback and He,[20] Brumback et al.,[21] and Brumback et al.[22] proposed to maximize the conditional likelihood that removes α_i by conditioning on sufficient statistics for α_i. This method yields consistent estimates for the unit-level covariate effects without requiring correct specification of the distribution of α_i, but does not provide estimates for the cluster-level covariate effects as they are conditioned out as well. Goetgeluk and Vansteelandt[23] developed a related, conditional method based on estimating equations that enables estimation of the cluster-level covariate effects in a two-step procedure. However, their method is restricted to the case where $h(\cdot)$ is either the identity or exponential function. When $h(\cdot)$ is not canonical or the cluster-level covariates effects are of interest, Neuhaus and Kalbfleisch[24] and Neuhaus and McCulloch[14] derived the partitioning method as a ''poor man's'' approximation to the conditional likelihood approach. The partitioning method decomposes the endogenous variable into between- and within-cluster components, i.e. $\beta_1 X_{ij} = \beta_{1B}\bar{X}_i + \beta_{1W}(X_{ij} - \bar{X}_i)$, where $\bar{X}_i$ is the ith cluster mean of X, and fits a GLMM utilizing both $(X_{ij} - \bar{X}_i)$ and $\bar{X}_i$ as separate predictors. The resulting estimate of β_{1W} is shown to be consistent for β_1 under the restrictive assumption that α_i is a linear function of $\bar{X}_i$ plus an independent Gaussian error term.[25–27] Also, as we can verify in the simulation studies of Neuhaus and McCulloch,[14] the estimate of the intercept is severely biased due to the presence of $\bar{X}_i$ in the model.

To generalize the partitioning method, Brumback et al.[25] and Brumback et al.[26] proposed a method that capitalizes on the full knowledge of the parametric form of $\mathbb{E}[\alpha_i|\mathbf{X}_i]$, where $\mathbf{X}_i = \{X_{i1}, \ldots, X_{in_i}\}^T$. Under the

assumption that $\psi(\mathbf{X}_i; \boldsymbol{\xi}) := \mathbb{E}[\alpha_i|\mathbf{X}_i]$ is a known, linear function in the parameter vector $\boldsymbol{\xi}$ and that $\nu_i := \alpha_i - \psi(\mathbf{X}_i; \boldsymbol{\xi})$ is i.i.d. Gaussian independent of $\mathbf{X}_i$, the authors proposed to fit a GLMM with both $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ as fixed parameters. In the special case where $\psi(\mathbf{X}_i; \boldsymbol{\xi}) = [1, \bar{X}_i]^T\boldsymbol{\xi}$, the method is numerically identical to the partitioning method.[14,24] However, this method is inapplicable when $\psi(\cdot)$ has no closed form and ν_i is still correlated with $\mathbf{X}_i$.

In this paper, we propose a new method that can be applied to logistic mixed-effects models for binary data, without the need for external IVs and without requiring a linear relationship between the endogenous variable and the cluster-level random effect. The main strategy is to apply the link-preserving imputation method of Paik and Sacco,[15] which was originally developed in the missing data context to impute regressors that are missing at random (MAR). Paik and Sacco[15] proved that when a binary outcome follows a generalized linear model (GLM) with logistic link and the missing regressor belongs to an exponential family, then the marginal distribution of the outcome without conditioning on the missing regressor is a GLM with logistic link as well with the missing regressor replaced with a function of completely observed regressors. Hence, the logistic link is preserved after marginalization. We extend this method to binary GLMMs that contain a continuous regressor that is both endogenous and missing. Our method can allow CSNI missingness in the endogenous variable, without requiring to correctly specify the missing mechanism.

3 Proposed method

In this section, we show that if data are collected in a clustered structure, we can develop consistent estimation procedure for logistic mixed-effects models under endogeneity without using external IVs. Suppose that Y_{ij} is a binary outcome of the jth unit ($j = 1, 2, \ldots, n_i$) in the ith cluster ($i = 1, 2, \ldots, K$). Let X_{ij} be an endogenous unit-level scalar regressor and $\mathbf{Z}_{ij}$ be a p-dimensional vector of exogenous regressors. Suppose that the outcome follows a GLMM with logistic link

$$\text{logit}\{P(Y_{ij} = 1|X_{ij}, \mathbf{Z}_{ij}, \alpha_i)\} = \beta_0 + X_{ij}\beta_1 + \mathbf{Z}_{ij}^T\boldsymbol{\beta_2} + \alpha_i \tag{2}$$

where $(\beta_0, \beta_1, \boldsymbol{\beta}_2^T)$ is a vector of unknown parameters and α_i's are independent and normally distributed random effects with mean 0 and variance D, correlated with X_{ij}. Our goal is to obtain consistent estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}_2^T)$.

The proposed strategy is to derive a marginal model for the outcome without the endogenous regressor using a link-preserving imputation of Paik and Sacco.[15] Following terminology of Chen et al.,[16] a marginal model is called link-preserving if the part of the linear predictor concerning the exogenous regressors is preserved under the same link function after marginalization. Thus under link-preserving imputation, the marginal mean has a form of replacing the endogenous regressor with some imputation value.

3.1 Construction of the marginal model

Our approach differs from the previous approaches in that instead of modeling $P(Y_{ij} = 1|X_{ij}, \mathbf{Z}_{ij})$, we model $P(Y_{ij} = 1|\mathbf{Z}_{ij}, \alpha_i)$. Let $f(\cdot)$ denote the conditional distribution of X given $\mathbf{Z}$, Y and α. Using Bayes's rule, we have

$$\log\frac{P(Y_{ij} = 1|\mathbf{Z}_{ij}, \alpha_i)}{P(Y_{ij} = 0|\mathbf{Z}_{ij}, \alpha_i)} = \log\frac{P(Y_{ij} = 1|X_{ij}, \mathbf{Z}_{ij}, \alpha_i)}{P(Y_{ij} = 0|X_{ij}, \mathbf{Z}_{ij}, \alpha_i)} + \log\frac{f(X_{ij}|\mathbf{Z}_{ij}, Y_{ij} = 0, \alpha_i)}{f(X_{ij}|\mathbf{Z}_{ij}, Y_{ij} = 1, \alpha_i)} \tag{3}$$

We also have

$$\exp(\beta_1 x) = \frac{P(Y_{ij} = 1|X_{ij} = x, \mathbf{Z}_{ij}, \alpha_i)}{P(Y_{ij} = 0|X_{ij} = x, \mathbf{Z}_{ij}, \alpha_i)}\frac{P(Y_{ij} = 0|X_{ij} = 0, \mathbf{Z}_{ij}, \alpha_i)}{P(Y_{ij} = 1|X_{ij} = 0, \mathbf{Z}_{ij}, \alpha_i)} = \frac{f(X_{ij} = x|\mathbf{Z}_{ij}, Y_{ij} = 1, \alpha_i)}{f(X_{ij} = x|\mathbf{Z}_{ij}, Y_{ij} = 0, \alpha_i)}\frac{f(X_{ij} = 0|\mathbf{Z}_{ij}, Y_{ij} = 0, \alpha_i)}{f(X_{ij} = 0|\mathbf{Z}_{ij}, Y_{ij} = 1, \alpha_i)} \tag{4}$$

where the first equality is due to logistic assumption in equation (2) and the second can be derived from equation (3). Thus, the last term of equation (3) can be rewritten as

$$\log\frac{f(X_{ij}|\mathbf{Z}_{ij}, Y_{ij} = 0, \alpha_i)}{f(X_{ij}|\mathbf{Z}_{ij}, Y_{ij} = 1, \alpha_i)} = -X_{ij}\beta_1 + \log\frac{f(X_{ij} = 0|\mathbf{Z}_{ij}, Y_{ij} = 0, \alpha_i)}{f(X_{ij} = 0|\mathbf{Z}_{ij}, Y_{ij} = 1, \alpha_i)} \tag{5}$$

Now, suppose that

$$X_{ij} = \gamma_0 + \mathbf{Z}_{ij}^T \boldsymbol{\gamma_1} + Y_{ij}\delta + \rho\alpha_i + e_{ij} \tag{6}$$

where $(\gamma_0, \boldsymbol{\gamma_1}^T, \delta, \rho)$ is a vector of unknown parameters, e_{ij}'s are independent and normally distributed with mean 0 and variance σ^2, and $\alpha_i \perp e_{ij}$. Note that when $D > 0$ with $\delta \neq 0$ or $\rho \neq 0$, X_{ij} is an endogenous regressor of model (2). When $\delta = 0$, equation (6) reduces to the X-model of Neuhaus and McCulloch.[14] However, when $\delta \neq 0$, $(X_{ij} - \bar{X}_i)$ is still correlated with α_i through $(Y_{ij} - \bar{Y}_i)$ in a nonlinear fashion, where $\bar{Y}_i$ is the ith cluster mean of Y. Hence, the partitioning method of Neuhaus and McCulloch[14] cannot be applied. Due to the distributional assumption in equation (6), the right-hand side of equation (4) can be further computed and gives

$$\exp(\beta_1 x) = \exp\left(\frac{\delta}{\sigma^2} x\right) \quad \text{for every } x \in \mathbb{R}$$

This leads to a relationship between the parameters from the $Y|X$ model and $X|Y$ model

$$\beta_1 = \frac{\delta}{\sigma^2} \tag{7}$$

Also, the right-hand side of equation (5) can be further computed to give

$$\log \frac{f(X_{ij}|\mathbf{Z}_{ij}, Y_{ij} = 0, \alpha_i)}{f(X_{ij}|\mathbf{Z}_{ij}, Y_{ij} = 1, \alpha_i)} = -X_{ij}\beta_1 + \frac{\delta}{\sigma^2}\left\{\frac{\mu_1(\mathbf{Z}_{ij}) + \mu_0(\mathbf{Z}_{ij})}{2} + \rho\alpha_i\right\} = -X_{ij}\beta_1 + \beta_1\left\{\frac{\mu_1(\mathbf{Z}_{ij}) + \mu_0(\mathbf{Z}_{ij})}{2} + \rho\alpha_i\right\} \tag{8}$$

where $\mu_y(\mathbf{Z}_{ij}) = \gamma_0 + \mathbf{Z}_{ij}^T \boldsymbol{\gamma_1} + y\delta,\ y = 1, 0$. The second equality is due to equation (7). Let $X_{ij}^* = \{\mu_1(\mathbf{Z}_{ij}) + \mu_0(\mathbf{Z}_{ij})\}/2 = \gamma_0 + \mathbf{Z}_{ij}^T \boldsymbol{\gamma_1} + \frac{\delta}{2}$. Combining equations (2), (3), and (8) leads to the marginal model

$$\text{logit}\{P(Y_{ij} = 1|\mathbf{Z}_{ij}, \alpha_i)\} = \beta_0 + X_{ij}^*\beta_1 + \mathbf{Z}_{ij}^T \boldsymbol{\beta_2} + (1 + \rho\beta_1)\alpha_i \tag{9}$$

The marginal probability $P(Y_{ij} = 1|\mathbf{Z}_{ij}, \alpha_i)$ without conditioning argument X_{ij} using a link-preserving imputation model (6) results in simply replacing X_{ij} in $P(Y_{ij} = 1|X_{ij}, \mathbf{Z}_{ij}, \alpha_i)$ with X_{ij}^*, and α_i with $b_i = (1 + \rho\beta_1)\alpha_i$. The internal IV X_{ij}^* explains some variability in X_{ij} not associated with the omitted variable term α_i. However, we still cannot apply standard GLMM methods to equation (9) because X_{ij}^* is linear in $\mathbf{Z}_{ij}$, rendering β_0, β_1, and $\boldsymbol{\beta_2}$ unidentifiable. This is a new problem which was not encountered in Paik and Sacco,[15] where the regressor is partially missing but not endogenous, and thus X_{ij} can be replaced only for the missing records. In our case, we need to replace X_{ij} with X_{ij}^* for all the records, whether or not they are missing.

3.2 Estimation

From equation (9), we have

$$\text{logit}\{P(Y_{ij} = 1|\mathbf{Z}_{ij}, \alpha_i)\} = \left(\beta_0 + \frac{\delta}{2}\beta_1 + \gamma_0\beta_1\right) + \mathbf{Z}_{ij}^T(\boldsymbol{\gamma_1}\beta_1 + \boldsymbol{\beta_2}) + b_i = \phi_0 + \mathbf{Z}_{ij}^T \boldsymbol{\phi_1} + b_i \tag{10}$$

where $\phi_0 = \beta_0 + \frac{\delta}{2}\beta_1 + \gamma_0\beta_1$ and $\boldsymbol{\phi_1} = \boldsymbol{\gamma_1}\beta_1 + \boldsymbol{\beta_2}$. Standard GLMM procedure produces consistent and asymptotically normal (CAN) estimates $\hat{\phi}_0$ and $\hat{\boldsymbol{\phi}}_{\mathbf{1}}$. To estimate β_0, β_1, and $\boldsymbol{\beta_2}$, we can borrow information from the imputation model (6). Estimating the parameters of equation (6) faces endogeneity problem, due to equation (2). However, unlike Y_{ij}, the conditional mean of X_{ij} is linear in the regressors, so we can use the Hausman–Taylor estimator.[12] We describe the whole procedure in Appendix 1. The resulting estimates $(\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_{\mathbf{1}}^T, \hat{\delta}, \hat{\sigma}^2)$ are CAN. Then from equations (7) and (10)

$$\hat{\beta}_0 = \hat{\phi}_0 - \left(\frac{\hat{\delta}}{2} + \hat{\gamma}_0\right)\hat{\beta}_1$$

$$\hat{\beta}_1 = \frac{\hat{\delta}}{\hat{\sigma}^2}$$
$$\hat{\boldsymbol{\beta}}_{\mathbf{2}} = \hat{\boldsymbol{\phi}}_{\mathbf{1}} - \hat{\boldsymbol{\gamma}}_{\mathbf{1}}\hat{\beta}_1 \tag{11}$$

which are CAN as well due to delta method. In summary, we estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}_{\mathbf{2}}^T)^T$ using the following three steps:

`STEP 1`: Compute the Hausman–Taylor estimates $(\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_{\mathbf{1}}^T, \hat{\delta}, \hat{\sigma}^2)$ of the parameters in equation (6).

`STEP 2`: Using standard GLMM procedures, compute estimates $\hat{\phi}_0$ and $\hat{\boldsymbol{\phi}}_{\mathbf{1}}$ of the parameters in equation (10).

`STEP 3`: Calculate $(\hat{\beta}_0, \hat{\beta}_1, \hat{\boldsymbol{\beta}}_{\mathbf{2}}^T)$ using equation (11).

The methodology can be extended to the case where the imputation model (6) includes a nonlinear term with respect to $\mathbf{Z}$ and an interaction term between Y and $\mathbf{Z}$. When there is an interaction term, the outcome model (2) should also include an interaction term between X and $\mathbf{Z}$. Derivation is analogous, so we skip the details.

4 Extension of the proposed method to allow missingness in X

In the methods introduced in section 2 that either use external instruments or the within-cluster deviations of X, endogeneity and missingness are two separate challenges that should be respectively dealt with. In our method, X_{ij}^* is a function of completely observed exogenous regressors, so the replacement resolves the endogeneity and missingness issue simultaneously.

Let r_{ij} denote a binary variable indicating whether X_{ij} is observed. Suppose that given the random effect and observed variables, the missingness is independent of X_{ij}

$$\pi_{ij} \equiv P(r_{ij} = 1 | X_{ij}, Y_{ij}, \mathbf{Z}_{ij}, \alpha_i) = P(r_{ij} = 1 | Y_{ij}, \mathbf{Z}_{ij}, \alpha_i) \tag{12}$$

Since r_{ij} depends on unobserved α_i, the missingness is nonignorable. Yuan and Little[17] called this missing mechanism cluster-specific non-ignorable (CSNI). Note that if r_{ij} depends only on Y_{ij} and $\mathbf{Z}_{ij}$, the missing mechanism is MAR. Therefore, equation (12) covers MAR missingness as a special case.

In our method, estimating $P(Y_{ij} = 1 | \mathbf{Z}_{ij}, \alpha_i)$ does not change when X_{ij} is missing. Estimating the imputation model (6) needs to take care of missing X_{ij}. Under equation (12), $E(X_{ij} | \mathbf{Z}_{ij}, Y_{ij}, \alpha_i, r_{ij} = 1) = E(X_{ij} | \mathbf{Z}_{ij}, Y_{ij}, \alpha_i, r_{ij} = 0)$, so the estimating procedure based on the records with observed X_{ij} yields CAN estimates of $(\gamma_0, \boldsymbol{\gamma}_{\mathbf{1}}^T, \delta, \sigma^2)$. Therefore, the proposed method can be easily extended to the missing X case by replacing estimation of $(\gamma_0, \boldsymbol{\gamma}_{\mathbf{1}}^T, \delta, \sigma^2)$ with complete record analysis. It is notable that even if the missingness is nonignorable, we do not require to specify the model for the missing probability.

5 Simulation studies

We conducted simulation studies to evaluate the finite sample properties of the proposed estimator. We set K, the number of clusters, and n_i, the number of units in one cluster, as $(K, n_i) = (50, 20)$ or $(100, 10)$. We generated Z_{ij} from `Uniform`$(0, 1)$, α_i from $N(0, D)$, Y_{ij} from (10), and X_{ij} from (6) with $\gamma_0 = -1, \gamma_1 = -1$, $\delta = 3, \sigma^2 = 1, \beta_0 = -1, \beta_1 = 3, \beta_2 = 1.5, D = 1$, and $\rho = 0.7$. Consequently, Y_{ij} follows equation (2). The observation indicator r_{ij} of X_{ij} was generated from Ber(π_{ij}) where π_{ij} satisfies equation (12) and

$$\text{logit}\{\pi_{ij}\} = \varphi_0 + \varphi_1 Y_{ij} + \varphi_2 Z_{ij} + \varphi_3 \alpha_i$$

with $\varphi_0 = 1$, $\varphi_1 = 1$, $\varphi_2 = -1$, and $\varphi_3 = -0.5$. The overall missing probability was 28.57%.

We compare the proposed method with the partitioning method[14] and the naive method, which ignores the presence of endogeneity. As an alternative to the proposed method, we also report the results of maximizing the following joint observed likelihood of X, Y, and r using the `mdhglm` package[28] with second-order Laplace approximation

$$\prod_{i=1}^{K} \int \prod_{j=1}^{n_i} \int f(Y_{ij} | Z_{ij}, \alpha_i)\, g(X_{ij} | Y_{ij}, Z_{ij}, \alpha_i) \pi_{ij}^{r_{ij}} (1 - \pi_{ij})^{1 - r_{ij}} k(\alpha_i) \mathrm{d}X_{ij,miss} \mathrm{d}\alpha_i \tag{13}$$

where $f(\cdot)$ is the marginal distribution (10) of $(Y|Z,\alpha)$, $g(\cdot)$ is the normal distribution (6) of $(X|Y,Z,\alpha)$, and $k(\cdot)$ is the normal distribution of α. We denote the results by HGLM (1). Note that unlike the proposed method, π_{ij} should be correctly specified. This is because due to dependence on α, π_{ij} cannot come out of the outer integration. Additionally, we maximize the joint likelihood specified by Neuhaus and McCulloch[14]

$$\prod_{i=1}^{K}\int\prod_{j=1}^{n_i}\int \tilde{f}(Y_{ij}|X_{ij},Z_{ij},\alpha_i)\tilde{g}(X_{ij}|Z_{ij},\alpha_i)\pi_{ij}^{r_{ij}}(1-\pi_{ij})^{1-r_{ij}}k(\alpha_i)\mathrm{d}X_{ij,miss}\mathrm{d}\alpha_i$$

where $\tilde{f}(\cdot)$ is the conditional distribution (2) of $(Y|X,Z,\alpha)$ and $\tilde{g}(\cdot)$ is the misspecified distribution of $(X|Z,\alpha)$ assuming that equation (6) holds with $\delta=0$. We denote the results by HGLM (2).

To compute the naive estimators, we applied a standard GLMM method to equation (2). When X was missing, we used only the complete records for the naive and partitioning methods. The R codes to compute the proposed estimators and their standard errors are given in the Supplementary Material. The variance of the proposed estimators was estimated using Jackknife variance. We calculated the 95% confidence intervals based on normality of the estimator.

Tables 1 and 2 show the bias, simulation variance, and coverage probability of each estimate, for full data and for data with missing regressor, respectively. We indicated significant bias in bold. The naive, partitioning, and HGLM (2) estimators show large bias compared to the proposed and HGLM (1) estimators. Since $\delta\neq 0$, the partitioning and HGLM (2) estimates of β_1 have larger bias than the proposed method. Also, the naive and partitioning estimates have large standard errors, especially when X is missing and the cluster size is small. The coverage probabilities of the proposed and HGLM (1) estimators are close to the nominal value in all cases.

Despite correct specification of the likelihood, HGLM (1) estimators appear to have significant bias and larger standard errors than the proposed method. HGLM (1) maximizes the joint likelihood (13) of X, Y, and r given Z. If the exact joint likelihood is available the resulting maximum likelihood estimator would be the most efficient. However, this joint likelihood involves a high-dimensional integration. The proposed method maximizes the marginal likelihood of $(Y|\mathbf{Z})$, which is lower-dimensional than HGLM (1) when covariates are missing. Even without missingness, the proposed method maximizes the marginal likelihood of $(Y|\mathbf{Z})$ instead of the joint

Table 1. Bias, standard error (S.E.), and coverage probability (C.P.) of the simulation estimates based on 500 samples, when X is fully observed.

	$(K,\ n_i)=(50,\ 20)$			$(K,\ n_i)=(100,\ 10)$		
Estimator	Bias	S.E.	C.P.	Bias	S.E.	C.P.
Proposed method						
$\hat{\beta}_0$	0.022	0.346	0.948	−0.014	0.322	0.952
$\hat{\beta}_1$	0.000	0.169	0.948	0.003	0.180	0.944
$\hat{\beta}_2$	−0.021	0.521	0.952	0.012	0.515	0.950
HGLM (1)						
$\hat{\beta}_0$	0.018	0.416	0.948	−0.011	0.399	0.960
$\hat{\beta}_1$	**0.027**	0.170	0.954	**0.068**	0.166	0.928
$\hat{\beta}_2$	**−0.094**	0.587	0.940	**−0.108**	0.594	0.954
Partitioning method						
$\hat{\beta}_0$	0.020	0.452	0.954	**−0.048**	0.434	0.970
$\hat{\beta}_1$	**0.160**	0.353	0.960	**0.237**	0.497	0.966
$\hat{\beta}_2$	0.055	0.772	0.958	**0.191**	0.770	0.958
HGLM (2)						
$\hat{\beta}_0$	**0.167**	0.534	0.886	**0.120**	0.461	0.926
$\hat{\beta}_1$	**0.100**	0.347	0.944	**0.060**	0.367	0.932
$\hat{\beta}_2$	**−0.361**	0.854	0.880	**−0.278**	0.803	0.918
Naive method						
$\hat{\beta}_0$	**−0.210**	0.441	0.922	**−0.274**	0.430	0.966
$\hat{\beta}_1$	**0.410**	0.358	0.868	**0.494**	0.516	0.900
$\hat{\beta}_2$	**0.483**	0.748	0.892	**0.606**	0.750	0.930

Table 2. Bias, standard error (S.E.), and coverage probability (C.P.) of the simulation estimates based on 500 samples, when X is missing.

	$(K, n_i) = (50, 20)$			$(K, n_i) = (100, 10)$		
Estimator	Bias	S.E.	C.P.	Bias	S.E.	C.P.
Proposed method						
$\hat{\beta}_0$	**0.032**	0.357	0.948	0.004	0.352	0.948
$\hat{\beta}_1$	0.000	0.194	0.954	0.011	0.216	0.948
$\hat{\beta}_2$	−0.014	0.570	0.952	0.020	0.589	0.946
HGLM (1)						
$\hat{\beta}_0$	**0.072**	0.347	0.946	**0.158**	0.358	0.934
$\hat{\beta}_1$	**−0.025**	0.192	0.940	**0.020**	0.200	0.950
$\hat{\beta}_2$	**−0.113**	0.578	0.948	**−0.109**	0.602	0.948
Partitioning method						
$\hat{\beta}_0$	**0.110**	0.542	0.940	−0.031	1.227	0.954
$\hat{\beta}_1$	**0.233**	0.472	0.970	**0.776**	3.416	0.936
$\hat{\beta}_2$	**0.296**	0.959	0.936	**0.616**	2.309	0.934
HGLM (2)						
$\hat{\beta}_0$	**0.348**	0.355	0.930	**0.324**	0.359	0.842
$\hat{\beta}_1$	**0.133**	0.479	0.886	**0.232**	0.450	0.920
$\hat{\beta}_2$	**−0.916**	0.504	0.538	**−0.793**	0.460	0.616
Naive method						
$\hat{\beta}_0$	**−0.086**	0.539	0.946	**−0.265**	1.009	0.954
$\hat{\beta}_1$	**0.508**	0.517	0.944	**1.081**	3.117	0.918
$\hat{\beta}_2$	**0.728**	0.937	0.896	**1.073**	2.138	0.894

Table 3. Estimates of β by proposed and naive method, with the standard errors (S.E.).

	Proposed		Proposed_CR		Naive	
Regressor	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	−2.812***	0.412	−2.783***	0.413	−3.051***	0.254
log Cost	0.031	0.043	0.031	0.043	0.061*	0.026
Age						
50–59	.	.	.	.	.	.
60–69	−0.069	0.086	−0.093	0.098	−0.094	0.059
70–79	−0.006	0.064	−0.008	0.071	−0.009	0.064
>80	0.145	0.088	0.127	0.099	0.131*	0.064
Sex						
Male	.	.	.	.	.	.
Female	−0.154***	0.033	−0.160***	0.038	−0.158***	0.035
Race						
White	.	.	.	.	.	.
Black	0.182*	0.087	0.189	0.113	0.192*	0.085
Hispanic	−0.062	0.041	−0.079*	0.037	−0.078	0.053
Others	0.068	0.066	0.082	0.067	0.080	0.068
Income						
High	−0.082	0.046	−0.078	0.050	−0.079	0.044
Medium	.	.	.	.	.	.
Low	−0.0001	0.080	−0.012	0.094	−0.011	0.054
Insurance						
Medicare	.	.	.	.	.	.
Medicaid	0.202**	0.062	0.195**	0.064	0.198**	0.067
Private	−0.459***	0.073	−0.497***	0.075	−0.499***	0.063
Others[a]	−0.174	0.124	−0.185	0.126	−0.183*	0.092

(continued)

Table 3. Continued

	Proposed		Proposed_CR		Naive	
Regressor	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
N.chro.cond.	0.147***	0.015	0.153***	0.016	0.153***	0.012
Length.Stay	0.031***	0.007	0.030***	0.007	0.028***	0.003
Bed size						
Small	0.226	0.554	0.238	0.707	0.240	0.143
Medium	.	.	.	.	.	.
Large	0.119	0.070	0.156**	0.054	0.160**	0.041
Ownership						
Profit	0.093	0.601	0.071	0.617	0.064	0.089
Non-Profit	.	.	.	.	.	.
Public	−0.014	0.083	−0.049	0.070	−0.043	0.046
Teaching						
Yes	0.055	0.076	0.020	0.058	0.020	0.053
No	.	.	.	.	.	.
OprMargin	−0.002	0.005	−0.004	0.005	−0.004	0.002

[a]Others include self-pay, no-charge, county indigent programs, charity care, etc. *$p < 0.05$, **$p < 0.01$, ***$p < 0.0001$.

Table 4. Estimates of β by partitioning method, HGLM (1), and HGLM (2), with the standard errors (S.E.).

	HGLM (1)		Partitioning		HGLM (2)	
Regressor	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	−3.292***	0.385	−0.097	1.268	−2.96***	0.239
log Cost	0.029	0.039	0.068*	0.027	0.047	0.024
Age						
50–59	.	.	.	.	.	.
60–69	−0.069	0.086	−0.093	0.059	−0.069	0.056
70–79	−0.006	0.064	−0.01	0.064	−0.007	0.061
>80	0.145	0.088	0.13*	0.064	0.146*	0.06
Sex						
Male	.	.	.	.	.	.
Female	−0.154***	0.033	−0.16***	0.035	−0.153***	0.032
Race						
White	.	.	.	.	.	.
Black	0.182*	0.088	0.188*	0.085	0.183*	0.078
Hispanic	−0.061	0.041	−0.078	0.053	−0.062	0.05
Others	0.068	0.066	0.086	0.068	0.067	0.066
Income						
High	−0.082	0.046	−0.071	0.044	−0.082	0.043
Medium	.	.	.	.	.	.
Low	0	0.08	−0.014	0.054	0	0.05
Insurance						
Medicare	.	.	.	.	.	.
Medicaid	0.202**	0.062	0.198**	0.067	0.202**	0.066
Private	−0.459***	0.073	−0.497***	0.063	−0.459***	0.058
Others[a]	−0.174	0.124	−0.182*	0.092	−0.174	0.091
N.chro.cond.	0.147***	0.015	0.153***	0.012	0.147***	0.011
Length.Stay	0.032***	0.007	0.028***	0.003	0.031***	0.003
Bed Size						
Small	0.225	0.555	0.224	0.142	0.224	0.152
Medium	.	.	.	.	.	.
Large	0.119	0.07	0.133**	0.043	0.122*	0.051

(continued)

Table 4. Continued

	HGLM (1)		Partitioning		HGLM (2)	
Regressor	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Ownership						
Profit	0.092	0.601	0.074	0.089	0.085	0.104
Non-Profit	.	.	.	.	.	.
Public	−0.014	0.083	−0.124*	0.058	−0.01	0.058
Teaching						
Yes	0.056	0.075	0.071	0.057	0.055	0.069
No	.	.	.	.	.	.
OprMargin	−0.002	0.005	−0.003	0.003	−0.002	0.003

[a]Others include self-pay, no-charge, county indigent programs, charity care, etc. *$p < 0.05$, **$p < 0.01$, ***$p < 0.0001$.

likelihood $(X, Y|\mathbf{Z})$, so that the proposed method seems less biased due to approximation. This may explain the smaller biases and standard errors for the proposed estimators over the HGLM (1) estimators.

We should also note that HGLM (1) requires the missing probability π_{ij} to be correctly specified, while the proposed method does not. In simulations, π_{ij} was correctly specified for HLGM (1). In practice, it is hard to identify the missing mechanism. Thus, we recommend using the proposed method.

6 Application to study on health care cost among San Diego inpatients

We applied the proposed method, naive method, partitioning method, HGLM (1), and HGLM (2) to the 2006 San Diego inpatient dataset. The response Y is the binary variable which takes value 1 if the patient experiences UR, and 0 otherwise. The endogenous variable X is the *logarithm* of health care cost in U.S. dollars ($). The hospital-level exogenous regressors, $\mathbf{Z}_1$, include Staffed beds size, Ownership (profit, public, or nonprofit), Educational status (teaching hospital or not) and Operational margin (OprMargin). The patient-level exogenous regressors, $\mathbf{Z}_2$, include Sex, Age, Insurance status, Income status, Race, Number of chronic conditions (N.chro.cond.), and Length of Stay (Length.Stay). We assume that Y follows a GLMM with logistic link as described in equation (2). In the data, X is missing up to 13.06% and the missing proportion varies widely among clusters, possibly representing a CSNI mechanism. To verify this, we fitted a logistic model with a cluster-level random effect using the indicator of whether the cost variable is observed as a binary outcome and all the observed characteristics as regressors. We conducted hypothesis testing on whether the variance component is zero via a likelihood ratio test when the null value lies on the boundary of the parameter space,[29] and rejected the null hypothesis ($p < 0.0001$). The Operational margin was also missing for one hospital and the average of the other 19 hospitals was imputed.

Table 3 shows the estimates of $\boldsymbol{\beta}$ by the proposed and naive method. As the difference between the two methods is due to both endogeneity and missingness, we also included the results of applying the proposed method to only complete records (Proposed_CR). Table 4 shows the same estimates by HGLM (1), partitioning method, and HGLM (2). Since HGLM (1) and HGLM (2) require to specify the functional form of the missing probability of X, we used a logistic GLMM with Y, $\mathbf{Z}_1$, $\mathbf{Z}_2$ as covariates. The proposed method and HGLM (1) have similar results, which may imply that the specification of the missing mechanism would be reasonable.

The proposed and HGLM (1) estimates of β_1 show that the effect of log Cost on the experience of UR is nonsignificant after adjusting for the unmeasured hospital-level confounders such as general service quality, competence of the hospital crew, and policies for discharged patients. The results are also reflected in the following plots (Figure 1), where the adjusted log Cost was computed as

$$(\text{adjusted log Cost}) = X - \hat{\gamma}_0 - \mathbf{Z}^T\hat{\boldsymbol{\gamma}}_\mathbf{1} - Y\hat{\delta} - \widehat{\rho\alpha}$$

where $\widehat{\rho\alpha}$ is the h-likelihood predictor of $\rho\alpha$. The HGLM (2) estimate of β_1 shows a nonsignificant effect as well, but with $p = 0.054$. The naive and partitioning estimates show a significant effect.

The effects of age and bed size in the previous hospital are also shown to be significant by the naive method, partitioning method, and HGLM (2), whereas the proposed method and HGLM (1) indicate a nonsignificant effect. The naive method ignoring endogeneity reports that the variance component of the random effects is not

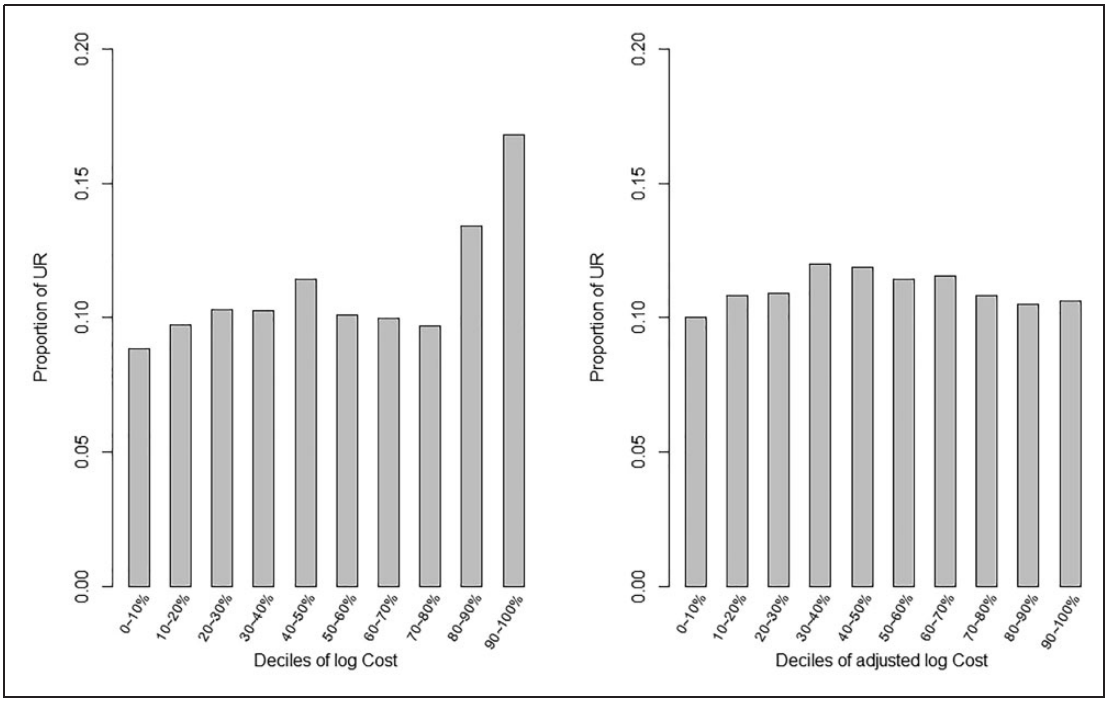

Figure 1. Barplots of the proportion of UR against deciles of log Cost (left) and adjusted log Cost (right).

significant ($p = 0.5$), while the proposed method shows otherwise ($p = 0.048$). We speculate that with the naive method, the effects of the unmeasured hospital condition cannot be distinguished from the effect of the endogenous health care cost variable.

The proposed method suggests that people who tend to experience UR are male, black, have relatively more chronic conditions and have stayed longer in the previous hospital. Strategies for better care coordination targeting people at risk for readmissions may help prevent UR.

7 Conclusion

We proposed a consistent and asymptotically normal parameter estimator of a logistic GLMM for binary outcome when one of the regressors is endogenous due to cluster-level omitted effects. The method does not require external IVs and can be implemented using existing software. The proposed method can also be applied to the case of missing data without requiring it to correctly specify the missing mechanism. The derivation of the method demonstrates that a clustered data structure can be exploited to draw valid analysis of multilevel data with correlated effects.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: G. S. Kim, Y. Lee, and M. C. Paik were supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIP) (No. 2016H1A2A1907857, No. 2019R1A2C1002408, No. 2017R1A2B4008956).

ORCID iD

Gi-Soo Kim https://orcid.org/0000-0002-2983-2367

Supplemental material

Supplemental material for this article is available online.

References

1. Kim H, Hung WW, Paik MC, et al. Predictors and outcomes of unplanned readmission to a different hospital. *Int J Qual Health Care* 2015; **27**: 513–519.
2. Kind AJH, Bartels C, Mell MW, et al. For-profit hospital status and rehospitalizations at different hospitals: an analysis of Medicare data. *AnnIntern Med* 2010; **153**: 718–727.
3. Jencks SF, Williams MV and Coleman EA. Rehospitalizations among patients in the Medicare Fee-for-Service Program. *N Engl J Med* 2009; **360**: 1418–1428.
4. James J. Medicare hospital readmissions reduction program. Health affairs, www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=102 (2013, accessed 02 August 2015).
5. Ebbes P, Bockenholt U and Wedel M. Regressor and random-effects dependencies in multilevel models. *Stat Neerl* 2004; **58**: 161–178.
6. Bowden RJ and Turkington DA. *Instrumental variables*. Cambridge: Cambridge University Press, 1984.
7. Amemiya T. The nonlinear two-stage least-squares estimator. *J Econom* 1974; **2**: 105–110.
8. Johnston KM, Gustafson P, Levy AR, et al. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Stat Med* 2008; **27**: 1539–1556.
9. Terza JV, Basu A and Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J Health Econ* 2008; **27**: 531–543.
10. Bound J, Jaeger DA and Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 1995; **90**: 443–450.
11. Mundlak Y. On the pooling of time series and cross section data. *Econometrica* 1978; **46**: 69–85.
12. Hausman JA and Taylor WE. Panel data and unobservable individual effect. *Econometrica* 1981; **49**: 1377–1398.
13. Kim JS and Frees EW. Multilevel modeling with correlated effects. *Psychometrika* 2007; **72**: 505–533.
14. Neuhaus JM and McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J R Stat Soc B* 2006; **68**: 859–872.
15. Paik MC and Sacco RL. Matched case-control data analyses with missing covariates. *J R Stat Soc C* 2000; **49**: 145–156.
16. Chen Q, Paik MC, Kim M, et al. Using link-preserving imputation for logistic partially linear models with missing covariates. *Comput Stat Data Anal* 2016; **101**: 174–185.
17. Yuan Y and Little RJA. Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *J R Stat Soc C* 2007; **56**: 79–97.
18. Andersen EB. Asymptotic properties of conditional maximum-likelihood estimators. *J R Stat Soc B* 1970; **32**: 283–301.
19. Lee Y, Nelder JA and Pawitan Y. *Generalized linear models with random effects: unified analysis via H-likelihood*, 2nd ed. London: Chapman and Hall, 2017.
20. Brumback BA and He Z. Adjusting for confounding by neighborhood using complex survey data. *Stat Med* 2011; **30**: 965–972.
21. Brumback BA, Dailey AB and Zheng HW. Adjusting for confounding by neighborhood using a proportional odds model and complex survey data. *Am J Epidemiol* 2012; **175**: 1133–1141.
22. Brumback BA, Cai Z, He Z, et al. Conditional pseudolikelihood methods for clustered ordinal, multinomial, or count outcomes with complex survey data. *Stat Med* 2013; **32**: 1325–1335.
23. Goetgeluk S and Vansteelandt S. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* 2008; **64**: 772–780.
24. Neuhaus JM and Kalbfleisch JD. Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998; **54**: 638–645.
25. Brumback BA, Dailey AB, Brumback LC, et al. Adjusting for confounding by cluster using generalized linear mixed models. *Stat Prob Lett* 2010; **80**: 1650–1654.
26. Brumback BA, Zheng HW and Dailey AB. Adjusting for confounding by neighborhood using generalized linear mixed models and complex survey data. *Stat Med* 2013; **32**: 1313–1324.
27. Brumback BA, Dailey AB, He Z, et al. Efforts to adjust for confounding by neighborhood using complex survey data. *Statistics in medicine* 2010; **29**(18): 1890–1899.
28. Lee Y, Molas M and Noh M. mdhglm: multivariate double hierarchical generalized linear models. R package version 1.6. https://CRAN.R-project.org/package=mdhglm (2016).
29. Zhang D and Lin X. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In: Diggle P, Gather U and Zeger S (eds) *Lecture notes in statistics vol. 192*. New York: Springer, 2008, pp. 19–36.

Appendix 1

Hausman–Taylor estimator

In this section, we derive the Hausman–Taylor estimator of $(\gamma_0, \boldsymbol{\gamma}_1^T, \delta, \sigma^2)$ from equation (6). Without loss of generality, let $\mathbf{Z}_{ij} = (\mathbf{Z}_{1i}^T, \mathbf{Z}_{2ij}^T)^T$ where $\mathbf{Z}_{1i}$ is a vector of cluster level regressors and $\mathbf{Z}_{2ij}$ is a vector of unit level regressors. Accordingly, we have $\boldsymbol{\beta}_2^T = (\boldsymbol{\beta}_{21}^T, \boldsymbol{\beta}_{22}^T)$ and $\boldsymbol{\gamma}_1^T = (\boldsymbol{\gamma}_{11}^T, \boldsymbol{\gamma}_{12}^T)$. Due to linearity, we can remove the $\rho\alpha_i$ term by substracting the clusterwise means from both sides of equation (6), and have

$$X_{ij} - \bar{X}_i = (Y_{ij} - \bar{Y}_i)\delta + (\mathbf{Z}_{2ij} - \overline{\mathbf{Z}}_{2i})^T \boldsymbol{\gamma}_{12} + e_{ij} - \bar{e}_i$$

where $\bar{X}_i$, $\bar{Y}_i$, $\overline{\mathbf{Z}}_{2i}$, and $\bar{e}_i$ are the ith cluster mean of X, Y, $\mathbf{Z}_2$, and e, respectively. Therefore, the within-cluster variations $(Y_{ij} - \bar{Y}_i)$ and $(\mathbf{Z}_{2ij} - \overline{\mathbf{Z}}_{2i})$ serve as internal instruments. The new error term $(e_{ij} - \bar{e}_i)$ has still mean 0 and since we know the correlation structure, we can apply the generalized least squares (GLS) method to estimate δ, γ_{12}, and σ^2.

Let

$$\mathbf{X}_i^c = \begin{pmatrix} X_{i1} - \bar{X}_i \\ X_{i2} - \bar{X}_i \\ \vdots \\ X_{i,n_i-1} - \bar{X}_i \end{pmatrix}, \quad \mathbf{Y}_i^c = \begin{pmatrix} Y_{i1} - \bar{Y}_i \\ Y_{i2} - \bar{Y}_i \\ \vdots \\ Y_{i,n_i-1} - \bar{Y}_i \end{pmatrix}, \quad \mathbf{Z}_{2i}^c = \begin{pmatrix} (\mathbf{Z}_{2i1} - \overline{\mathbf{Z}}_{2i})^T \\ (\mathbf{Z}_{2i2} - \overline{\mathbf{Z}}_{2i})^T \\ \vdots \\ (\mathbf{Z}_{2i,n_i-1} - \overline{\mathbf{Z}}_{2i})^T \end{pmatrix}$$

Then, $\mathbf{X}_i^c$ follows a multivariate normal distribution

$$\mathbf{X}_i^c \sim N(\mathbf{Y}_i^c\delta + \mathbf{Z}_{2i}^c\boldsymbol{\gamma}_{12}, \quad \sigma^2\boldsymbol{\Sigma}_i)$$

where $\boldsymbol{\Sigma}_i$ is an $(n_i - 1) \times (n_i - 1)$ matrix with $\left(1 - \frac{1}{n_i}\right)$ for the diagonals and $\left(-\frac{1}{n_i}\right)$ for the off-diagonals. Multiplying both sides by $\mathbf{L}_i = \boldsymbol{\Sigma}_i^{-\frac{1}{2}}$, we have

$$\mathbf{L}_i\mathbf{X}_i^c \sim N((\mathbf{L}_i\mathbf{Y}_i^c)\delta + (\mathbf{L}_i\mathbf{Z}_{2i}^c)\gamma_{12}, \quad \sigma^2\mathbf{I}_{n_i-1})$$

Then, ordinary least squares procedure yields consistent and normal estimates $\hat{\delta}$, $\hat{\gamma}_{12}$, and $\hat{\sigma}^2$.

As for γ_0 and $\boldsymbol{\gamma}_{11}$, we can fit a linear mixed-effects model (LMM) on

$$(X_{ij} - Y_{ij}\delta - \mathbf{Z}_{2ij}^T\boldsymbol{\gamma}_{12}) = \gamma_0 + \mathbf{Z}_{1i}^T\boldsymbol{\gamma}_{11} + \rho\alpha_i + e_{ij}$$

with $\hat{\delta}$ and $\hat{\boldsymbol{\gamma}}_{12}$ plugged in the left-hand side. As the endogenous Y_{ij} is moved to the outcome, standard methods produce CAN estimates $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}_{11}$.